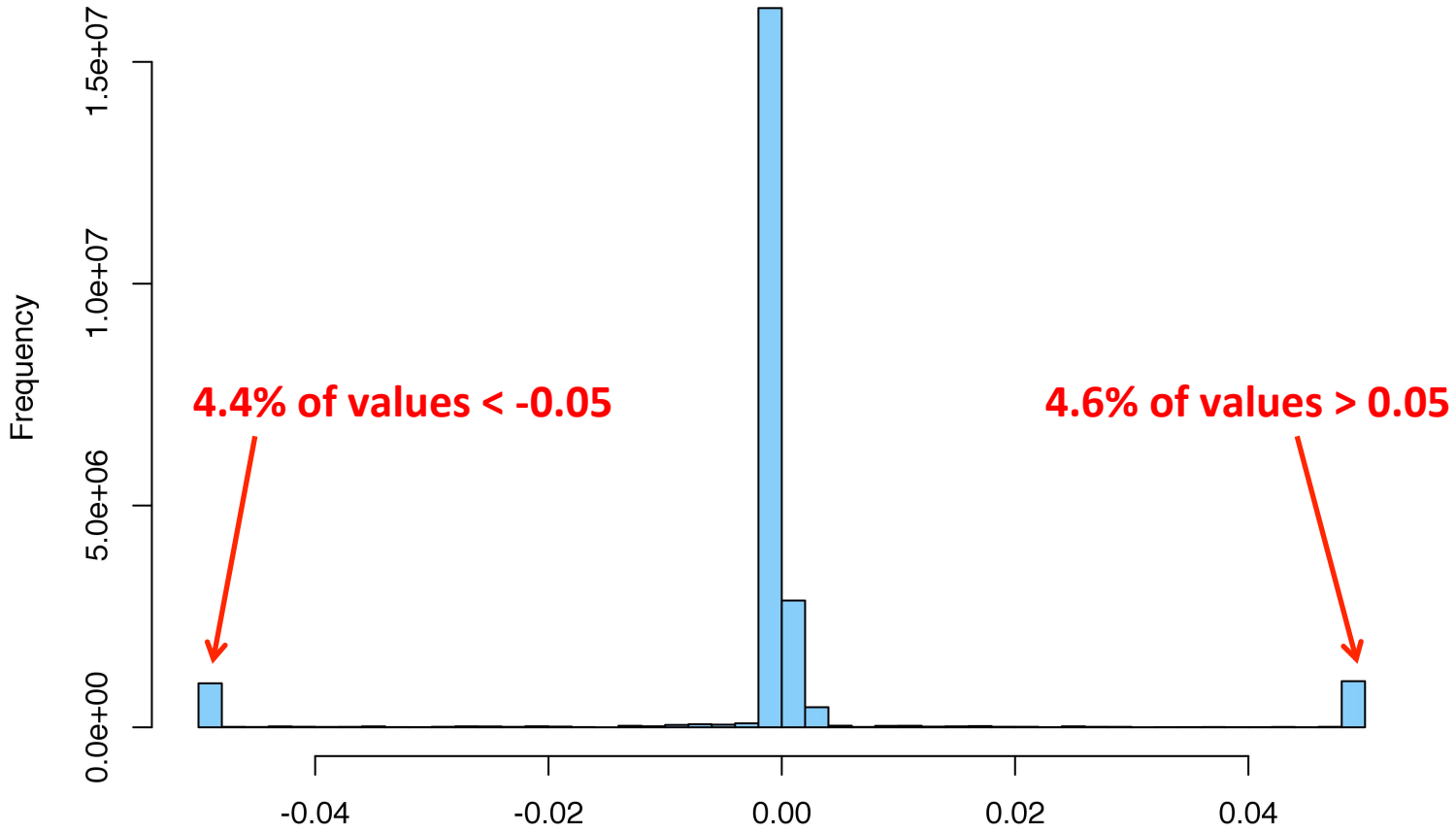
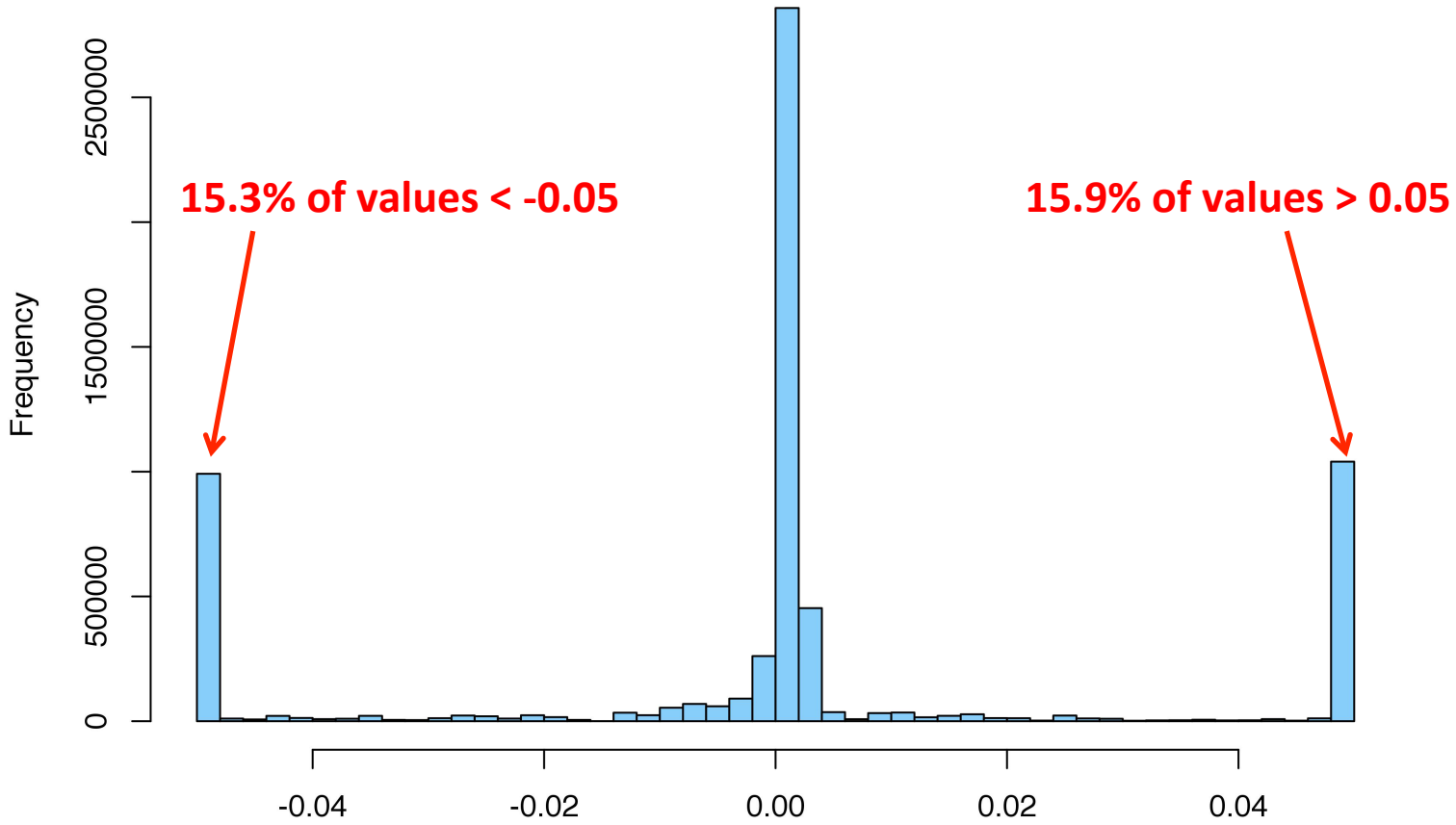
Distribution of all fractional errors*



