# HGSVC Functional Analysis
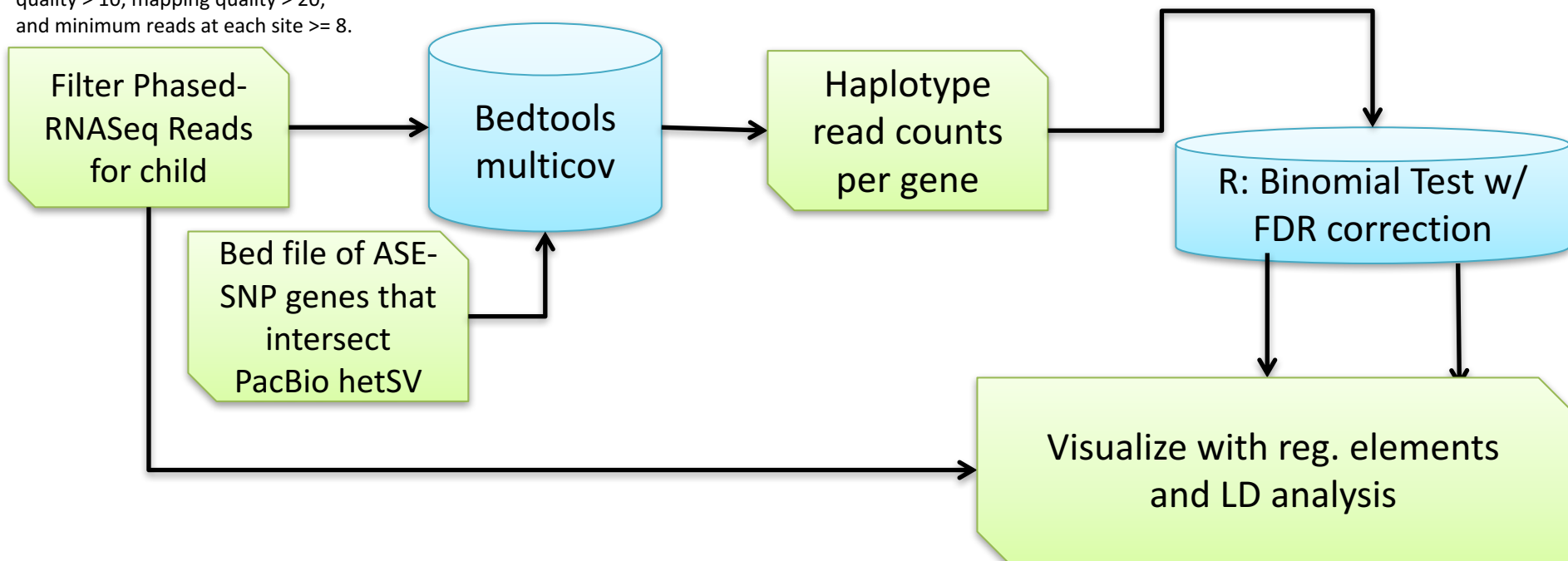
Conor Nodzak, Jia Wen, Mindy Shi

UNC Charlotte

May 1, 2017

# SV ASE Analysis

Quality filtering: only use uniquely mapping reads with NM <= 6, base quality > 10, mapping quality > 20, and minimum reads at each site >= 8.

Filter Phased-RNASeq Reads for child

Bedtools multicov

Haplotype read counts per gene

Bed file of ASE-SNP genes that intersect PacBio hetSV

R: Binomial Test w/ FDR correction

Visualize with reg. elements and LD analysis

(*FDR at 5% )

| Trio Child | SNP ASE Genes Tested | SV ASE Genes | ASE SVs | ASE INS | ASE INS Exon / Intron | ASE DEL | ASE DEL Exon / Intron | ASE INV | ASE INV Exon / Intron |
|---|---|---|---|---|---|---|---|---|---|
| HG00514 | 250 | 91 | 94 | 54 | | 40 | | | |
| HG00733 | 302 | 101 | 97 | 55 | | 42 | | | |
| NA19240 | 227 | 80 | 85 | 41 | | 44 | | | |

