Initial analysis of the macrophage spondyloarthritis RNA-seq data

Colin Dewey

July 28, 2016

Table of Contents

[Overview of the data 1](#_Toc457481351)

[Examination of possible biases 1](#_Toc457481352)

[Mapped reads 4](#_Toc457481353)

[Verifying the sex of the samples 7](#_Toc457481354)

[Gene differential expression analysis 7](#_Toc457481355)

[Sample clustering 8](#_Toc457481356)

[Principal component analysis (PCA) 8](#_Toc457481357)

[Hierarchical clustering 10](#_Toc457481358)

[Highly-expressed genes 11](#_Toc457481359)

[Examination of PC2 12](#_Toc457481360)

[Differentially expressed genes 16](#_Toc457481361)

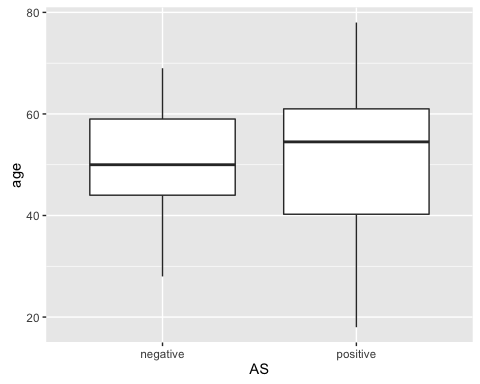
[Genes of interest 16](#_Toc457481362)

# Overview of the data

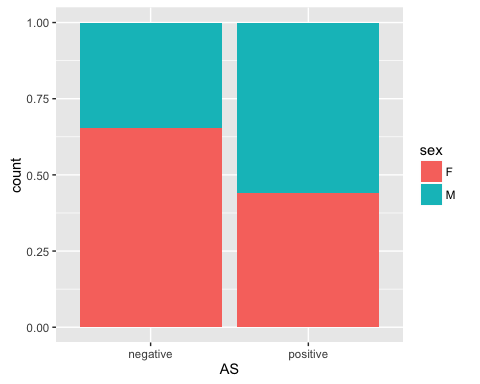
The RNA-seq dataset consisted of 234 samples, which were derived from 79 patients. Most patients had three samples corresponding to the three treatment types: NT, Curdlan, LPS. However, 3 patients were missing at least one sample: MC111, MC112, MC130. Out of the 79 patients, 51.9% were female, 50.6% were from the UW population, and 63.3% were diagnosed with spondyloarthritis.

## Examination of possible biases

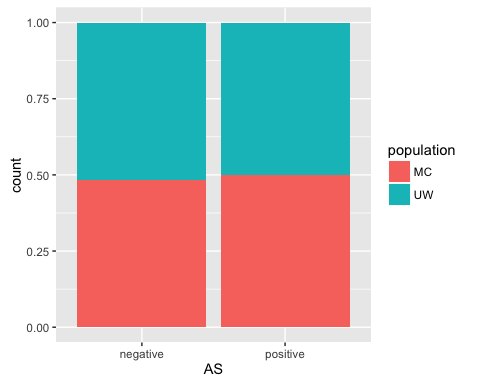
I first examined whether there were any biases in age, population, or gender between the cases and controls.



Although the cases were generally older, this difference was not statistically significant (p = 0.71, Mann-Whitney test).



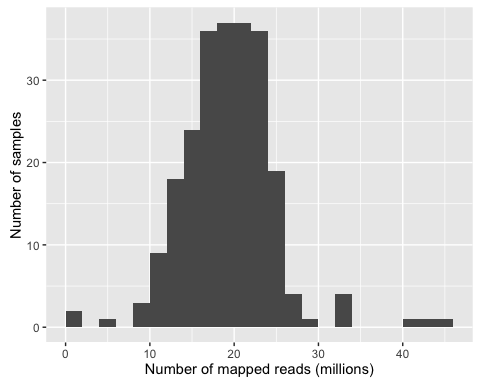
Although the cases where enriched for males relative to the controls, this difference was not statistically significant (p = 0.1, Fisher's exact test).



