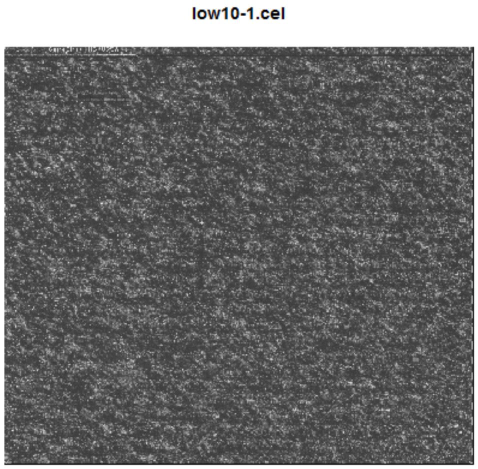
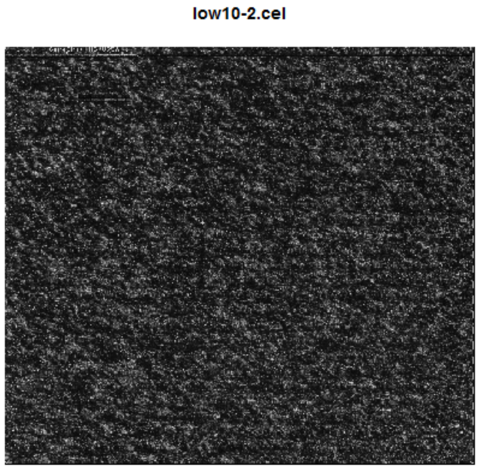
Ben Siranosian
PHP2620
Homework 3

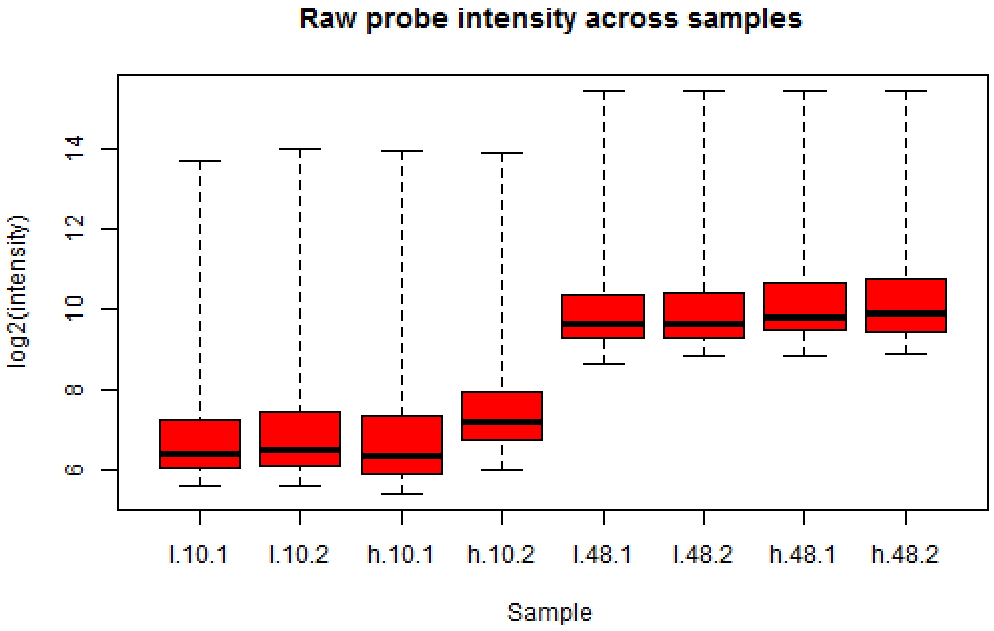
Exploratory analysis of *estrogen* microarray data

First, I checked the raw images of the microarrays for systematic quality problems. No problems were evident from the raw images, although the first replicate low at 10h is a much lighter image (shown compared to the second replicate, which is representative of the rest of the samples, for reference). However, normalization should eliminate any issue this could present.



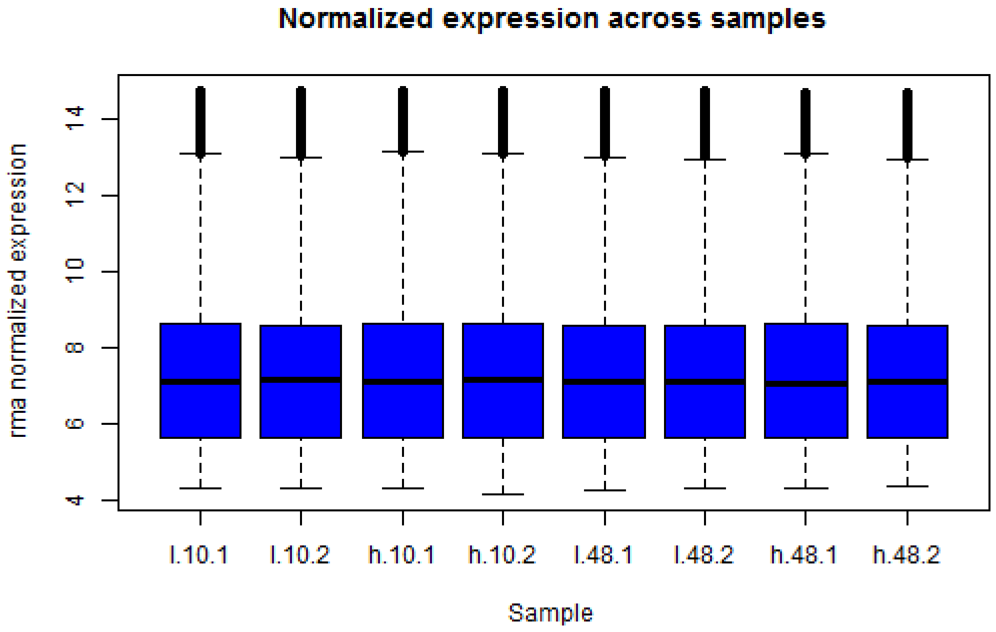


To check the need for normalization, I visualized the raw intensities of each sample as a boxplot. There is a significant difference between the distribution of intensity values of samples at 10 and 48 hours. This violates the normal assumption that most genes remain unchanged in a differential expression experiment and shows the need for normalization.

