Supplementary: Santiago & Potashkin (2015) reanalysis

Paul Pavlidis and Lilah Toker

Contents

1 Introduction

Santiago and Potashkin 2015 (S&P) report that HNF4A and PTBP1 are good blood-based biomarkers of early-stage Parkinson's disease. Their conclusions were based on meta-analysis of four microarray studies from the literature, augmented by analysis of a protein interaction network, and by qPCR analysis of these genes in two additional cohorts.

The analysis we report here is intended to detail some of the issues we spotted on reading their paper, and which are summarized in our letter to the PNAS editors. Our main conclusion is that HNF4A and PTBP1 are not substantiated as differentially expressed in PD (Parkinson's disease) compared to HC (healthy controls) when the four data sets analyzed by S&P are treated with very basic quality control measures.

The main points we are covering here are:

- The confounded batch effect in one data set (GSE22491; the HC and PD samples were run on different days) that seems to be dominating the expression pattern shown in Figure 1 of S&P
- The inclusion of pooled samples as if they were additional independent biological replicates in two of the data sets
- A second data set that has a notable batch effect that is also partly confounded with diagnosis (GSE18838)
- The weak differences in expression (if any) for the cited markers in all but GSE22491
- The possibility that HNF4A is expressed at very low or even negligible levels in blood. We bring this up because S&P reported that the HNF4A protein was "not identified" in blood, and we immediately noticed that in at least some data sets the levels of HNF4A are at what we would call "background noise" levels.

2 Methods and Data

In general we are trying to follow the approach of S&P, but using R instead of INMEX (our attempt to replicate the results with INMEX are described in the Appendix), and with different quality control. One notable possible difference from S&P is that for GSE18838, we used data reprocessed from CEL files (by Gemma, which uses the affy "power tools" provided by Affymterix) because the data provided in GEO has been mangled by GeneSpring. We have not been able to determine which data S&P used for that data set (it might have come from Gemma because they mention Gemma in their methods). Another difference is that for GSE54536, the data provided has negative values, which become missing values after log-transformation. S&P do not discuss how they dealt with this but their Figure 1 implies they used the output of INMEX (the heatmap in the paper comes from INMEX; again see the Appendeix for more discussion). We chose to follow what S&P says they did, and log-transform the data independently, but after adding a constant to make all values in GSE54536 the data set positive.

We are using the experimental design files downloaded from Gemma (edited slightly to make them more R-friendly). We have spot-checked these to make sure samples are not mixed up etc. We have also consulted the source publications for further checking and corroboration. In Gemma, batches are defined by the scan date extracted from raw data files, if available from GEO. Samples with the same scan day are considered a batch. There is of course some arbitrariness to doing this if the batches are in consecutive days, but the data here don't have particularly complicated patterns so determining scan batches is fairly unambiguous. For GSE22491, we rechecked the scan dates in the raw data manually. GSE54536 is the only data set for which we lacked batch information.

For quality control (QC) we use sample correlation heatmaps as a quick way to spot oddities, as well as histograms of expression values. Inspection of gender markers is also a useful QC to confirm the design as reported in the source papers. Gender markers tend to show bimodal expression differences in data sets of mixed gender; e.g. XIST is not expressed in males. The caveat is that we do not have gender annotations for all of the data sets, but we were able to use this to identify meta-data issues with two of the studies. In addition, we interpret the lower mode of expression of XIST and other gender markers as one estimate of background noise.

The source R code and data files used to generate this supplement will be made freely available at pavlab.chibi.ubc.ca.

3 Quality control, data preparation and exploration

Referring to the original publications, the GEO records, and the provided meta-data, there are some initial observations about the designs of the studies:

- The authors of GSE18838 note that two samples were outliers: CT#3 and PD#19. CT#3 is included in the data set, PD#19 is not. We suspect that they mean a different control sample was removed.

- GSE22491 (Mutez et al.) has a batch confound in the design: all the HC samples were scanned on April 27 2007, while all the PD samples were scanned on May 7 2007. It also has a pooled sample and an asymptomatic carrier of LRRK2, plus some of the individuals are related (thus perhaps not to be considered independent, but this is a minor issue). We removed the pool but retained the asymptomatic carrier (despite S&P saying that "only samples from PD patients and controls were analyzed", which argues it should have been removed). The sample which is the pool is presumed to be C8, since there is no C8 in Table 1 of Mutez et al. This is GSM558686 and we removed it.
- GSE6613 was scanned in 14 batches spread over a ~1.5-year period, starting in March 2003 and ending June 2004
- GSE54536 has two pooled samples: GSM1318551 ("RNA pool from PD patients 1-5") (sic - there are only 4 PD patients) and GSM1318556 ("RNA pool from PD controls 1-5"). These were retained in the analysis of S&P but we cannot see any justification for retaining them, so we are removing them. No batch information is available.

