Unbiased Estimation of Parent-of-Origin Effects Using RNA-seq Data from Human

Vasyl Zhabotynsky[1], Wei Sun[2], Kaoru Inoue[1], Terry Magnuson[1] and Mauro Calabrese[1]

1 University of North Carolina, Chapel Hill

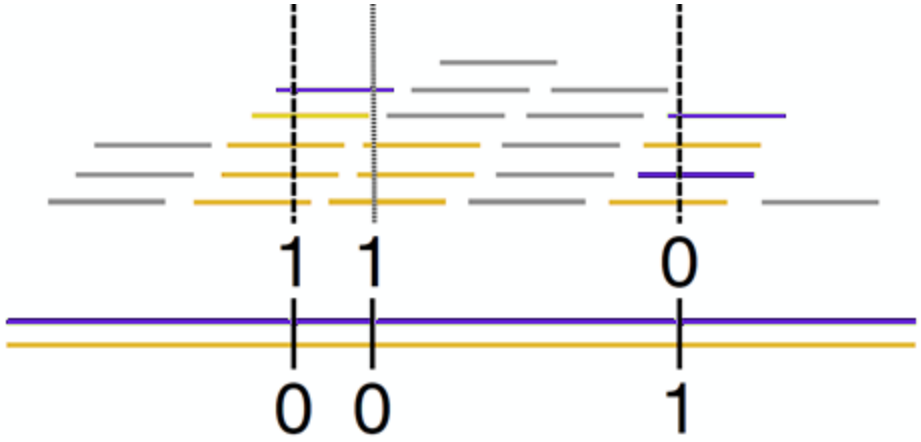
2 Fred Hutchinson Cancer Research Center

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Department of Biostatistics

Total and Allele Specific RNA-seq reads

- Can quantify total reads mapped to a gene
- Fraction of the reads that overlaps a SNP can be attributed to one of the parents



Note: Estimating allele-specific reads for each SNP separately creates a potential for double-counting

RNA-seq reads (cont)

Consider 4 individuals with a gene expression associated with A allele to be half of the expression from B allele and additional parental imbalance with paternal allele producing 3 times more reads

We can see both genetic and parent-of-origin effects in both

Mat|Pat	Mat	Pat	Total
A|A	100	300	400
A|B	100	600	700
B|A	200	300	500
B|B	200	600	800

allele-specific (AS) reads:

- B > A
- Pat > Mat

and total expression:

- Total (B|B) > Total (A|A)
- Total (A|B) > Total (B|A)

Incorporating Parent-of-origin Effect

Define the two alleles of a candidate eQTL as A_1 and A_2 and genotype in the i'th individual as g_i.

Account for parent-of-origin effect by distinguishing A_1A_2 genotype as having haplotypes h_{i1},h_{i2} harboring A_1,A_2 alleles respectively and A_2A_1 for which h_{i1},h_{i2} harbor A_2 and A_1 allele

Model allele-specific reads from first allele n_{i1} (n_i=n_{i1}+n_{i2}) by a Beta Binomial distribution as

$$n_{i1} \sim f_{BB}(n_{i1}; n_i, \pi_i, \varphi), \quad \log\left[\pi_i/(1-\pi_i)\right] = b_0 z_i + b_1 x_i,$$

where

$$x_i = \begin{cases} 1 & h_{i1} \text{ is from the paternal} \\ -1 & h_{i1} \text{ is from the maternal} \end{cases} \qquad z_i = \begin{cases} 0 & \text{if } g_i = A_k A_k,\ k = 1,2 \\ 1 & \text{if } g_i = A_2 A_1 \\ -1 & \text{if } g_i = A_1 A_2 \end{cases}$$