# Gene Intersected with SV Exhibits ASE
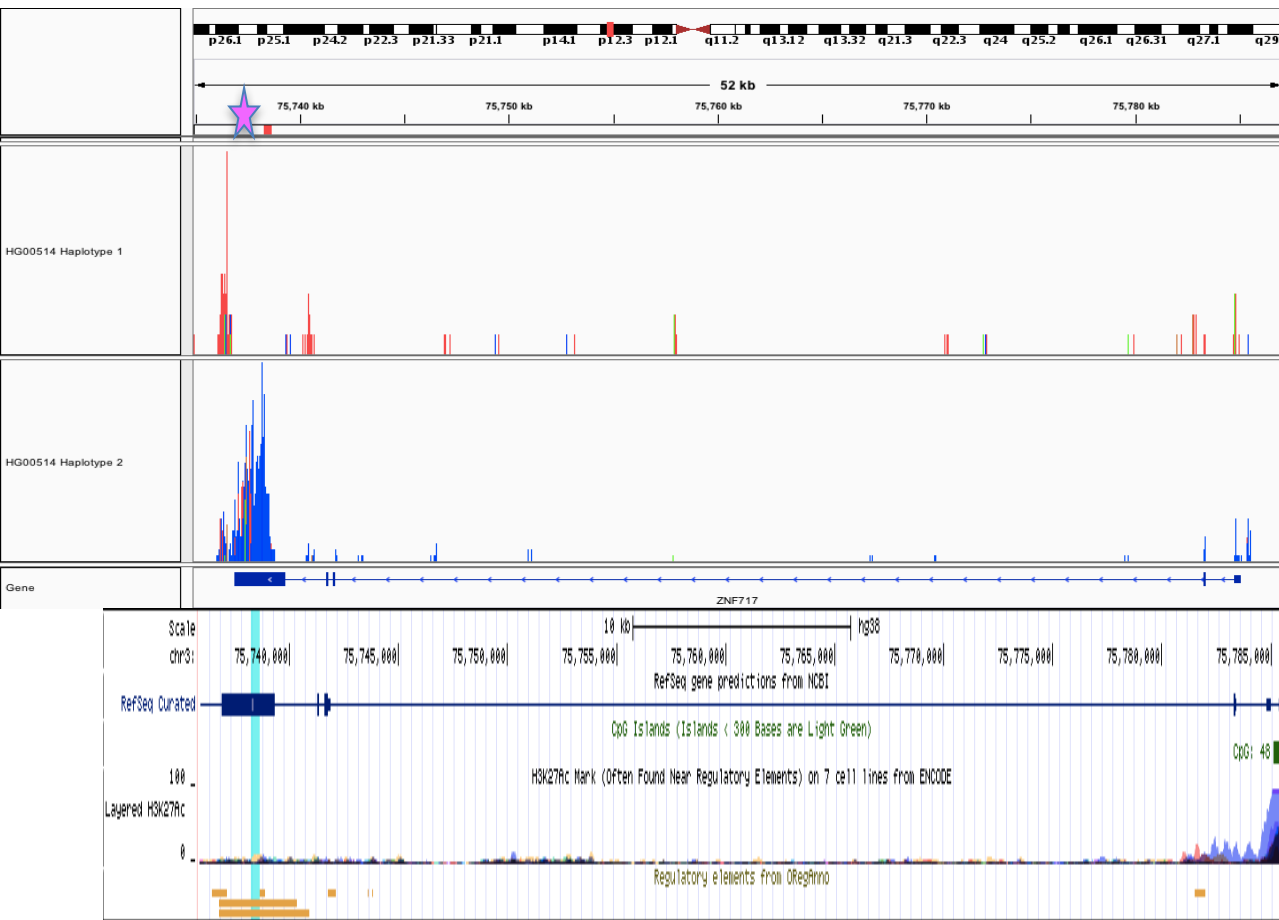


IGV plot of coverage for phased RNASeq reads for gene intersecting hetSV in HG00514.
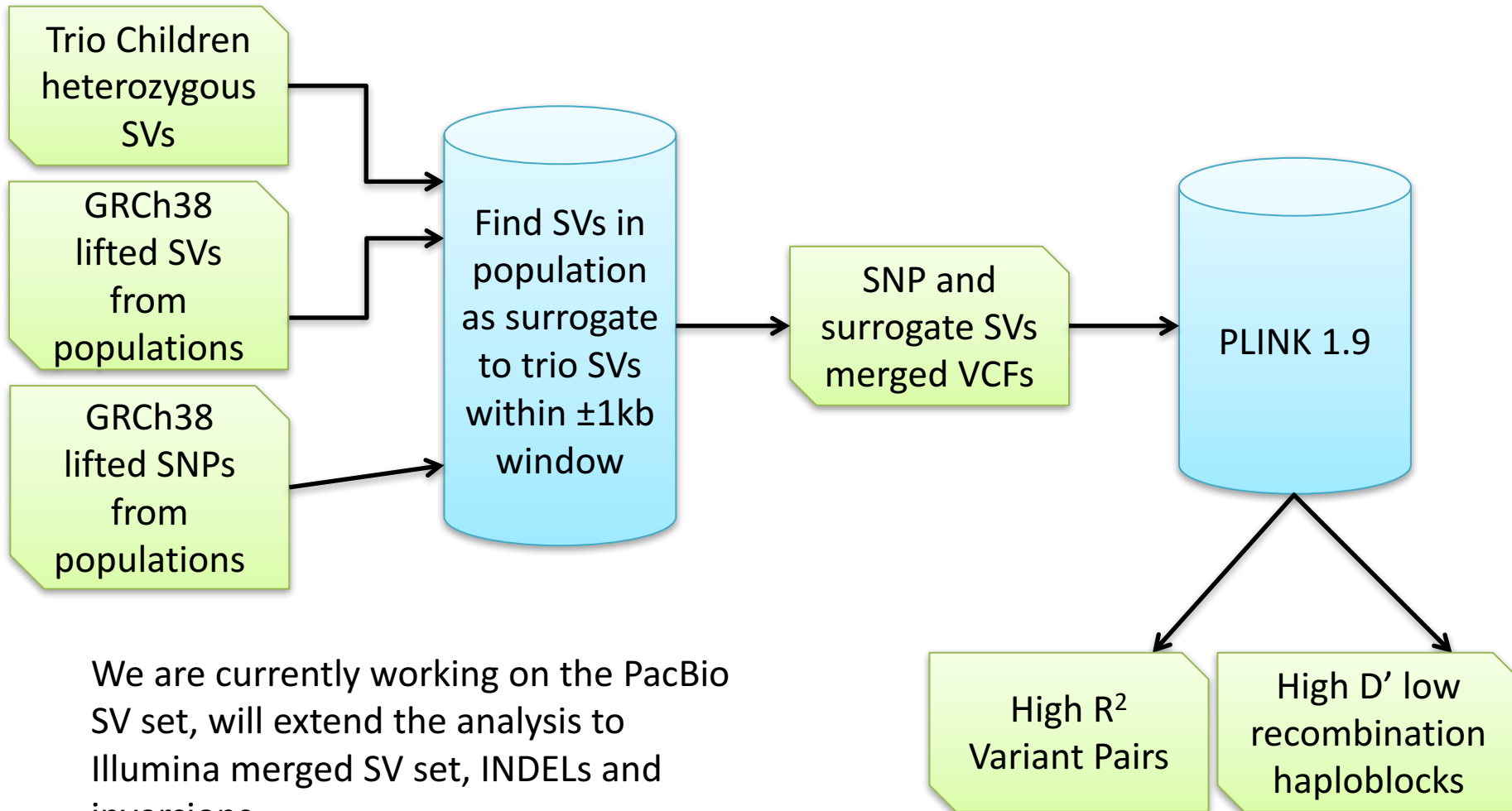
Haplotype 1 reads in red.
Haplotype 2 reads in blue.

Regulatory elements tracks For ZNF717 shows multiple transcription factors binding sites (orange bars) overlap with deletion.

Blue highlight denotes the location of deletion (typed (1|0), 422 bps) at exon 5 of ZNF717. Read count gene level ratio: 184:687, adjusted p-value = 1.683e-68.
Denotes location of ASE SNPs, 641bp away from SV.