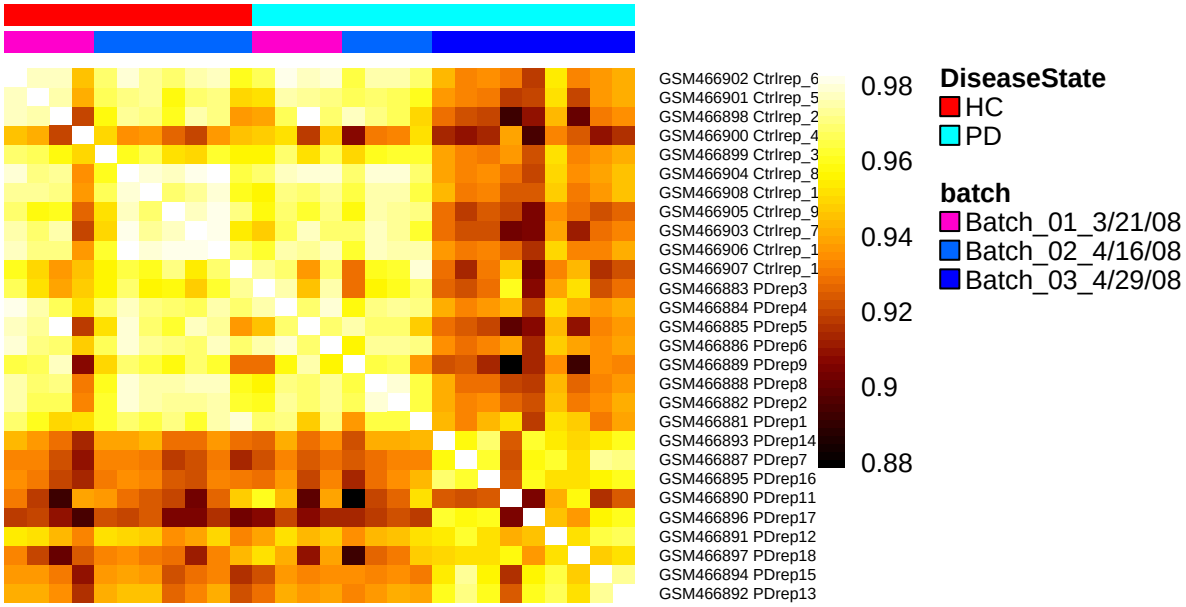
3.1 Sample correlation heatmaps before normalization

In these images the PD status and 'batches' (where available) are indicated by colored bars in the margin. At this point the "pool" samples are retained.

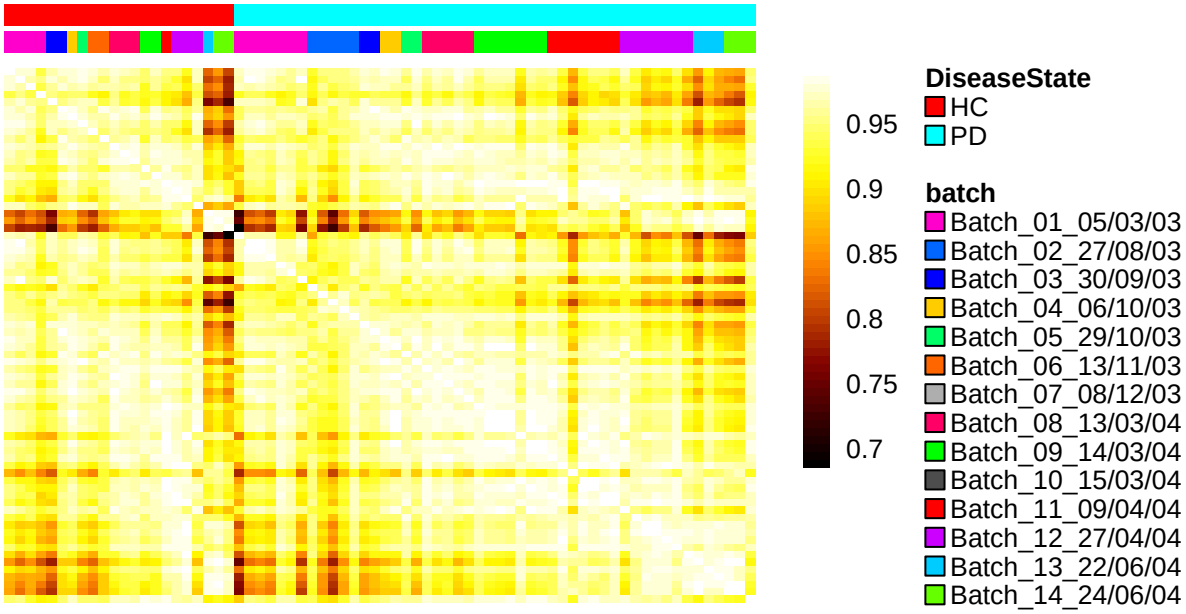
Observations:

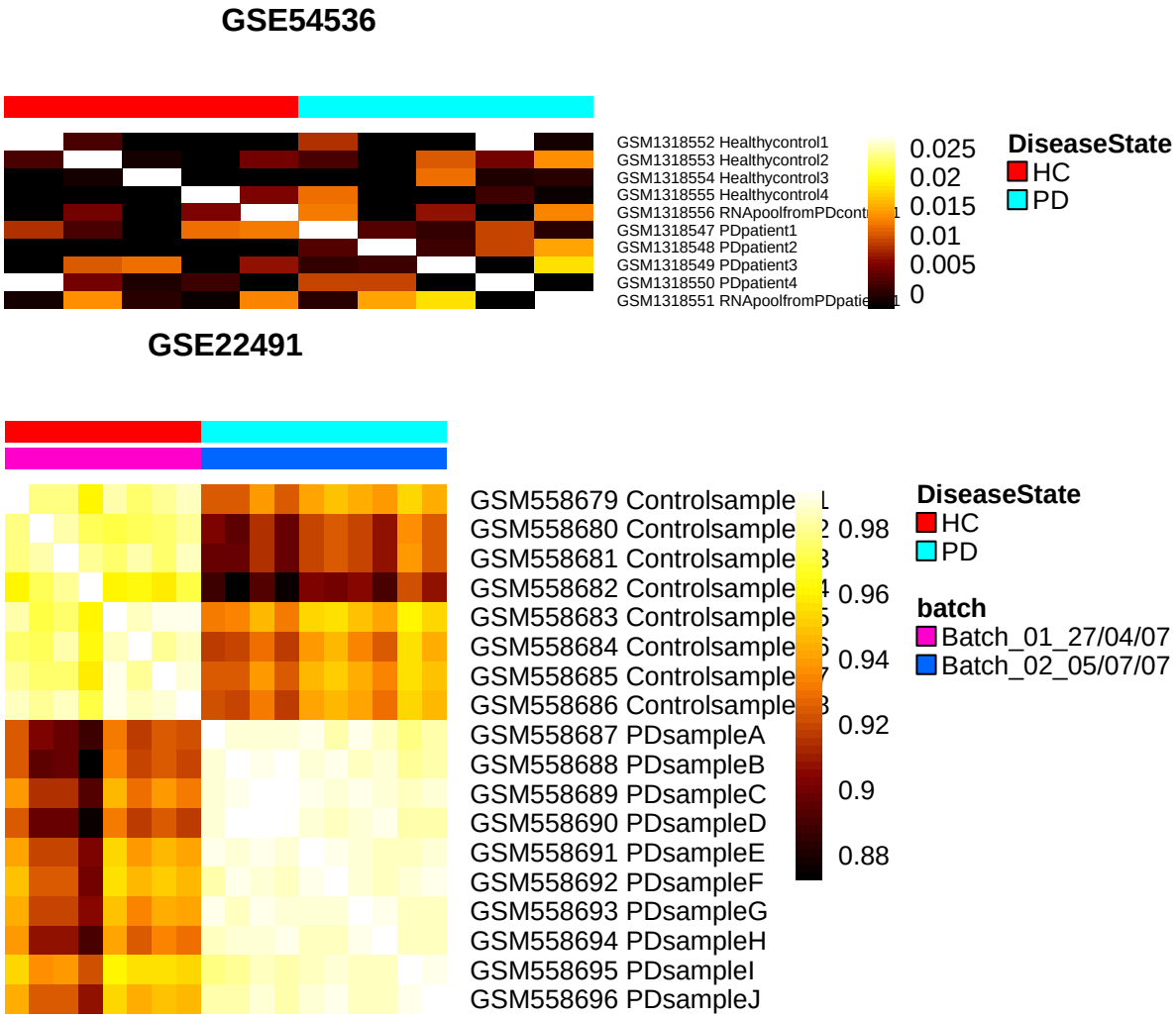
- GSE18838 shows a correlation pattern that reveals a likely batch effect, which is partly confounded with the conditions (the offending batch run on April 29 2008 are all PD samples). Comparison to Figure 4 in Shehadeh et al. strongly suggests that this is responsible for the genes they report as markers in that paper. Sample CT3, which was cited as an outlier by Sheahded et al., does not appear to be an outlier on this basis; possibly they mean some other control. Control 4 and PD17 are the ones that look like outliers but not very bad.
- The block pattern in the heatmap for GSE22491 is concordant with the PD vs. HC grouping, but unfortunately also with "batch". This confound is what renders this data set unusable in our opinion.
- GSE6613 looks like it has a batch effect as well but at least the batches are reasonably balanced across the conditions.
- GSE54536 has all sample correlations near zero. The authors of this data set do not provide enough information for us to understand what they did, because the "non-normalized" data looks reasonable (though containing an outlier; see Appendix; we could not use that data, unfortunately).