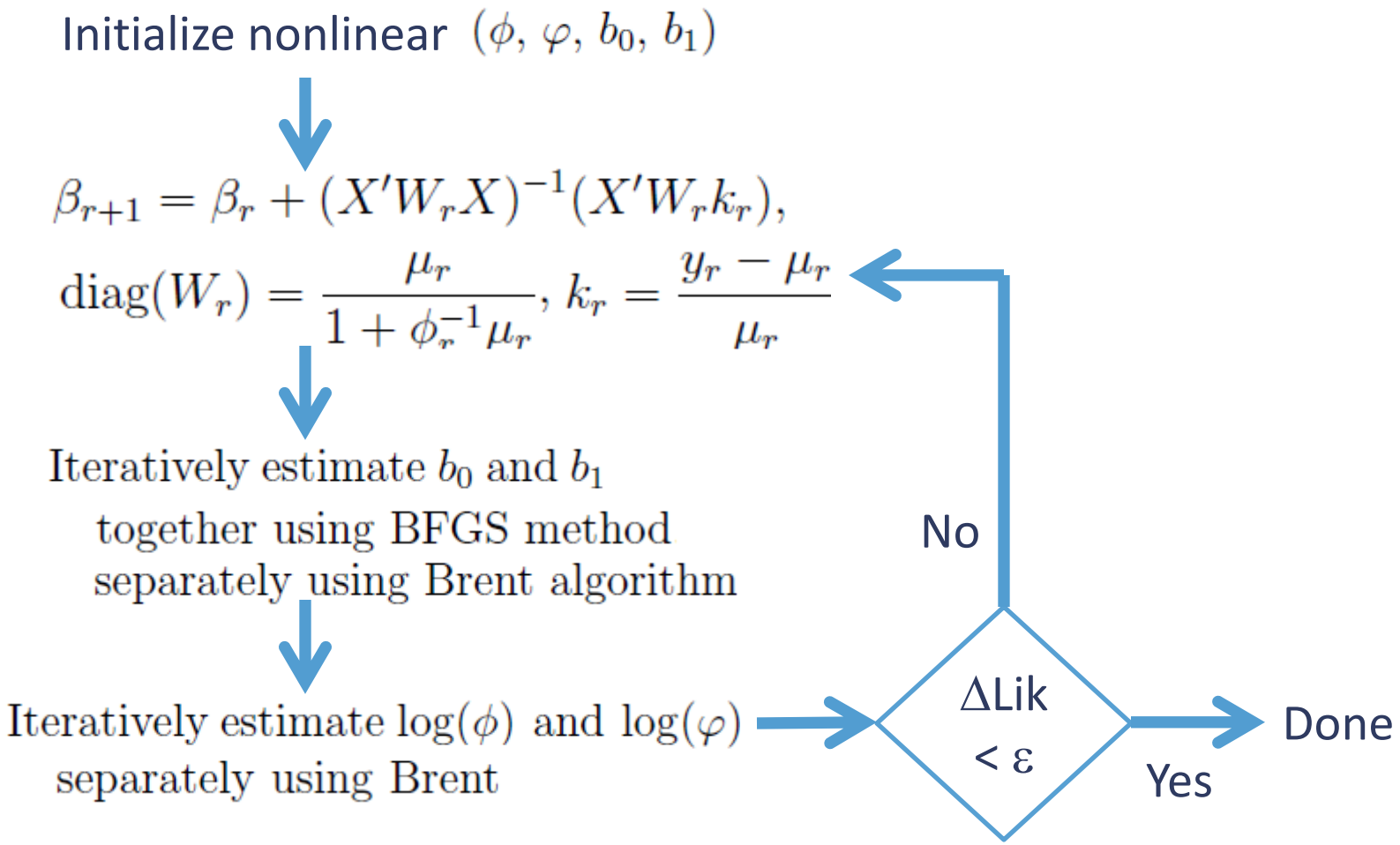
Total Read counts

Total read counts are to be modeled with Negative Binomial with mean structure accounting for covariates such β_k such as read depth, dominance, sex, and described above genetic and parent-of-origin effects by η_{ij} as:

$$\eta_i = \begin{cases} 0 & g_i = A_1A_1 \\ \log\{1+\exp(b_0+x_ib_1)\} - \log\{1+\exp(x_ib_1)\} & g_i = A_1A_2, A_2A_1 \\ b_0 & g_i = A_2A_2 \end{cases}$$