There was little difference between the population frequencies of the cases and controls.

## Mapped reads

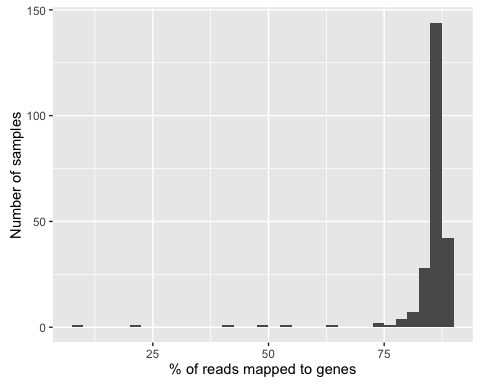
To get an idea of the amount of read data generated for each sample, I plotted a histogram of the number of reads mapping to genes in each sample.



The median number of mapped reads was 19.4 million, which is generally an adequate number for gene differential expression (DE) analysis. However, there were some outlier samples with small numbers of mapped reads, which could be detrimental to the DE analysis. I decided to filter out 2 samples that had fewer than 5 million mapped reads. These samples are listed in the table below.

|  |  |
| --- | --- |
| sample\_id | num\_mapped\_reads |
| MC111Curdlan | 1546316 |
| MC135Curdlan | 1793749 |

I also examined the percentage of raw reads from each sample that were successfully mapped to a gene. A histogram of the percentage of reads mapping to genes in each sample is given below.

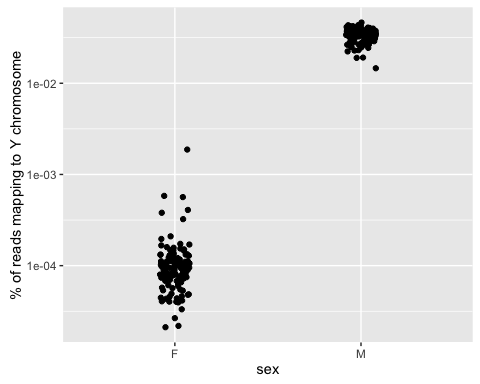


The vast majority of the samples have good read mapping rates, however there are some clear outliers samples here which could resprent samples with poor-quality RNA or that have other technical issues. To be conservative, I decided to initially filter out any sample that a read mapping rate of less than 80%. This threshold resulted in the filtering of 13 samples, which are listed in the table below.

|  |  |
| --- | --- |
| sample\_id | pct\_mapped\_reads |
| MC135Curdlan | 9.2 |
| UW039LPS | 21.5 |
| UW024Curdlan | 40.0 |
| UW012NT | 49.6 |
| MC122Curdlan | 53.5 |
| MC121Curdlan | 64.7 |
| UW025LPS | 74.2 |
| MC123Curdlan | 74.9 |
| UW015LPS | 75.7 |
| MC125LPS | 78.0 |
| MC135NT | 78.0 |
| UW015Curdlan | 78.1 |
| MC114NT | 79.8 |

## Verifying the sex of the samples

As a way of verifying the sex of the samples, I computed the percentage of reads of each sample that were mapped to genes on the Y chromosome. Ideally, for female samples, we would not observe any reads mapping to genes on the Y chromosome. However, in practice, due to technical reasons, we observe a small number of reads mapping to Y in female samples. Below is a plot of these percentages, stratified by the sex of the patient from which the samples were derived.



After correction of the previously identified sample swaps, the samples cleanly separate by sex in this analysis.

