Homework Problem 1

BATCH	N_SAMPLES	N_CANCER	N_NORMAL
1	11	11	0
2	18	14	4
3	4	0	4
4	5	0	0
5	19	15	0

Key problems:

Imbalanced distribution:

> Batches 1, 3, 5 are entirely composed of either cancer or normal samples -> It creates a strong association between batch and biological variables.
>
> This can lead to confounding -> batch effects are misinterpreted as biological signals.

Small sample sizes:

> Batches 3, 4 have very few samples -> It reduce statistical power and increase the risk of overfitting.

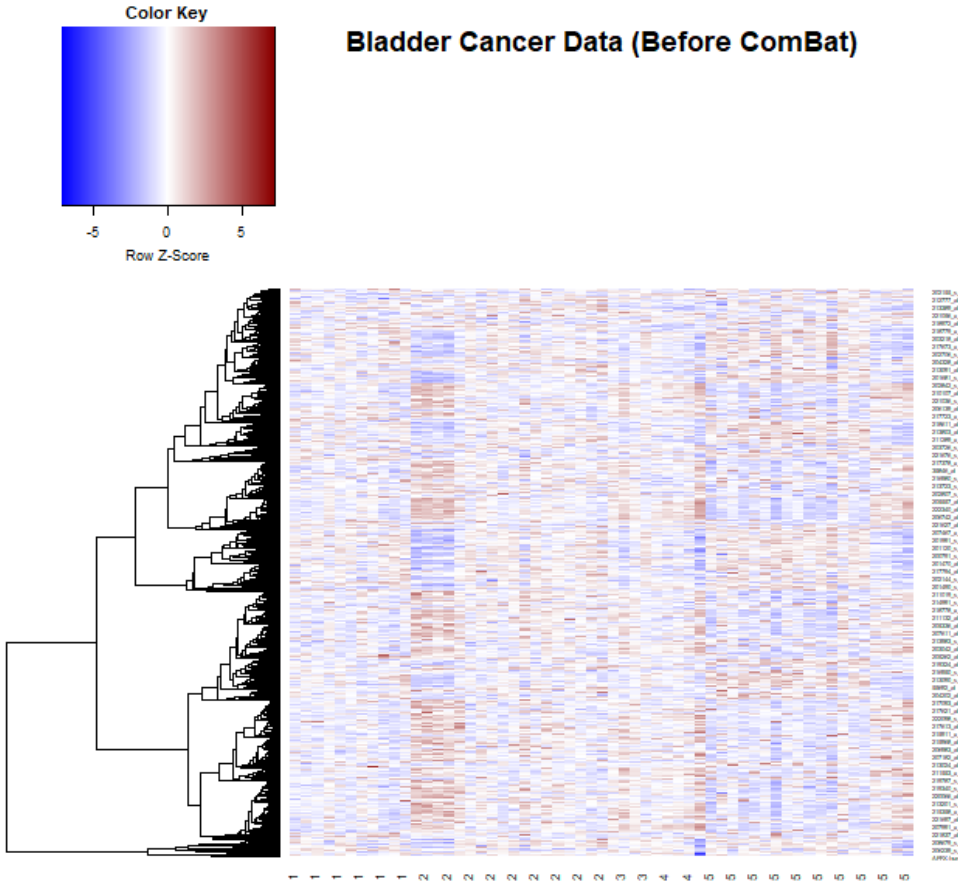
Missing or mislabeled data:

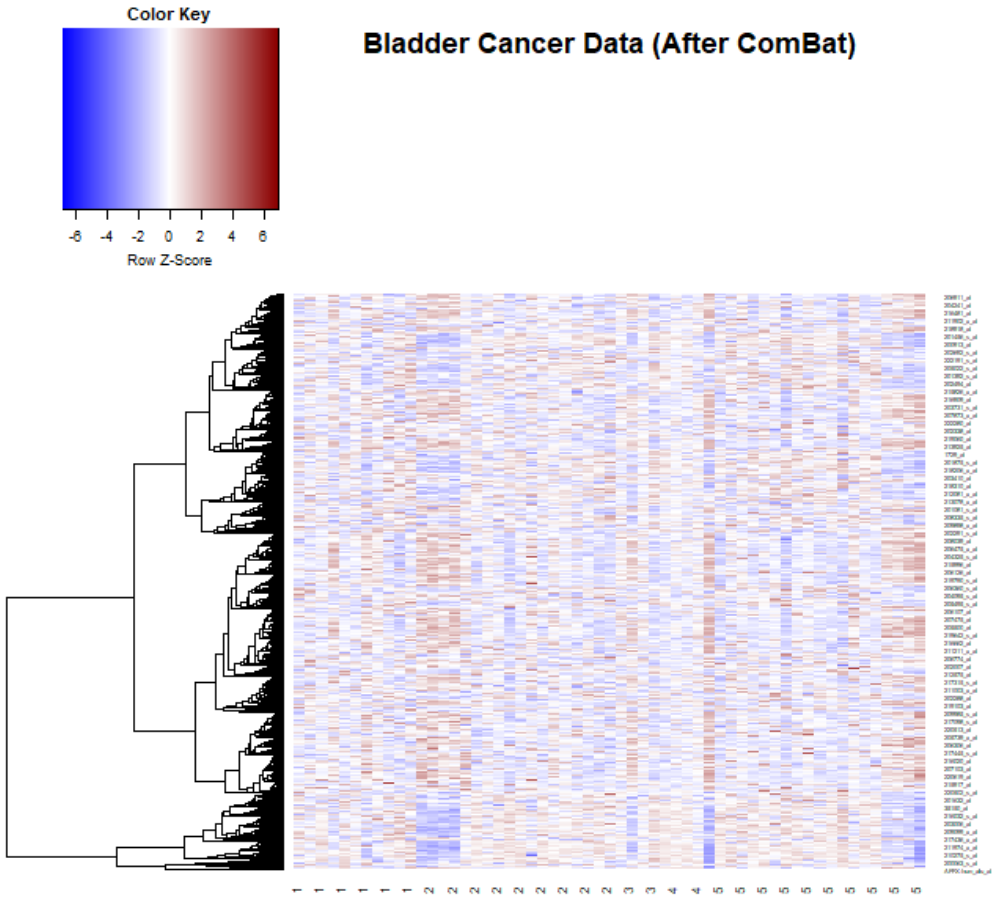
> Batch 4 contains no cancer or normal samples -> It suggests potential issues with data labeling or sample collection.

Batch effects masking biological signals:

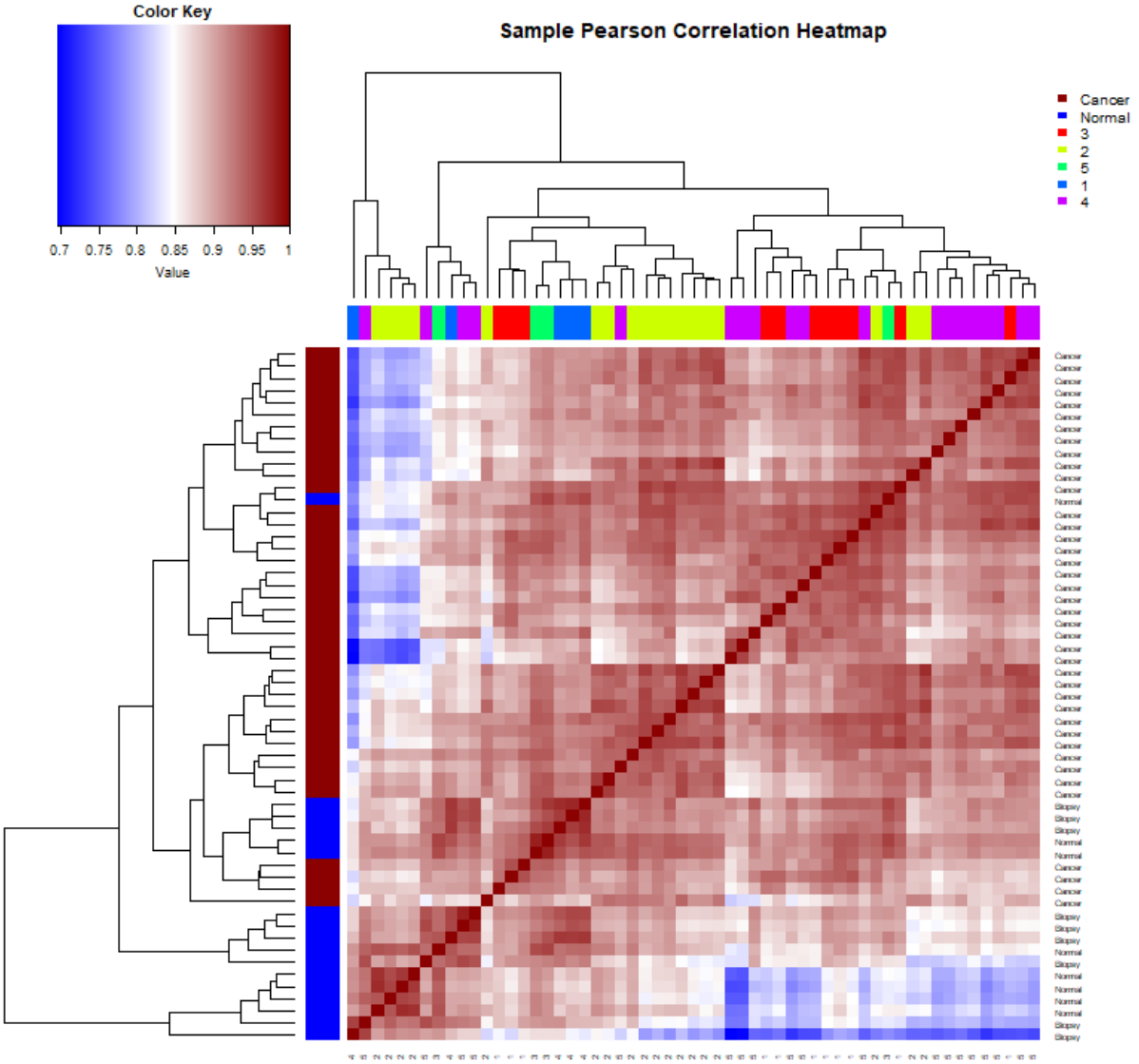
> The strong association between batch and cancer status means that batch effects could mask true biological differences between cancer and normal samples.