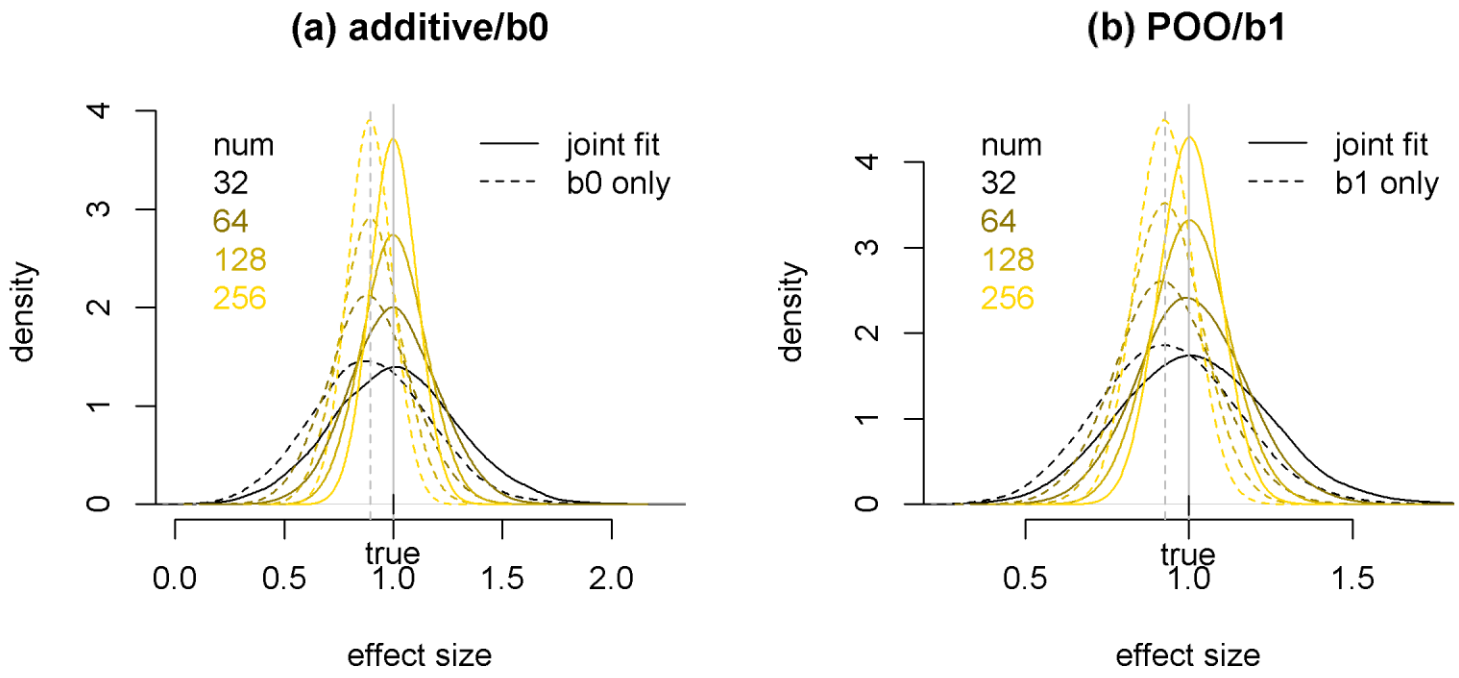
so that $$\log(\mu_i) = \sum_{k=1}^{p} \beta_k c_{ik} + \eta_i$$