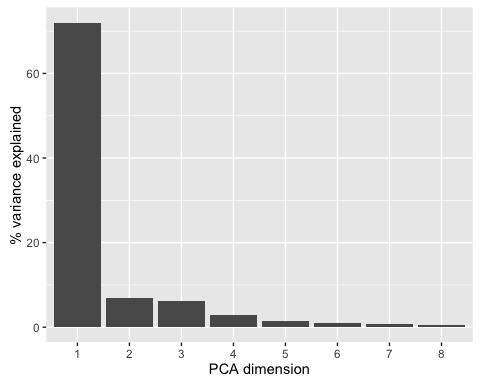
# Gene differential expression analysis

I next performed differential expression analyses on these data. I used all samples except for the 14 samples with either low mapped read counts or low percentage of mapped reads. In total, 220 samples were used in this analysis.

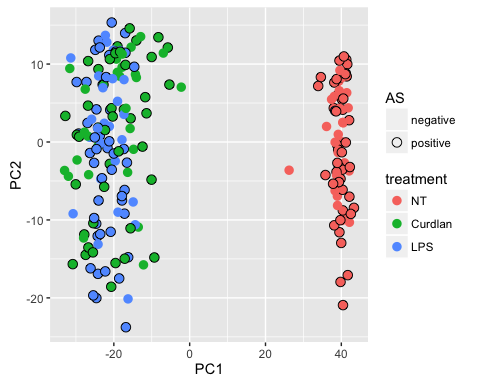
## Sample clustering

### Principal component analysis (PCA)

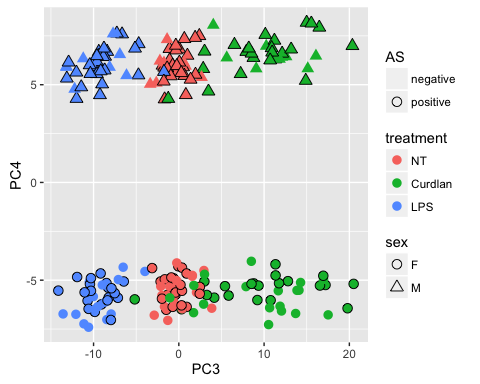
I first used PCA on the gene expression profiles of each sample as a way to (1) visualize how the samples cluster in low dimension and (2) identify the factors that are contributing to the most variation in the gene expression levels. The plot below shows how much variance in the gene expression profiles is explained by the first eight dimensions of the PCA. From this plot, it looks like the first four dimensions are explaining most of the variance in these data.



I then examined how each of the first four principal components associate with the various experimental factors. In the scatterplot below, each point corresponds to a single sample and the coordinates of the points are given by the first two principal components. Here we can see that PC1, which explains the most variance in these data, corresponds to effects of treatment. PC2 is less clear, although the outliers in PC2 (say those with PC2 < -15), are somewhat enriched for samples corresponding to patients who are AS positive.