# LD and Haploblock Workflow

Trio Children heterozygous SVs

GRCh38 lifted SVs from populations

GRCh38 lifted SNPs from populations

Find SVs in population as surrogate to trio SVs within ±1kb window

SNP and surrogate SVs merged VCFs

PLINK 1.9

High $R^2$ Variant Pairs

High D' low recombination haploblocks

We are currently working on the PacBio SV set, will extend the analysis to Illumina merged SV set, INDELs and inversions.
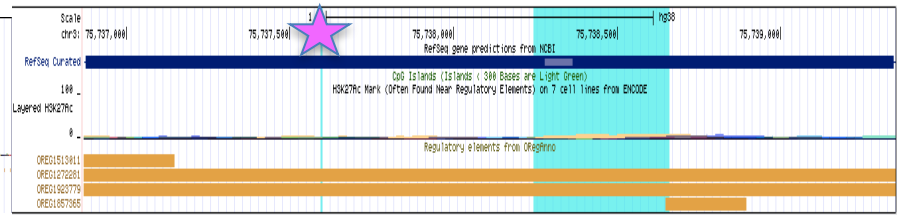* Genotyping SVs and INDELs later will remove the need of surrogate SVs

# Does ASE Extend to Haploblocks?



R² Color Key
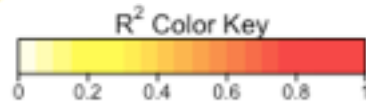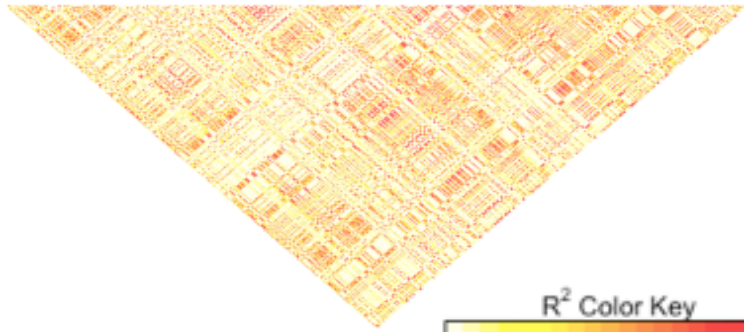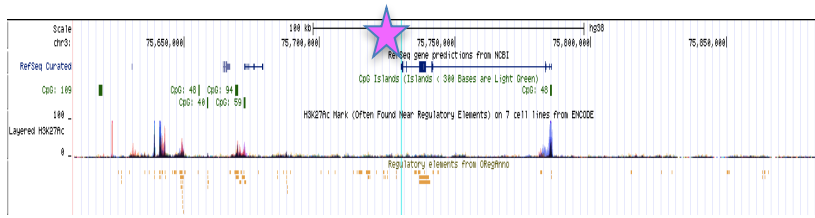
R² heatmap for CHS variants
± 100kb of ZNF717.

Lack of strong LD pattern would
suggest the ASE effect is not the
result of a haplotype effect.

R² heatmap for exon 5 of ZNF717
The wider blue highlight shows
coordinates of 422bp deletion.

⭐ shows position of ASE-SNP

Population variants show high LD surrounding
the location of ASE variants, less so for the
rest of the region.

There might be some local LD structure but
not informative yet as the ASE SNP and SV
don't have surrogate in the populations.

5

# Imprinted Genes

- The imprinted gene database
  - http://geneimprint.com/site/genes-by-species
- 13 imprinted genes are found to overlap with SNP ASE genes
  - ZDBF2
  - PLAGL1
  - MEST
  - SNURF
  - **SNRPN**
  - **NAA60**
  - CPA4
  - **AIM1**
  - ZFAT
  - **RB1**
  - GNAS
  - SGCE
  - PEG10
- 5 genes in YRI and CHS trios show that ASE pattern is consistent with imprinted pattern.
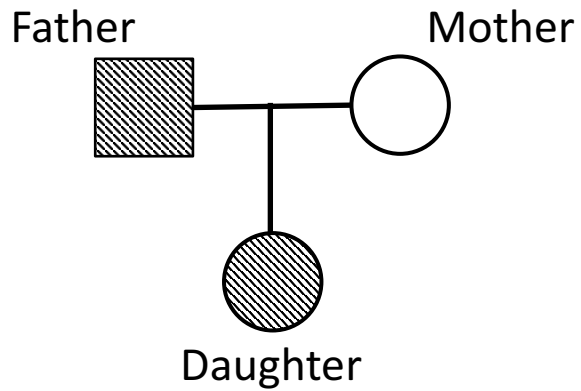- 4 ASE imprinted genes are SV ASE genes.

Five genes in YRI and CHS trios show that ASE pattern is consistent with imprinted gene pattern.
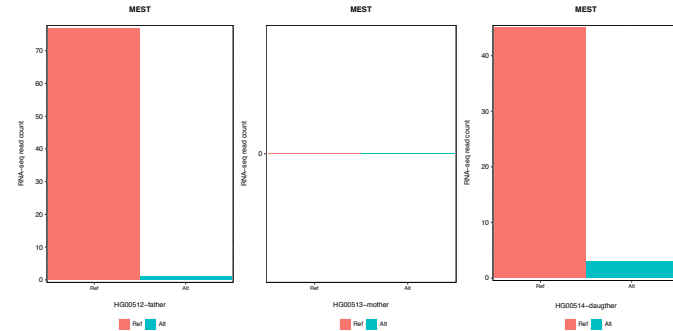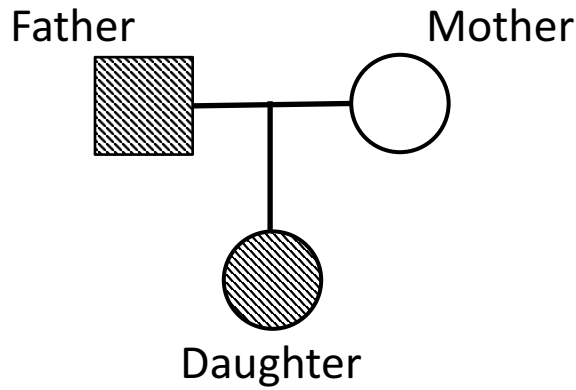