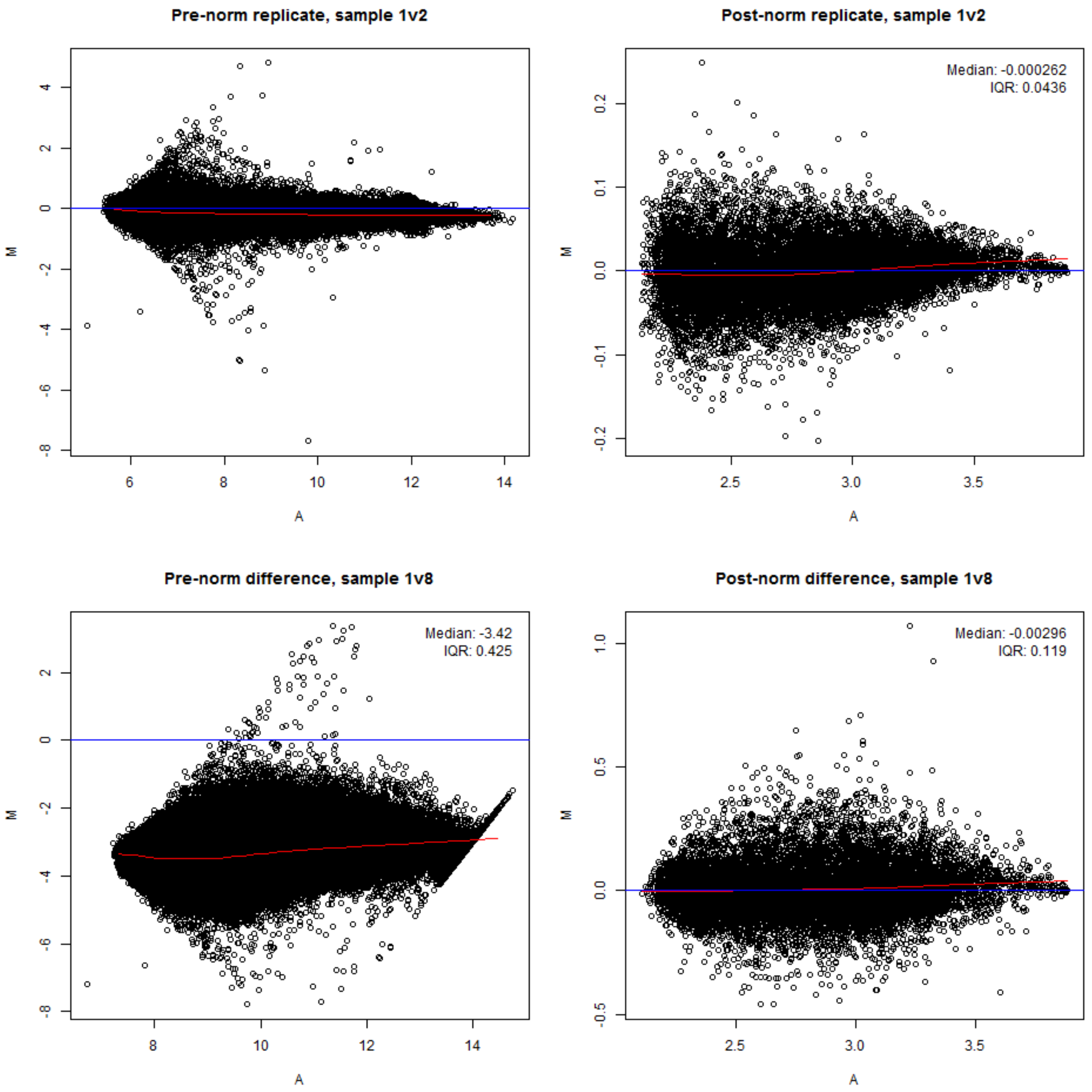


Ben Siranosian
PHP2620
Homework 3

To normalize the data, I used The Robust Multi- Array Analysis (RMA) method. RMA does a global background adjustment, normalizes by quantile normalization and uses median polish to convert probe level data to expression measures. My first check on the quality of normalization was to repeat the boxplot above. Clearly the normalization worked – the data from each sample is on a level playing field for our downstream analysis.



I also visualized MvA plots of raw and normalized expression data. The first row of figures shows the comparison of a replicate pre and post normalization. The second row shows the comparison of two different samples pre and post normalization. The need for normalization is most evident in the bottom left figure – the median is far below zero. In the control, normalization dramatically shrunk the vertical spread of the points while keeping the median at zero (look at the scale). In the comparison of different samples, normalization brought the median of the data to zero and also greatly decreased the vertical spread.



Differential expression in *estrogen* microarray data

In the following comparisons, I chose to filter genes with a fold change of greater than 1.5 and rank them by their measure of significance (p-value). In my opinion, differential expression is a binary term – it either happens between the samples or it doesn't. Filtering at a fold change of 1.5 eliminates genes that might have a significant p-value but would be uninteresting from a DE perspective. Sorting by p-value highlights the genes that are most significant, given that they are differentially expressed. In this case, the top genes all have small FDR values (the maximum in the first four tables is smaller than 0.02), eliminating the need to filter or sort on this quantity.

