BTrecASE a flexible and powerful method for *cis*-eQTL detection

Elena Vigorito, Chris Wallace

September 20, 2018

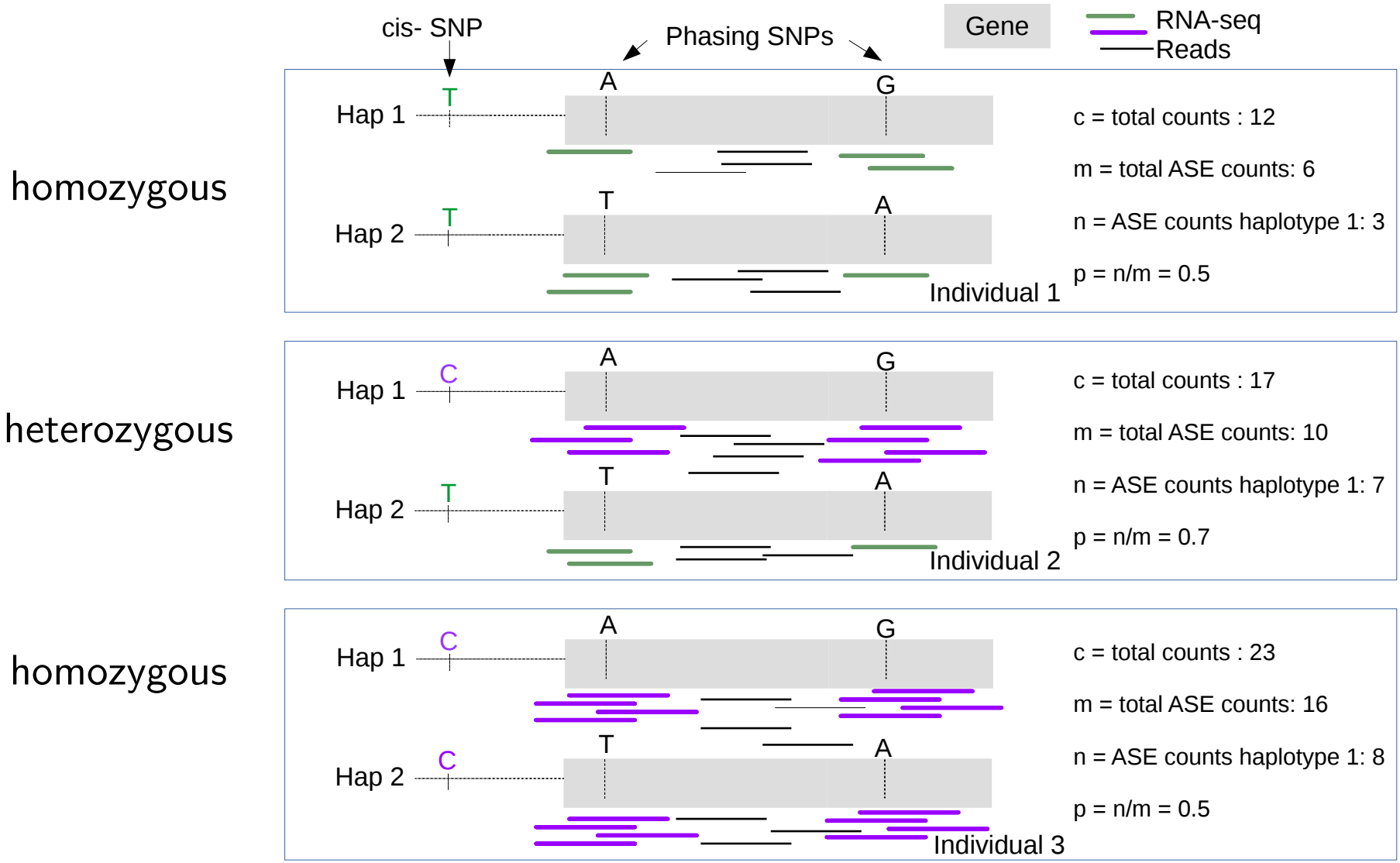
Most common methods to detect eQTL

- Linear regression of gene expression against genotype adjusting by covariates that affect gene expression.
  Advantages $\Rightarrow$ computationally fast and works for cis and trans eQTL.
  Disadvantages $\Rightarrow$ signal/noise can be low, may require big sample sizes.

- For cis-eQTLS: Power can be gained by comparing gene expression between the two chromosomes within individuals, called *allele specific expression*.

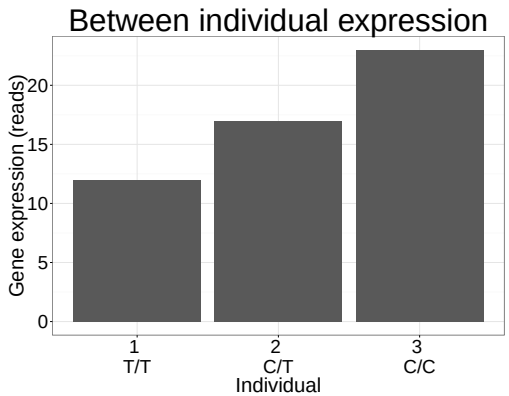
- All require genotypes!

Allele specific expression

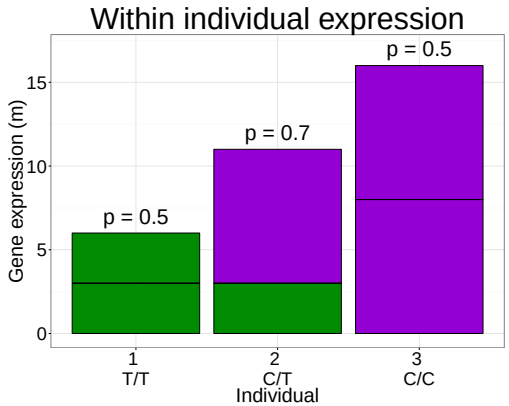


- We look for allelic imbalance within heterozygous individuals

Connecting between individual variation and within individual variation



$$\frac{\mu_{C/C}}{\mu_{T/T}} = \frac{\pi}{1-\pi}$$



Within individual expression

Advantages $\Rightarrow$ controls non-genetic factors.

Disadvantages $\Rightarrow$ Only applies to cis-eQTL and heterozygous individuals, requires haplotype reconstruction.

Our contribution

- Developed a Bayesian formulation of the TrecASE, to properly allow for uncertain haplotype phase, and deliver better flexibility to examine cell-state related effects.

- Extended the model to allow genotype imputation for cis-SNP based on the genotype of the feature-SNPs. Useful when only RNA-seq data is available (no genotypes).

Bayesian formulation of TrecASE

Why Bayesian?

- We can incorporate prior knowledge for the magnitude of b_{AI}
- It provides flexibility to model b_{AI} according to individual characteristics (disease activity, measured cell composition)

Priors:

$b_0 \sim N(6, 4)$; $b_1 \sim Cauchy(0, 2.5)$, $log(library.size)$

$b_{AI} \sim N(0, 0.54)$ when genotype for cisSNP is known

[*ie allelic effect exceeds 1.7-fold change with probability 0.05*]

$b_{AI} \sim N(0, 0.2)$ when genotype for cisSNP is unknown

[*ie allelic effect exceeds 1.2-fold change with probability 0.05*]