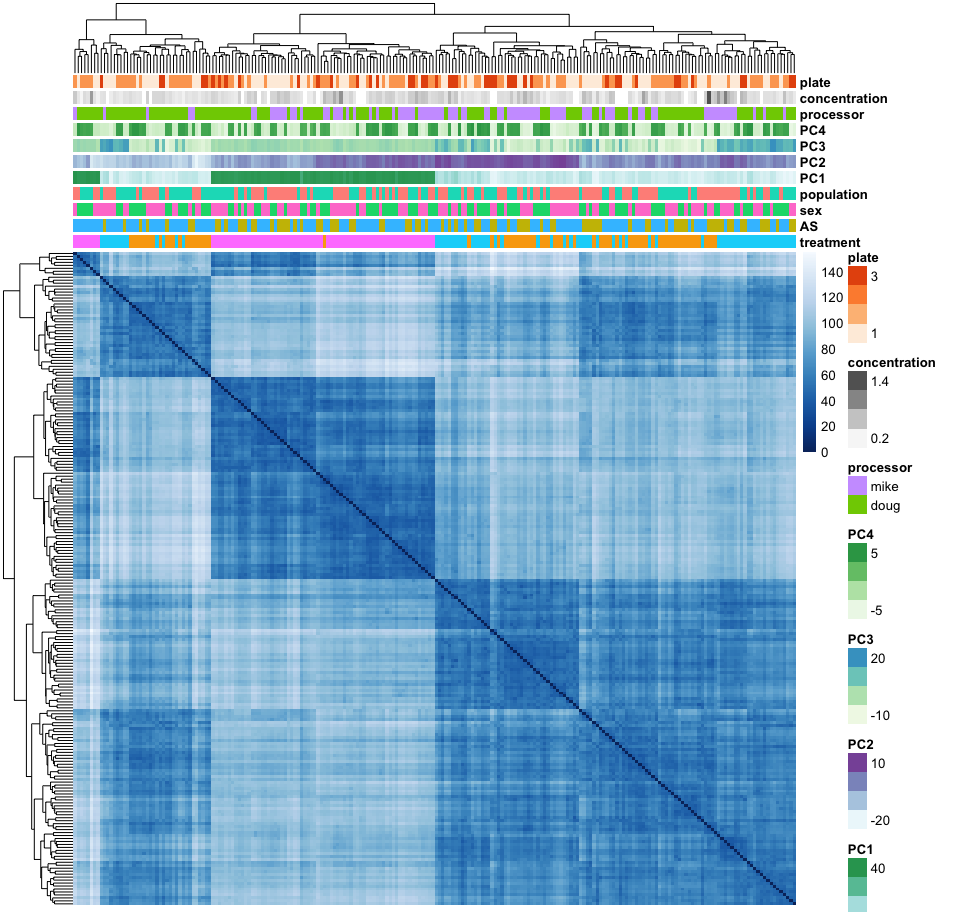


The scatterplot below similarly examines PC3 and PC4. PC3 nicely separates out the samples according to treatment type. One can see a few samples for which the treatment may not have been applied successfully (e.g., the green Curdlan sample mixed in with the blue LPS samples in the lower-left corner). PC4 is clearly the dimension in which the sexes are separated.