Estrogen vs Control at hour 10

PROBE	LOG FC	P-VALUE	FDR
39642_AT	2.94	4.74E-09	3.13E-05
910_AT	3.11	4.96E-09	3.13E-05
31798_AT	2.80	1.03E-07	3.51E-04
41400_AT	2.38	1.11E-07	3.51E-04
40117_AT	2.56	1.47E-07	3.58E-04
1854_AT	2.51	1.95E-07	3.58E-04
39755_AT	1.68	2.05E-07	3.58E-04
1824_S_AT	1.91	2.27E-07	3.58E-04
1126_S_AT	1.78	4.12E-07	5.78E-04
1536_AT	2.66	5.80E-07	7.32E-04
981_AT	1.82	6.46E-07	7.42E-04
33252_AT	1.74	8.86E-07	9.20E-04
1505_AT	2.40	9.48E-07	9.20E-04
34363_AT	-1.75	1.14E-06	1.03E-03
1884_S_AT	2.80	1.26E-06	1.06E-03
36134_AT	2.49	1.50E-06	1.19E-03
37485_AT	1.61	1.99E-06	1.48E-03
239_AT	1.57	4.07E-06	2.66E-03
38116_AT	2.32	4.09E-06	2.66E-03
35249_AT	2.22	5.18E-06	3.12E-03

Estrogen vs Control at hour 48

Nine of the genes from the table below overlap with the first. They are highlighted with yellow.

PROBE	LOG FC	P-VALUE	FDR
910_AT	3.86	8.27E-10	1.04E-05
31798_AT	3.60	1.28E-08	7.63E-05
1854_AT	3.34	1.81E-08	7.63E-05
38116_AT	3.76	8.12E-08	2.51E-04
38065_AT	2.99	1.12E-07	2.51E-04
39755_AT	1.77	1.36E-07	2.51E-04
1592_AT	2.30	1.39E-07	2.51E-04
41400_AT	2.24	1.81E-07	2.75E-04
33730_AT	-2.04	1.96E-07	2.75E-04
1651_AT	2.97	2.39E-07	3.02E-04
38414_AT	2.02	2.66E-07	3.05E-04
1943_AT	2.19	3.72E-07	3.69E-04
40117_AT	2.28	3.80E-07	3.69E-04
40533_AT	1.64	4.94E-07	4.45E-04
39642_AT	1.61	6.71E-07	5.18E-04
34851_AT	1.96	7.51E-07	5.18E-04
1824_S_AT	1.64	7.95E-07	5.18E-04
35995_AT	2.76	8.32E-07	5.18E-04
893_AT	1.54	8.43E-07	5.18E-04
40079_AT	-2.41	8.62E-07	5.18E-04

Control at hour 48 vs Control at hour 10

PROBE	LOG FC	P-VALUE	FDR
AFFX-CREX-5_AT	-6.83	3.11E-10	2.70E-06
AFFX-CREX-3_AT	-6.53	4.28E-10	2.70E-06
AFFX-BIODN-5_AT	-3.73	2.82E-08	1.18E-04
AFFX-BIOB-M_AT	-3.41	4.14E-08	1.31E-04
AFFX-BIODN-3_AT	-2.49	1.60E-07	4.04E-04
39581_AT	-2.67	3.89E-07	6.76E-04
AFFX-BIOC-3_AT	-2.99	3.91E-07	6.76E-04
37014_AT	-1.52	4.28E-07	6.76E-04
2004_AT	-2.06	1.00E-06	1.40E-03
AFFX-BIOC-5_AT	-2.05	1.57E-06	1.98E-03
34363_AT	-1.65	1.78E-06	2.00E-03
38065_AT	-2.11	1.92E-06	2.00E-03
40071_AT	-1.73	2.11E-06	2.00E-03
33730_AT	1.51	2.25E-06	2.00E-03
32597_AT	-1.52	2.37E-06	2.00E-03
AFFX-BIOB-3_AT	-2.61	8.18E-06	5.44E-03
33899_AT	-1.58	1.13E-05	6.54E-03
38116_AT	-2.03	1.14E-05	6.54E-03
1651_AT	-1.78	1.40E-05	7.40E-03
40079_AT	1.69	1.46E-05	7.40E-03

Estrogen at hour 48 vs Estrogen at hour 10

Eleven of the genes from the table below overlap with the third. They are highlighted with yellow.

PROBE	LOG FC	P-VALUE	FDR
AFFX-CREX-5_AT	-7.51	1.39E-10	1.47E-06
AFFX-CREX-3_AT	-7.01	2.33E-10	1.47E-06
AFFX-BIODN-5_AT	-4.41	7.09E-09	2.98E-05
AFFX-BIOB-M_AT	-3.79	1.70E-08	5.36E-05
AFFX-BIOC-3_AT	-4.08	2.98E-08	7.52E-05
1197_AT	-2.58	5.43E-08	1.00E-04
AFFX-BIODN-3_AT	-2.83	5.56E-08	1.00E-04
AFFX-BIOC-5_AT	-2.43	3.98E-07	6.27E-04
39642_AT	-1.58	7.93E-07	1.11E-03
40071_AT	-1.90	9.82E-07	1.17E-03
2004_AT	-2.06	1.02E-06	1.17E-03
AFFX-BIOB-3_AT	-3.20	1.62E-06	1.70E-03
AFFX-BIOB-5_AT	-3.42	2.28E-06	2.06E-03
39581_AT	-2.04	3.45E-06	2.90E-03
35934_AT	-1.55	1.54E-05	9.03E-03
36274_AT	-1.72	2.28E-05	1.11E-02
32755_AT	-1.68	2.77E-05	1.21E-02
31792_AT	-1.60	3.17E-05	1.25E-02
859_AT	-1.67	5.07E-05	1.82E-02
35977_AT	-1.66	6.21E-05	1.96E-02

Interaction between treatment and time

There is not strong evidence to support an interaction between treatment and time in this dataset. The top 10 most significant DE genes are reported below. The FDR value for these genes is very high compared to the other contrasts. The fold changes are also much lower on average. However, a FDR of <0.2 for 10 genes might be useful in a preliminary analysis to select targets for experimental validation.