MEST, a paternal allele expressed gene, shows inherited trend in CHS trio in RNAseq data.
The shade denotes the imprinted pattern.
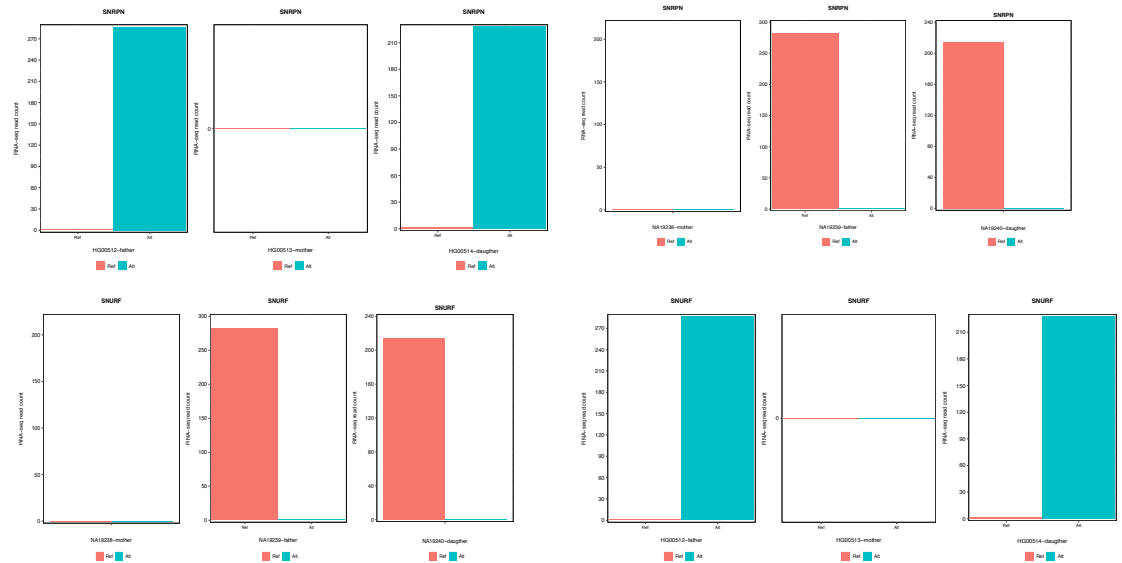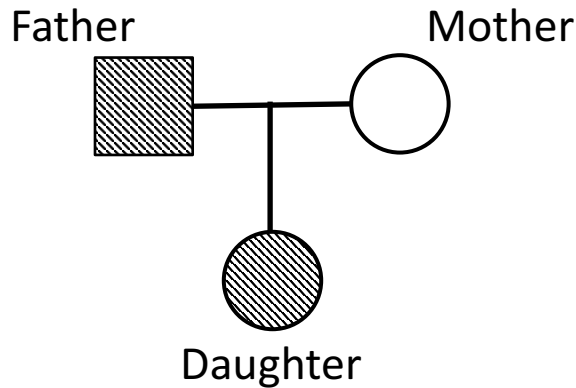


SNURF and **SNRPN**, paternal allele expressed genes, show inherited trend in YRI and CHS trio in RNAseq data.
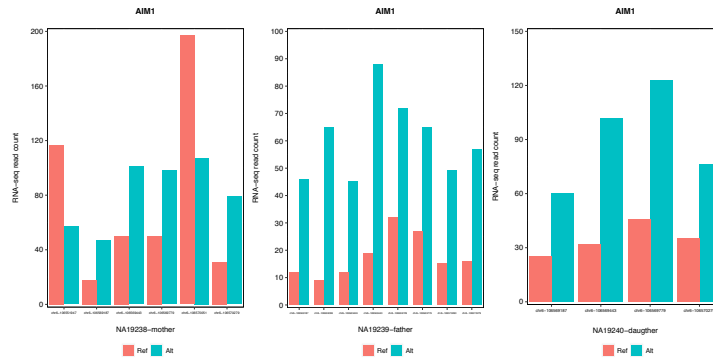
**AIM1**, a paternal allele expressed gene, shows the inherited trend in YRI trio RNAseq data.
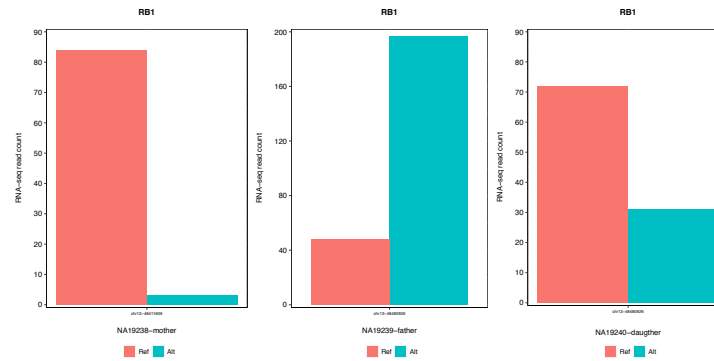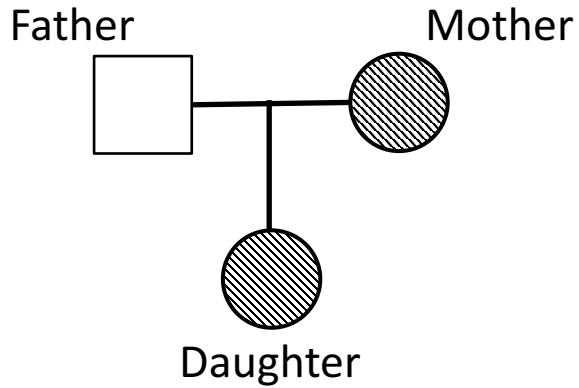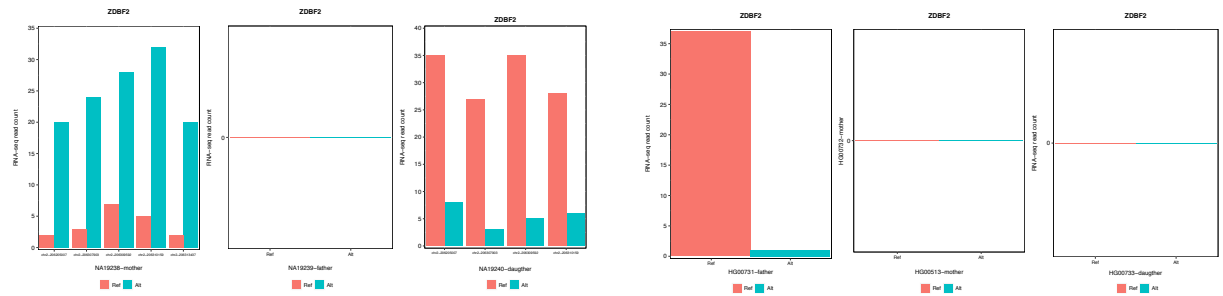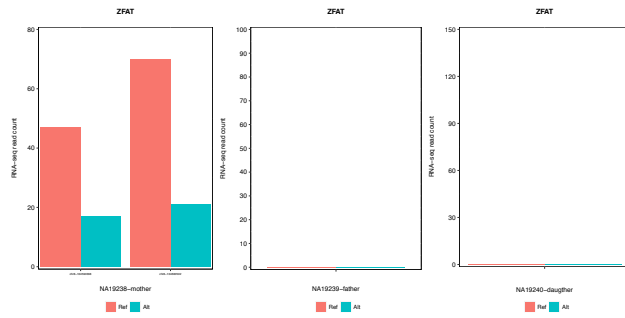