Algorithm details

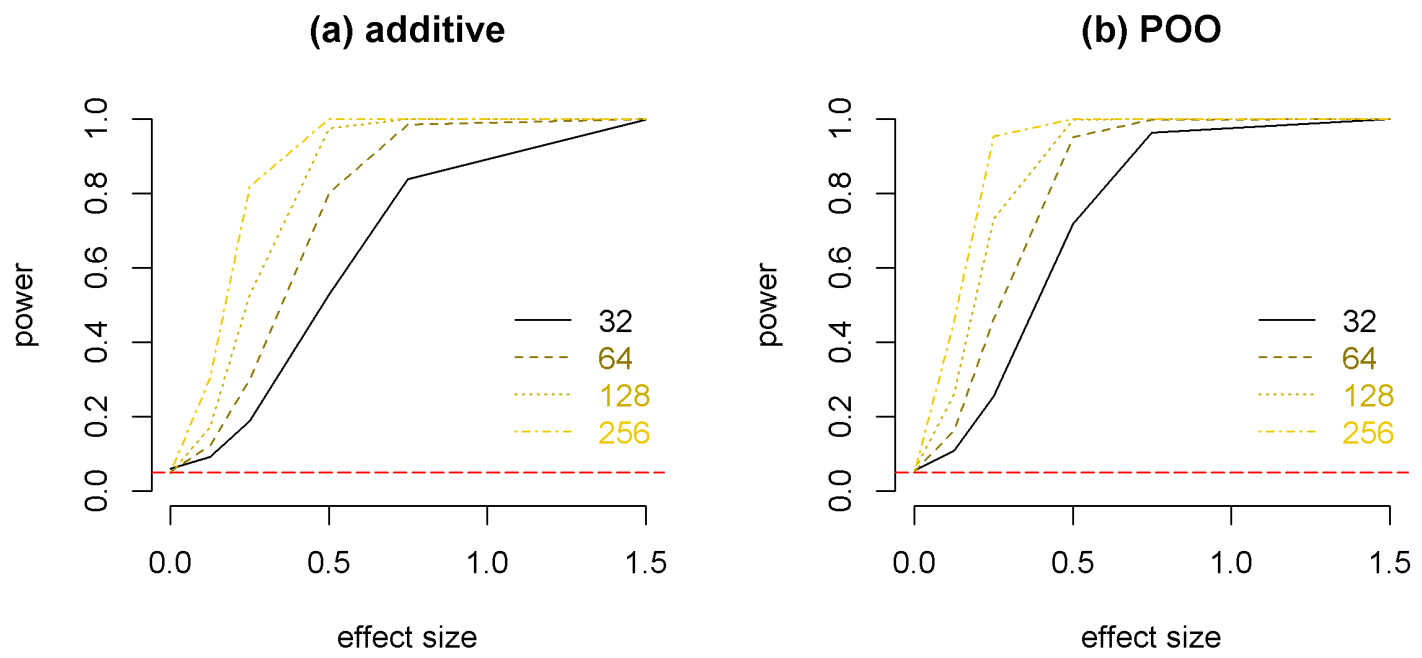


Simulations

- Select sample sizes 32 (close to dataset we have), 64, 128, 256
- Over-dispersion: BB ¼, NB ¾
- Mean Total Read Count 250 & 10% of reads to be AS