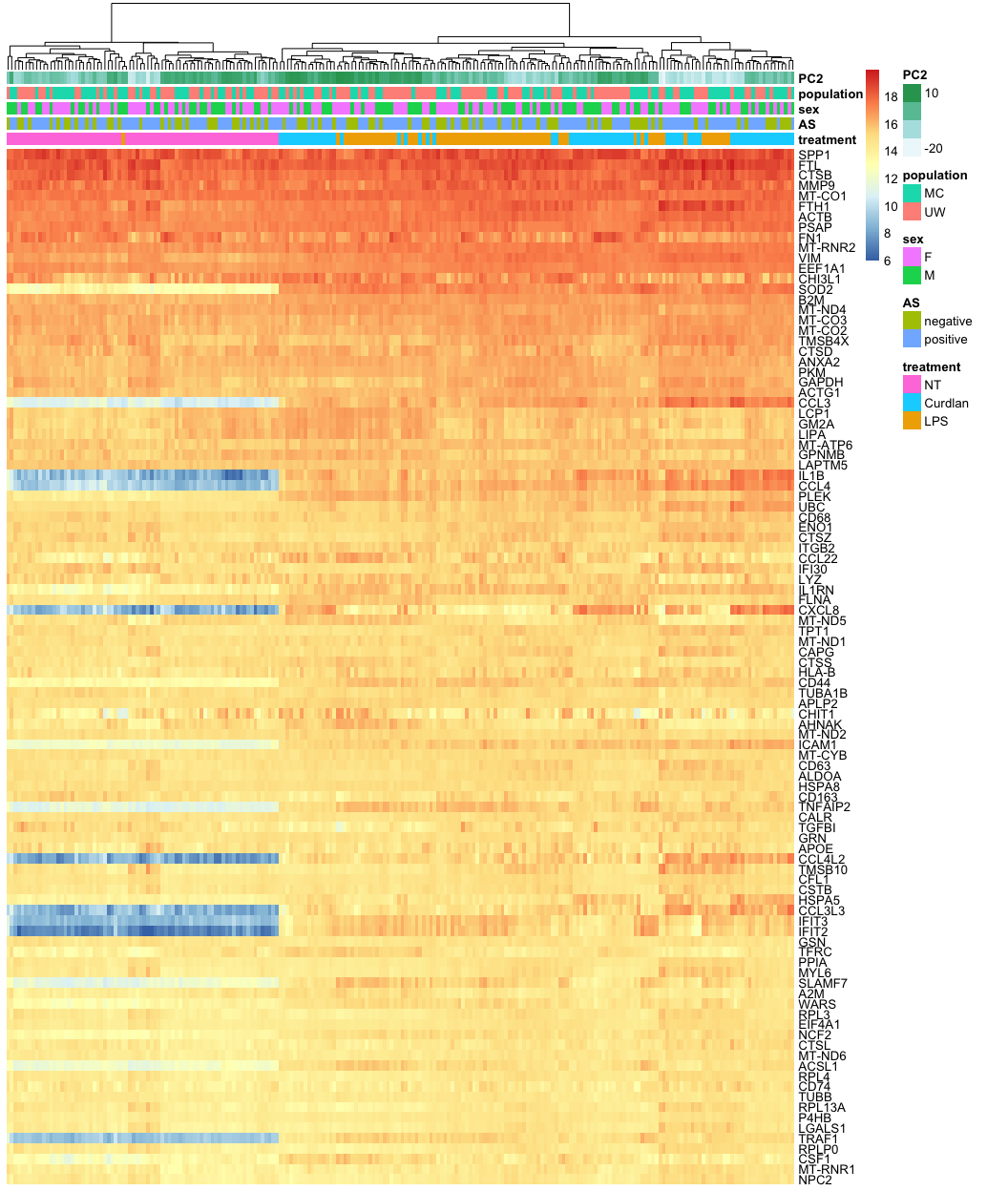
### Hierarchical clustering

It also useful to visualize the samples via a distance heatmap with the samples clustered in a hierchical structure. In the plot below, the rows and columns of the heatmap correspond to samples and the intensity of the heatmap value for a pair of samples corresponds to how similar their gene expression profiles look (i.e., dark blue indicates to very similar samples and white indicates very disssimilar samples). The rows and columns of the heatmap are ordered (in the same way) by a hierchical clustering (see tree structure at top and left). The colored tracks at the top of the plot show the experimental factors and the values of the first four PCs associated with the sample in each column. This clustering confirms that the major division of the samples is treated vs. untreated, and that samples are also clustering somewhat by treatment type (Curdlan or LPS). However, it is apparent that whatever effect is represented by PC2 is also playing a major role. At the left of the plot there is a clear set of outlier samples with very low values of PC2, from all treatment types. In addition, some of the higher-level divisions of the samples appear to be determined by the magnitude of the PC2 value.



