ASE: allele-specific expression

Outline
1. Definition of ASE
2. Detecting ASE (introductory case)
3. Applications and prevalence of ASE
4. Important ASE considerations
   (a) Variant calling
   (b) Mapping bias
   (c) Many variants in a gene
5. ASE tools
6. GeneiASE – a tool to detect genes with ASE from RNA-seq data

[1] Definition of allele-specific expression (ASE)

Adding another layer to transcriptome complexity...

One gene can produce many different transcripts...

Adding another layer to transcriptome complexity...

...and each gene is present on two chromosomes.
=> it has two alleles

Allele, definition

An allele is the variant form of a given gene (or locus). Sometimes, different alleles can result in different observable phenotypic traits, such as different pigmentation. /.../

If both alleles at a gene (or locus) on the homologous chromosomes are the same, they and the organism are homozygous with respect to that gene (or locus). If the alleles are different, they and the organism are heterozygous with respect to that gene (or locus).

https://en.wikipedia.org/wiki/Allele
Allele-specific expression, definition

An imbalance in transcription between the maternal and paternal alleles at a locus.
- I.e., a deviation from the expected 50/50 ratio of transcription from the two alleles of a diploid organism.
- Can be assessed within a single individual (Present also when ploidy >2, e.g., plants)

Other events may also be “allele-specific”, e.g.
- transcription factor binding
- DNA backbone methylation
- X-chromosome inactivation in female mammals

[2] Detecting ASE

Detecting allele-specific expression using RNA-seq data

- RNA-seq reads provide the sequence of a transcript
- ... which enables the determination of the allelic origin of the reads overlapping with the SNV

Detecting allele-specific expression using RNA-seq data

General outline:
1. Map the RNA-seq reads
2. Count the reads that map to either allele
3. Calculate effect size and p-value

Detecting allele-specific expression, definition

genomic DNA -> transcript (e.g. mRNA)
Detected allele-specific expression using RNA-seq data

1. Map the RNA-seq reads

<table>
<thead>
<tr>
<th>Paternal allele (a)</th>
<th>Maternal allele (A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGTTCTTTAATTAGC...</td>
<td>AGTTCTTTAATTAGC...</td>
</tr>
<tr>
<td>...AGTCTTC...C...AATTAGC...</td>
<td></td>
</tr>
</tbody>
</table>

Reads – 10x coverage of the locus

<table>
<thead>
<tr>
<th>Paternal allele (a)</th>
<th>Maternal allele (A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGTTCTTTAATTAGC...</td>
<td>AGTTCTTTAATTAGC...</td>
</tr>
<tr>
<td>...AGTCTTC...T...AATTAGC...</td>
<td></td>
</tr>
</tbody>
</table>

Mapped reads

<table>
<thead>
<tr>
<th>Paternal allele (a)</th>
<th>Maternal allele (A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGTTCTTTAATTAGC...</td>
<td>AGTTCTTTAATTAGC...</td>
</tr>
<tr>
<td>...AGTCTTC...T...AATTAGC...</td>
<td></td>
</tr>
</tbody>
</table>

Reads – 10x coverage of the locus

2. Count the reads

<table>
<thead>
<tr>
<th>Paternal allele (a)</th>
<th>Maternal allele (A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGTTCTTTAATTAGC...</td>
<td>AGTTCTTTAATTAGC...</td>
</tr>
<tr>
<td>...AGTCTTC...C...AATTAGC...</td>
<td></td>
</tr>
</tbody>
</table>

3x AGTTCTTTAATTAGC...

7x AGTTCTTTAATTAGC...

3 reads mapped to paternal allele

7 reads mapped to maternal allele

In total 10 reads mapped to the locus

3. Calculate effect size and p-value

Effect size: (other definitions possible)

\[ \text{ASE}_{\text{effect}} = \frac{c_{\text{alt}}}{c_{\text{alt}} + c_{\text{ref}}} - 0.5 \]

i.e., the fraction of counts mapped to alternative allele minus 0.5 =>

- if no ASE then \( \text{ASE}_{\text{effect}} = 0 \)
- range of \( \text{ASE}_{\text{effect}} \) is \([-0.5, 0.5]\)

P-value: Use binomial with \( p=0.5 \) (assuming 50/50 transcription)

Our example from previous slide:

\[ \text{Effect size } = -0.2 \]

\[ \text{P-value: } \text{binomial test for deviation from 50/50 distribution between alleles (in R):} \]

\[ > \text{pbetbinom}(3, \text{size}=10, \text{prob}=0.5) \]

\[ > 0.171875 \]

