1) Process conglomerated RPKM data file to generate RPKM counts for each tissue on relevant samples

2) Filter on >=10 individuals with >0.1 RPKM and raw read counts greater than 6 -- ie, Genes must:

- have at least 10 samples with
    - RPKM > 0.1 and
    - raw read counts greater than 6"

→ **GTEx is applying other (unlisted) criteria -- Had to do ultimately enforce that my data matched the GTEx data (in terms of samples & genes)**

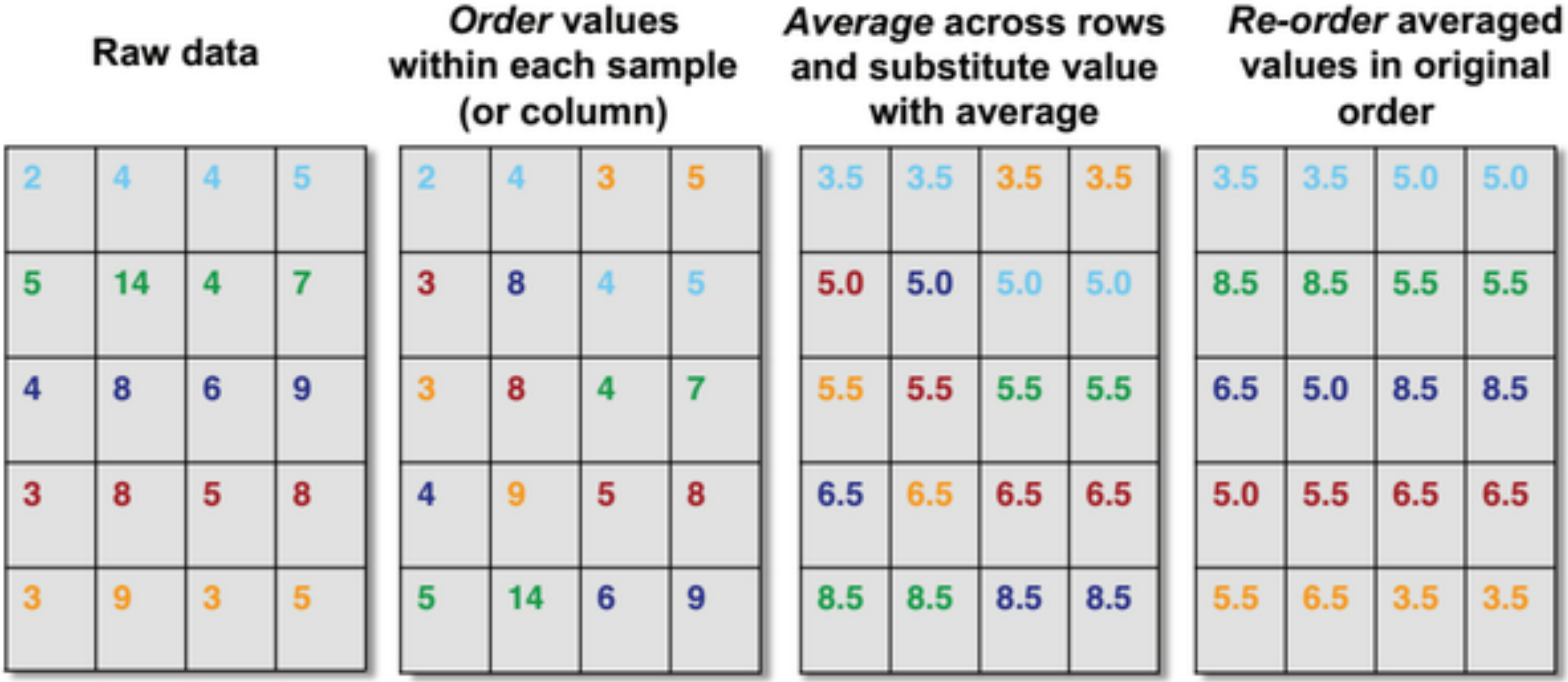
3) Quantile normalization

4) Inverse quantile normalization

5) Compare calculated expression values w/those reported by GTEx

3) *“Quantile normalization was performed within each tissue to bring the expression profile of each sample onto the same scale.”*