GSE18838



GSE6613



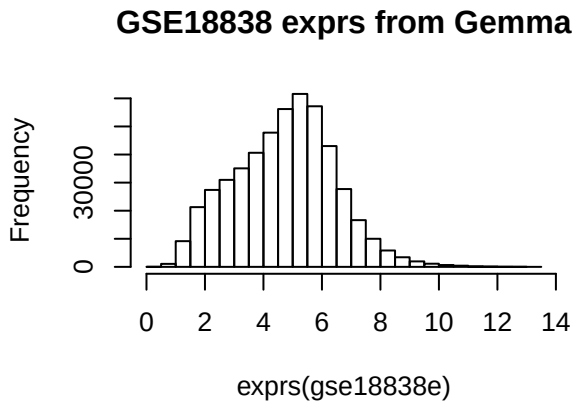


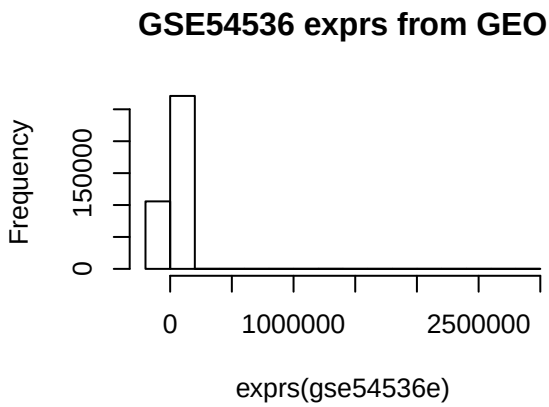
Note that we are not doing anything about the batch effects at the moment. For GSE18838, the batch effect is partly confounded with the diagnosis, but it is more informative to carry it along for now.

3.2 Rescaling and normalization

S&P report that they first log2 transformed, then quantile normalized the data.

The following uses the data scaled as it was first found (all from GEO Series Matrix files except for GSE18838).