⇒ Not significant in this particular example

⇒ If coverage was 30x (9+21 reads) instead of 10x (3+7), then p-value < 0.03

eQTL vs. ASE

**eQTL**

- Inter-individual differences in expression
- Modest effects
- Large number of SNP-gene combinations
- Many samples needed
- May use microarrays for gene expression
- Genotyping required

**ASE**

- Sufficient power with a single individual
- Identical cellular environment for the two chromosomes
- No association to regulatory region
- Must use RNA-seq for gene expression


10 individuals genotyped
To infer a changed protein, the SNV must be
- in coding region
- non-synonymous

Possible to detect if you also have information about non-transcribed variants (e.g., from whole-genome DNA sequencing or SNP-array).

Possible to detect if you have expression measured from both normal and tumor tissue (in the same individual).

Genes with significant ASE (% of genes with heterozygous variant).

(a) Variant detection
(b) Mapping bias
(c) Many variants in a gene
[4] Important ASE considerations: (a) Variant detection

Variant detection

Variant = a position in the genome that is different from another genome.

- Homozygous variant: the two alleles are identical to each other
- Heterozygous variant: the two alleles are different
- "Ref." = the allele is the same as for the reference genome
- "Alt." = alternate = the allele is different from the reference genome
- SNV is one type of variant, others include insertion, deletion, ...

Variant detection = detecting what variants are present in a sample:

1. Variant calling – any position with evidence of an alternative base
2. Variant prioritization – define reliable variants with high confidence

Typically performed based on genomic DNA data, from

- Microarrays (e.g. Illumina Omni 2.5M)
- Sequencing (e.g. whole-genome re-sequencing or exome sequencing)

Variant detection from sequencing data

Start by map the reads.

<table>
<thead>
<tr>
<th>Paternal allele (a)</th>
<th>Maternal allele (A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>...AGTCTTC3ATTAGC...</td>
<td>...AGTCTTC3ATTAGC...</td>
</tr>
</tbody>
</table>

Reads ~ 10x coverage of the locus

...AGCTTTCTAATTAGC...
...AGCTTTCTAATTAGC...
...AGCTTTCTAATTAGC...
...AGCTTTCTAATTAGC...
...AGCTTTCTAATTAGC...
...AGCTTTCTAATTAGC...
...AGCTTTCTAATTAGC...
...AGCTTTCTAATTAGC...

Variant detection from sequencing data

OK, piece of cake?

<table>
<thead>
<tr>
<th>Paternal allele (a)</th>
<th>Maternal allele (A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>...AGTCTTC3ATTAGC...</td>
<td>...AGTCTTC3ATTAGC...</td>
</tr>
</tbody>
</table>

Mapped reads

...AGCTTTCTAATTAGC...
...AGCTTTCTAATTAGC...
...AGCTTTCTAATTAGC...
...AGCTTTCTAATTAGC...
...AGCTTTCTAATTAGC...
...AGCTTTCTAATTAGC...
...AGCTTTCTAATTAGC...
...AGCTTTCTAATTAGC...

Variant detection from sequencing data

This is what we actually have:

<table>
<thead>
<tr>
<th>Reference sequence</th>
<th>...AGTCTTC3ATTAGC...</th>
</tr>
</thead>
</table>

Mapped reads

...AGCTTTCTAATTAGC...
...AGCTTTCTAATTAGC...
...AGCTTTCTAATTAGC...
...AGCTTTCTAATTAGC...
...AGCTTTCTAATTAGC...
...AGCTTTCTAATTAGC...
...AGCTTTCTAATTAGC...
...AGCTTTCTAATTAGC...

=> need to detect the variant positions in the reference sequence

Variant detection from sequencing data

Standard: GATK (DePristo et al., 2011) or Samtools – works on any mapped sequencing data.