→ This makes the distributions (for different samples) similar in terms of statistical properties



Raw data

2	4	4	5
5	14	4	7
4	8	6	9
3	8	5	8
3	9	3	5

***Order* values within each sample (or column)**

2	4	3	5
3	8	4	5
3	8	4	7
4	9	5	8
5	14	6	9

***Average* across rows and substitute value with average**

3.5	3.5	3.5	3.5
5.0	5.0	5.0	5.0
5.5	5.5	5.5	5.5
6.5	6.5	6.5	6.5
8.5	8.5	8.5	8.5

***Re-order* averaged values in original order**

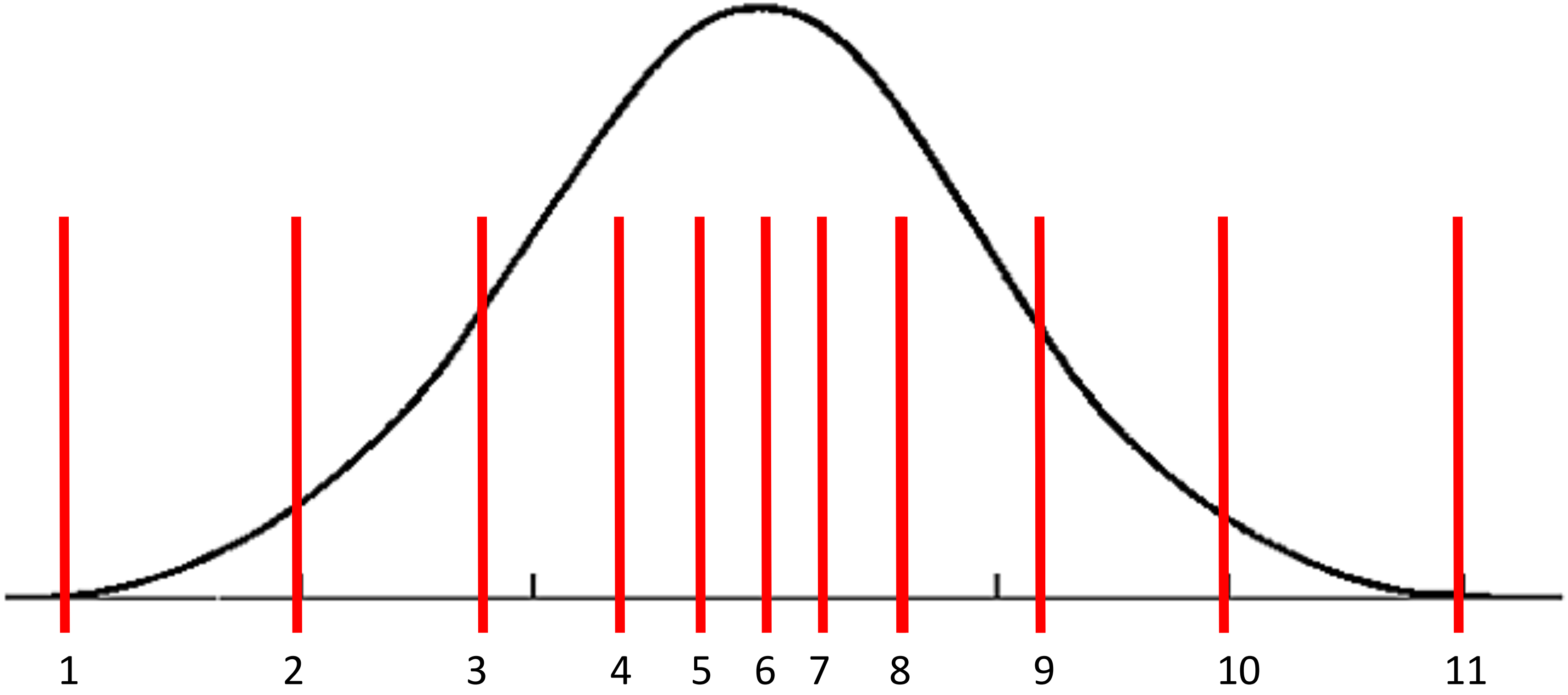
3.5	3.5	5.0	5.0
8.5	8.5	5.5	5.5
6.5	5.0	8.5	8.5
5.0	5.5	6.5	6.5
5.5	6.5	3.5	3.5

Genes (RPKM values)

Samples

Adapted from RA Irizarry, Dec 18 2014 (web post)

4) "To protect from outliers, inverse quantile normalization was performed for each gene, mapping each set of expression values to a standard normal."