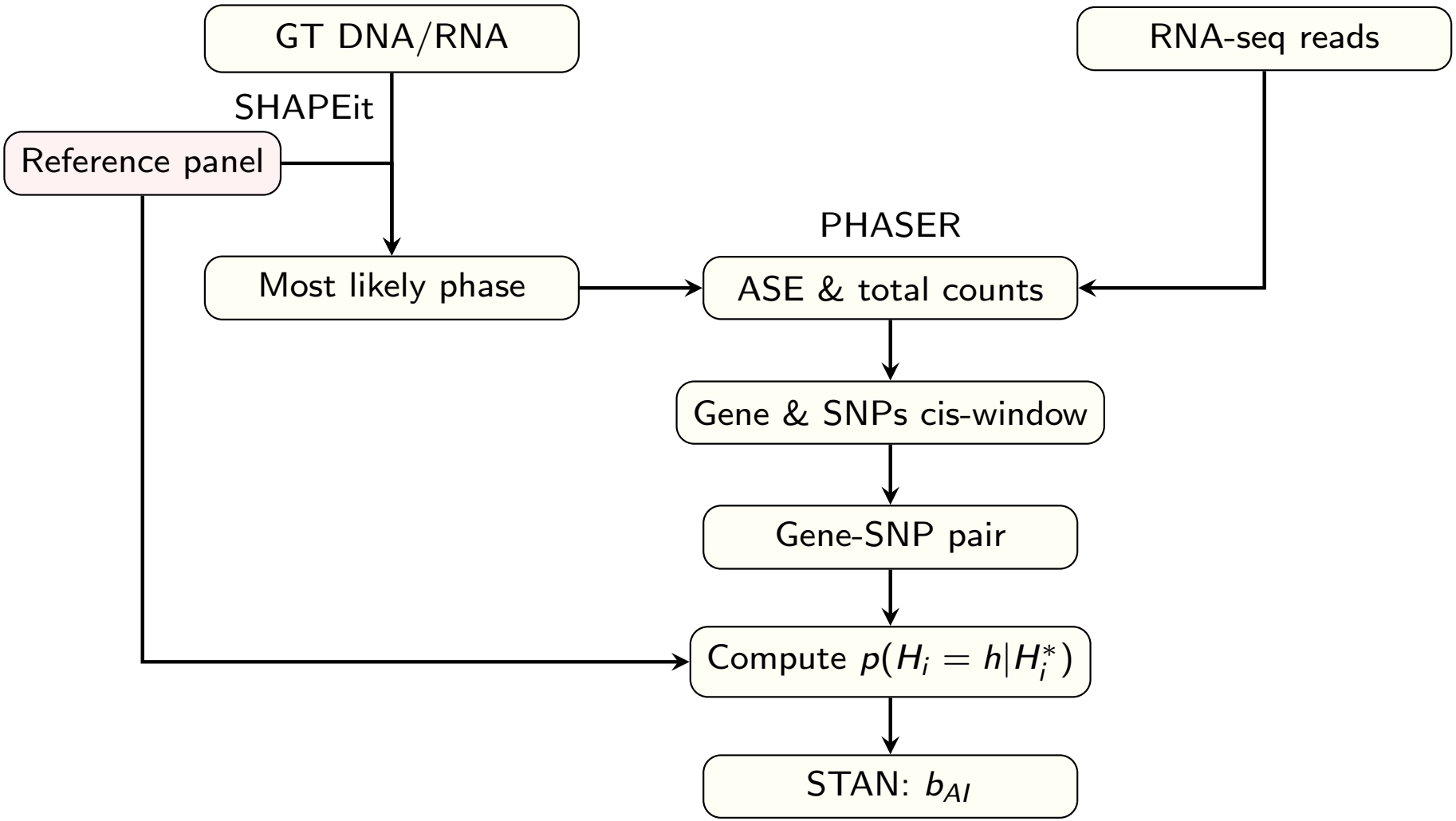
$\theta \sim \Gamma(1, 0.01)$, $\phi \sim \Gamma(1, 0.01)$

Model assumptions:

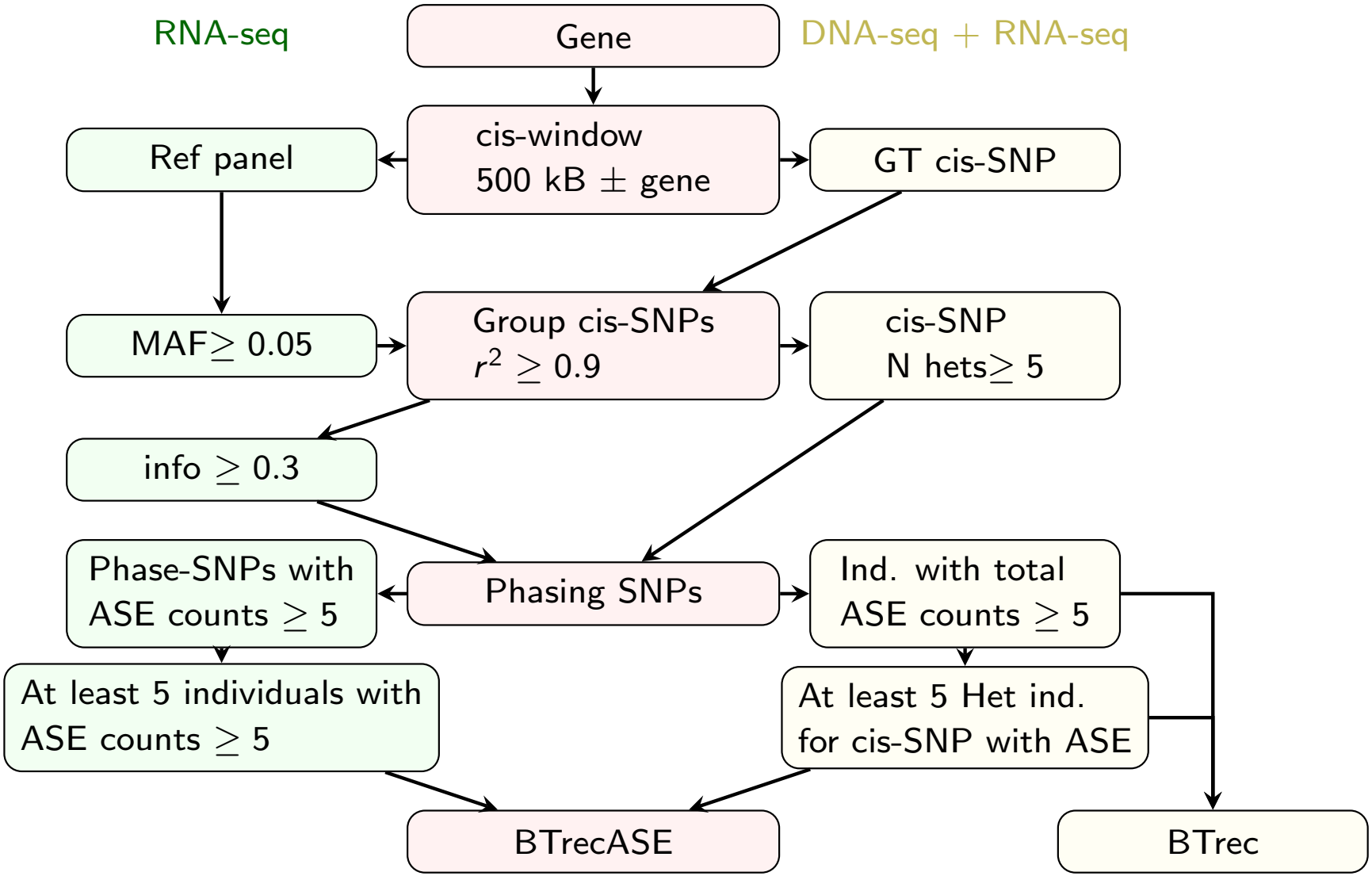
No genotype error (from DNA or RNA)

Samples haplotye frequencies are similar to Reference panel (ethnicity)