PROBE	LOG FC	P-VALUE	FDR
33730_AT	-1.55	2.81E-05	0.121
38414_AT	1.57	3.01E-05	0.121
34851_AT	1.72	3.15E-05	0.121
1651_AT	2.16	4.62E-05	0.121
39642_AT	-1.32	4.80E-05	0.121
34363_AT	1.43	7.93E-05	0.159
40079_AT	-1.81	1.16E-04	0.159
38065_AT	1.73	1.25E-04	0.159
1945_AT	2.38	1.32E-04	0.159
757_AT	-1.37	1.49E-04	0.159

Control genes

67 control genes were tested for in the whole experiment. Control genes were significantly differentially expressed in both the control and estrogen time comparison, with 8 and 9 genes present in the top 20 list for each, respectively. All control genes in the top 20 list have a negative fold change, meaning they were added in higher quantities in the 10h samples. Most of the control genes are relatively constant through the 8 samples: 55 of the 67 have a variance of less than 1. Two control genes have a variance greater than 10: AFFX-CreX-3_at and AFFX-CreX-5_at. These two genes appear at the top of the time effect DE tables and were definitely spiked inconsistently across samples.

Measure of significance

It's easy to see that there are more than 20 significantly DE genes in most of these comparisons. FDR would be a good measure to define the total number of interesting genes. An FDR cutoff would be determined by the next step in the analysis. If costly experimental validation is necessary, a strict FDR might be applied so that few false discoveries move to the next step. On the other hand, if a bioinformatics analysis is done, a less strict FDR might be applied to capture the maximum number of truly DE genes. The table below shows the number of genes to be called significant at each FDR cutoff for each comparison.

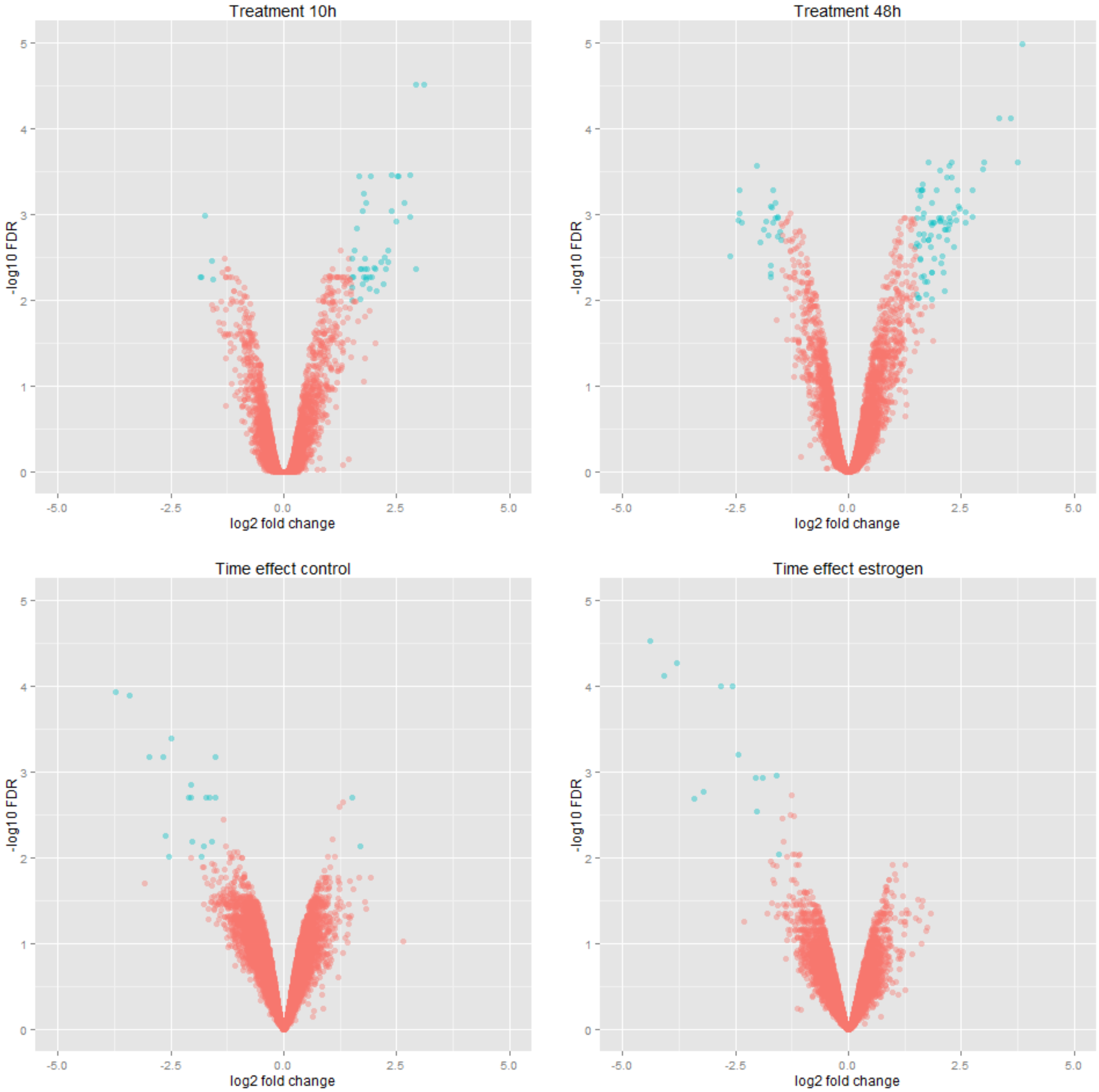
FDR CUTOFF	TREATMENT 10H	TREATMENT 48H	TIME EFFECT CONTROL	TIME EFFECT ESTROGEN	INTERACTION
0.001	13	36	8	8	0
0.01	113	250	34	25	0
0.05	326	638	590	172	0
0.10	457	957	1856	489	0
0.20	745	1382	3782	1412	19

