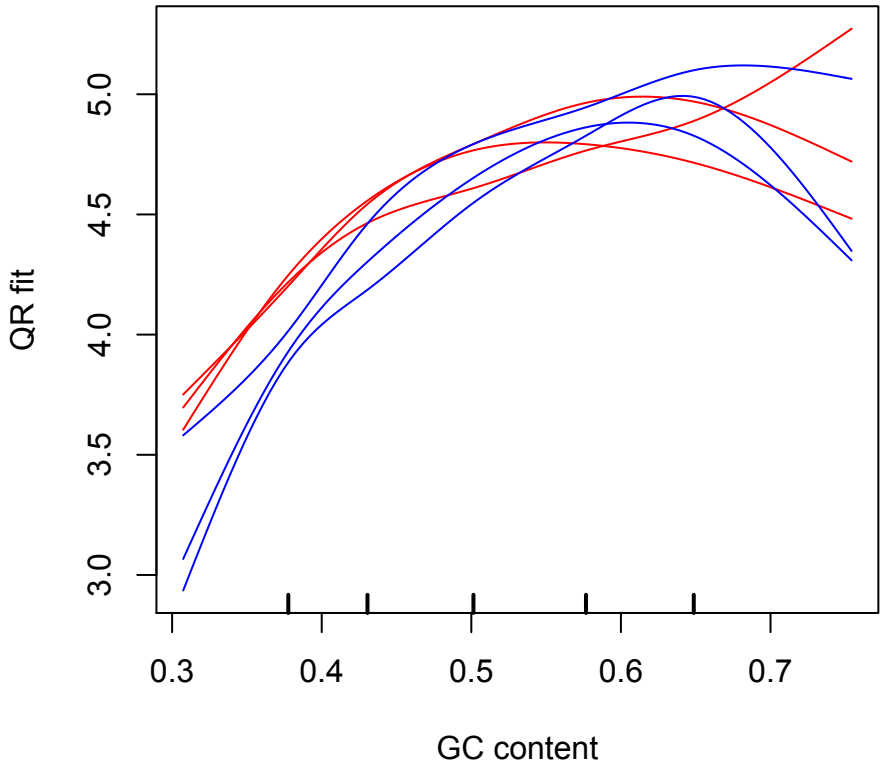
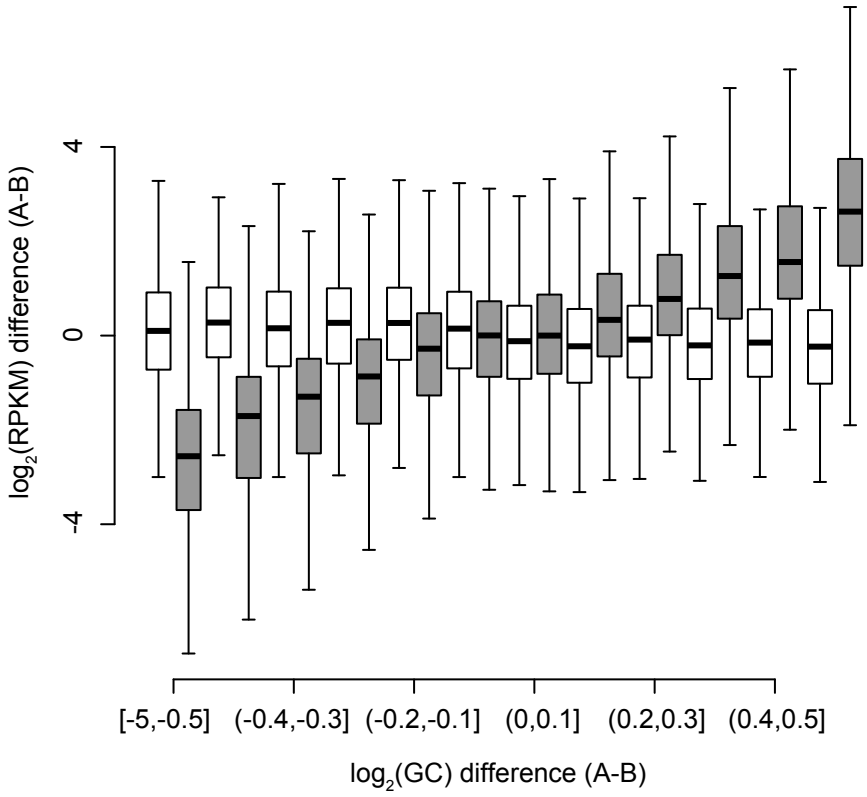
Supplementary material for Removing technical variability in RNA-seq data using conditional quantile normalization

Kasper D. Hansen[1], Rafael A. Irizarry[1], and Zhijin Wu[2]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA
[2]Department of Community Health, Section of Biostatistics, Brown University



Supplementary Figure 1. The effect of GC content for sample in the Pai data. The estimated effect of GC content for the 6 samples in the Pai data is shown (red is female, blue is male). Otherwise as Figure 3(c). This is a small sample dataset using primary tissue samples as opposed to cell lines. We observe very different effect of GC content for the 6 samples.



Supplementary Figure 2. The effect of GC content on each side of a gene. Shown are data from two samples from the Montgomery data (NA11918, NA12761) also depicted in Figure 1(b), using the same set of genes (ie. genes with average $\log_2$ RPKM values greater than 2). Each gene was split into two halves of equal length, labeled "A" and "B". For each half we obtained GC content of the gene and computed $\log_2$ RPKM. The difference depicted along the x-axis is the difference between $\log_2$ of the GC content of the two halves. The difference depicted along the y-axis is the difference between $\log_2$ of the RPKM of the two halves. For one of the two samples we observe that if two gene halves have very different GC content, they will have different RPKM values, reflecting the relationship between GC content and RPKM depicted in Figure 1(b). If a gene is inherently highly or lowly expressed, we would expect both halves of the gene to have the same expression level, despite possible differences in GC content. We conclude that the effect of GC content on RPKM is likely to be a technical artifact.