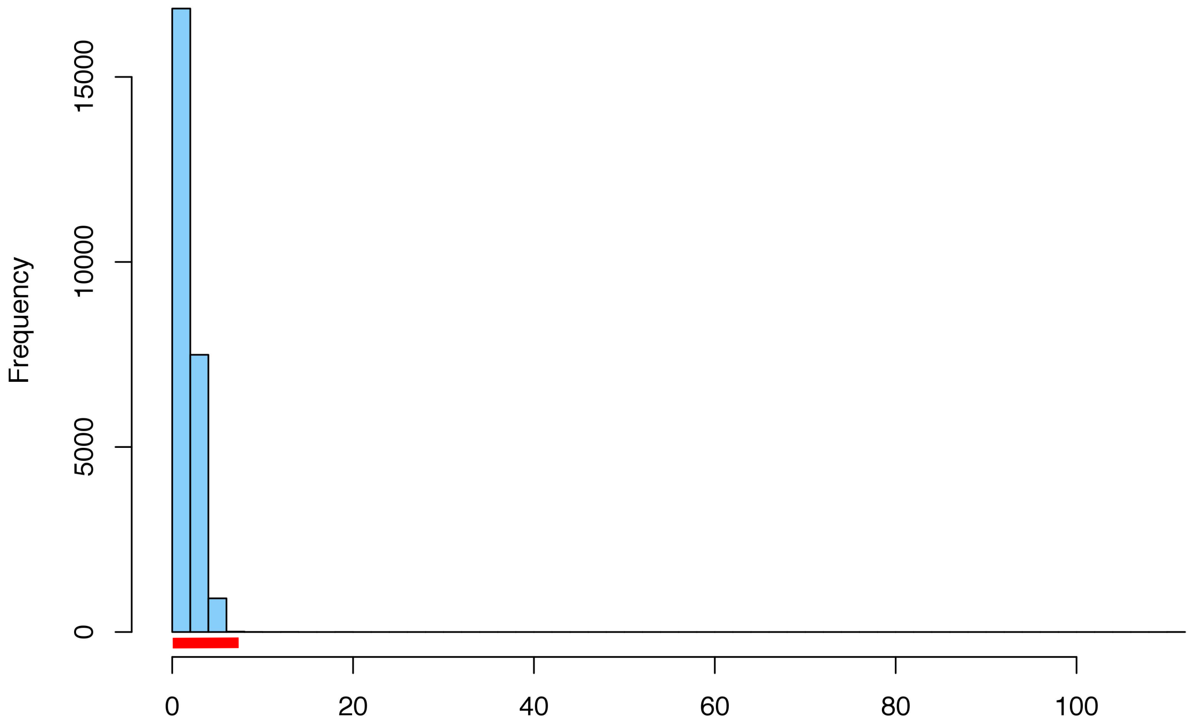


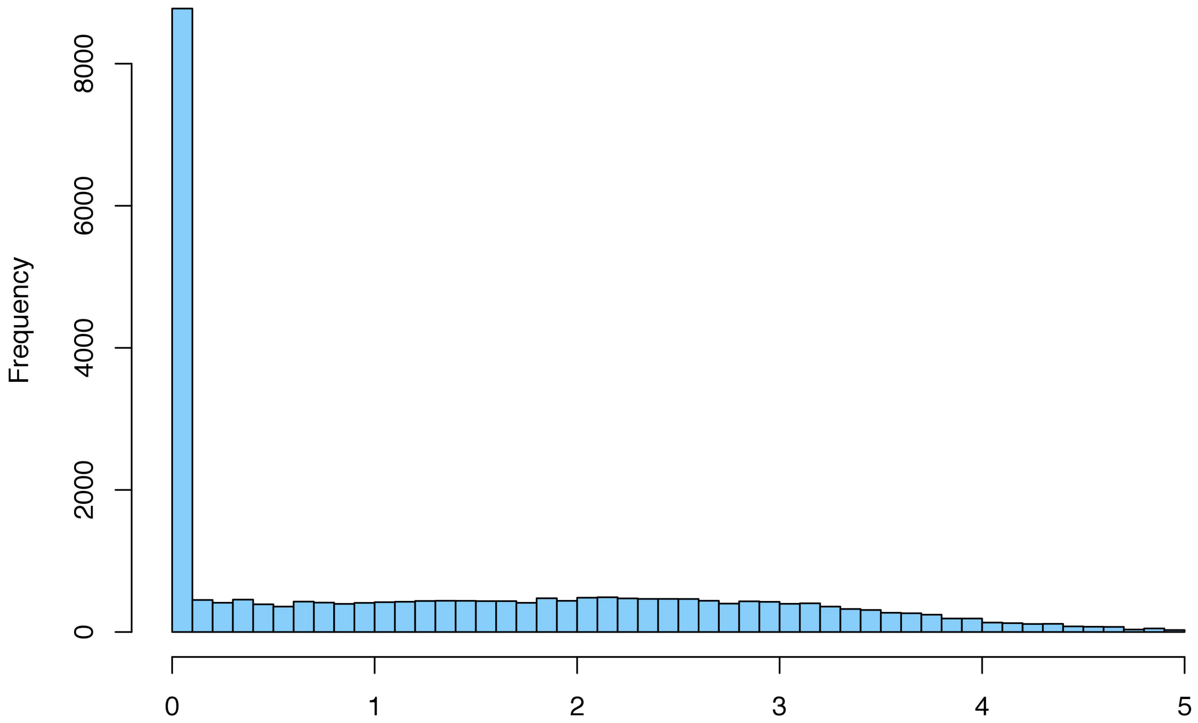
Toy example for 1 gene with 11 samples (ie, expression values)

→ split integral of standard normal up into 12 equal-area slices