**RB1**, a maternal allele expressed gene, shows the inherited trend in YRI trio RNAseq data.

Other genes show that ASE pattern is inconsistent with imprinted gene pattern.
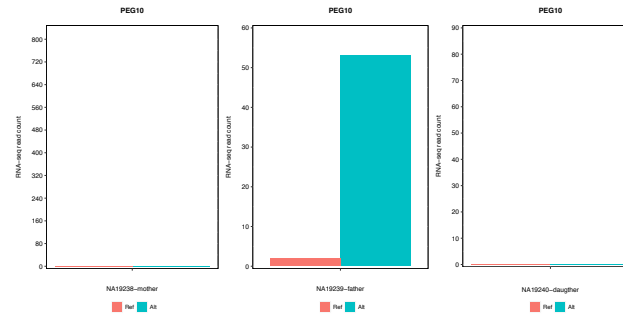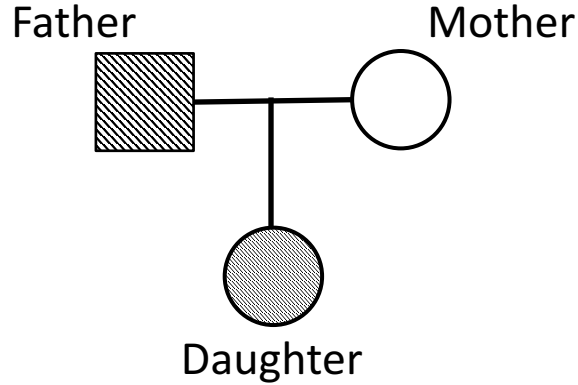


PEG10, a paternal allele expressed gene, is ASE gene in father, but not in mother or child for YRI trio.
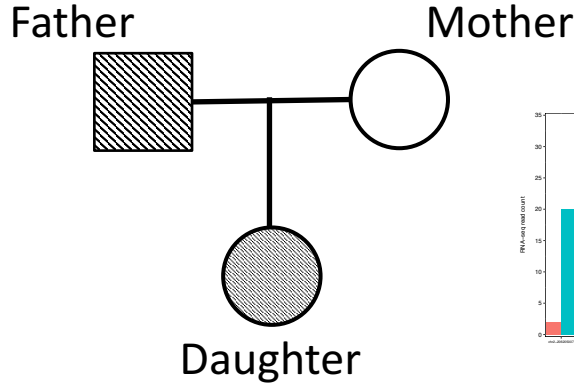


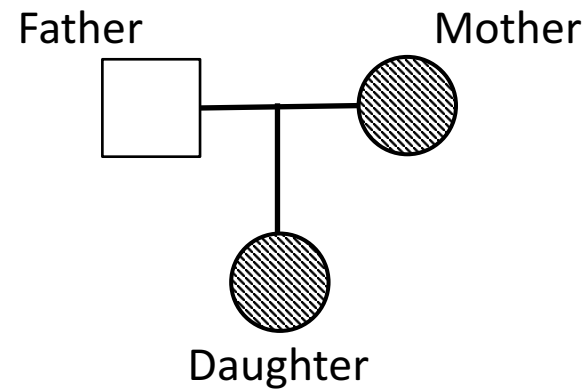ZDBF2, a paternal allele expressed gene, is ASE gene in mother and child, but not in father for YRI trio.
ZDBF2 is ASE gene in father, but not in mother or child for PUR trio.

Father · Mother
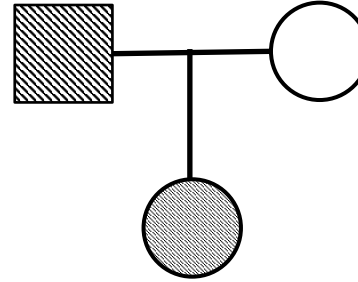
ZFAT, a paternal allele expressed gene, and GNAS, a isoform dependent allele expressed gene, are ASE genes in mother, but not in father or child for YRI trio.
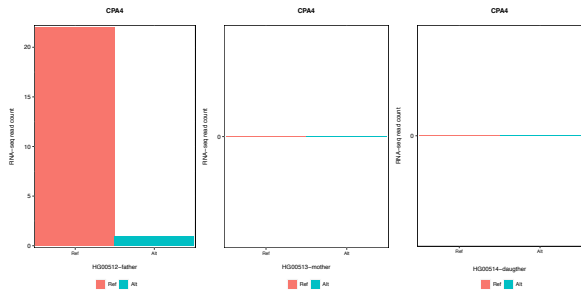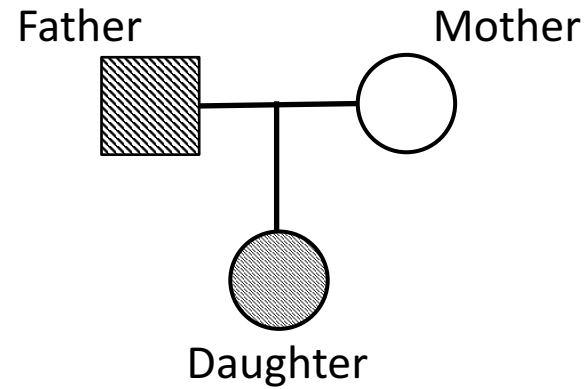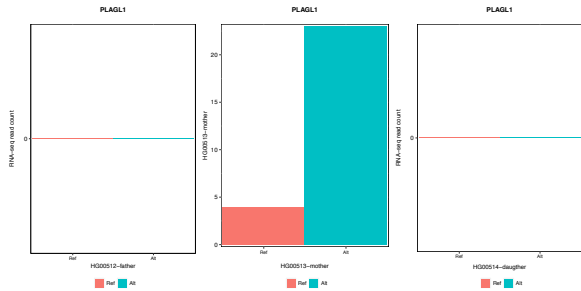
Daughter

Father · Mother
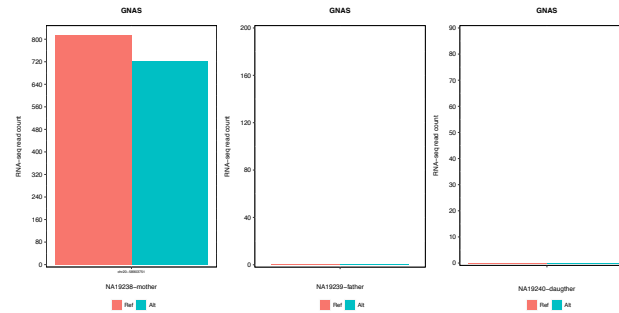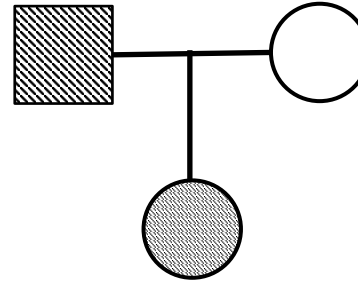
Daughter