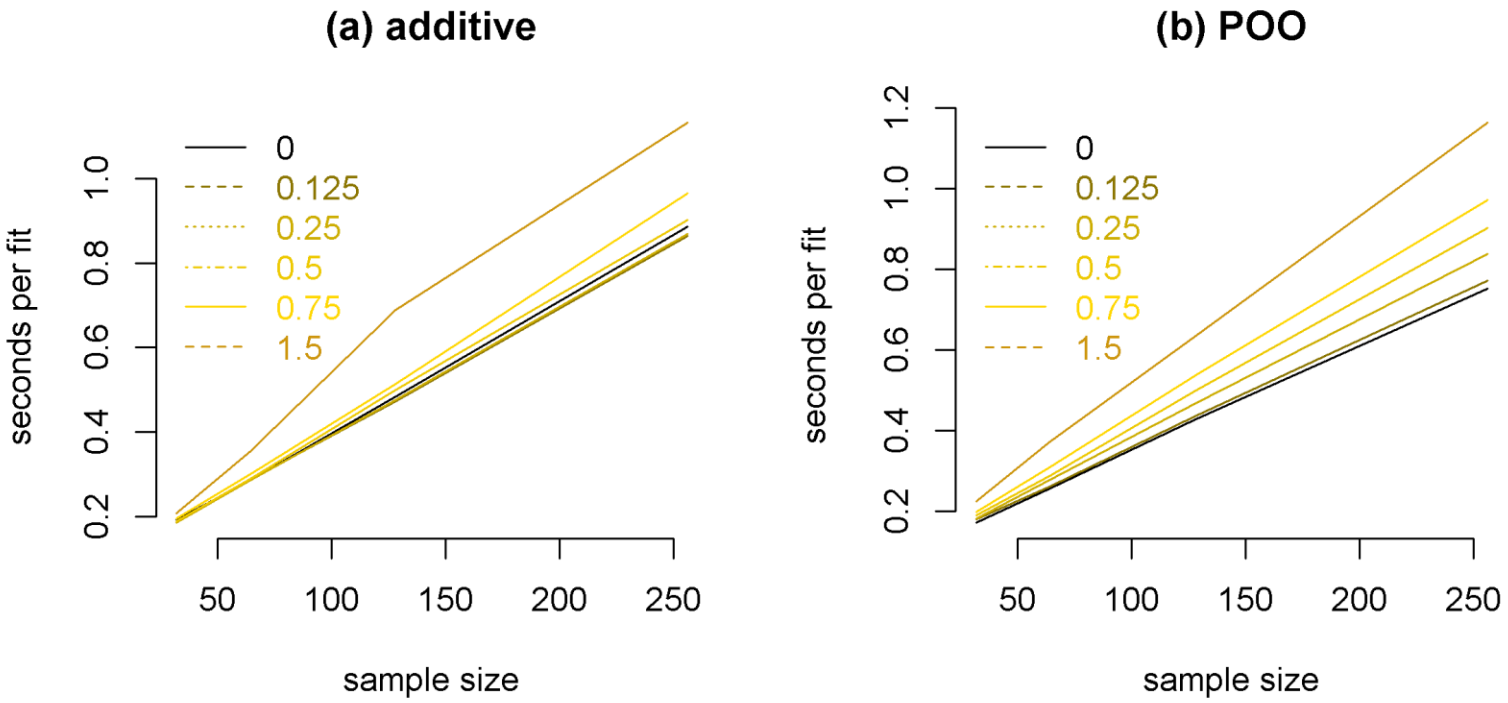


Simulations. Power



- Even at smaller sample size we get power of around 80% for one fold change (equivalent to effect size 0.693) in both effects.
- We observe higher power in parent-origin effect: ASE can be used to quantify genetic effect only if eQTL is heterozygous

Timing



Current implementation scales well with increasing sample size.