Ben Siranosian
PHP2620
Homework 3

Summary

The addition of estrogen to breast cancer cells has a noticeable effect on gene expression. When analyzing data at the strict cutoff of FDR < 0.01, estrogen appears to effect gene expression in a time dependent manner. 113 DE genes are found when comparing treatment to control at 10h; this quantity increases more than twofold to 250 after 48h. Comparing a single treatment type at 10h and 48h had a less noticeable effect, only 34 and 25 DE genes were found, respectively. In each time effect comparison, 9 of the genes found at FDR < 0.01 were control "AFFX" genes, while no controls were present in the top hits for the treatment comparisons.

Below, I present volcano plots for the first four comparisons. Genes with log fold change > 1.5 and FDR < 0.01 are highlighted in blue.



Ben Siranosian
PHP2620
Homework 3

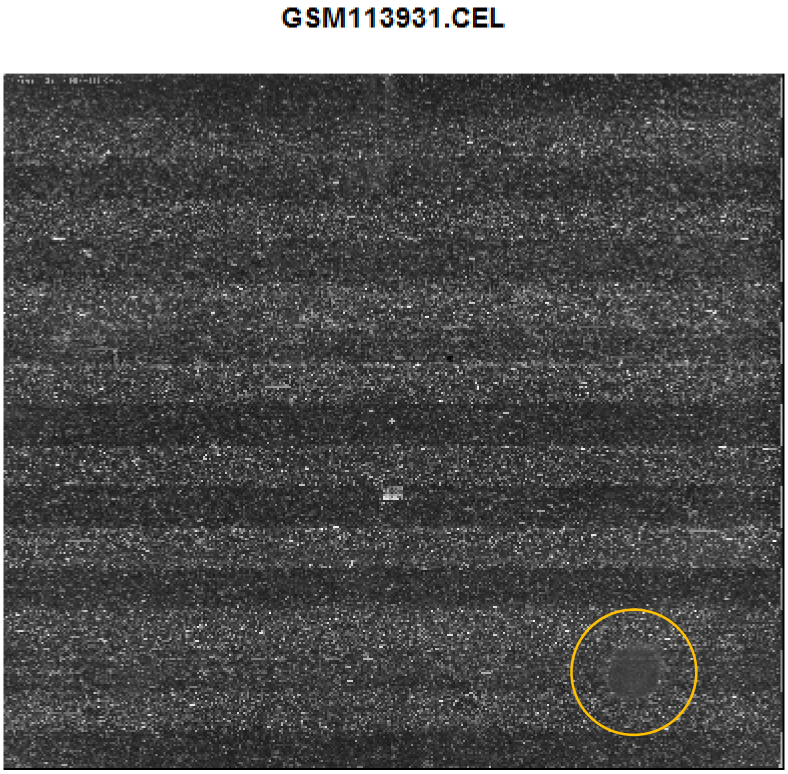
GDS2938 analysis

Experimental design

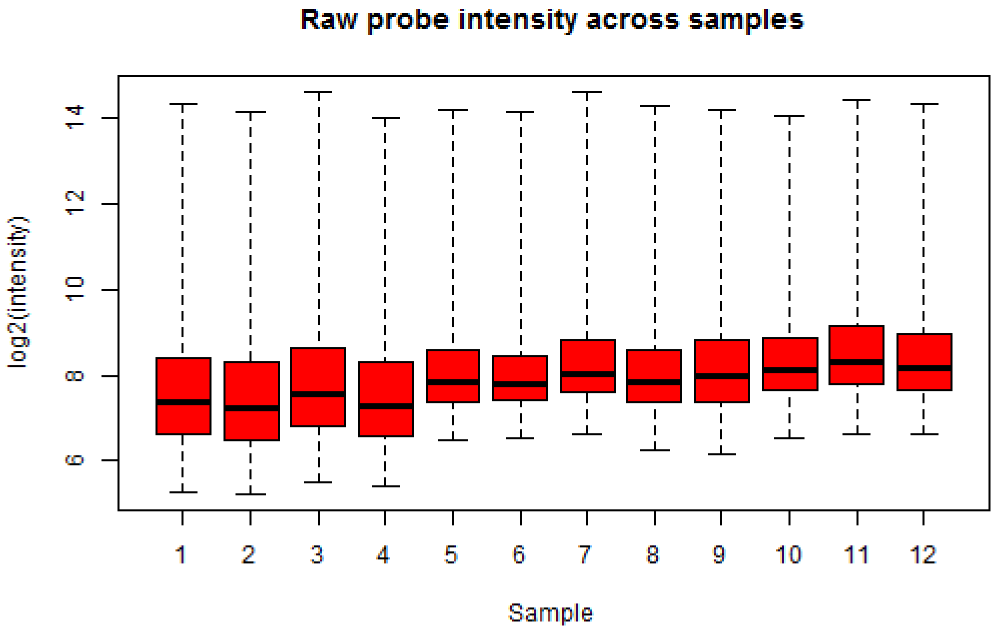
This data is from an experiment that studied transcription in thyroid epithelial cells (TECs). TECs are usually resistant to the Fas-mediated apoptosis pathway but treatment with interferon-gamma (IFN-gamma) and IL-1beta can overcome the resistance. cDNA microarrays were used to asses transcriptional changes (especially in apoptosis genes) in TECs under treatment with IFN-gamma, IL-1beta, or both.

