Team Abracadabra- Sam Fenske and Sarah Mughal

1. What is the necessity of normalizing the microarray data?
    a. Normalization plays an important role in the beginning stages of microarray data analysis. It is the process by which sources of systematic or non-biological variation are removed/ minimized from samples that would otherwise affect the gene expression levels. Sources of variation may arise from discrepancy in replicate slides, scanning conditions, hybridization conditions, as well as different technicians conducting lab work.

2. A comma-separated file containing the normalized gene expression values. Write out the expression data to a CSV file using the exprs() and write.csv() functions, where rows are the probe sets and columns are the samples.

```
# normalizing data (step 4) ------------------------------

rma(raw)

# CSV file for normalized data (step 4) ------------------

nraw <- exprs(nraw)
write.csv(nraw, file="nraw.csv")
```

3. Visualize the analysis report from the QC method you used and explain it.
    a. To find any correlation between gene samples, our team used the arrayQualityMetrics package. The .html report produced as a result was then further analyzed to determine the presence of outliers, which were found to be GSM800742 (array 1), GSM800751 (array 10), GSM800754 (array 13), GSM800758 (array 17), and GSM800775 (array 34). These outliers are visible throughout the figures and boxplots in the produced report. In figure 1, which plots the distance between arrays, array 1 and 10 are apparent outliers. In figure 10, which plots Normalized Unscaled Standard Error, array 1 and 34 were determined to be outliers. In figure 16, which plots spatial distribution of M, arrays 13 and 17 were determined to be outliers. After normalization using RMA, the number of outliers was reduced by 2 as seen in the arrayQualityMetric reports below.

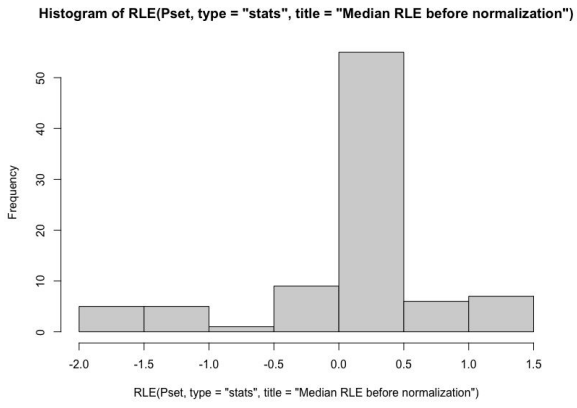
	array	sampleNames	*1	*2	*3	*4	*5	*6	sample	ScanDate
☑	1	GSM800742_chip_array_C06N-H.CEL	x			x			1	08/01/08 12:23:56
☐	2	GSM800743_chip_array_C11N-H.CEL							2	08/01/08 12:35:17
☐	3	GSM800744_chip_array_C24N-H.CEL							3	08/07/08 17:15:10
☐	4	GSM800745_chip_array_C27N-H.CEL							4	08/01/08 13:52:28
☐	5	GSM800746_chip_array_C28N-H.CEL							5	08/01/08 14:03:46
☐	6	GSM800747_chip_array_C30N-H.CEL							6	08/01/08 15:00:05
☐	7	GSM800748_chip_array_C31N-H.CEL							7	08/01/08 15:11:16
☐	8	GSM800749_chip_array_C32N-H.CEL							8	08/01/08 15:22:42
☐	9	GSM800750_chip_array_C33N-H.CEL							9	08/01/08 15:33:53
☑	10	GSM800751_chip_array_C35N-H.CEL	x	x					10	08/07/08 12:17:15
☐	11	GSM800752_chip_array_C36N-H.CEL							11	08/07/08 12:29:11
☐	12	GSM800753_chip_array_C38N-H.CEL							12	08/07/08 12:40:31
☑	13	GSM800754_chip_array_C41N-H.CEL					x		13	08/07/08 12:51:39
☐	14	GSM800755_chip_array_C42N-H.CEL							14	08/07/08 13:48:28
☐	15	GSM800756_chip_array_C44N-H.CEL							15	08/07/08 13:59:57
☐	16	GSM800757_chip_array_C45N-H.CEL							16	08/07/08 14:11:08
☑	17	GSM800758_chip_array_C47N-H.CEL					x		17	07/12/07 12:09:45
☐	18	GSM800759_chip_array_C06T-H.CEL							18	10/18/07 14:39:39
☐	19	GSM800760_chip_array_C11T-H.CEL							19	10/18/07 14:50:56
☐	20	GSM800761_chip_array_C24T-H.CEL							20	10/18/07 16:09:16
☐	21	GSM800762_chip_array_C27T-H.CEL							21	10/18/07 16:32:30
☐	22	GSM800763_chip_array_C28T-H.CEL							22	10/18/07 16:43:48
☐	23	GSM800764_chip_array_C30T-H.CEL							23	10/19/07 12:00:32
☐	24	GSM800765_chip_array_C31T-H.CEL							24	10/19/07 12:11:51
☐	25	GSM800766_chip_array_C32T-H.CEL							25	10/19/07 12:22:53
☐	26	GSM800767_chip_array_C33T-H.CEL							26	10/19/07 12:34:04
☐	27	GSM800768_chip_array_C35T-H.CEL							27	10/19/07 13:26:46
☐	28	GSM800769_chip_array_C36T-H.CEL							28	10/19/07 13:38:32
☐	29	GSM800770_chip_array_C38T-H.CEL							29	10/19/07 13:50:06
☐	30	GSM800771_chip_array_C41T-H.CEL							30	10/19/07 14:01:29
☐	31	GSM800772_chip_array_C42T-H.CEL							31	10/19/07 14:57:14
☐	32	GSM800773_chip_array_C44T-H.CEL							32	10/19/07 15:08:27
☐	33	GSM800774_chip_array_C45T-H.CEL							33	10/19/07 15:19:46
☑	34	GSM800775_chip_array_C47T-H.CEL				x			34	10/19/07 15:31:03