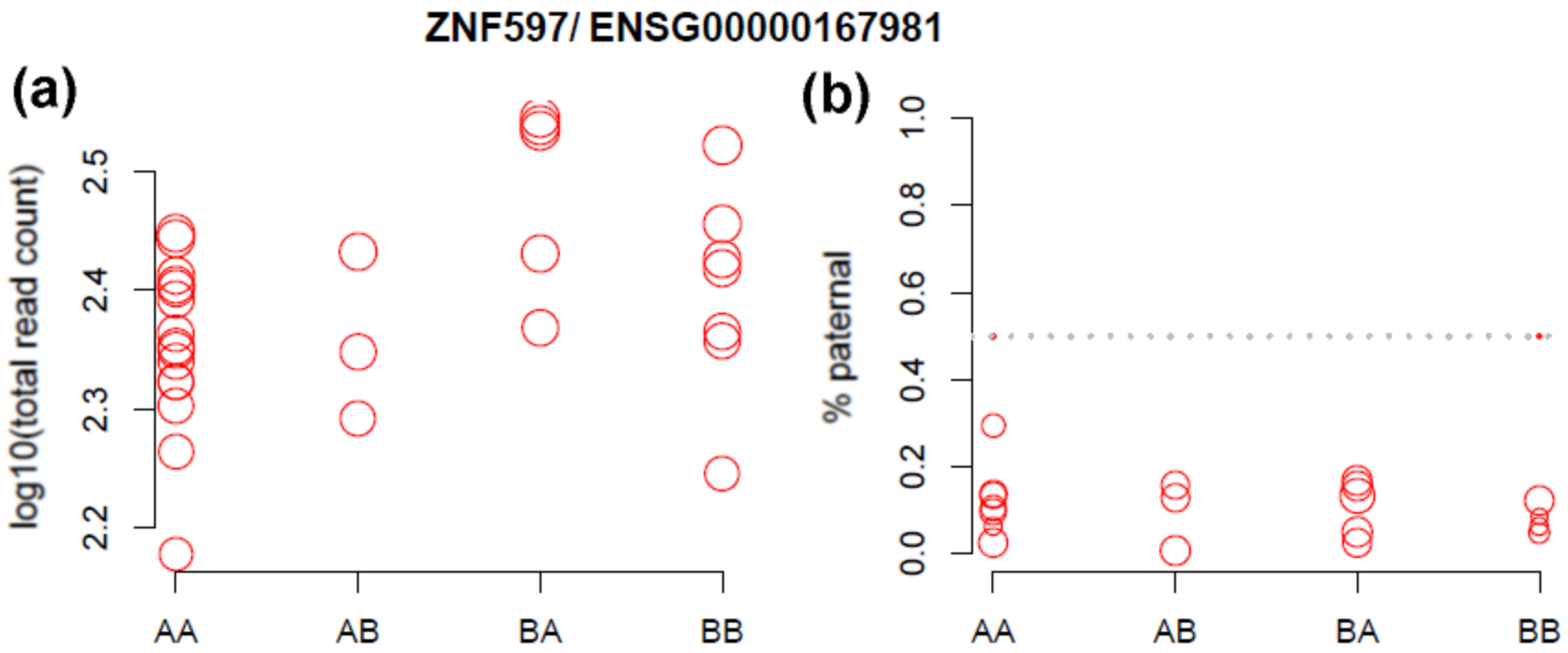
Data Collection and Processing

- Collected 30 HapMap Caucasian samples (15 females + 15 males); mapped with Tophat2 using hg38 reference
- Each of these samples as well as their parents are genotyped in the HapMap project; phased and imputed against 1000 Genomes reference panel
- Reads with at least one heterozygous SNP were classified to one of two parents
- Candidate *cis*-acting eQTLs were obtained from analysis of 227 European samples from Geuvadis consortium:
- 12,386 candidate genes with enough allele-specific counts and no strong *trans*-eQTL were identified

