AlleleSNP Introduction

1. Background

Allele-specific effects (ASE) are the variations within a single individual, such as differences in chromatin signatures, DNA methylation and gene expression, that are related to the different alleles of a SNP (Birney et al. 2010; Rozowsky et al. 2011). When a heterozygous SNP shows ASE of epigenetic signatures, such as transcription factor (TF) binding or histone modifications, it is a strong indication of its functionality because it shows that within the same cellular environment, the two SNP alleles can behave differently (**Fig 1**). ASE can be identified through examination of the NGS data: for example, we can collect a TF ChIP-seq reads that mapped to a certain SNP, if the number of reads that contain the reference allele and the alternate allele are imbalanced, it might indicate that the SNP might play a critical role in regulating the binding affinity of the TF (**Fig 2**).

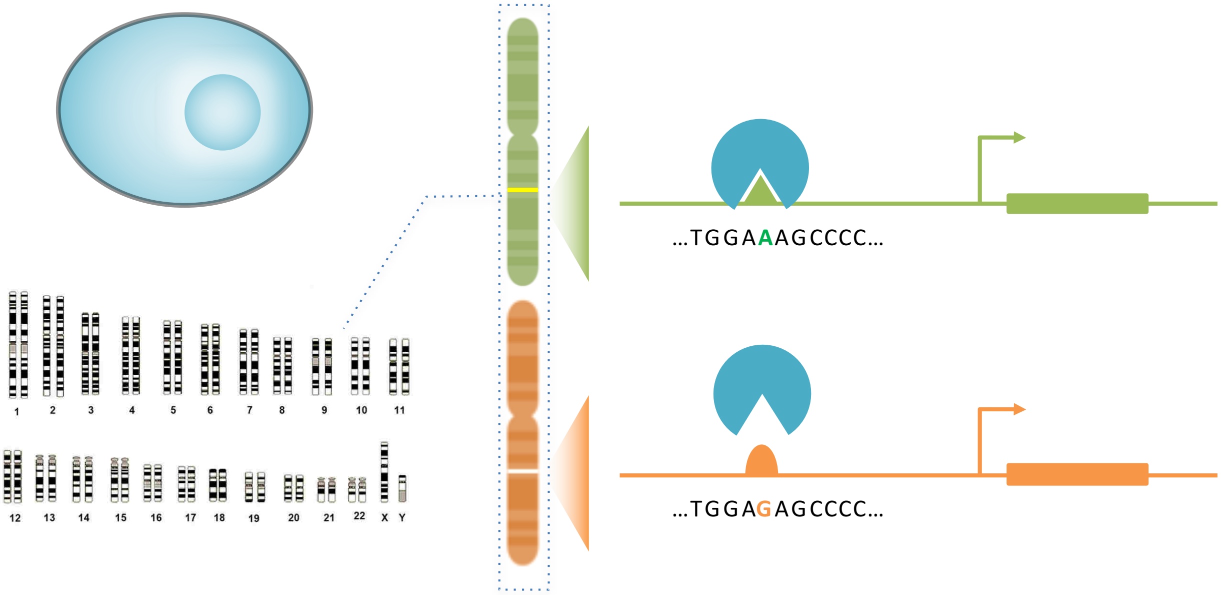

Figure 1. Schematic plot of allele-specific effects

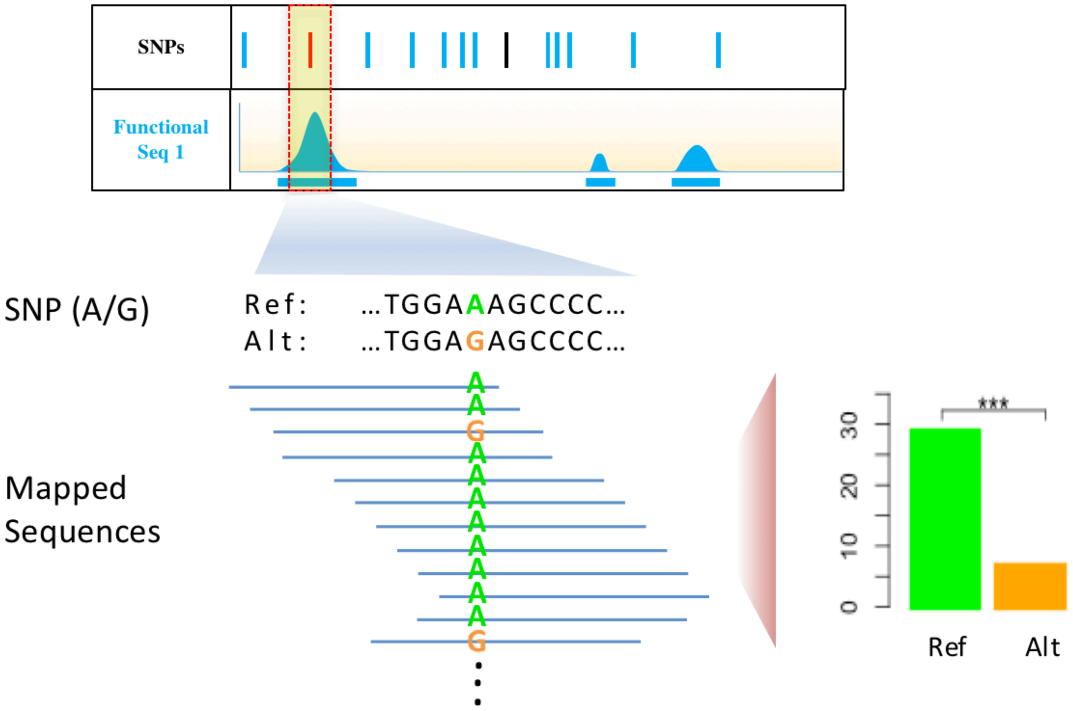

Figure 2. Schematic plot of ASE identification through NGS data

2. Installation

```
if (!require(AlleleSNP)) {
        library(devtools)