SGCE, a paternal allele expressed gene, is ASE gene in father, but not in mother or child for YRI trio.
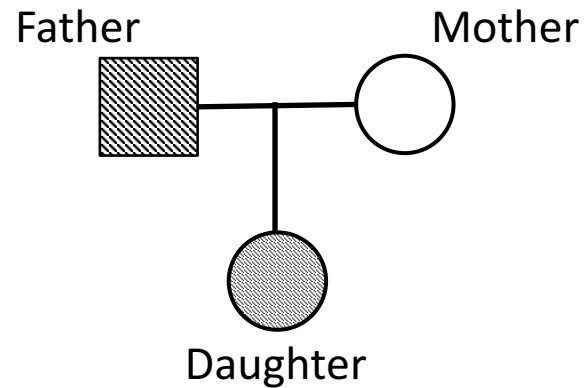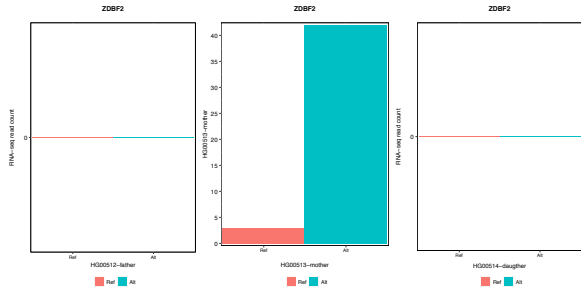
Daughter
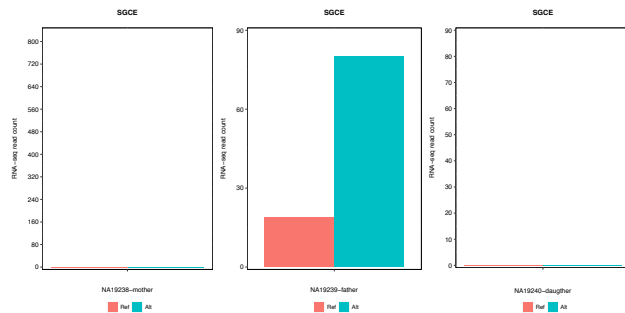
- CPA4, a maternal allele expressed gene, is only ASE gene in father, but not for mother and child in YRI trio.



- PLAGL1, a paternal allele expressed gene, is only ASE gene in mother, but not for father and child in YRI trio.



- ZDBF2, a paternal allele expressed gene, is only ASE gene in mother, but not for father and child in YRI trio.

# Investigation on the HG00731 SNP ASE
## (too few ASE SNPs compared with other samples)

# RNAseq Bam Files

| | | |
|---|---|---|
| SRS000212.gsnap_GRCh38Primary.20150922.YRI.mRNA.bam | 2015-10-22 04:36 | 5.9G |
| SRS000212.gsnap_GRCh38Primary.20150922.YRI.mRNA.bam.bai | 2015-10-22 07:06 | 6.6M |
| SRS000213.gsnap_GRCh38Primary.20150922.YRI.mRNA.bam | 2015-10-22 04:45 | 6.6G |
| SRS000213.gsnap_GRCh38Primary.20150922.YRI.mRNA.bam.bai | 2015-10-22 07:06 | 6.7M |
| SRS000214.gsnap_GRCh38Primary.20150922.YRI.mRNA.bam | 2015-10-22 06:17 | 5.3G |
| SRS000214.gsnap_GRCh38Primary.20150922.YRI.mRNA.bam.bai | 2015-10-22 07:06 | 6.6M |
| SRS008631.gsnap_GRCh38Primary.20150922.CHS.mRNA.bam | 2015-10-22 06:23 | 6.0G |
| SRS008631.gsnap_GRCh38Primary.20150922.CHS.mRNA.bam.bai | 2015-10-22 07:06 | 6.6M |
| SRS008632.gsnap_GRCh38Primary.20150922.CHS.mRNA.bam | 2015-10-22 06:32 | 5.3G |
| SRS008632.gsnap_GRCh38Primary.20150922.CHS.mRNA.bam.bai | 2015-10-22 07:06 | 6.6M |
| SRS008633.gsnap_GRCh38Primary.20150922.CHS.mRNA.bam | 2015-10-22 06:43 | 6.7G |
| SRS008633.gsnap_GRCh38Primary.20150922.CHS.mRNA.bam.bai | 2015-10-22 07:06 | 7.0M |
| SRS008745.gsnap_GRCh38Primary.20150922.PUR.mRNA.bam | 2015-10-22 06:51 | 5.1G |
| SRS008745.gsnap_GRCh38Primary.20150922.PUR.mRNA.bam.bai | 2015-10-22 07:06 | 6.5M |
| SRS008746.gsnap_GRCh38Primary.20150922.PUR.mRNA.bam | 2015-10-22 06:58 | 6.4G |
| SRS008746.gsnap_GRCh38Primary.20150922.PUR.mRNA.bam.bai | 2015-10-22 07:06 | 7.0M |
| SRS008747.gsnap_GRCh38Primary.20150922.PUR.mRNA.bam | 2015-10-22 07:06 | 8.1G |
| SRS008747.gsnap_GRCh38Primary.20150922.PUR.mRNA.bam.bai | 2015-10-22 07:06 | 7.2M |

PUR trios

- Double check STAR_WASP pipeline and re-run the whole pipeline for HG00731 in PUR trio.
- GSNAP bam files posted on 1000GP:
- http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/working/20151026_strand_specific_mRNA/
- HG00731 has the smallest bam file among 9 individuals.