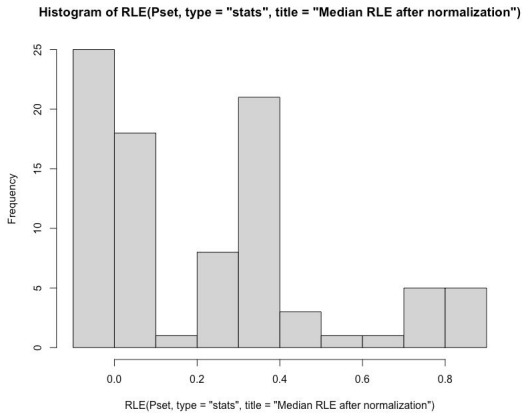
Raw Data

	array	sampleNames	*1	*2	*3	sample	ScanDate
☑	1	GSM800742_chip_array_C06N-H.CEL		x		1	08/01/08 12:23:56
☐	2	GSM800743_chip_array_C11N-H.CEL				2	08/01/08 12:35:17
☐	3	GSM800744_chip_array_C24N-H.CEL				3	08/07/08 17:15:10
☐	4	GSM800745_chip_array_C27N-H.CEL				4	08/01/08 13:52:28
☐	5	GSM800746_chip_array_C28N-H.CEL				5	08/01/08 14:03:46
☐	6	GSM800747_chip_array_C30N-H.CEL				6	08/01/08 15:00:05
☐	7	GSM800748_chip_array_C31N-H.CEL				7	08/01/08 15:11:16
☐	8	GSM800749_chip_array_C32N-H.CEL				8	08/01/08 15:22:42
☐	9	GSM800750_chip_array_C33N-H.CEL				9	08/01/08 15:33:53
☐	10	GSM800751_chip_array_C35N-H.CEL				10	08/07/08 12:17:15
☑	11	GSM800752_chip_array_C36N-H.CEL	x			11	08/07/08 12:29:11
☐	12	GSM800753_chip_array_C38N-H.CEL				12	08/07/08 12:40:31
☐	13	GSM800754_chip_array_C41N-H.CEL				13	08/07/08 12:51:39
☐	14	GSM800755_chip_array_C42N-H.CEL				14	08/07/08 13:48:28
☐	15	GSM800756_chip_array_C44N-H.CEL				15	08/07/08 13:59:57
☐	16	GSM800757_chip_array_C45N-H.CEL				16	08/07/08 14:11:08
☐	17	GSM800758_chip_array_C47N-H.CEL				17	07/12/07 12:09:45
☐	18	GSM800759_chip_array_C06T-H.CEL				18	10/18/07 14:39:39
☐	19	GSM800760_chip_array_C11T-H.CEL				19	10/18/07 14:50:56
☐	20	GSM800761_chip_array_C24T-H.CEL				20	10/18/07 16:09:16
☐	21	GSM800762_chip_array_C27T-H.CEL				21	10/18/07 16:32:30
☐	22	GSM800763_chip_array_C28T-H.CEL				22	10/18/07 16:43:48
☐	23	GSM800764_chip_array_C30T-H.CEL				23	10/19/07 12:00:32
☐	24	GSM800765_chip_array_C31T-H.CEL				24	10/19/07 12:11:51
☐	25	GSM800766_chip_array_C32T-H.CEL				25	10/19/07 12:22:53
☐	26	GSM800767_chip_array_C33T-H.CEL				26	10/19/07 12:34:04
☐	27	GSM800768_chip_array_C35T-H.CEL				27	10/19/07 13:26:46
☐	28	GSM800769_chip_array_C36T-H.CEL				28	10/19/07 13:38:32
☐	29	GSM800770_chip_array_C38T-H.CEL				29	10/19/07 13:50:06
☐	30	GSM800771_chip_array_C41T-H.CEL				30	10/19/07 14:01:29
☐	31	GSM800772_chip_array_C42T-H.CEL				31	10/19/07 14:57:14
☐	32	GSM800773_chip_array_C44T-H.CEL				32	10/19/07 15:08:27
☐	33	GSM800774_chip_array_C45T-H.CEL				33	10/19/07 15:19:46
☑	34	GSM800775_chip_array_C47T-H.CEL	x	x		34	10/19/07 15:31:03