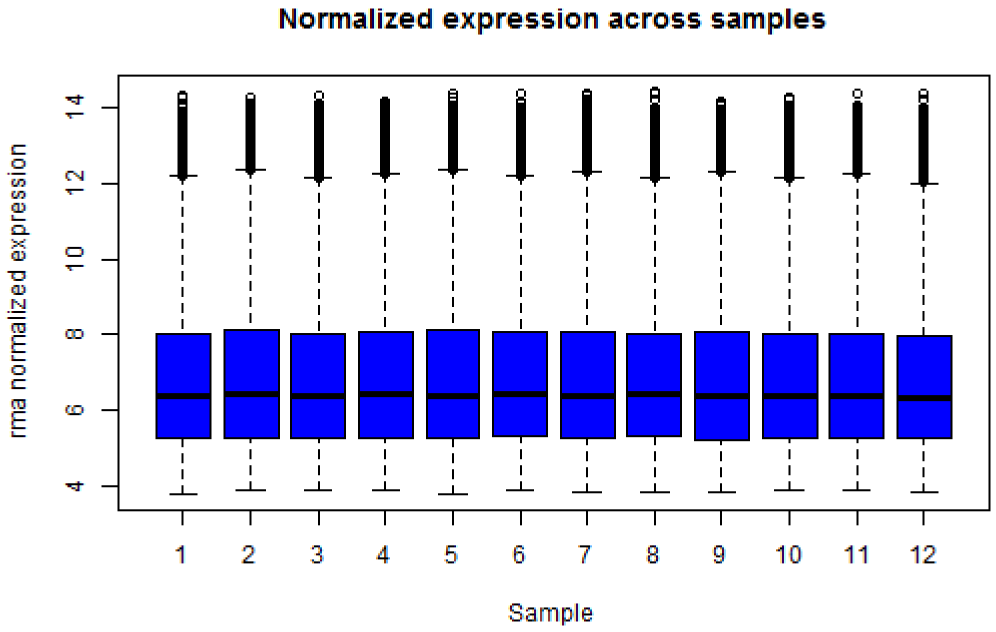
Preprocessing

First, I checked the raw images of the microarrays. In general, the images are not as clear and error-free as the estrogen dataset. Horizontal bands are visible in all of the images. Additionally, some images had visible artifacts, as highlighted below.



To check for the need for normalization, I used a boxplot (red) to compare the range of raw intensities across samples. There are clear differences in the distribution of intensities between samples, indicating the need for normalization of this dataset. Once again, I chose to use RMA normalization to convert the raw probe intensities to gene expression levels. I constructed the same boxplot (blue) after normalization – the medians and quartiles are equal, indicating that the normalization did a good job.





Differential expression

Overall, the differential expression in this experiment is not nearly as clear as the estrogen data. I dropped the requirement for genes to have a fold change above 1.5 because it was too stringent. In the following tables I present the 10 most significant genes for each condition, sorted by p-value, regardless of whether they should actually be called “significant.”

Ben Siranosian
PHP2620
Homework 3

IFN-gamma treatment

This treatment appears to have a significant effect on two genes, 209459_s_at and 209460_at. All other genes are too lowly differentially expressed or too likely to be false positives to be interesting.

PROBE	LOG FC	P-VALUE	FDR
209459_S_AT	2.46	4.32E-08	9.63E-04
218501_AT	0.89	8.84E-06	0.072
209460_AT	2.03	9.65E-06	0.072
221815_AT	0.75	5.33E-05	0.297
44790_S_AT	0.83	9.25E-05	0.406
213258_AT	0.96	1.11E-04	0.406
207620_S_AT	0.61	1.28E-04	0.406
208613_S_AT	-1.02	1.46E-04	0.406
222173_S_AT	-1.32	2.64E-04	0.552
212298_AT	0.93	2.83E-04	0.552

IL1-beta treatment

Results for this treatment are completely inconclusive. They suggest that no expression level is significantly changed by IL1-beta treatment. The FDR value is too high in all cases for a change to be considered significant. No genes overlap between this list and the list for IFN-gamma. The FDR values here are so bad, I was sure there was a problem with my analysis. However I checked it over and it seems to be correct… is there something I'm doing wrong in my code?

PROBE	LOG FC	P-VALUE	FDR
218573_AT	-1.71	6.03E-05	0.9999
210946_AT	0.56	1.99E-04	0.9999
222258_S_AT	0.74	2.48E-04	0.9999
218624_S_AT	-0.51	4.27E-04	0.9999
203932_AT	2.15	5.21E-04	0.9999
201659_S_AT	-0.52	7.28E-04	0.9999
200800_S_AT	-1.06	7.66E-04	0.9999
209126_X_AT	0.31	8.66E-04	0.9999
212543_AT	0.75	1.01E-03	0.9999
220253_S_AT	-0.68	1.12E-03	0.9999

IFN-Gamma and IL1-beta treatment

These results are just as bad as the IL1-beta treatment. Whereas IFN-gamma treatment alone had a significant effect on some genes, adding IL1-beta eliminated these changes. The FDR value is too high in all cases for a change in expression to be considered significant change. Additionally, the fold change values are much lower than would be necessary to call a change interesting.