# ASE SNPs Identified by GSNAP_WASP and STAR_WASP
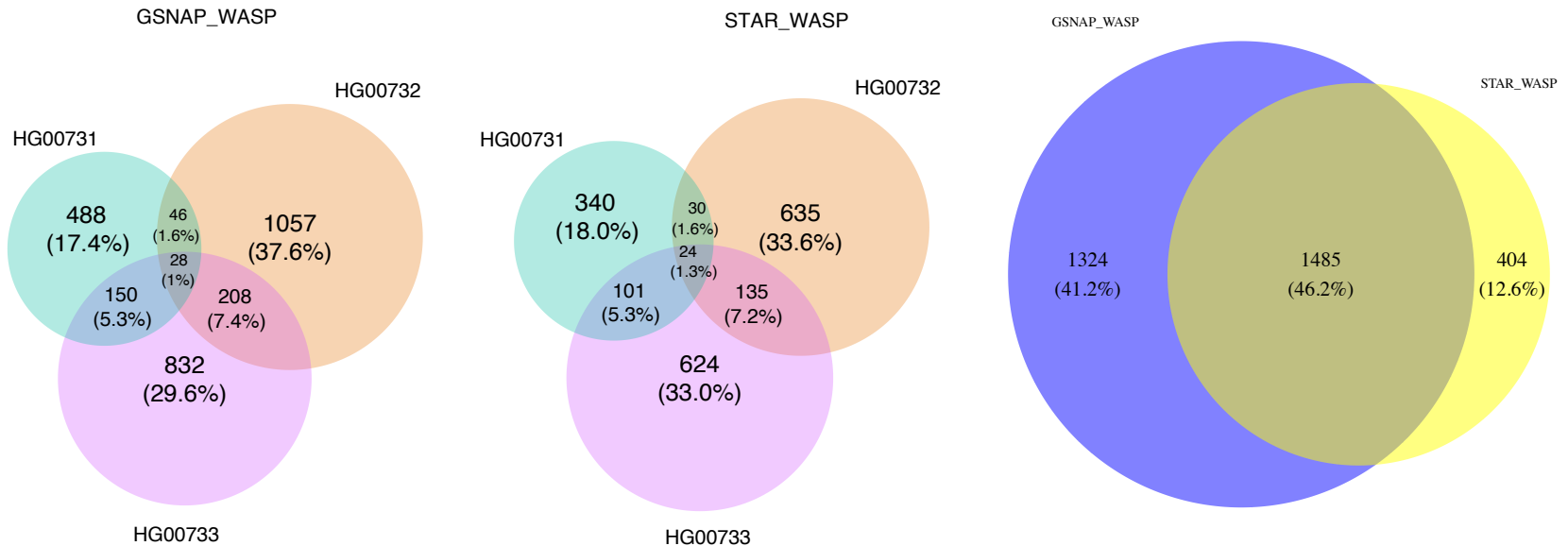


- Use bam files generated from GSNAP, and then follow the same ASE pipeline based on WASP.
- Get similar results for the father HG00731 in PUR trio.
- Most of ASE SNPs (79%) detected by STAR_WASP are verified by GSNAP_WASP.

# Read Statistics for GSNAP_WASP and STAR_WASP

| Mapping aligner | Sample ID | Reads from Bam file generated from STAR (before WASP correction) | WASP Corrected and Merged reads | WASP Removed-DUP Reads | Quality filter reads (NM <=6, MAPQ >20) | #ASE SNP |
|---|---|---|---|---|---|---|
| | NA19238 | 63773306 | 62518123 (98.03%) | 46816839 (73.41%) | 42704787 (66.96%) | 1070 |
| | NA19239 | 72247769 | 70765692 (97.95%) | 52450123 (72.60%) | 47632127(65.93%) | 963 |
| | NA19240 | 59056967 | 57866475 (97.98%) | 38968803 (65.99%) | 35346667 (59.85%) | 595 |
| STAR_WASP | HG00512 | 65900095 | 64557187 (97.96%) | 41748455 (70.76%) | 37725994 (57.25%) | 574 |
| | HG00513 | 58441553 | 57444527 (98.29%) | 42829019 (73.29%) | 38926345 (66.61%) | 723 |
| | HG00514 | 68885118 | 67639427 (98.19%) | 48743774 (70.76%) | 44239830 (64.22%) | 621 |
| | HG00731 | 56111065 | 55154167 (98.29%) | 41209699 (73.44%) | 37459826 (66.76%) | 495 |
| | HG00732 | 69864294 | 68548657 (98.12%) | 49066517 (70.23%) | 44494785 (63.69%) | 824 |
| | HG00733 | 91792573 | 90264811 (98.34%) | 65715517 (71.59%) | 59860057 (65.21%) | 884 |
| | HG00731 | 55796450 | 54513439 (97.70%) | 41134114 (73.72%) | 39516227 (70.82%) | 712 |
| GSNAP_WASP | HG00732 | 69046530 | 67229581 (97.37%) | 48616769 (70.41%) | 46646732 (67.56%) | 1339 |
| | HG00733 | 91464107 | 89269383 (97.60%) | 65717889 (71.85%) | 63382445 (69.30%) | 1218 |

PUR trio

- We think that due to the least number of reads for the PUR father HG00731, our pipeline detects the least number of ASE SNPs for HG00731.
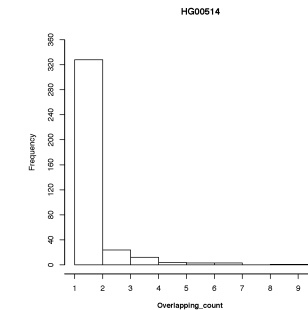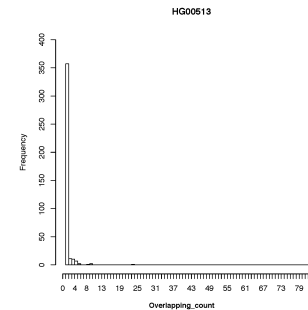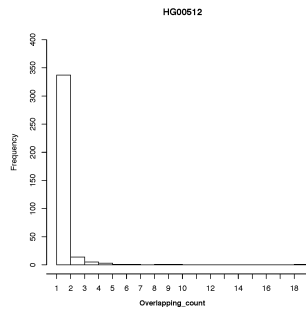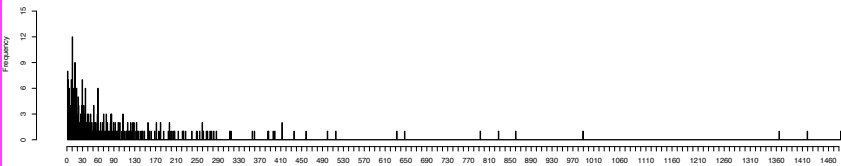
# ASE SNPs Overlapped with ASE Genes
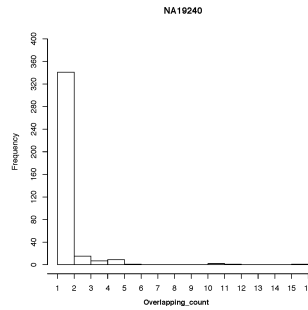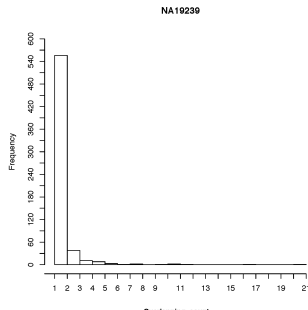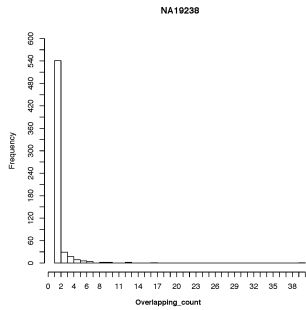