GATK scores the SNVs by taking into account a number of characteristics, including:

- Sequencing depth (coverage)
- Mapping quality
- Position bias (base quality)

Specific RNA-seq based tools:

- Colliread – Le Bras et al., 2016
- RVboost – Wang et al., 2014
- ACCUSA2 – Picchiotta et al., 2013

GATK the most widely used, even for RNA-seq.
Variant detection – VCF, Variant Call Format

VCF is a text file format ("flat text"). Example VCF output from GATK:

```
##fileformat=VCFv4.1
...
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1_NA12878 [SAMPLE1_BLALBA] ...
1 873762 . T C 5231.78 PASS [ANNOTATIONS] GT:AD:DP:GQ:PL 0/1:173,141:282:99:255,0,255 ...
1 899282 rs28548431 C T 71.77 PASS [ANNOTATIONS] GT:AD:DP:GQ:PL 0/1:1,3:4:26:103,0,26 ...
1 974165 rs9442391 T C 29.84 LowQual [ANNOTATIONS] GT:AD:DP:GQ:PL 0/1:14,4:14:61:61,0,255...
...
```

| GT: the genotype of this sample at this site (0/0, 0/1, 1/1, 1/2, ...), 0=ref., 1=alt. |
| AD: allele depths, i.e., the number of reads that support each of the reported alleles |
| GQ: quality of assigned genotype (max=99) |

Full specification of VCF file format: http://samtools.github.io/hts-specs/

Variant detection – which variants to use (prioritization)?

- What sequencing depth?
- Heterozygous
- Other criteria

Filtering of known variants:
- Keep only variants in dbsNP?

Mapping bias

Reference genome variants ("ref.") have an advantage in the mapping.

Maternal allele: `ATCGAATGAAGCTCATTGGATCAGAT` (ref.)

Paternal allele: `ATCGAATGAAGCTATTGGATCAGAT` (alt.)

Reference: `ATCGAATGAAGCTATTGGATCAGAT`

The paternal allele read will map with a lower mapping quality. In case of sequencing error or poor base quality at another position, this might push the mapping quality of the paternal allele read below the threshold, and the read will be discarded.

Mapping bias example in real data

Heterozygous variants (alt/ref) mapped to reference genome.

- X-axis, alternate allele fraction [0, 1]
- Y-axis, density

(Data from 16 RNA-seq experiments; Variants detected with RNA-seq data or SNP array.)

Mapping bias – ways to get around it in ASE detection

- Masking variants ([A,C,G,T]=N)
  - You lose information.
- Construct all possible versions of the genome from existing variants
  - Can soon generate a prohibitive amount of genome versions.
- Map reads to diploid genome (or transcriptome)
  - Requires that you either have or construct the diploid genome (or transcriptome) of the individual.
- Modify the binomial probability p to reflect the mapping bias
  - Requires simulation to properly modify p.
Mapping bias – ways to get around it in ASE detection

Masking variants ([A,C,G,T]xN)
You lose information.

Construct all possible versions of the genome from existing variants
Can soon generate a prohibitive amount of genome versions.

Map reads to diploid genome (or transcriptome)
Requires that you either have or construct the diploid genome (or transcriptome) of the individual. E.g., Turro et al. (2011) (transcriptome).

Modify the binomial probability p to reflect the mapping bias.
Requires simulation to properly modify p. E.g., Montgomery et al. (2010).

[4] Important ASE considerations:
(c) Many variants in a gene

Many variants in a gene
More than one variant within a gene is common:

Diploid genome

Allele-specific gene expression (mRNAs)

2 different "haploisoforms"

RNA-seq reads

Possible combinations (haplotypes) from RNA-seq reads:
• A + T and G + C
• A + C and G + T

But: RNA-seq does not necessarily capture the relation between the two SNVs.

Possible combinations (haplotypes) from RNA-seq reads:
• A + T and G + C
• A + C and G + T

But can our RNA-seq reads provide the phase?

Many variants in a gene – phasing

Phasing = deciding which alleles that are on the same chromosomal homologue

RNA-seq contains information that there are two heterozygous SNVs.

Variants detected from RNA-seq reads:
• A/G at one locus
• T/C at one locus

Many variants in a gene – phasing

Phasing = deciding which alleles that are on the same chromosomal homologue

RNA-seq contains information that there are two heterozygous SNVs.

Variants detected from RNA-seq reads:
• A/G at one locus
• T/C at one locus

Many variants in a gene

RNA-seq contains information that there are two heterozygous SNVs.

Variants detected from RNA-seq reads:
• A/G at one locus
• T/C at one locus

Many variants in a gene

RNA-seq contains information that there are two heterozygous SNVs.

Variants detected from RNA-seq reads:
• A/G at one locus
• T/C at one locus
Phasing is useful but not necessary to detect ASE

Phasing information typically achieved by sequencing the genomes of the parents of the subject. Direct haplotype sequencing also possible.