→ Then re-assign the original expression values (from previous step) to these 11 values from the standard normal distribution

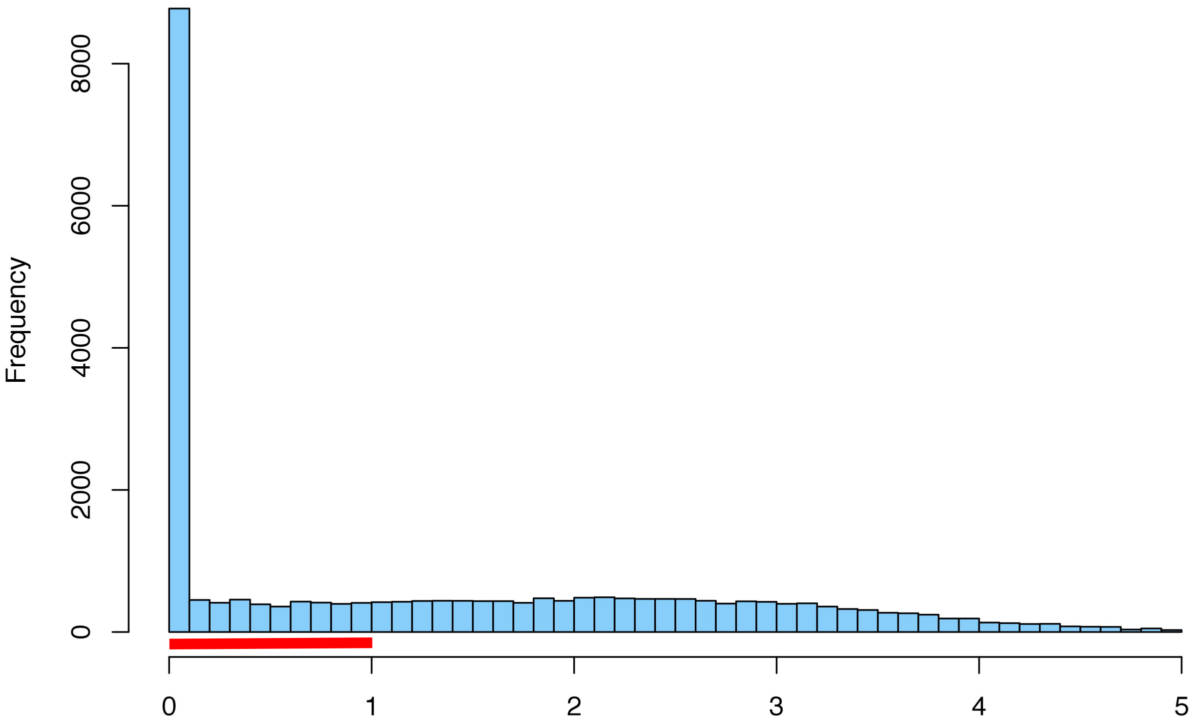
5) Compare calculated expression values w/those reported by GTEx