YRI trio
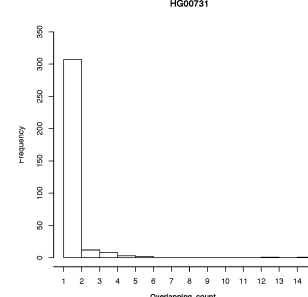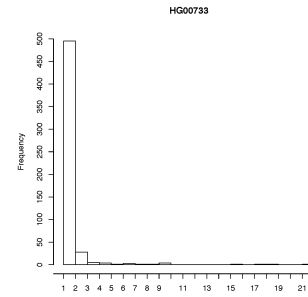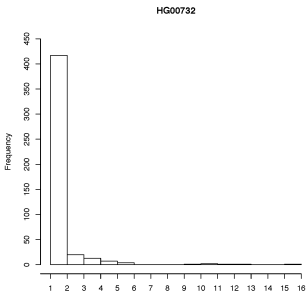
PUR trio

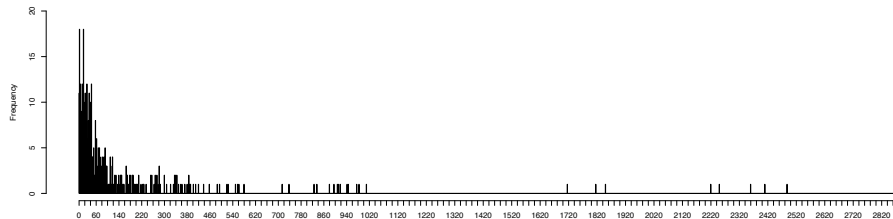CHS trio

No significant distribution difference among all individuals.

PUR trio

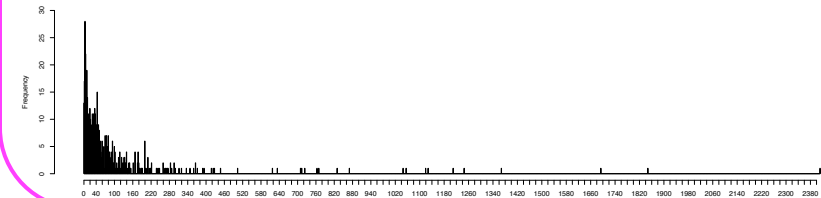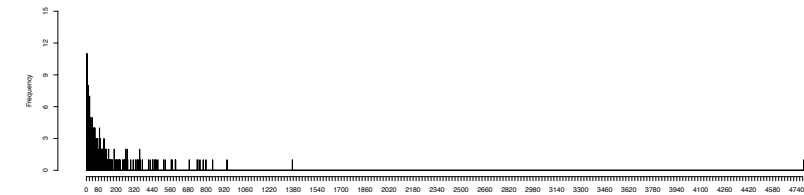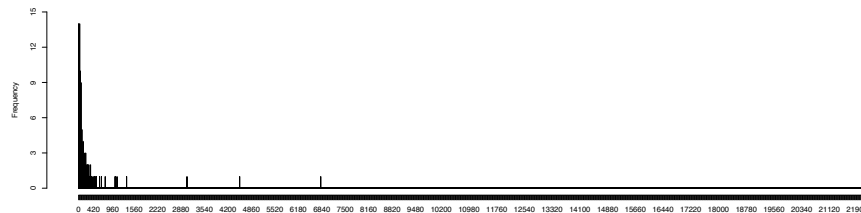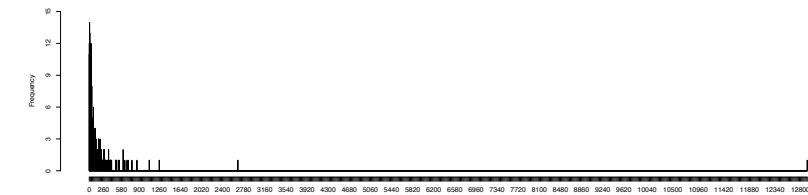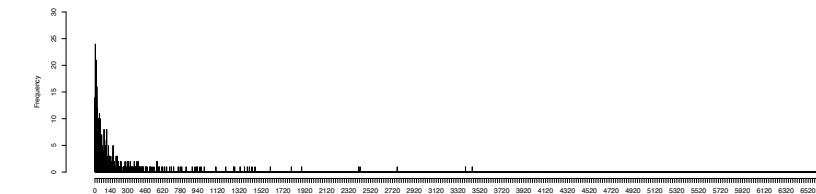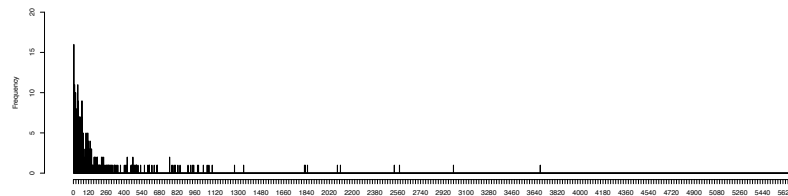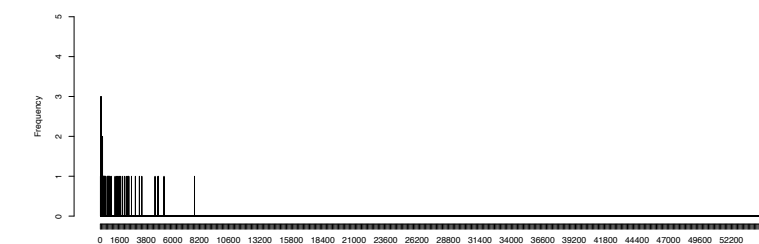Het-SNPs overlapped with ASE genes

# Het-SNPs Overlapped with Protein Coding Genes