Running Bayesian TrecASE on real data



Workflow Bayesian TrecASE for a Gene



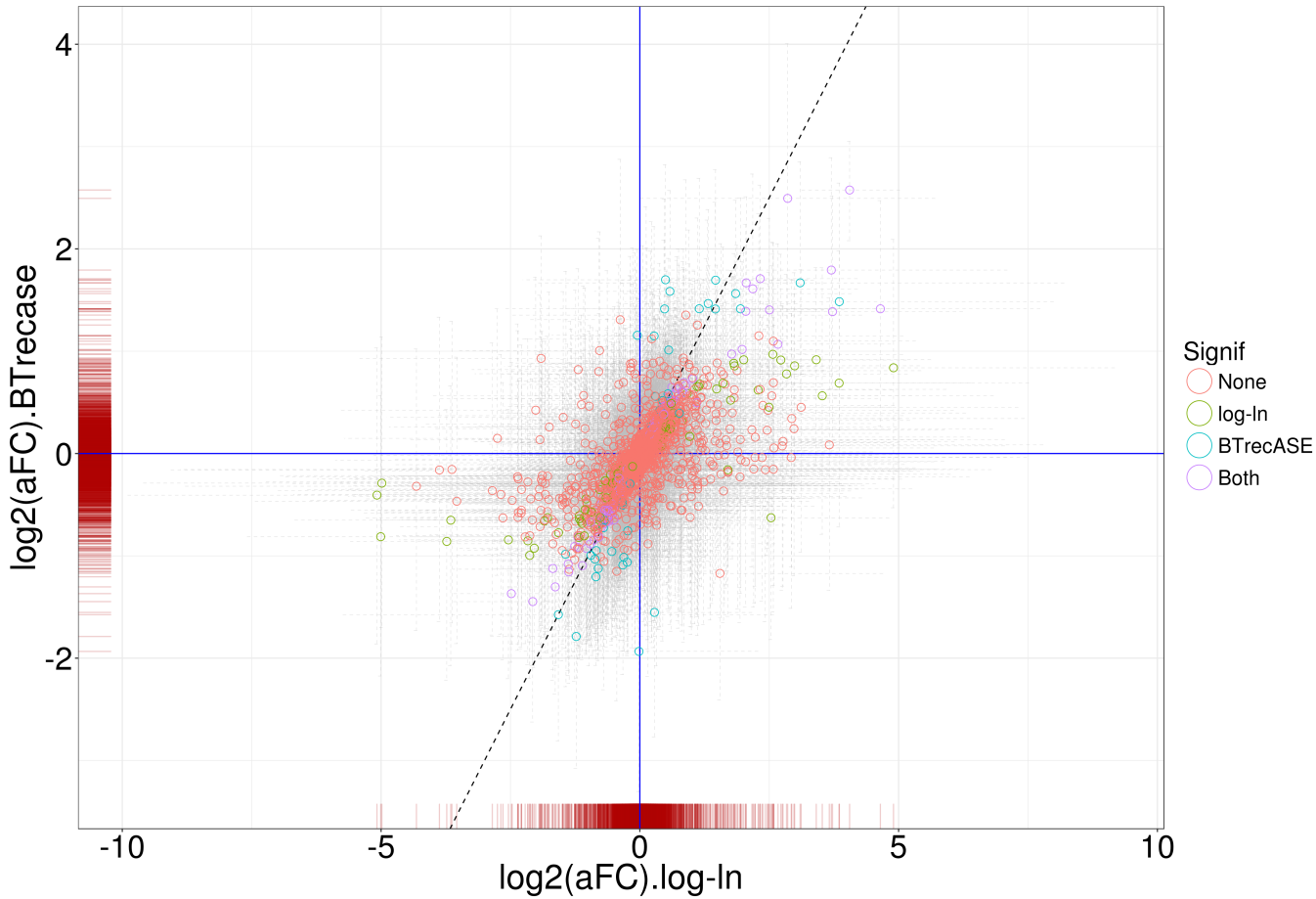
We can use ASE with or without DNAseq

PEAC is one more step if we want to use all data

Testing Bayesian TrecASE on real data

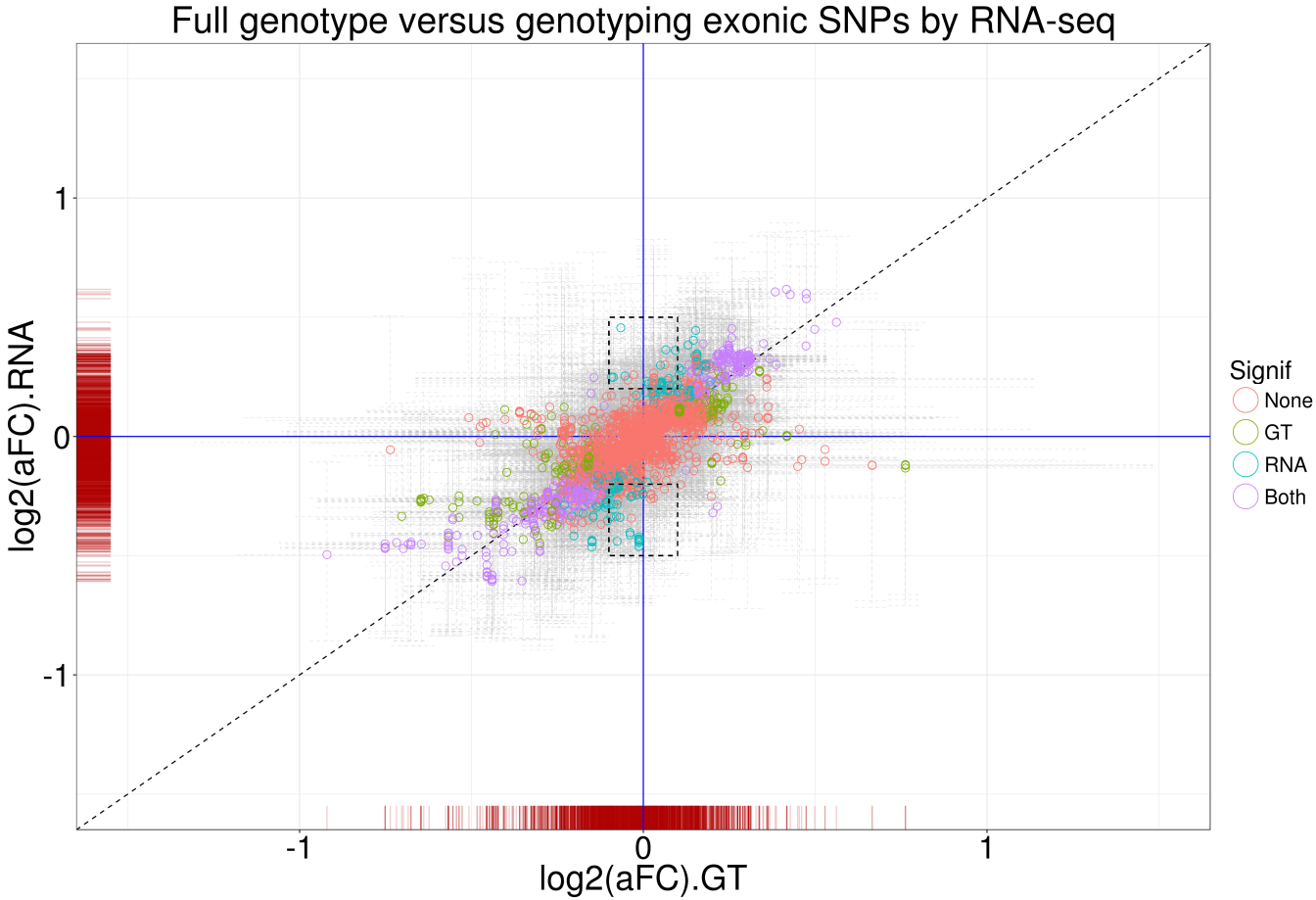
- Samples: 85 GEUVADIS individuals of EUR ancestry.
- Reference panel: 1000K genome project phase 3 (2004 individuals)
- Assessing performance of Bayesian TrecASE:
  - Genes on chromosome 22
  - Compare direction of effect estimates when running BTrecASE with genotypes versus log-linear model
  - Compare running BTrecASE with genotypes versus imputing genotype cisSNP based on the genotype of feature SNPs

Log-linear model versus Bayesian TrecASE



- Good concordance in the direction of effects
- Estimates are not directly comparable

Bayesian TrecASE with genotype cisSNP or imputation

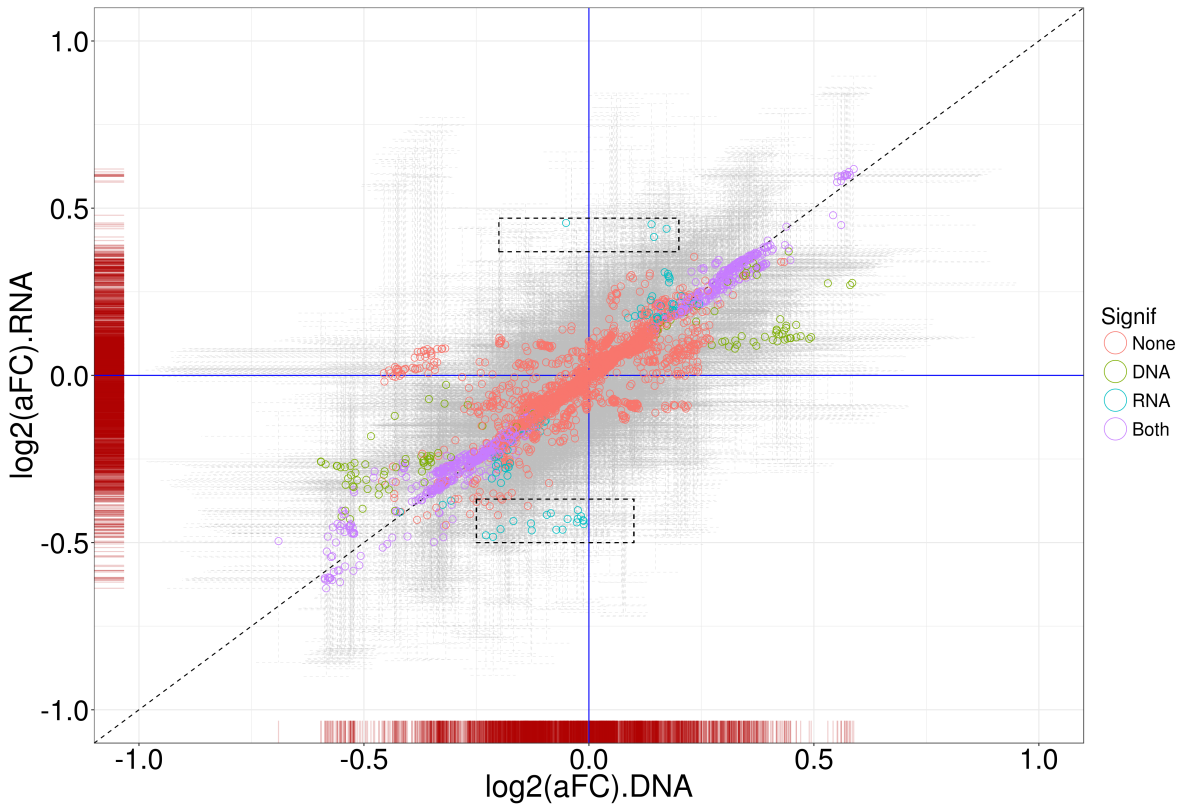


- Only cis-SNPs with info ≥ 0.3 are being compared.
- We expect many false positives but we are concerned about false negatives.