Data Analysis

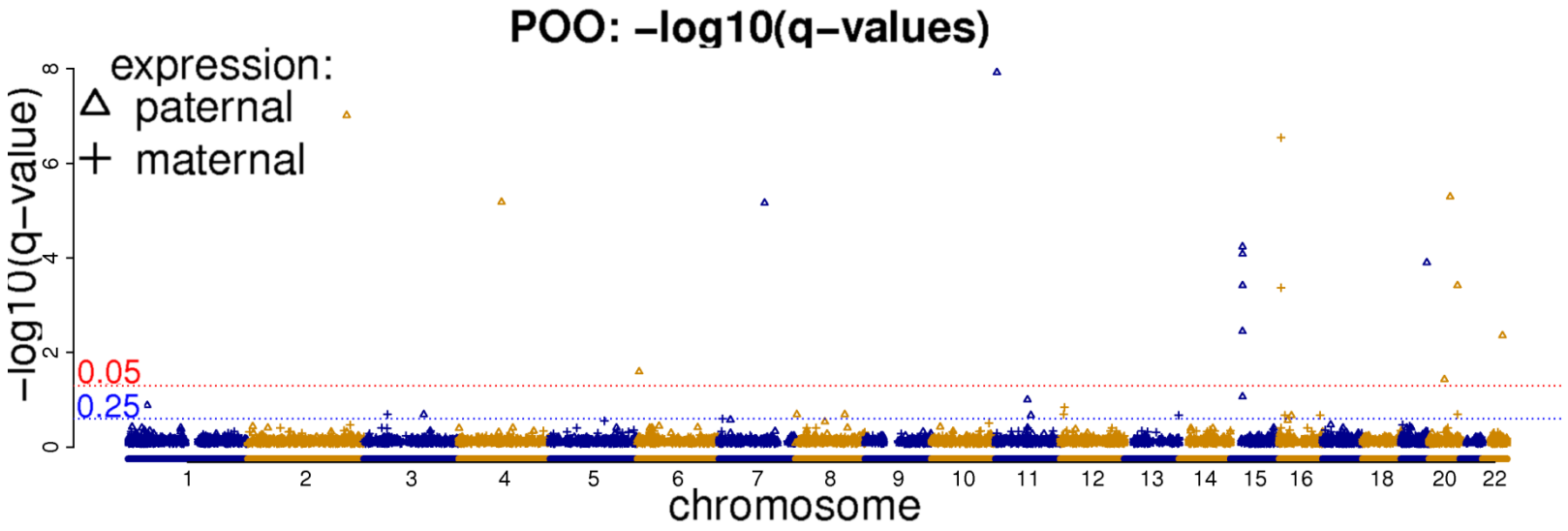
- For the total read counts we fitted the model with read-depth covariate and 3 batches by month of data collection (3 batches with 10 samples per batch)
- We found 16 genes with significant imprinting effects (q-value<0.05), out of which 6 were novel.
- 14 of 16 genes had higher paternal expression
- At FDR 0.25 we identified 15 more genes, 12 of which likely missed the cutoff due to power – 4 had smaller effect size, 8 had low allele-specific counts
- For those 12 genes 8 had higher paternal expression

A Gene with Parent-of-Origin Effect



BA means that B is maternal haplotype and A is paternal haplotype. This gene is clearly maternally expressed looking at both total and allele-specific counts.

Data Analysis (cont)



- Overall non-random distribution of parental imprinting:
  - Fisher test for a chromosome level same parent imprinting 0.02 (for q-val<0.25) or 0.04 (for q-val<0.5)
- Also, testing each chromosome separately (for q-val<0.25) we get a statistically significant result for chromosome 16.

Data Analysis: Known Imprinting

- 32 of known imprinted genes could be tested in our dataset.
- 10 were found to be significant (q-value < 0.05) by our method. For several other genes we observed signal of imprinting, but it was too weak to produce significant q-value.
- Overall we observed that even for insignificant results those with smaller q-values tend to have estimated imprinting direction matching with reported imprinting direction
- Genes classified in the database as “predicted imprinting” weren’t replicated in our analysis.

Summary

- We provide an extension to existing methods that would allow joint modeling of genetic and parent-of-origin effects for human RNA-seq data.
  - Method achieves better power by combining total and allele-specific counts.
  - The method we implemented in human data is capable of discovering parent-of-origin effects consistent with known imprinted genes
- Thanks to collaborators on this project:
  - Wei Sun (advisor), Kaoru Inoue, Terry Magnuson and Mauro Calabrese