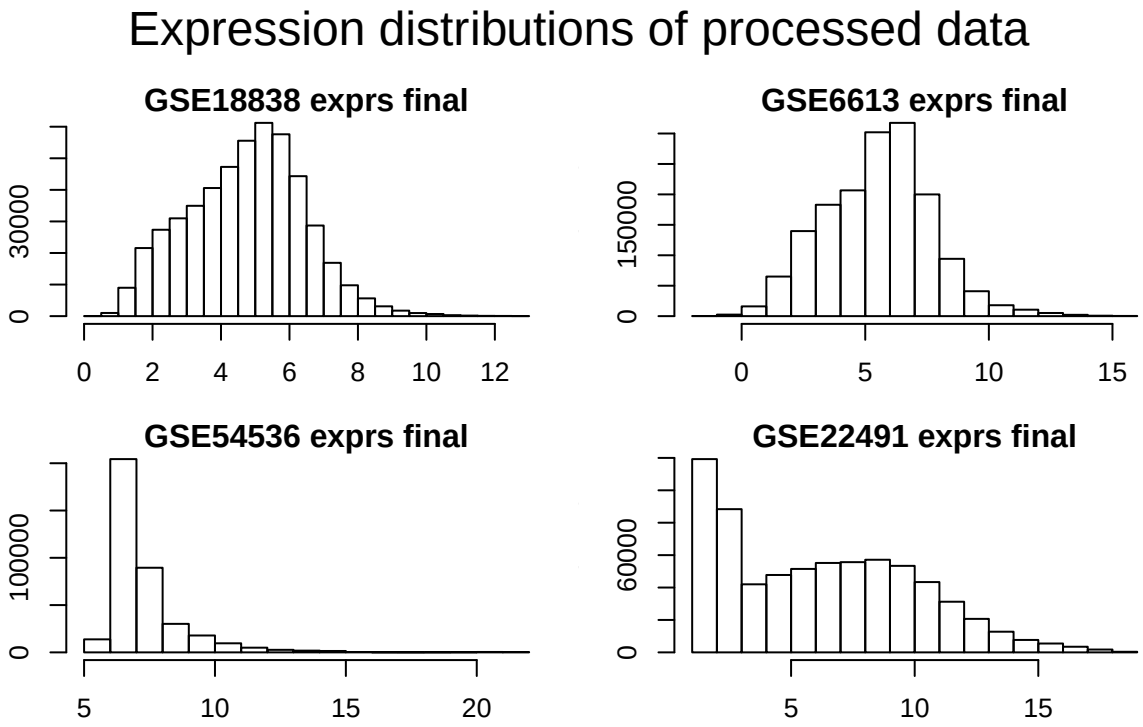


Observations from the above histograms:

- The processed data for GSE18838 offered via GEO had been run through Genespring, which means the values are log ratios per-gene (something like a Z-score); this is reflected in the distribution.
- GSE22491 is not on a log scale so we must log-transform and quantile normalize.
- GSE6613 is not on a log scale so we must log-tranforme and quantile normalize
- GSE54536 has values ranging from -91 to ~2.9 million. See below for how we treated the negative values.

3.3 Log-transforming and quantile normalizing.

This was reasonably straightforward except for GSE54536 (and that GSE18838 was replaced with reprocessed data). Because GSE54536 was not already log-transformed, and contains non-positive values, we checked with the authors of INMEX to see what they do. It emerged that their approach causes undesirable distortions so we used a simpler approach, simply adding a constant to foce all value positive. See the Appendix for a look at an alternative but ultimately unusable version of GSE54536. The new distributions of expression values are plotted in the next figure.