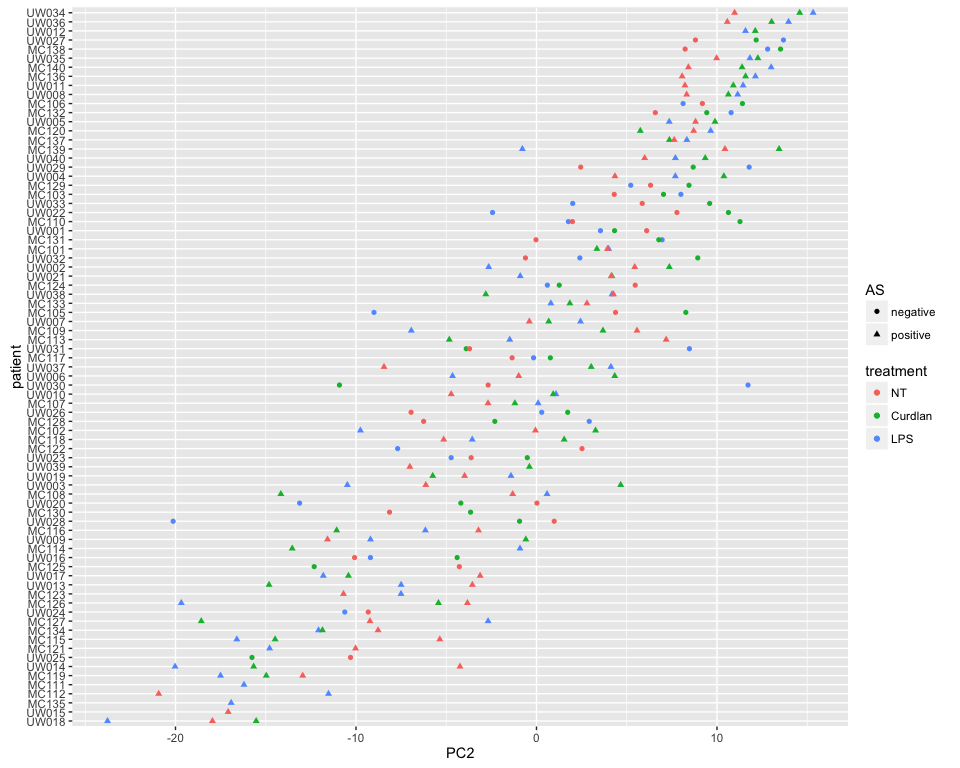
### Highly-expressed genes

As a final way of visualizing the samples and how they cluster with each other, we can examine the expression profiles of the most highly-expressed genes. The heatmap below gives the expression profiles for the 100 genes with highest mean expression level. The rows correspond to genes, the columns correspond to samples, and the heatmap value gives the expression level (in units of log2 normalized read counts) of a given gene in a given sample. The samples are again hierarchically clustered. From this plot, there is a clear signature of the untreated samples with a dozen or so genes that are markedly lower in expression than in the treated samples. However, again, we see that the samples with very low PC2 value are big outliers in these expression profiles.

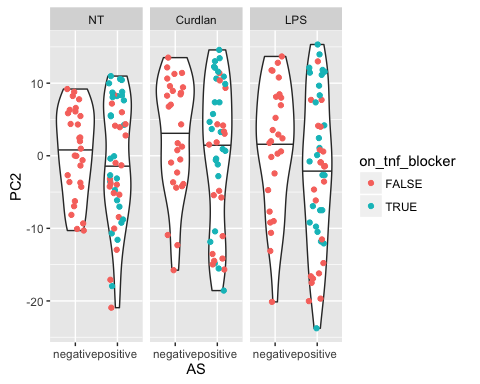


### Examination of PC2

PC2 is a bit problematic because it is does not appear to be highly associated with any measured experimental factors, although samples that are outliers in PC2 tend to be from AS positive patients. I dug a bit deeper into PC2 to see if I can find any hints as to what effect(s) it might represent. I first examined whether samples from the same patient tended to have similar values of PC2, which would indicate that PC2 has something to do with a property of the patient, or the initial pool of macrophages extracted from that patient. The plot below visualizes the values of PC2 for the samples grouped by patient, with the patients ordered by the mean PC2 value of its samples. From this plot it appears to be the case that samples from the same patient tend to have PC2 values of similar magnitude, although some patients have samples with wildly different PC2 values. It is also possibly informative that patients with high mean sample PC2 value have much lower variance in PC2 across the treatment types.