Quality control

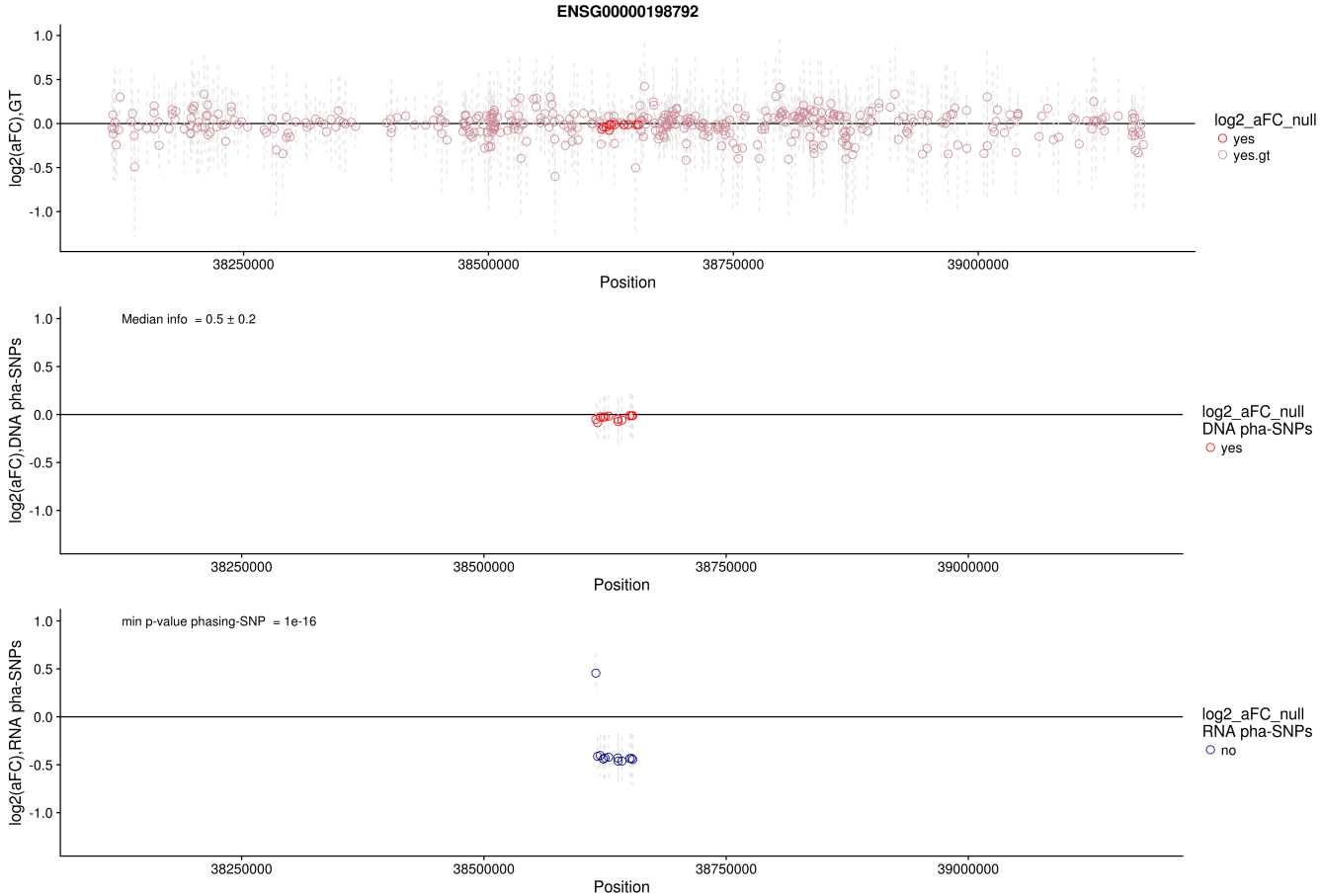
False positives may arise from RNA-genotyping errors

We compare BTrecASE with imputation of cis-SNP from DNA versus RNA genotyping of feature-SNPs



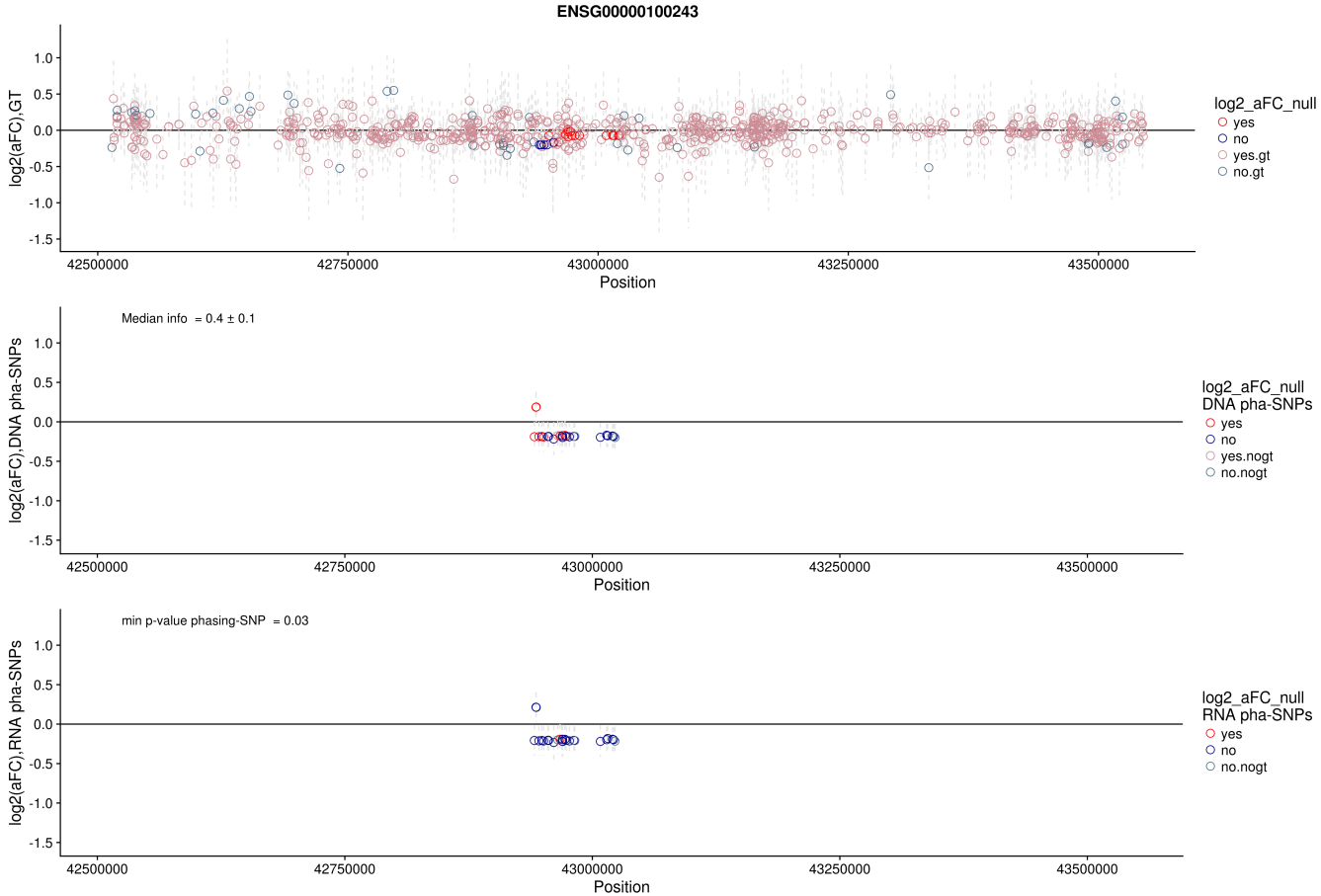
For each feature-SNP we test if the proportion of het individuals in our sample is significantly different from the reference panel