        install_github("foreverycc/AlleleSNP_Package")
}
library(AlleleSNP)
```

3. Identify ASE by bam files

The first way to identify ASE is through examination of bam file.

For each bam file (ChIP-seq, DNase-seq, ATAC-seq, FAIRE-seq, etc.), AlleleSNP will search for heterozygous given SNPs, then extract the number of reads that contain either the reference or the alternate allele, and finally perform statistical tests (**Fig 2**).

3.1 Input SNP format

```
assnp_dir = .libPaths()
index_snp_file = paste0(assnp_dir, "/AlleleSNP/extdata/input_snps/input_snp_example2.
csv")
read.csv(index_snp_file, header = F)
```

```
##            V1  V2
## 1   rs1051730 EUR
## 2  rs10937405 ASN
## 3   rs8034191 EUR
## 4   rs8042374 EUR
## 5   rs9387478 ASN
## 6    rs402710 EUR
## 7 rs139852726 EUR
## 8   rs7741164 ASN
```

3.2 Input bam files

```
bam_dir = paste0(assnp_dir, "/AlleleSNP/extdata/sample/A549/bam_files")
list.files(bam_dir)
```

```
##  [1] "A549_H3K27ac.bam"      "A549_H3K27ac.bam.bai"  "A549_H3K4me1.bam"
##  [4] "A549_H3K4me1.bam.bai"  "A549_H3K4me3.bam"      "A549_H3K4me3.bam.bai"
##  [7] "A549_H3K9ac.bam"       "A549_H3K9ac.bam.bai"   "A549_PolII.bam"
## [10] "A549_PolII.bam.bai"
```

