Question 1
How many genes are left after removing genes with zero expression in all samples?
23,259

Question 2
How many genes are left after removing genes where 20 or more samples have cpm < 1?
14,464

Question 3
What is the range of library sizes (min, max)?
(11,464,513, 20,839,468)

Question 4
What is the range of library sizes (min, max) after normalization?
(13,671,808.8746, 22,214,610.4758)

Question 5
Compare the two library size bar charts you made. How did the distribution of library sizes change after normalization? Briefly discuss why it is important to normalize your RNA-seq data.
After normalization, the range and variation of library sizes was reduced. Normalization is important because it can reduce bias caused by artifacts of sample preparation and measurement.

Question 6
What are the top ten differentially expressed genes according to your FLD analysis? (Copy and paste your function's output.) Do these genes make sense given the tissue and groups in the experiment?
Highest FLD:

```
AK000953: 1.018487695
RAB30:    0.988243073247
DBNDD1:   0.957612220352
CTDSPL:   0.915162216643
GRB14:    0.885166790019
YTHDC2:   0.851677433058
MYLK4:    0.842607579064
ABCG1:    0.746580252103
FNDC5:    0.739653965436
```

BC010186: 0.737282716401
Yes - many of these genes are involved with muscles and growth signaling, which makes sense given that the samples come from muscle tissue and were collected before and after a year-long exercise regimen.

Question 7
Does your result point toward one gene with large effect or many genes with small effects? Does RNA-seq expression data always give researchers a clear answer?
The result points toward many genes with small effects. RNA-seq expression data does NOT always give a clear answer. From Heather: The lesson from this study is that any systemic metabolic consequences from skeletal muscle are the result of many genes of small effect

rather than a handful of big effect changes in gene expression.

Question 8
How does the study design of this experiment relate to the assumptions made when studying gene expression data?
We assume that gene expression changes relate to something interesting happening biologically. In this study, there were positive changes in systemic insulin and glucose levels in individuals at risk of developing type-2 diabetes. By comparing the two groups before and after the exercise program, we assume that differences in gene expression are related to the health benefits derived.

Question 9
If you were going to spend time and money following up on one of these top ten genes, what would be your candidate and why? (There could be many correct answers.)
DBNDD1. Dystrobrevin is a protein that binds to dystrophin in the costamere of skeletal muscle cells (wikipedia).

Question EC.1
What do you expect to see?
I expect to see the before group and the after group form their own branches of the dendrogram, or that before and after of the same sample to be most closely related.

Question EC.2
What did you actually see? If you did not find what you expected, what sorts of variation could account for this?
I saw no particular order in the dendrogram. Biological and technical variation could account for this.