I also examined the relationship of a sample's PC2 value to AS status, treatment, and whether the patient was on a TNF blocker medication. This is vizualized with the violin plots below (the horizontal bar in each "violin" is the median for that group). As we had seen hints of before, AS positive samples tended to have lower values of PC2, and this held up across all treatment types. However, none of these differences are statistically significant at a 0.05 significance level (using the non-parametric Wilcoxon test). There are hints from these plots that the patients on TNF blockers generally have samples with higher PC2 values, but there were no statistically significant differences found.



Below is a table of the patients whose untreated samples had extreme (<-15) values of PC2.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| patient\_id | AS | sex | age | age\_onset | HLA\_B27 | severity | f\_history | tnf\_block | PC2 |
| MC112 | positive | M | 57 | 34 | positive | severe | no | FALSE | -20.9 |
| UW018 | positive | F | 56 | 30 | positive | mild | no | TRUE | -18.0 |
| UW015 | positive | F | 54 | 51 | negative | none | no | FALSE | -17.1 |

Some of the treated samples also had extreme (<-15) values of PC2. These are listed in the table below.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| patient\_id | treatment | AS | sex | age | age\_onset | HLA\_B27 | severity | f\_history | tnf\_block | PC2 |
| UW018 | LPS | positive | F | 56 | 30 | positive | mild | no | TRUE | -23.8 |
| UW028 | LPS | negative | M | 44 | NA | negative | NA | no | FALSE | -20.1 |
| UW014 | LPS | positive | F | 55 | 26 | positive | mild | yes | FALSE | -20.0 |
| MC126 | LPS | positive | M | 56 | 21 | positive | none | no | FALSE | -19.7 |
| MC127 | Curdlan | positive | F | 39 | 21 | positive | none | no | TRUE | -18.6 |
| MC119 | LPS | positive | M | 50 | 29 | positive | severe | no | FALSE | -17.5 |
| MC135 | LPS | positive | M | 75 | 21 | positive | none | yes | FALSE | -16.9 |
| MC115 | LPS | positive | F | 64 | 20 | positive | none | no | FALSE | -16.6 |
| MC111 | LPS | positive | M | 69 | 30 | positive | severe | yes | FALSE | -16.2 |
| UW025 | Curdlan | negative | F | 59 | NA | negative | NA | no | FALSE | -15.8 |
| UW014 | Curdlan | positive | F | 55 | 26 | positive | mild | yes | FALSE | -15.7 |
| UW018 | Curdlan | positive | F | 56 | 30 | positive | mild | no | TRUE | -15.5 |

The following patients had multiple samples with extreme values of PC2.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| patient\_id | AS | sex | age | age\_onset | HLA\_B27 | severity | f\_history | tnf\_block | num\_extreme\_samples |
| UW018 | positive | F | 56 | 30 | positive | mild | no | TRUE | 3 |
| UW014 | positive | F | 55 | 26 | positive | mild | yes | FALSE | 2 |

## Differentially expressed genes

I performed two different differential gene expression analyses. The first analyzed the untreated samples only and assessed whether there were any differences in gene expression due to sex, population, or AS status. The second analyzed the effects of the two treatments and whether either treatment had effects that were AS-specific. These current analyses do not take into account HLA-B27 status or whether patients were on TNF blockers. The table below summarizes the numbers of genes that were detected to be differentially expressed with respect to each effect analyzed.