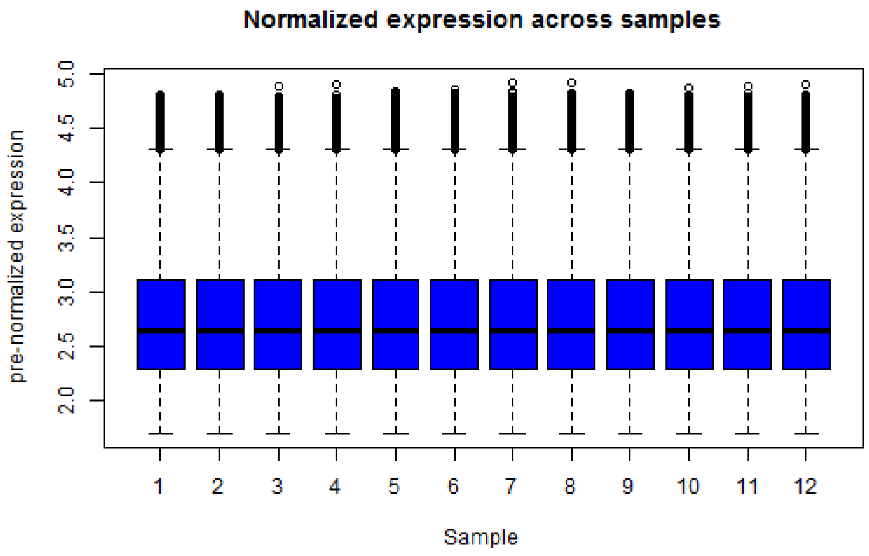
PROBE	LOG FC	P-VALUE	FDR
201143_S_AT	-0.46	1.37E-03	0.9993
202676_X_AT	-0.37	2.01E-03	0.9993
210138_AT	-1.08	2.14E-03	0.9993
207431_S_AT	-0.53	2.79E-03	0.9993
207927_AT	0.41	2.92E-03	0.9993
206385_S_AT	-1.68	3.33E-03	0.9993
205357_S_AT	0.33	3.63E-03	0.9993
212223_AT	-0.62	3.78E-03	0.9993
202250_S_AT	-0.28	3.90E-03	0.9993
204089_X_AT	-0.58	3.96E-03	0.9993

Summary

Overall, I gathered very little from the analysis of this dataset. However, in the paper (http://dx.doi.org/10.1210/en.2007-0126) the authors present several genes that are differentially expressed in table 1. No genes from my list of IFN-gamma match their list of DE genes. I'm becoming more confident that I messed up something in the analysis. Unfortunately, I don't have the time to correct it. When you read this, could you check my code and let me know what I am doing wrong? I'd appreciate it.

Repeat using pre-normalized data

To test if the problems I was experiencing above were the result of coding errors or a truly uninteresting sample, I used the pre-normalized data available from GEO. My first test was to repeat the boxplot and ensure the data were in fact normalized. The box plot below confirms this assumption – the medians and quartiles are consistent across all 12 samples.



I then repeated the analysis from above (conversion to expression set, fit using eBayes, topTable reporting). This time, many genes were found significant! I report the top 10 genes for IFN-gamma, IL1-beta, and both treatments, sorted by p-value.

IFN-gamma treatment

Several statistically significant DE genes were uncovered for IFN-gamma treatment. However, the fold change values are not as high as we would like to be sure of true differential gene expression. When a cutoff was applied to only select genes with fold change greater than 1.5 (like the estrogen experiment), only 5 genes were identified at FDR < 0.05.

PROBE	GENE TITLE	LOG FC	P-VALUE	FDR
204269_AT	pim-2 oncogene	0.80	2.32E-08	5.18E-04
206421_S_AT	serpin peptidase inhibitor, clade B (ovalbumin), member 7	1.23	8.36E-08	9.31E-04
209459_S_AT	4-aminobutyrate aminotransferase	0.88	1.50E-07	9.97E-04
209460_AT	4-aminobutyrate aminotransferase	0.87	1.79E-07	9.97E-04
204490_S_AT	CD44 molecule (Indian blood group)	0.63	3.84E-07	1.71E-03
218506_X_AT	glyoxylate reductase 1 homolog (Arabidopsis)	0.55	4.73E-07	1.76E-03
219558_AT	ATPase type 13A3	0.72	8.45E-07	2.69E-03
213425_AT	wingless-type MMTV integration site family, member 5A	0.79	1.12E-06	2.70E-03
206569_AT	interleukin 24	1.68	1.20E-06	2.70E-03
212297_AT	ATPase type 13A3	0.58	1.21E-06	2.70E-03

IL1-beta treatment

Similar results were found for IL1-beta treatment. When a fold change cutoff of 1.5 was applied, only 3 genes were found at FDR < 0.05. None of the top 10 genes for IL1 appear in the top 10 for IFN. When the list is extended to 20 genes, 1 appears in both.

PROBE	GENE TITLE	LOG FC	P-VALUE	FDR
202531_AT	interferon regulatory factor 1	1.24	8.64E-09	1.81E-04
209545_S_AT	receptor-interacting serine-threonine kinase 2	0.71	1.99E-08	1.81E-04
217478_S_AT	major histocompatibility complex, class II, DM alpha	0.75	3.10E-08	1.81E-04

213537_AT	major histocompatibility complex, class II, DP alpha 1	0.78	3.24E-08	1.81E-04
212671_S_AT	major histocompatibility complex, class II, DQ alpha 2	1.63	4.79E-08	2.13E-04
203932_AT	major histocompatibility complex, class II, DM beta	0.85	7.05E-08	2.47E-04
210029_AT	indoleamine 2,3-dioxygenase 1	2.04	7.75E-08	2.47E-04
209474_S_AT	ectonucleoside triphosphate diphosphohydrolase 1	-0.83	9.93E-08	2.77E-04
209312_X_AT	major histocompatibility complex, class II, DR beta 5	1.04	1.36E-07	3.38E-04
217362_X_AT	major histocompatibility complex, class II, DR beta 6 (pseudogene)	0.95	2.10E-07	4.67E-04