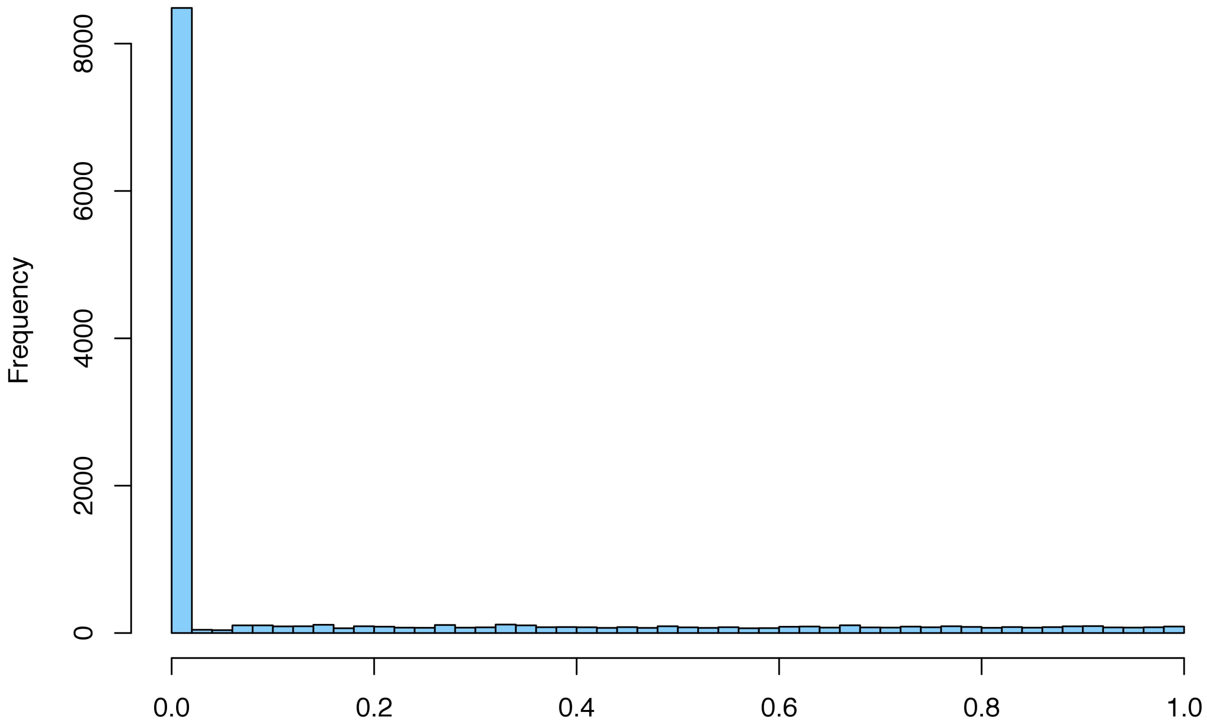




Error for a given gene ≡ ⟨ |fract_error| ⟩
where:
|fract_error| = abs { known– calculated / known }
and the mean ⟨ ⟩ is taken over all 96 samples