$$\text{Fractional\_error} = [\text{RPKM}_{calculated} - \text{RPKM}_{true}] / \text{RPKM}_{true}$$

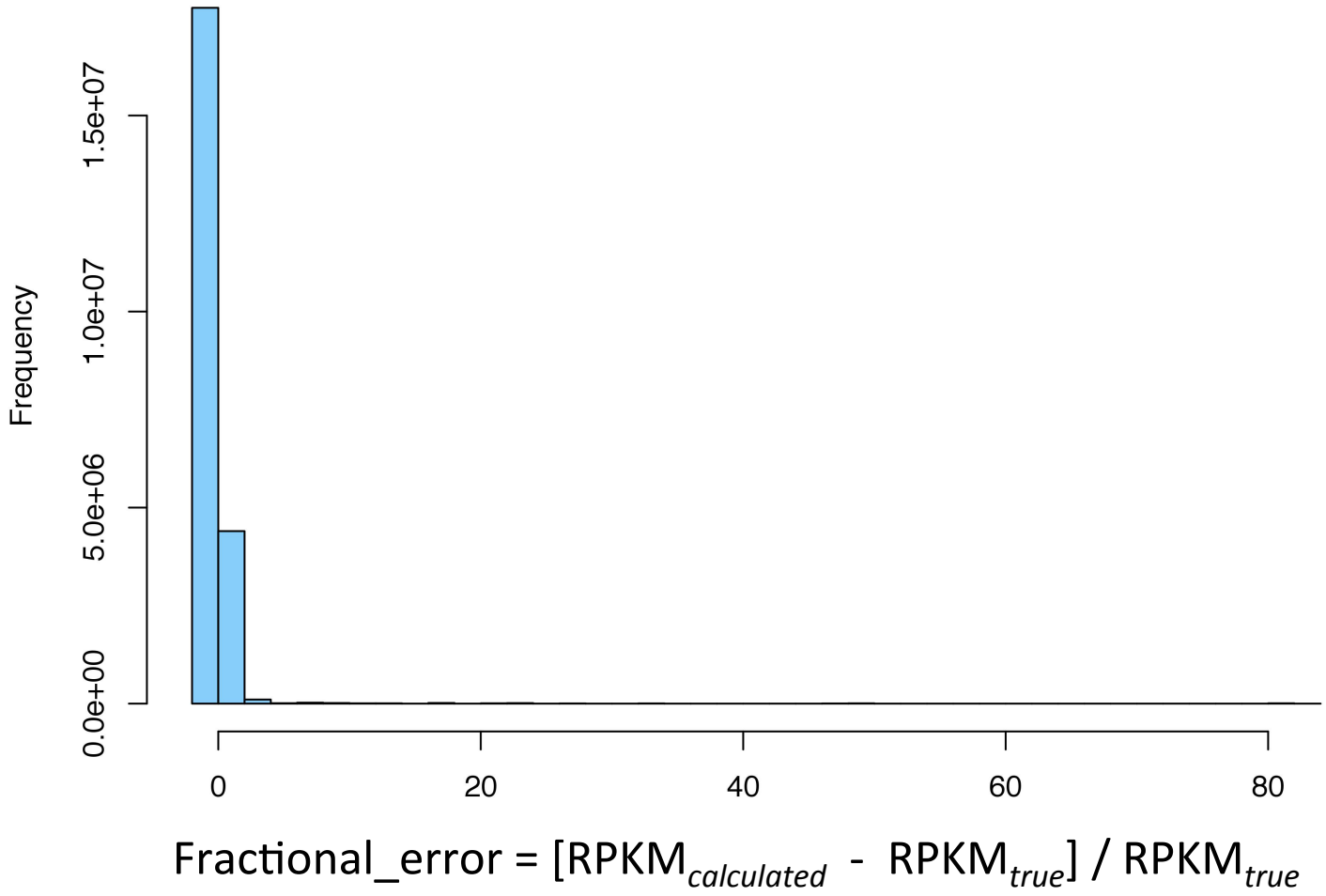
*for single-exon genes

Distribution of all *non-zero* fractional errors

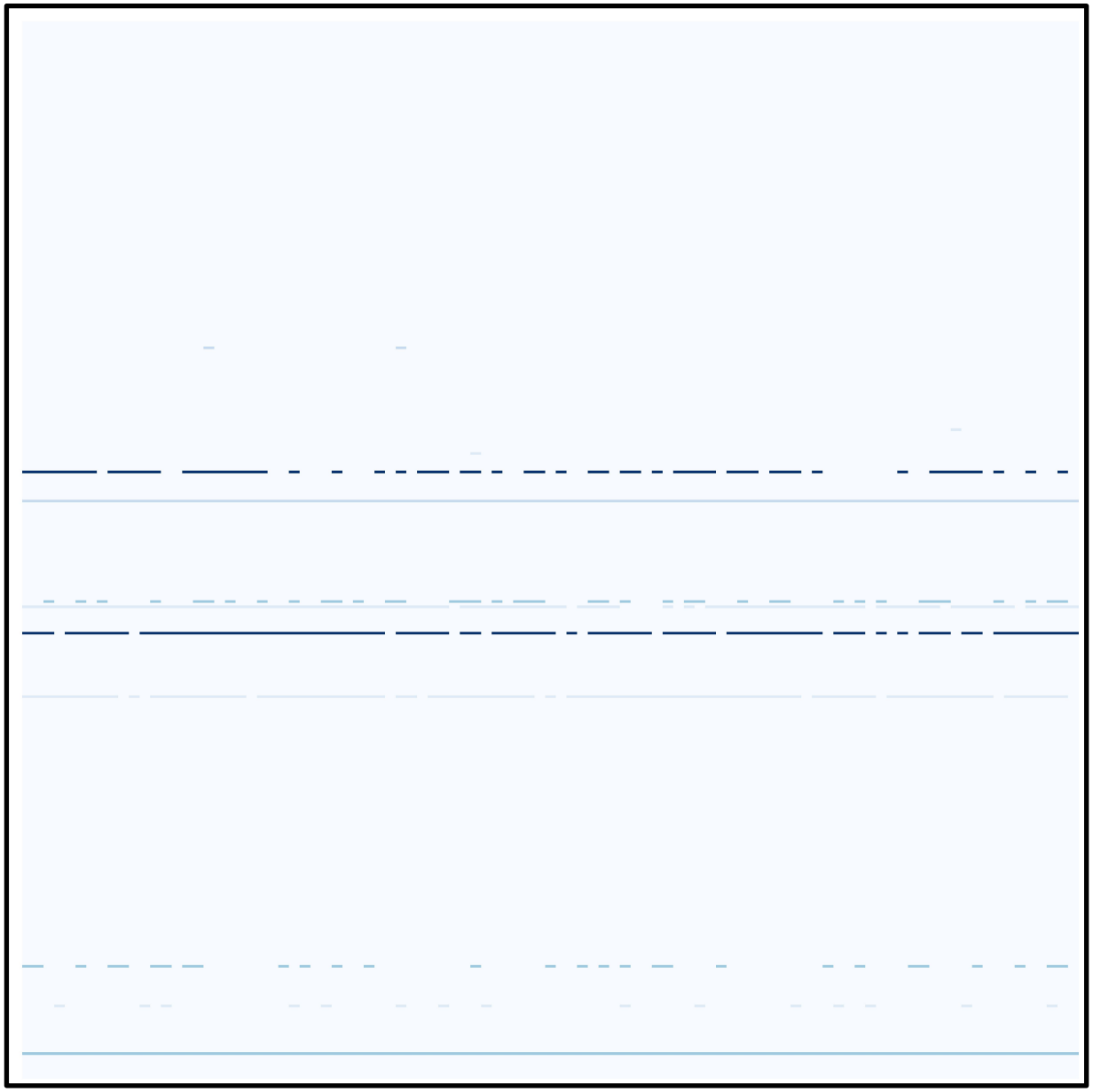


Fractional_error = [$RPKM_{calculated}$ - $RPKM_{true}$] / $RPKM_{true}$

The substantial errors in calculated values are very large *over*-estimates



Specific genes are the culprits (not samples)



Potential causes of outlier error rates

- single-exon genes (often associated w/pgenes) → duplications → low mappability scores
- low mappability scores? → check using intersect bed w/encode in UCSC Genome Browser. Genome browser has a track for mappability → first download and check w/intersect bed?
- exon lengths from same genome build? In any case GTEx is reporting in GENE read counts
- With BAM file as input, GTEx uses RNA-SeQC: "Expression levels were produced at the gene and exon level in RPKM units using RNA-SeQC"
  → Black box & confounding factors (GC bias, mapability, uniqueness, etc)

Misc Notes

- Strange that *processed* read counts data are not available at the GTEx Portal
- BAM files not available to re-compute RPKM from RNA-SeqC
- GTEx: tophat/bowtie, though will be STAR 2.4.2a in v7 (CommonMind= STAR)
- PsychENCODE currently processing all to be uniform?