Bayesian TrecASE gene level genotype error



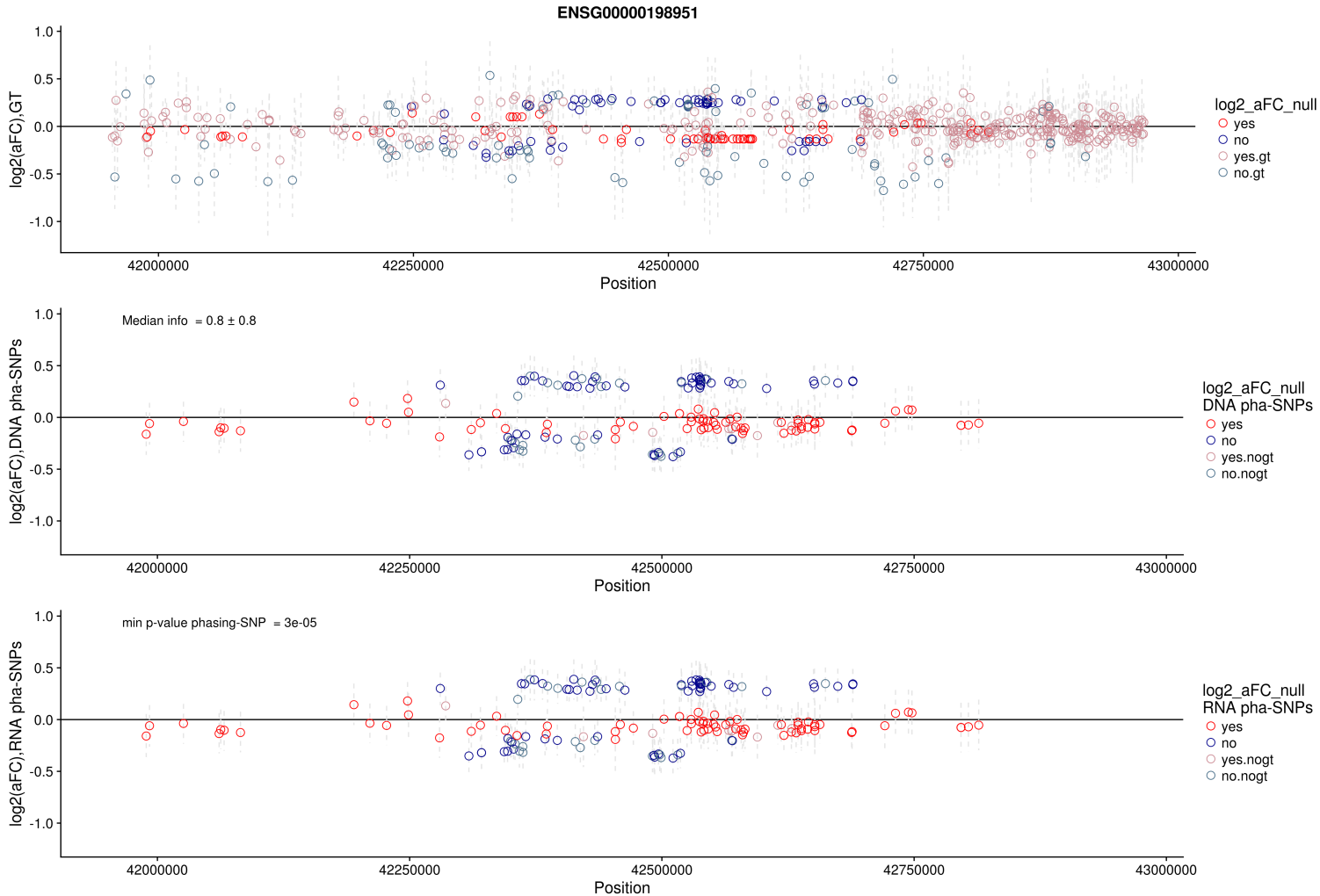
- Only cis-SNPs with info ≥ 0.3 are being compared.
- We flag feature-SNPs with genotype frequencies not consistent with reference panel

Bayesian TrecASE gene level imputation error



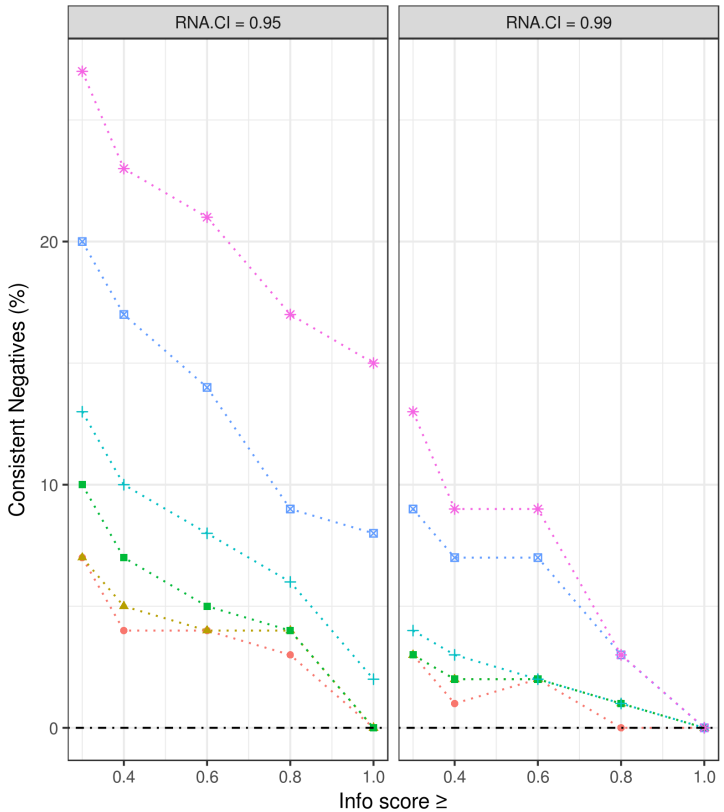
- Only cis-SNPs with info ≥ 0.3 are being compared.
- We flag feature-SNPs with genotype frequencies not consistent with reference panel

Bayesian TrecASE gene level good example

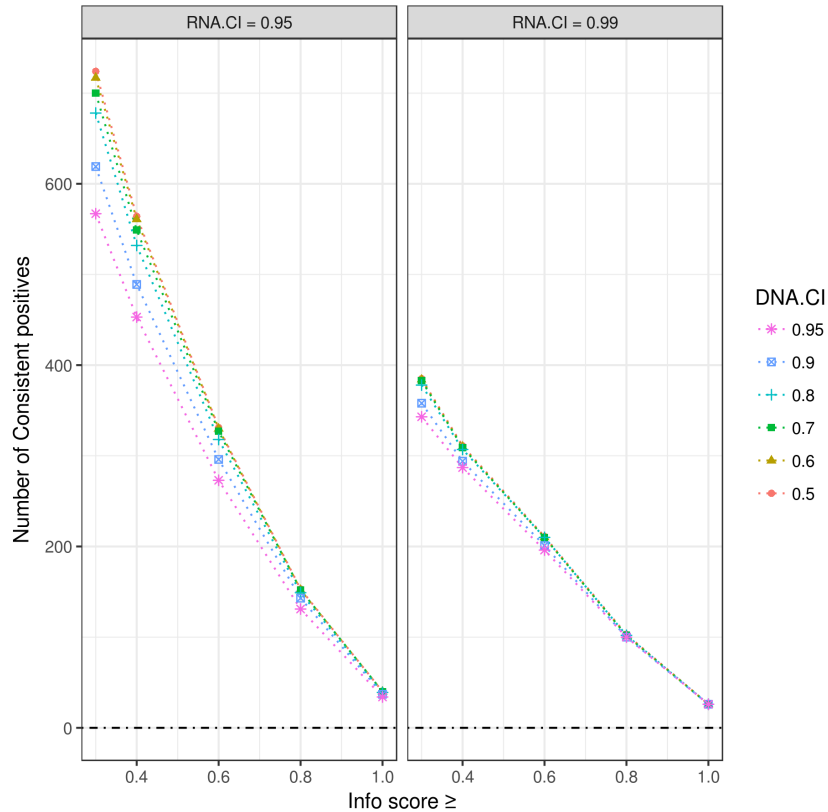


Effect of imputation quality on error rate

Consistent negatives (%)