5) Compare calculated expression values w/those reported by GTEx



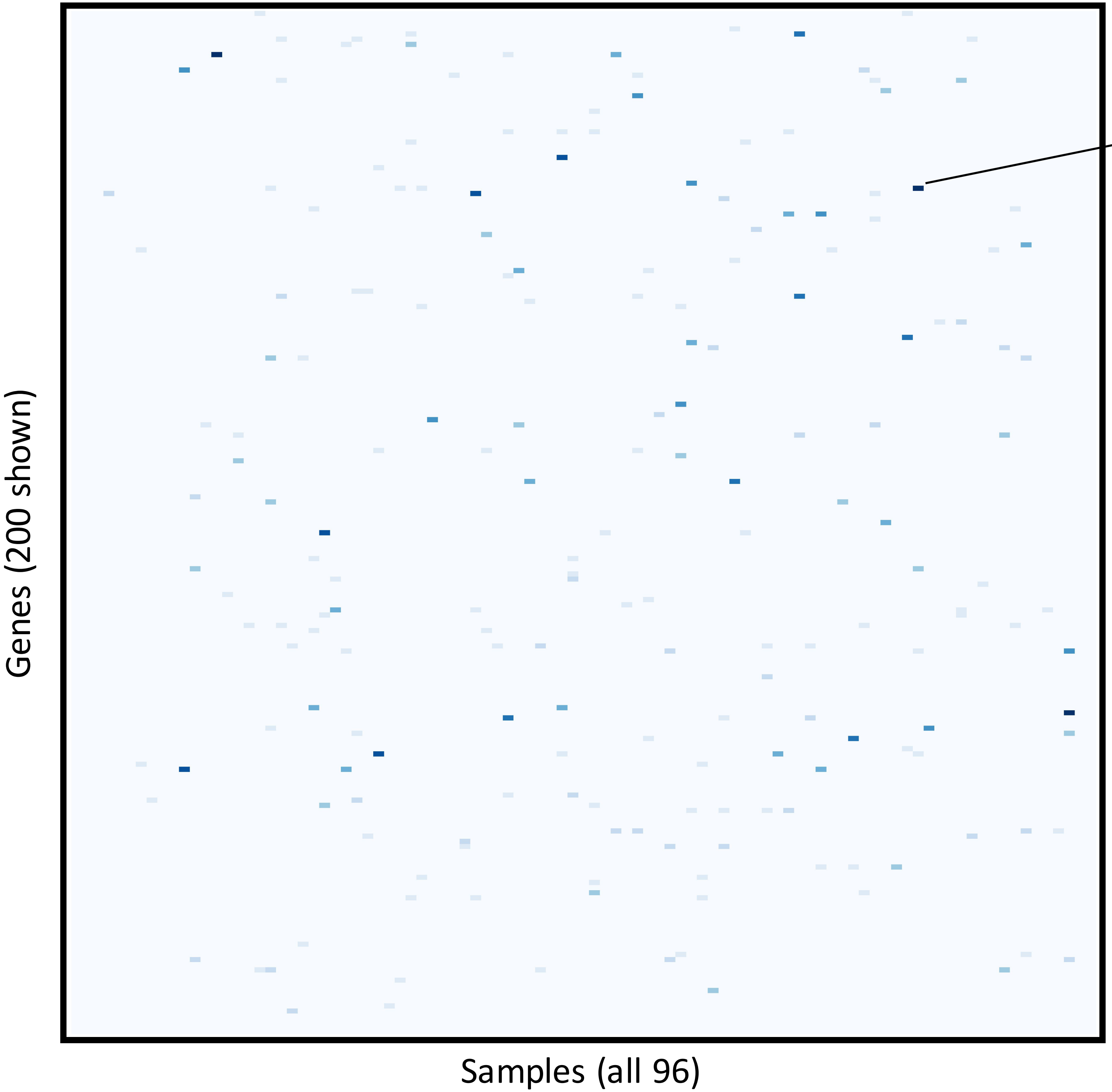


Error for a given gene ≡ ⟨ |fract_error| ⟩

where:

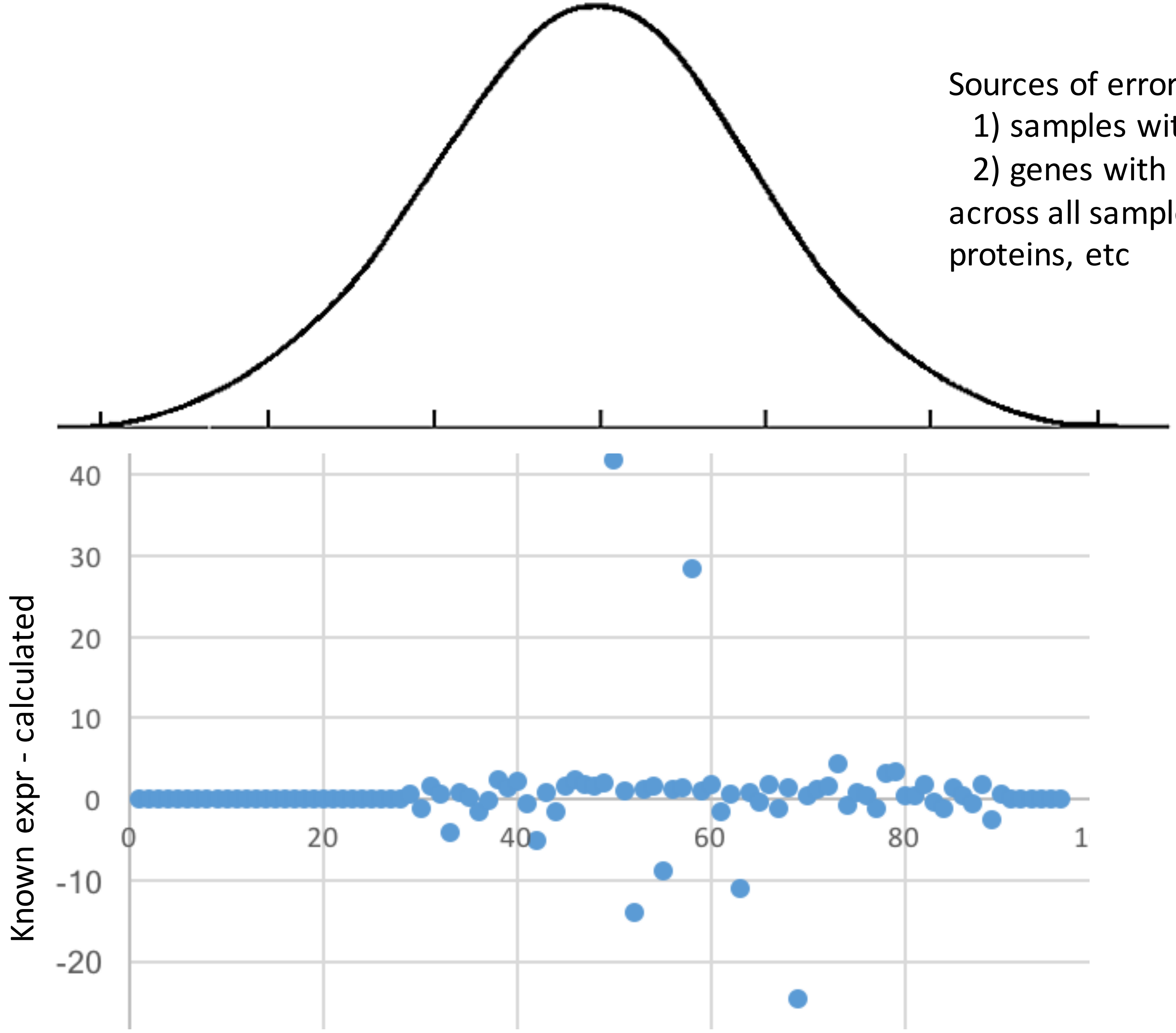
|fract_error| = abs { known– calculated / known }

and the mean ⟨ ⟩ is taken over all 96 samples



|fract_error|
= abs { known– calculated / known }

Max value in this heatmap: 180



Sources of error:

1) samples with expression values close to 0

2) genes with extremely high expression values across all samples (not shown) -- ex: ribosomal proteins, etc