If you don’t know the phase (and for most RNA-seq data sets, you don’t):
- Try to infer it from (a) your RNA-seq data – possible, but typically only partial phasing
- (b) existing population data (LD) – not applicable on new variants
- Disregard from it and calculate ASE anyway

Phasing
- reduces mapping bias
- enables the detection of haploloisoform expression (isoforms representing the two homologous chromosomes)
- but is not necessary to detect ASE in genes with >1 SNV

ASE tools

A list of tools that can detect ASE, given specified input data:
- cisASE – paired genomic+transcriptomic data, Liu et al., 2016
- MutRSeq – nonsynonomous SNVs from RNA-seq data, Fu et al., 2016
- GeneiASE – unphased RNA-seq data, Edsgård et al., 2016
- ASE-TIGAR – parental data required, bayesian, Nariai et al., 2016
- ASEQ – paired genomic+transcriptomic data, Romanel et al., 2015
- MBASED – phased or unphased RNA-seq data, Mayba et al., 2015
- Allim – parental data required, Pandey et al., 2013
- MMSEQ – attempts haploloisoform identification, Turro et al., 2011
- Stelly) – requires phased data, Skelly et al., 2011
- AlleleSeq – requires genomic sequence, Rozowsky et al., 2011
- AlleleDB – database for ASE etc. of 1000genomes, Chen et al., 2016

ASE tools – where only RNA-seq data from a single individual is required.
- cisASE – paired genomic+transcriptomic data, Liu et al., 2016
- MutRSeq – nonsynonomous SNVs from RNA-seq data, Fu et al., 2016
- GeneiASE – unphased RNA-seq data, Edsgård et al., 2016
- ASE-TIGAR – parental data required, bayesian, Nariai et al., 2016
- ASEQ – paired genomic+transcriptomic data, Romanel et al., 2015
- MBASED – phased or unphased RNA-seq data, Mayba et al., 2015
- Allim – parental data required, Pandey et al., 2013
- MMSEQ – attempts haploloisoform identification, Turro et al., 2011
- AlleleSeq – requires phased data, Skelly et al., 2011
- AlleleDB – database for ASE etc. of 1000genomes, Chen et al., 2016

GeneiASE

GeneiASE detects genes with significant ASE, in single individuals and based only on RNA-seq data. Haplotype information (phasing) is not needed.

Data required:
- RNA-seq data

Pre-processing required:
- Mapping and quality control of reads
- Variant detection (e.g., GATK)
- Filter variants if desired
- Allele counts for variants extracted into custom input text file

Availability:
- https://github.com/edsgard/geneiase [GNU GPL3 license]
GeneiASE

The situation:
- unphased data
- non-uniform effect within gene
- technical variability

The GeneiASE solution:
1. For each gene, loop over all its SNVs and their 2x1 matrix of read counts
2. Calculate a test statistic (s) for each SNV, based on read counts
3. Combine the test statistics for the SNVs within a gene to test statistic for entire gene (g)
4. Resample from parametric null SNV model (estimated from DNA data) 10^5 times, calculate the resulting distribution of gene test statistic (g^0).
5. Compare g to g^0 and calculate a p-value for gene i.

GeneiASE – calculate SNP-based gene-wise test statistic

1. Count reads for each SNV in a gene; add pseudo counts if required
2. Calculate SNV test statistic s, based on absolute value of effect size, eff.
3. Calculate gene test statistic g using Stouffer-Liptak method; k is number of SNVs in gene
4. Resample from parametric null SNV model (estimated from DNA data) 10^5 times, calculate the resulting distribution of gene test statistic (g^0).
5. Compare g to g^0 and calculate a p-value for gene i.

GeneiASE – null model, and gene-wise p-value calculation

1. Estimate SNV null model parameters
2. DNA based estimate of the technical variability
3. For each gene (gene i):
   1. Sample allele counts from null SNV model (Random effect model)
   2. Calculate k SNV test statistic s, based on absolute value of effect size
   3. Calculate gene test statistic g using Stouffer-Liptak method
   4. Repeat 1-3 k times (default: 100)
   5. Calculate p-value for gene i

Running GeneiASE – input

Static ASE
- geneiase -cvt -i cvt.test.input.tab
- geneiase -icd -i icd.test.input.tab

Condition-dependent ASE
- geneiase -cd -i cd.test.input.tab

Running GeneiASE – results

The number of genes with significant (FDR<0.05) ASE as detected by GeneiASE from 16 RNA-seq samples (primary white blood cells).
Thank you for your attention

contact: olofem@kth.se