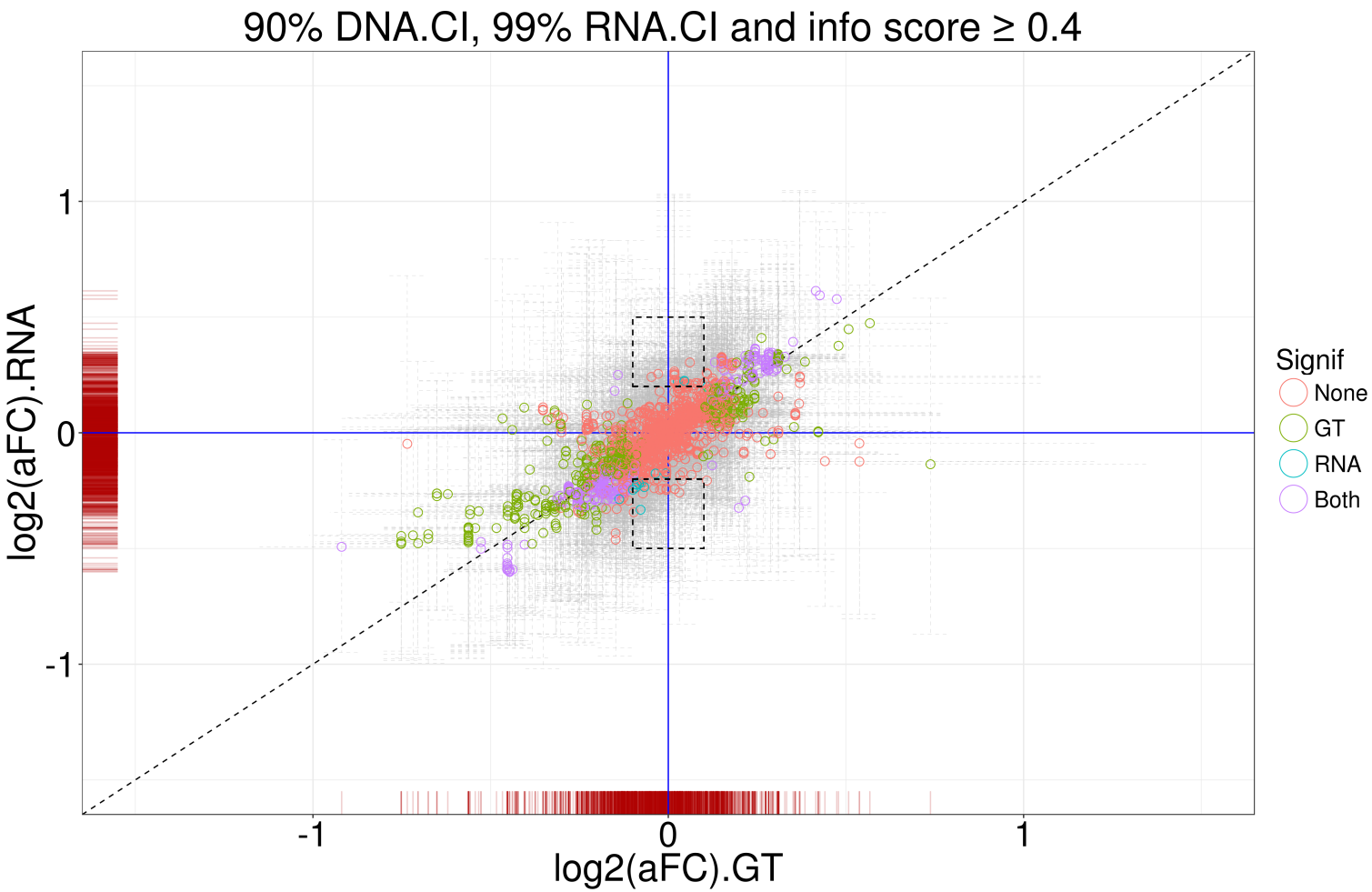


Consistent Positives

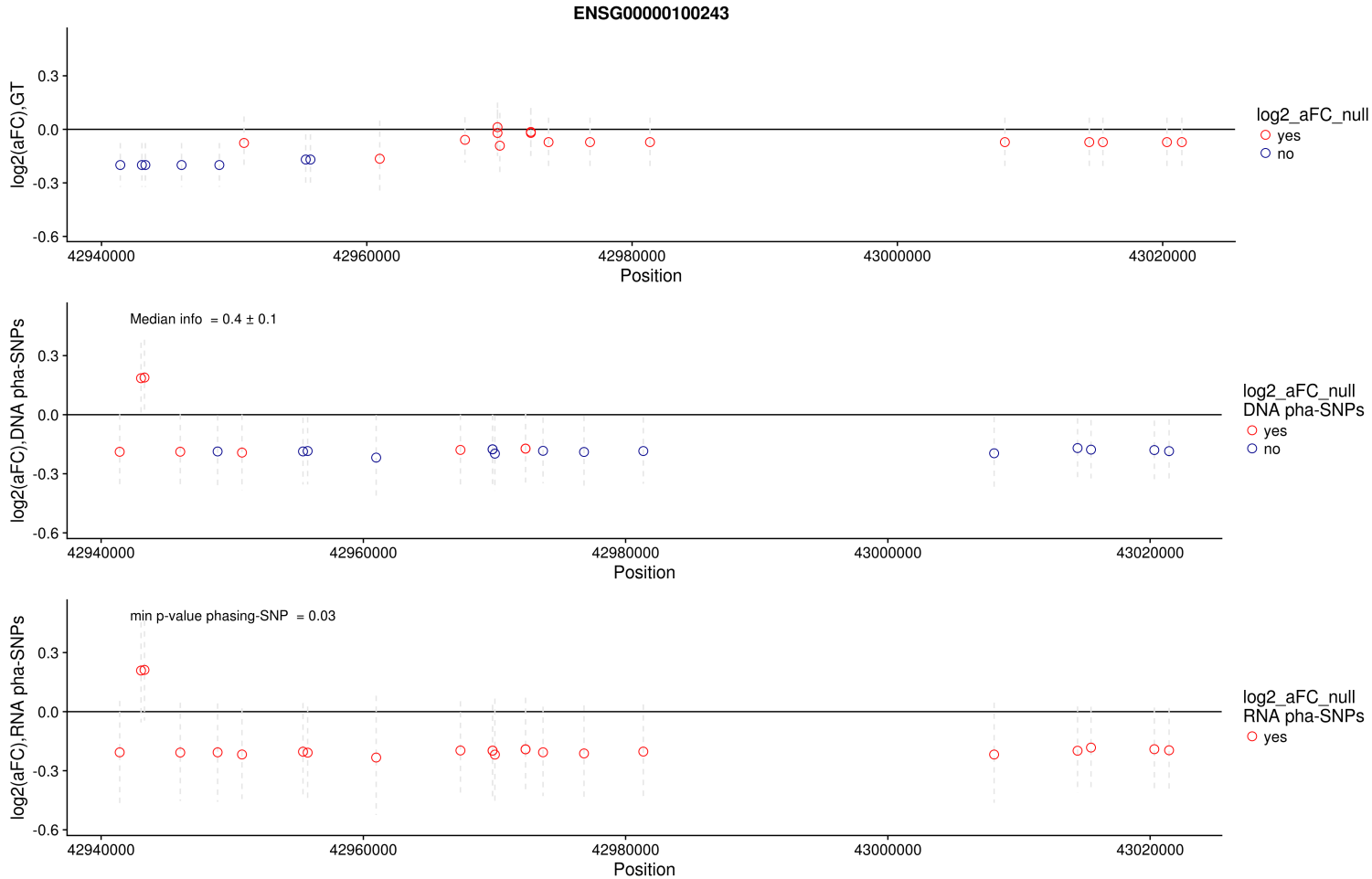


Associations with min.p.fsnp $\leq 10^{-8}$ were excluded.