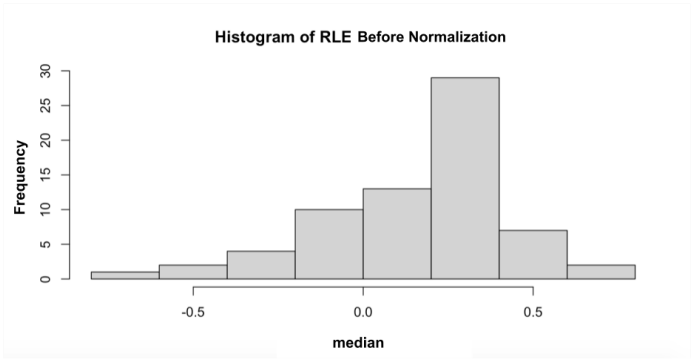
Normalized Data

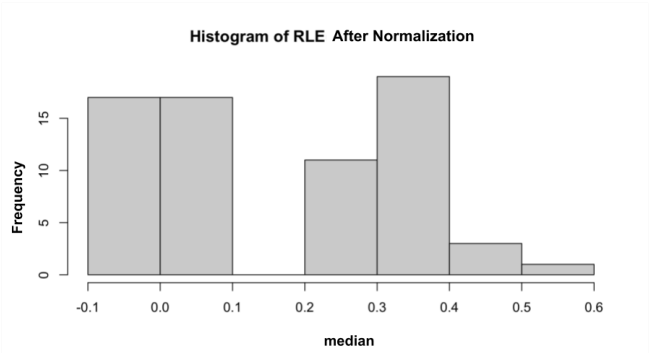
4. A histogram of median RLE scores.
    a. Using the affyBatch data that reads the CEL file downloaded from NCBI via the readAffy() method, the arrayQualityMetrics() method may provide relative log expression (RLE) and normalized unscaled standard error (NUSE) plots.
        i. dataset2 <- ReadAffy(celfile.path = "GSE32323_RAW")
        ii. Pset <- fitPLM(dataset2, normalize = FALSE)
        iii. hist(RLE(Pset,type='stats'))