3.3 Get allele-specific binding events by examination of the bam files

```
get_assnp_byBam(index_snp_file = index_snp_file, bam_dir = bam_dir, sample_name = "A5
49_singleBam")
```

```
##          rsID   biofeature ref alt ref_rmdup alt_rmdup genotype_singleBam
## 1   rs3813570   A549_PolII  49  74        21        26               TRUE
## 2     rs31490 A549_H3K4me3  25  16        14        11               TRUE
## 3     rs27996 A549_H3K4me3  30  21        15        13               TRUE
## 4  rs57064725 A549_H3K27ac   9  15         7         9               TRUE
## 5  rs57064725 A549_H3K4me3  23  30        17        15               TRUE
## 6    rs684513 A549_H3K4me3  46  38        18        15               TRUE
## 7   rs3813570 A549_H3K4me3  54  51        20        22               TRUE
## 8    rs503464   A549_PolII  33  31        21        16               TRUE
## 9  rs59133824 A549_H3K4me3  78  78        29        28               TRUE
## 10 rs59683676 A549_H3K4me3  77  77        29        28               TRUE
##    genotype_sample genotype_vcf genotype_final biofeature_overlap_names
## 1               NA           NA           TRUE                       NA
## 2               NA           NA           TRUE                       NA
## 3               NA           NA           TRUE                       NA
## 4               NA           NA           TRUE                       NA
## 5               NA           NA           TRUE                       NA
## 6               NA           NA           TRUE                       NA
## 7               NA           NA           TRUE                       NA
## 8               NA           NA           TRUE                       NA
## 9               NA           NA           TRUE                       NA
## 10              NA           NA           TRUE                       NA
##    biofeature_overlap_num biofeature_overlap ref_count alt_count ref_cnv
## 1                      NA                 NA        NA        NA       1
## 2                      NA                 NA        NA        NA       1
## 3                      NA                 NA        NA        NA       1
## 4                      NA                 NA        NA        NA       1
## 5                      NA                 NA        NA        NA       1
## 6                      NA                 NA        NA        NA       1
## 7                      NA                 NA        NA        NA       1
## 8                      NA                 NA        NA        NA       1
## 9                      NA                 NA        NA        NA       1
## 10                     NA                 NA        NA        NA       1
##    alt_cnv  p.val.raw  p.val.cnv p.val.cnv.bh p.val.cnv.bonf
## 1        1 0.03004691 0.03004691    0.3004691      0.3004691
## 2        1 0.21102360 0.21102360    0.7420078      1.0000000
## 3        1 0.26243754 0.26243754    0.7420078      1.0000000
## 4        1 0.30745625 0.30745625    0.7420078      1.0000000
## 5        1 0.41010272 0.41010272    0.7420078      1.0000000
## 6        1 0.44520467 0.44520467    0.7420078      1.0000000
## 7        1 0.84537032 0.84537032    1.0000000      1.0000000
## 8        1 0.90065325 0.90065325    1.0000000      1.0000000
## 9        1 1.00000000 1.00000000    1.0000000      1.0000000
## 10       1 1.00000000 1.00000000    1.0000000      1.0000000
```

4. Identify ASE by sample

If you have more information of a sample, you could also identify ASE is through integrating multiple data types of a sample.

Here AlleleSNP provided a way to integrate bam files, peak files, and vcf files together to identify ASE. This mode may identify heterozygous SNPs that are not called using single bam file. But it also requires more prerequisite work, such as peak calling, vcf calling.

4.1 In the sample mode, we can incorporate more types of data, including:

- bam files
- peak files (.bed)
- vcf files

```
sample_dir = paste0(assnp_dir, "/AlleleSNP/extdata/sample/A549")
for (dir in list.files(sample_dir, full.names = T)) {
        cat (dir, "\n")
        for (file in list.files(dir)) {
                cat (file, "\n")
        }
}
```

```
##
/Library/Frameworks/R.framework/Versions/3.3/Resources/library/AlleleSNP/extdata/samp
le/A549/bam_files
## A549_H3K27ac.bam
## A549_H3K27ac.bam.bai
## A549_H3K4me1.bam
## A549_H3K4me1.bam.bai
## A549_H3K4me3.bam
## A549_H3K4me3.bam.bai
## A549_H3K9ac.bam
## A549_H3K9ac.bam.bai
## A549_PolII.bam
## A549_PolII.bam.bai
##
/Library/Frameworks/R.framework/Versions/3.3/Resources/library/AlleleSNP/extdata/samp
le/A549/peak_files
## A549_H3K27ac_peaks.bed
## A549_H3K4me1_peaks.bed
## A549_H3K4me3_peaks.bed
## A549_H3K9ac_peaks.bed
## A549_PolII_peaks.bed
##
/Library/Frameworks/R.framework/Versions/3.3/Resources/library/AlleleSNP/extdata/samp
le/A549/vcf_files
## A549_ChIPseq_GATK.vcf
## A549_ChIPseq_Samtools.vcf
## A549_WGS_GATK.vcf
```