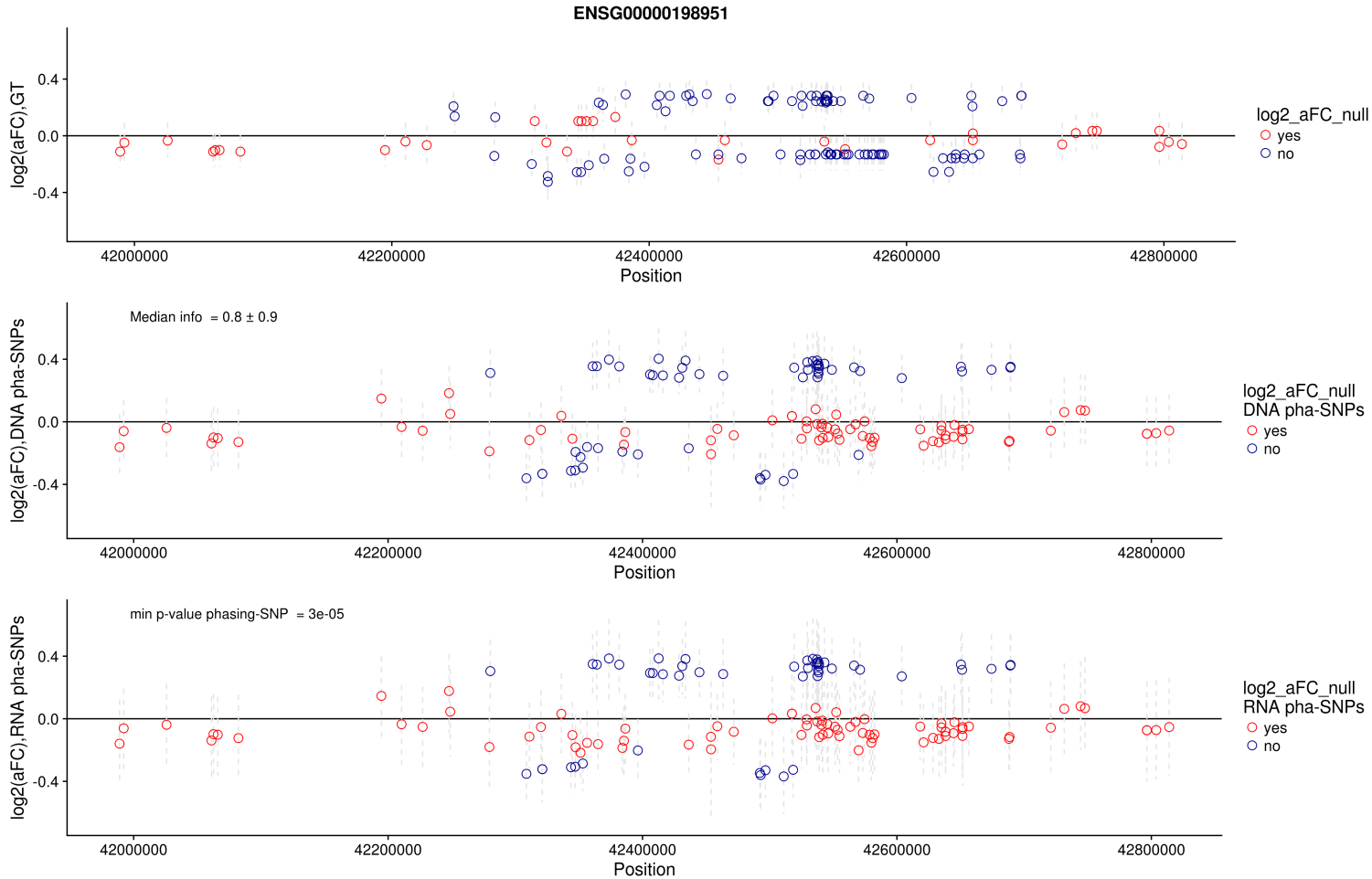
Bayesian TrecASE adjusted cut-offs



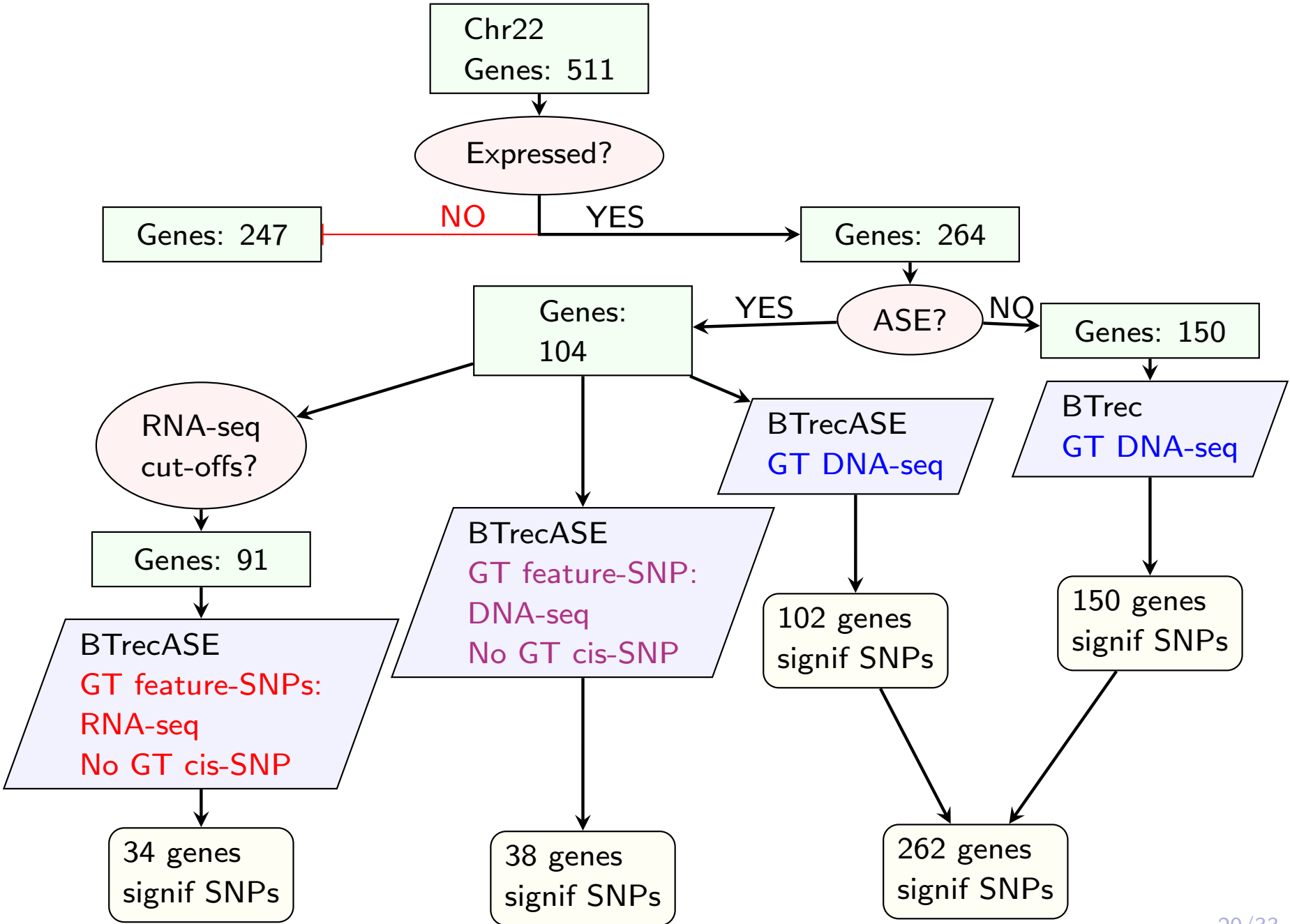
Bayesian TrecASE adjusted cut-offs



Bayesian TrecASE adjusted cut-offs



Summary of associations



Questions about PEAC data

- Scientific: starting with RA patients
  - eQTL synovium/ blood pre-treatment
  - eQTL treatment effect?
  - eQTL by cell types?
  - eQTL by pathotype?
  - future eQTL by disease?
- Samples: I am using HospitalNumber (HN) as individual ID

Overview RA, Genentech 3 batches

Tissue	Reads	Timepoint	QCd.GT	N
Blood	Paired	6mth	YES	9
Blood	Paired	Baseline	NO	8
Blood	Paired	Baseline	YES	50
Blood	Single	6mth	NO	3
Blood	Single	6mth	YES	41
Blood	Single	Baseline	YES	10
Synovium	Paired	6mth	NO	8
Synovium	Paired	6mth	YES	46
Synovium	Paired	9mth (3mths late)	NO	1
Synovium	Paired	Baseline	NO	10
Synovium	Paired	Baseline	YES	84

Pre-treatment:

- 44 paired-end reads blood samples have same HN as Synovium
- 9 single-end reads blood samples have same HN as Synovium

RA, Pre-treatment, Genentech 3 batches, Paired-end?

Model choice?

- No GT cis-SNP : all samples
  - For calling variants I need for each fastq sample:
    <individual_id><sample_name><sequencer_id><flowcell_id><lane_number>
    <adpater_seq>< batch, any other covariate ??>
  - Pool high quality RNA-seq data for each individual regardless tissue or timepoint
  - Which samples can be combined?
- Array GT: (method needs to be adjusted)

RA, Pre-treatment, Genentech 3 batches, Paired-end?

Phasing may be improved by using RNA-seq from same individual

- Benefit in pooling samples per individual, applies to GT and noGT

Ethnicity: synovium

No GT for 8 NA, 1 White Euro and 1 Bangladesh

Ethnicity	N	Ethnicity	N	Ethnicity	N	Ethnicity	N
Indian	3	Asian	10	Black Cari	1	Caribbean	1
White Brit	24	Black	6	White	2	white	2
Caucasian	6	Black Afri	4	Asian-Pakistani	1	Filipino	1
Black Brit	5	Afro-Carib	3	African	1	Asian-Japa	1
Bangladesh	7	Sudanese	1	White Euro	2	Pakistani	2
NA	8	white brit	1	White other	1	Mixed Brit	1

RA, Pre-treatment

Samples that failed alignment, any problem before?

SampleID..QMUL.ID.only.	SampleID..QMUL.or.Genentech.	HospitalNumber
QMUL2010054	QMUL2010054	9247079
QMUL2011003	SAM9103834	9265033
QMUL2011001	SAM9185509	QE902125

Batch	Reads	Timepoint
Ambry	Paired	Baseline
GenentechBatch1	Paired	Baseline
GenentechBatch1	Paired	Baseline

Tissue	Diagnosis	Ethnicity
Synovium	RA	Caribbean
Synovium	RA	asian
Blood	RA	Black

Duplicated sample: which one to choose?

SampleID..QMUL.or.Genentech.	SampleID..QMUL.ID.only.	Batch
SAM20389187	QMUL2013093	GenentechBatch2
SAM24297997	QMUL2013093	GenentechBatch3

Tissue	Reads	Timepoint
Synovium	Paired	6mth
Synovium	Paired	6mth

QCd.GT	HospitalNumber
YES	WH1326872
YES	WH1326872

Genentech missing data?

Missing data: Diagnosis, Hospital Number, Timepoint

Tissue	Reads	Batch	QCd.GT	N
Synovium	Paired	GenentechBatch3	NO	23
Blood	Paired	GenentechBatch2	NO	19
Blood	Paired	GenentechBatch1	NO	1
HealthySynovium	Paired	GenentechBatch3	NO	8
HealthySynovium	Paired	GenentechBatch2	NO	4

RA UA, Ambry: split or repeated samples?