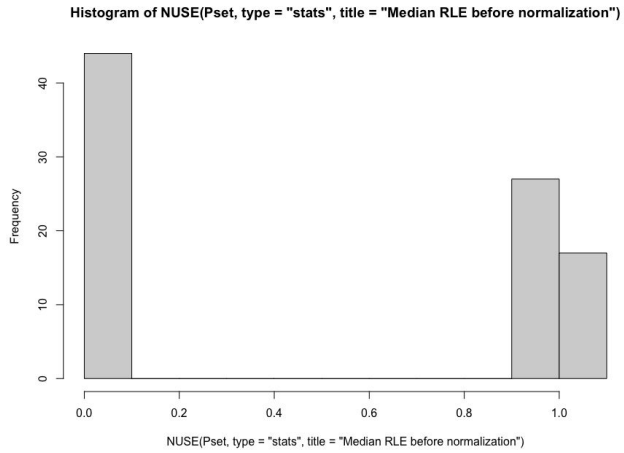


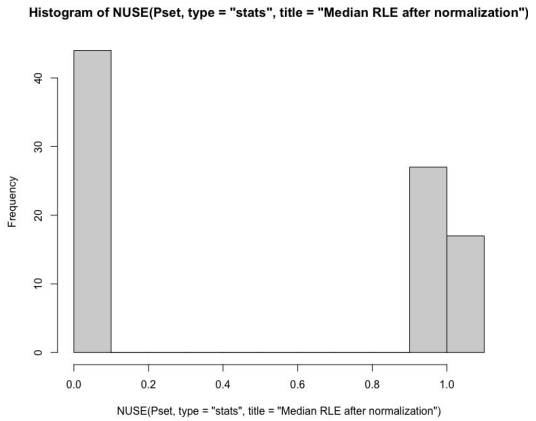
Histograms without cell line data:



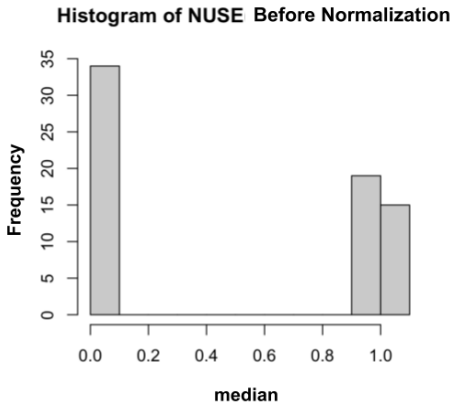


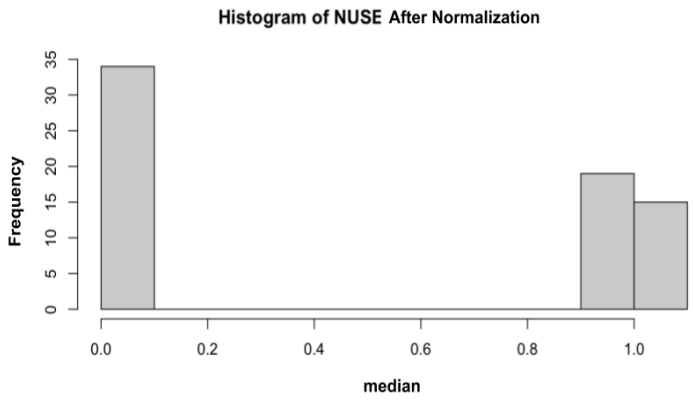
5. Another histogram of median NUSE scores.



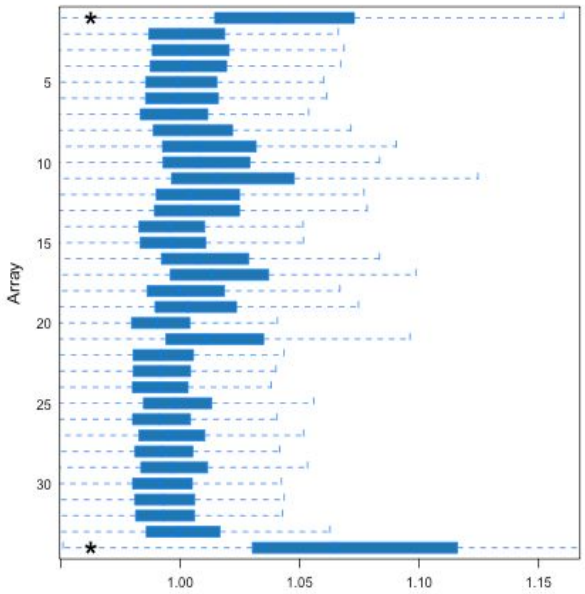


Histograms without cell line data:

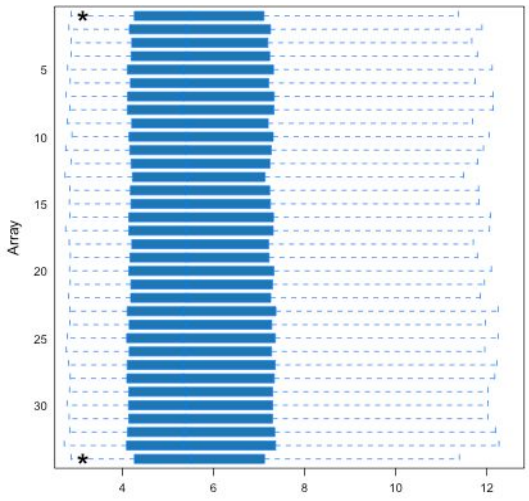




6. Interpretation of plots.
    a. Arrays 34-44 may be problematic in the RLE plot because, compared to all of the other arrays, their boxes are centered much further from zero and are much more spread out- with error bars that span the whole graph.
    b. The same group of arrays (34-44) stand out in the NUSE graph. Ideally, the boxes are centered about 1, but that group is much more offset than the rest of the boxes, which are all centered near 1. This could mean that those arrays are of lower quality.
7. Compare your corrected and normalized results with each other.
    a. Array intensity distribution before normalization

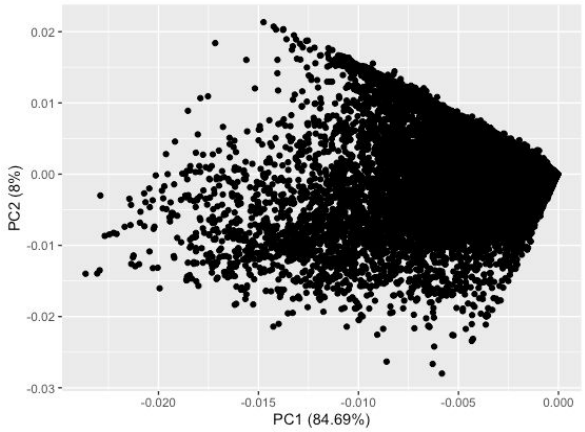


    b. Array intensity distribution after normalization (rma())

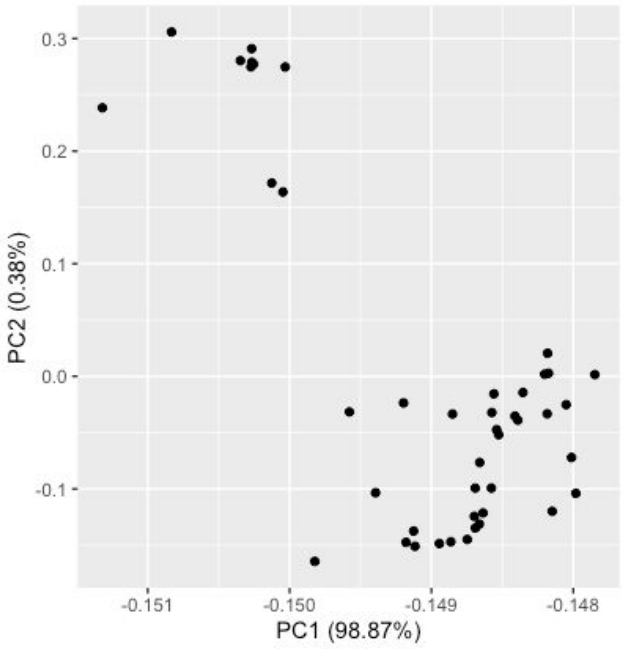


- The array intensity distribution shows a spread of signal intensities for each array. A consistent and uniform dataset will have arrays of similar width and position. The distribution before normalization shows arrays of similar width, but arrays 1 and 34 are off center compared to all the others. The distribution of the normalized data shows boxes of uniform width, all centered about the same point.

8. What is principal component analysis? Compare your PCA results with each other.
    a. Principal component analysis is a mathematical procedure that transforms correlated variables into a smaller number of uncorrelated variables. It is used to minimize the amount of information and simplify the complexity of high-dimensional data whilst retaining trends and patterns.
    b. Pca before normalization



    a. Pca after normalization


0.3
0.2
0.1
0.0
-0.1
PC2 (0.38%)
-0.151
-0.150
-0.149
-0.148
PC1 (98.87%)