Homework Problem 2ś

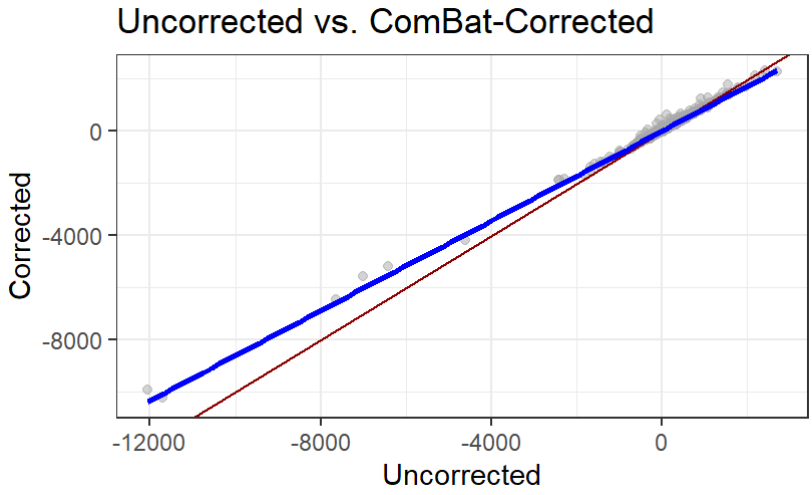


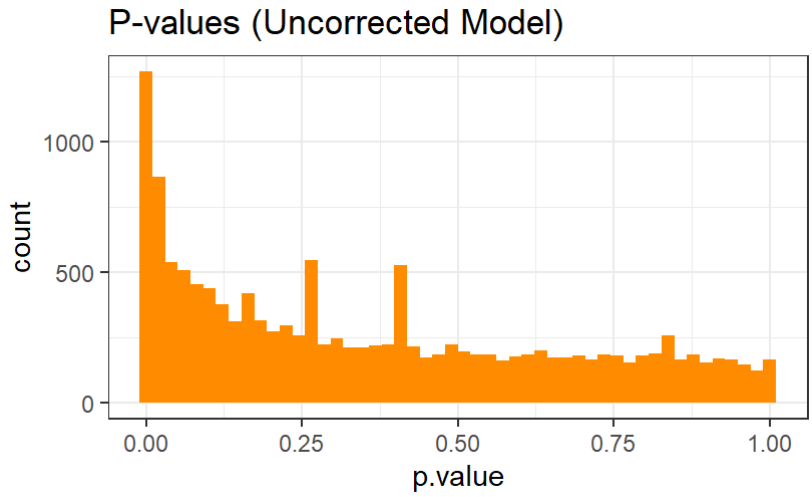


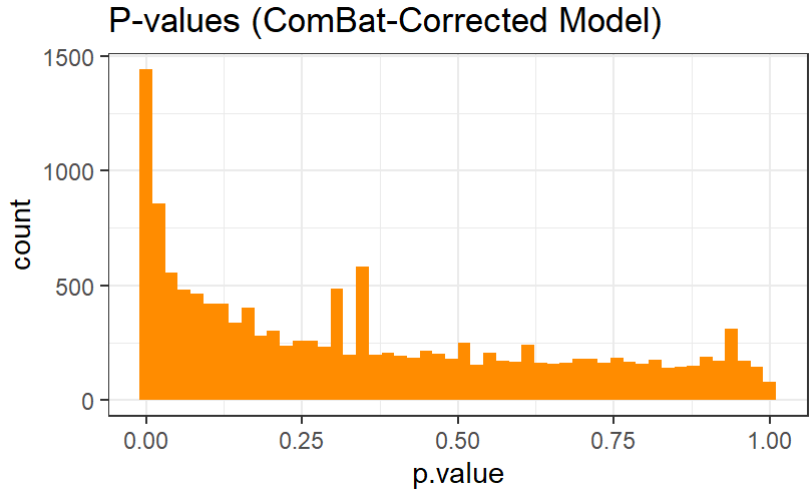
Homework Problem 3



Homework Problem 4







Homework Problem 5