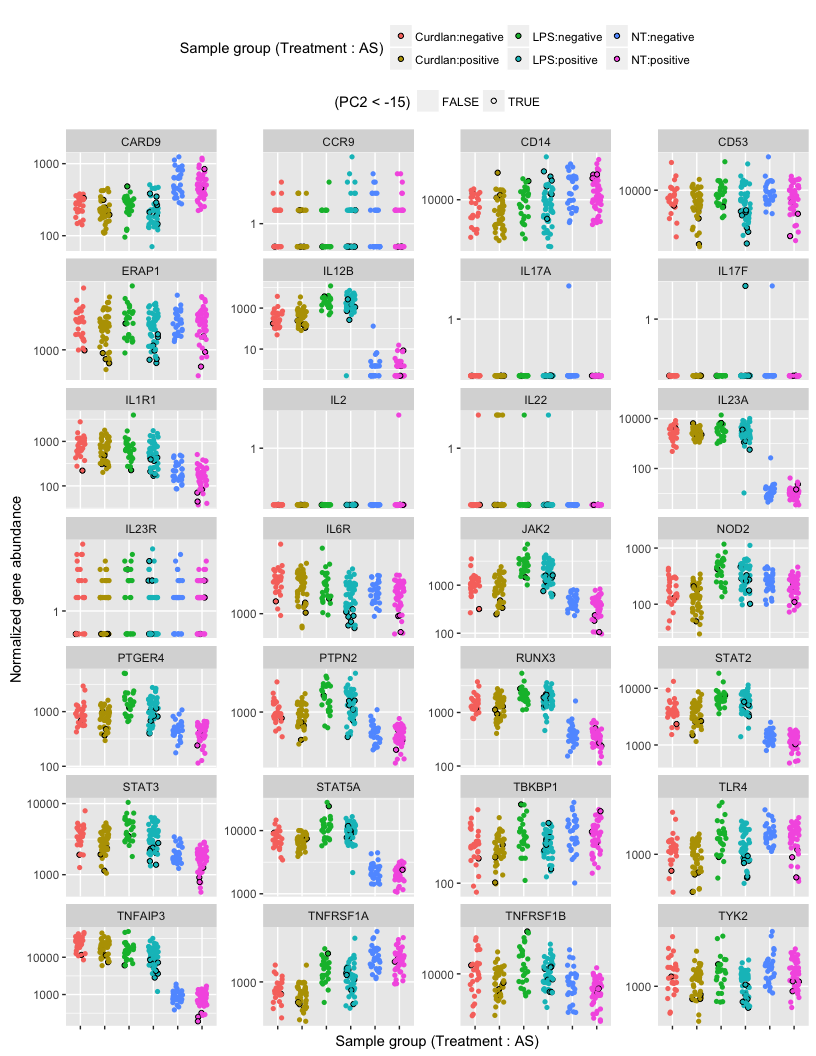
Number of DE genes by effect. The number of DE genes is further broken down by those that are up, down, up with a fold change >= 2, or down with a fold change >= 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| effect | DE | DE up | DE down | DE up 2x | DE down 2x |
| AS | 0 | 0 | 0 | 0 | 0 |
| sex | 43 | 25 | 18 | 23 | 3 |
| population | 3 | 2 | 1 | 1 | 1 |
| Curdlan | 12517 | 6350 | 6167 | 2334 | 1144 |
| LPS | 13394 | 6712 | 6682 | 2367 | 1701 |
| AS-specific Curdlan | 0 | 0 | 0 | 0 | 0 |
| AS-specific LPS | 0 | 0 | 0 | 0 | 0 |

It is clear that the treatments had major effects on gene expression. Unfortunately, there was little effect found due to AS status, both in the untreated and treated samples. It is possible that PC2 effect seen through PCA has something to do with AS status but that the number samples with a strong PC2 were in the minority and thus did not sway the differential expression analysis. In the individual gene plots shown the next section, I have higlighted the points corresponding to samples with strong PC2 signal to give some insight into what is going on in these samples.

### Genes of interest

Below are plots of the measured abundances for the genes of interest that you sent to me, with samples broken into groups by treatment and AS status. I have additionally highlighted samples with strong PC2 signal (PC2 < -15). It is clear that for several of these genes, although there is little effect observed due to AS status when examining all samples, the samples with strong PC2 signal appear to be behaving quite differently.



You were particularly interested in the expression pattern of TNFAIP3, and so I have provided a large version of its plot below. Looking at the untreated samples, it is clear that the samples with strong PC2 signal (which are all AS positive) have very different expression of TNFAIP3.