4.2 Get allele-specific binding events by integrating all the data of the sample

```
get_assnp_bySample(index_snp_file = index_snp_file, sample_name = "A549", sample_dir
= sample_dir)
```

```
##          rsID   biofeature ref  alt ref_rmdup alt_rmdup genotype_singleBam
## 1   rs3813570   A549_PolII  49   74        21        26               TRUE
## 2  rs57064725 A549_H3K27ac   9   15         7         9               TRUE
## 3    rs684513 A549_H3K4me3  46   38        18        15               TRUE
## 4     rs27996 A549_H3K4me3  30   21        15        13               TRUE
## 5  rs57064725 A549_H3K4me3  23   30        17        15               TRUE
## 6     rs31490 A549_H3K4me3  25   16        14        11               TRUE
## 7   rs3813570 A549_H3K4me3  54   51        20        22               TRUE
## 8    rs503464   A549_PolII  33   31        21        16               TRUE
## 9  rs59133824 A549_H3K4me3  78   78        29        28               TRUE
## 10 rs59683676 A549_H3K4me3  77   77        29        28               TRUE
##    genotype_sample genotype_vcf genotype_final biofeature_overlap
## 1             TRUE         TRUE           TRUE               TRUE
## 2             TRUE         TRUE           TRUE               TRUE
## 3             TRUE         TRUE           TRUE               TRUE
## 4             TRUE         TRUE           TRUE               TRUE
## 5             TRUE         TRUE           TRUE               TRUE
## 6             TRUE         TRUE           TRUE               TRUE
## 7             TRUE         TRUE           TRUE               TRUE
## 8             TRUE         TRUE           TRUE               TRUE
## 9             TRUE         TRUE           TRUE               TRUE
## 10            TRUE         TRUE           TRUE               TRUE
##    biofeature_overlap_num
## 1                       4
## 2                       4
## 3                       4
## 4                       3
## 5                       4
## 6                       4
## 7                       4
## 8                       4
## 9                       4
## 10                      4
##                                                  biofeature_overlap_names
## 1    A549_H3K27ac_peaks,A549_H3K4me3_peaks,A549_H3K9ac_peaks,A549_PolII_peaks
## 2    A549_H3K27ac_peaks,A549_H3K4me3_peaks,A549_H3K9ac_peaks,A549_PolII_peaks
## 3    A549_H3K27ac_peaks,A549_H3K4me3_peaks,A549_H3K9ac_peaks,A549_PolII_peaks
## 4                     A549_H3K27ac_peaks,A549_H3K4me3_peaks,A549_H3K9ac_peaks
## 5    A549_H3K27ac_peaks,A549_H3K4me3_peaks,A549_H3K9ac_peaks,A549_PolII_peaks
## 6  A549_H3K27ac_peaks,A549_H3K4me1_peaks,A549_H3K4me3_peaks,A549_H3K9ac_peaks
## 7    A549_H3K27ac_peaks,A549_H3K4me3_peaks,A549_H3K9ac_peaks,A549_PolII_peaks
## 8    A549_H3K27ac_peaks,A549_H3K4me3_peaks,A549_H3K9ac_peaks,A549_PolII_peaks
## 9    A549_H3K27ac_peaks,A549_H3K4me3_peaks,A549_H3K9ac_peaks,A549_PolII_peaks
## 10   A549_H3K27ac_peaks,A549_H3K4me3_peaks,A549_H3K9ac_peaks,A549_PolII_peaks
##    ref_count alt_count ref_cnv alt_cnv  p.val.raw  p.val.cnv p.val.cnv.bh
## 1         48        36    1501    1558 0.03004691 0.05533592    0.5533592
## 2         29        39    1501    1558 0.30745625 0.35532555    0.8862402
## 3         45        46    1501    1558 0.44520467 0.35803028    0.8862402
## 4         27        15    1754     929 0.26243754 0.40944408    0.8862402
## 5         29        39    1501    1558 0.41010272 0.49624781    0.8862402
## 6         30         8    1754     929 0.21102360 0.67303236    0.8862402
## 7         48        36    1501    1558 0.84537032 0.70660674    0.8862402
## 8         24        38    1501    1558 0.90065325 0.78816140    0.8862402
## 9         42        55    1501    1558 1.00000000 0.88474339    0.8862402
## 10        40        53    1501    1558 1.00000000 0.88624024    0.8862402
##    p.val.cnv.bonf
## 1       0.5533592
## 2       1.0000000
## 3       1.0000000
## 4       1.0000000
## 5       1.0000000
## 6       1.0000000
## 7       1.0000000
## 8       1.0000000
## 9       1.0000000
## 10      1.0000000
```