SampleID..QMUL.ID.only.	Batch	Tissue	Reads	Diagnosis	Timepoint
QMUL2011071	Ambry	Synovium	Paired	UA	Baseline
QMUL2011071	Ambry	Synovium	Paired	UA	Baseline
QMUL2011071	Ambry	Synovium	Paired	UA	Baseline
QMUL2011071	Ambry	Synovium	Paired	UA	Baseline
QMUL2011071	Ambry	Synovium	Paired	UA	Baseline
QMUL2011071	Ambry	Synovium	Paired	UA	Baseline
QMUL2010009	Ambry	Synovium	Paired	RA	Baseline
QMUL2010009	Ambry	Synovium	Paired	RA	Baseline
QMUL2010009	Ambry	Synovium	Paired	RA	Baseline
QMUL2010009	Ambry	Synovium	Paired	RA	Baseline

bastion.path

/mnt/volume/PEAC/RNAseq/RawData/Ambry/FASTQ/QMUL201171/QMUL2011071_TTAGGC_L003_R1.fastq.gz
/mnt/volume/PEAC/RNAseq/RawData/Ambry/FASTQ/QMUL201171/QMUL2011071_TTAGGC_L003_R2.fastq.gz
/mnt/volume/PEAC/RNAseq/RawData/Ambry/FASTQ/QMUL201171/QMUL2011071_TTAGGC_L007_R1.fastq.gz
/mnt/volume/PEAC/RNAseq/RawData/Ambry/FASTQ/QMUL201171/QMUL2011071_TTAGGC_L007_R2.fastq.gz
/mnt/volume/PEAC/RNAseq/RawData/Ambry/FASTQ/QMUL201171/QMUL_2011_71_TTAGGC_L003_R1.fastq.gz
/mnt/volume/PEAC/RNAseq/RawData/Ambry/FASTQ/QMUL201171/QMUL_2011_71_TTAGGC_L003_R2.fastq.gz

/mnt/volume/PEAC/RNAseq/RawData/Ambry/FASTQ/QMUL20109/QMUL201009_CGATGT_L004_R1.fastq.gz
/mnt/volume/PEAC/RNAseq/RawData/Ambry/FASTQ/QMUL20109/QMUL201009_CGATGT_L004_R2.fastq.gz
/mnt/volume/PEAC/RNAseq/RawData/Ambry/FASTQ/QMUL20109/QMUL_2010_9_CGATGT_L003_R1.fastq.gz
/mnt/volume/PEAC/RNAseq/RawData/Ambry/FASTQ/QMUL20109/QMUL_2010_9_CGATGT_L003_R2.fastq.gz

Anything else I am missing???

Existing statistical models for cis-eQTL (1)

- RASQUAL (Kumasaka et al, 2016)

$$L(\pi, \lambda, \theta, \delta, \varphi) \propto \prod_{i=1}^{N} \sum_{G_i} p(G_i)\, p_{NB}(c_i | G_i; \pi, \lambda, \theta)$$

$$x \prod_{l=1}^{L} \sum_{D_{il}} p(D_{il} | G_i)\, p_{BB}(c_{il}^{(1)} | c_{il}, D_{il}; \pi, \delta, \varphi, \theta)$$

π, expected proportion of allele specific expression for alternative haplotype

G_i, genotype of cis-SNP for individual i

c_i, total counts for individual i

D_{il}, diplotype for cis-SNP and each feature SNP for ind. i

$c_{il}^{(1)}$, counts at each feature SNP for alt. haplotype for ind. i

Existing statistical models for cis-eQTL (2)

- TrecASE (Sum, 2012 and Hu et al 2015)

$$L(\mathbf{b}, b_{AI}, \phi, \theta | c, n, m, G, \mathbf{X}) = \prod_{i=1}^{N} g_{TR}(c_i; \mathbf{b}, b_{AI}, \phi, \mathbf{X}) \times \prod_{i:m_i>0} p_{BB}(n_i, m_i; \pi, \theta)$$

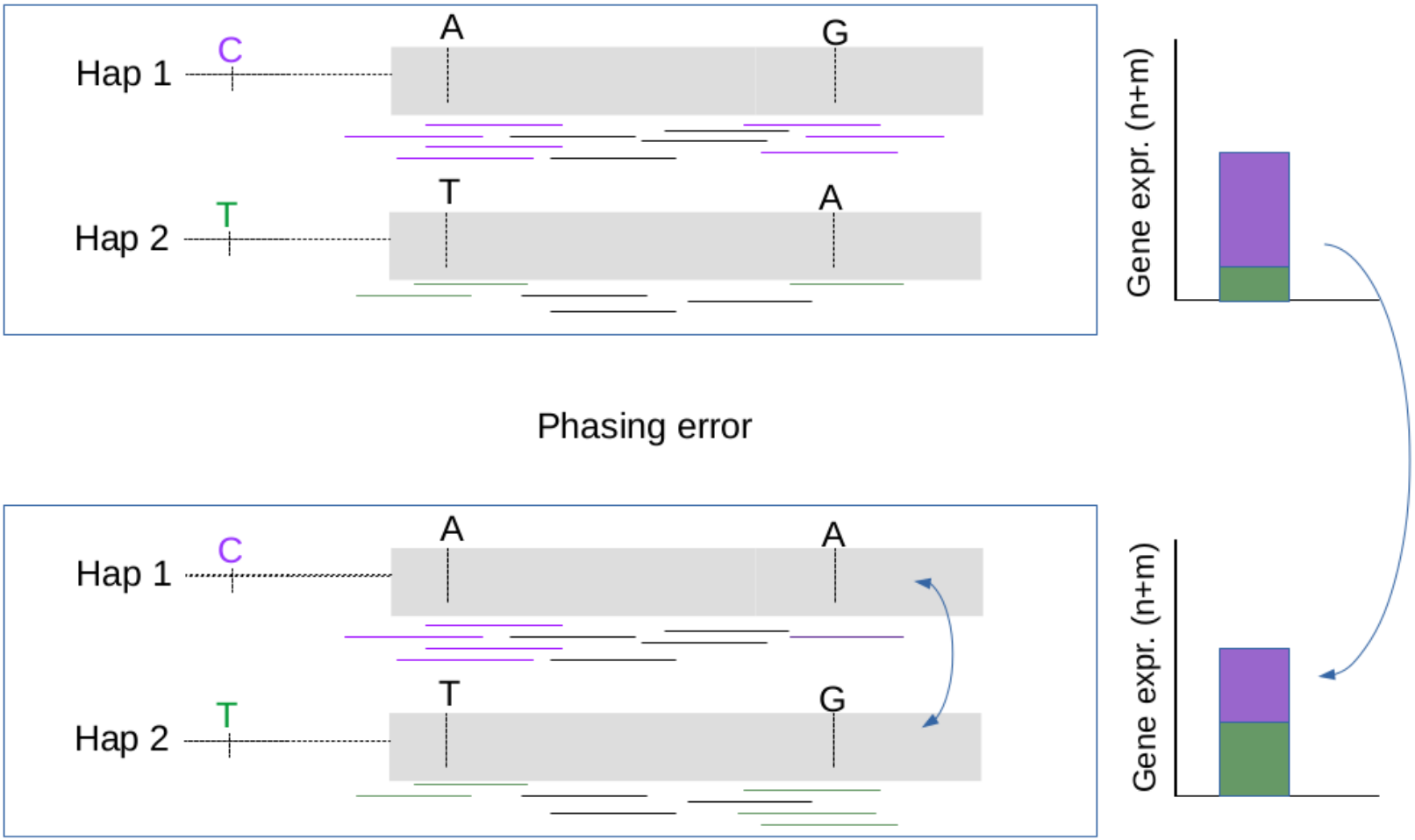
$$\pi = \begin{cases} \text{logit}(b_{AI}) & G_i \text{ heterozygous at cis-SNP} \\ 0.5 & G_i \text{ homozygous at cis-SNP} \end{cases}$$

n_i = Allele specific counts for alternative haplotype for ind i

m_i = Total allele specific counts for ind i

Assumes known genotypes and phase

Effect of phasing error on eQTL effect



Reformulation of TrecASE: phasing uncertainty

Phasing:

n_i and m_i are dependent on the unobserved phased haplotypes, H_i.

We observed unphased genotypes at those variants, $H_i^\star$.

We use external data on haplotype frequencies to estimate $P(H_i = h|H_i^\star)$, then incorporate phase uncertainty by writing $n_i(h)$ and $m_i(h)$ to indicate dependence on H_i:

$$L_2(b_{AI}, \theta|\mathbf{n}, \mathbf{m}, \mathbf{H}^\star) = \prod_{i=1}^{N} \sum_h P(H_i = h|H_i^\star)\, f_{BB}\left(n_i(h); \pi, \theta, m_i(h)\right)$$