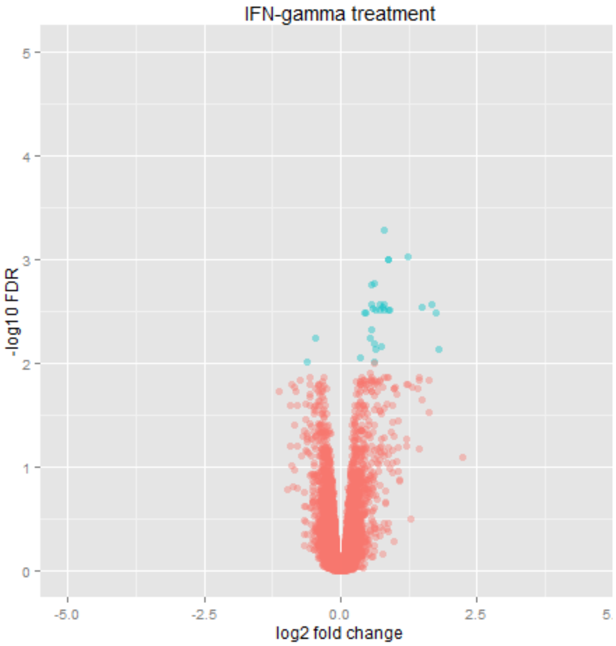
Both treatments

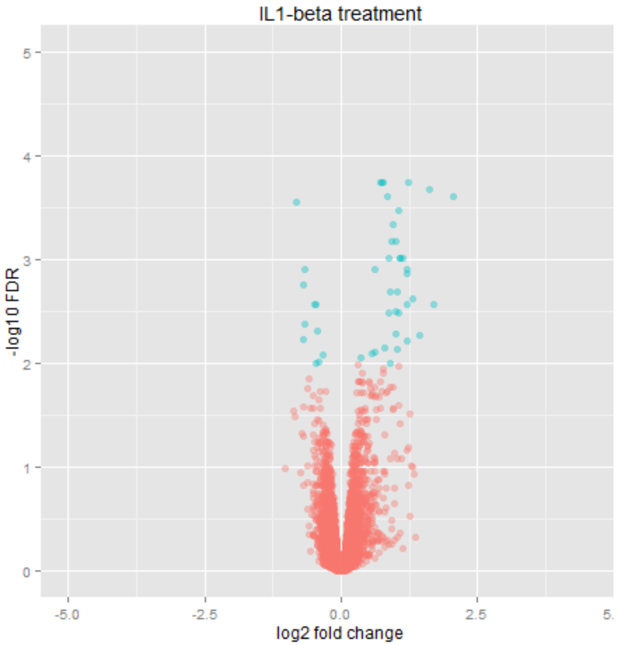
Treatment with both IFN-gamma and IL1-beta had a much stronger effect on gene expression. The best way to see this is to look at the volcano plot below. When a fold change cutoff of 1.5 was applied, 22 genes were significant at FDR < 0.01, in stark contrast to the single treatments. Treating with both IFN-gamma and IL1-beta increased the number of significant DE genes as well as the amount these genes are differentially expressed. When looking at the top 20 genes, 3 from this list are present in the IFN-gamma list and 7 are present in the IL1-beta list.

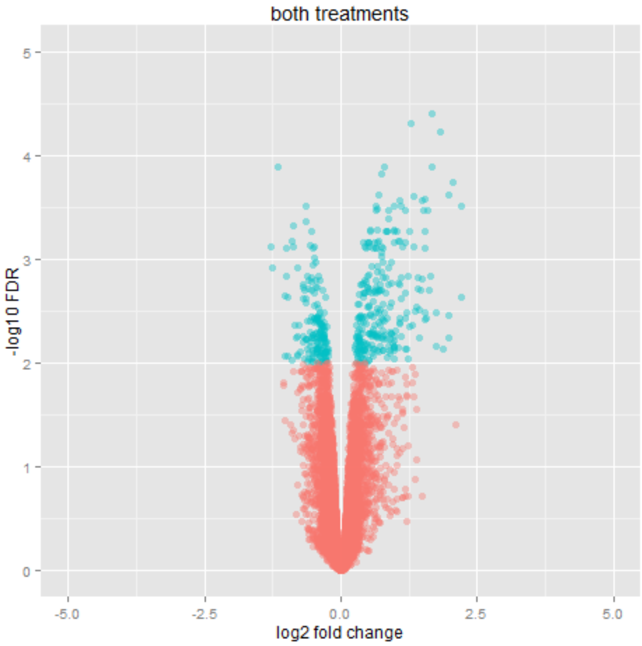
PROBE	**GENE TITLE**	**LOG FC**	**P-VALUE**	**FDR**
209545_S_AT	receptor-interacting serine-threonine kinase 2	1.07	2.59E-10	5.77E-06
206421_S_AT	serpin peptidase inhibitor, clade B (ovalbumin), member 7	1.67	3.56E-09	3.96E-05
202531_AT	interferon regulatory factor 1	1.27	6.68E-09	4.96E-05
1405_I_AT	chemokine (C-C motif) ligand 5	1.82	1.08E-08	6.00E-05
222288_AT	unknown	-1.15	2.90E-08	1.29E-04
212671_S_AT	major histocompatibility complex, class II, DQ alpha 2	1.66	3.85E-08	1.30E-04
205518_S_AT	cytidine monophospho-N-acetylneuraminic acid hydroxylase, pseudogene	0.79	4.08E-08	1.30E-04
204269_AT	pim-2 oncogene	0.74	5.47E-08	1.52E-04
210029_AT	indoleamine 2,3-dioxygenase 1	2.05	7.40E-08	1.83E-04
214038_AT	chemokine (C-C motif) ligand 8	1.98	1.17E-07	2.40E-04

Ben Siranosian
PHP2620
Homework 3

To visualize the differentially expressed genes, I created volcano plots for each treatment. Genes highlighted in blue are significant at FDR < 0.01. Clearly, using the pre-normalized data solved the problem I was initially having with this dataset. Was RMA not appropriate for this? Did the artifacts I observed in the image throw off my analysis? I could also test different normalization methods (such as GCRMA) to see if they produced similar results.


IFN-gamma treatment
-log10 FDR
log2 fold change


IL1-beta treatment
-log10 FDR
log2 fold change


both treatments
-log10 FDR
log2 fold change