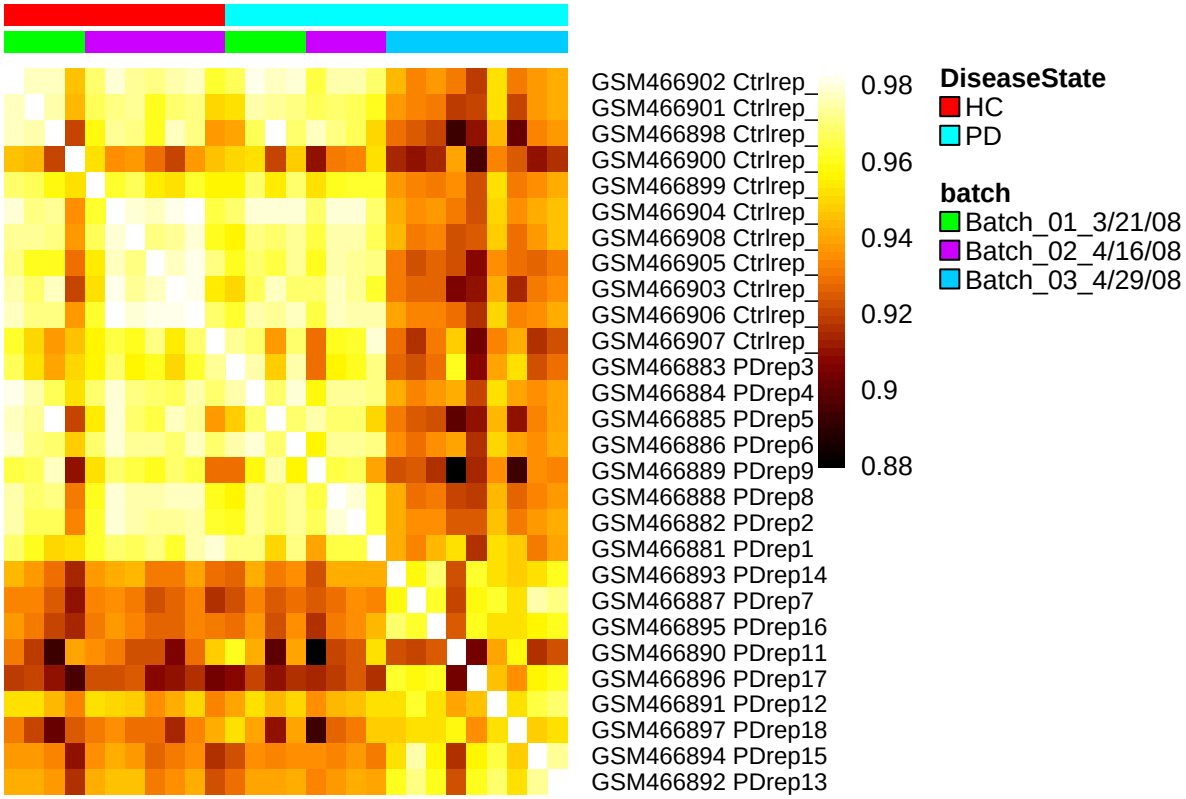


3.4 Updated sample correlation heatmaps

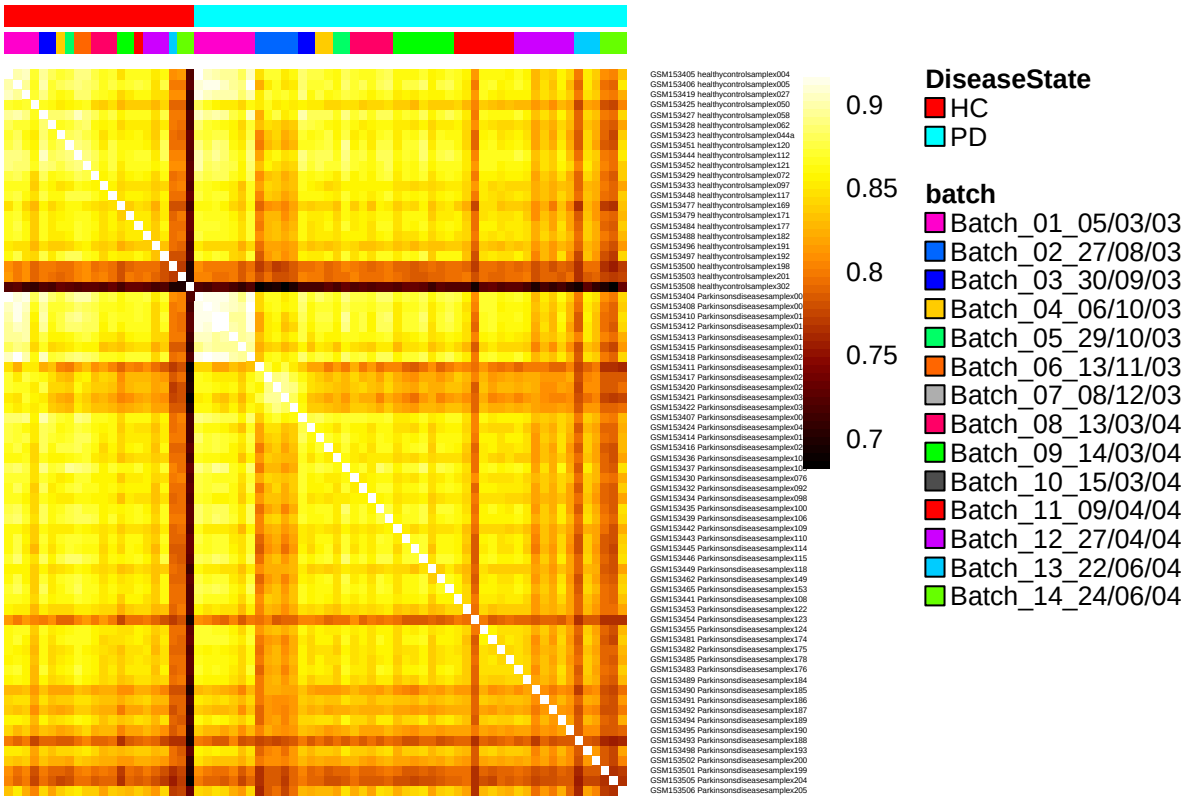
There are a few new observations

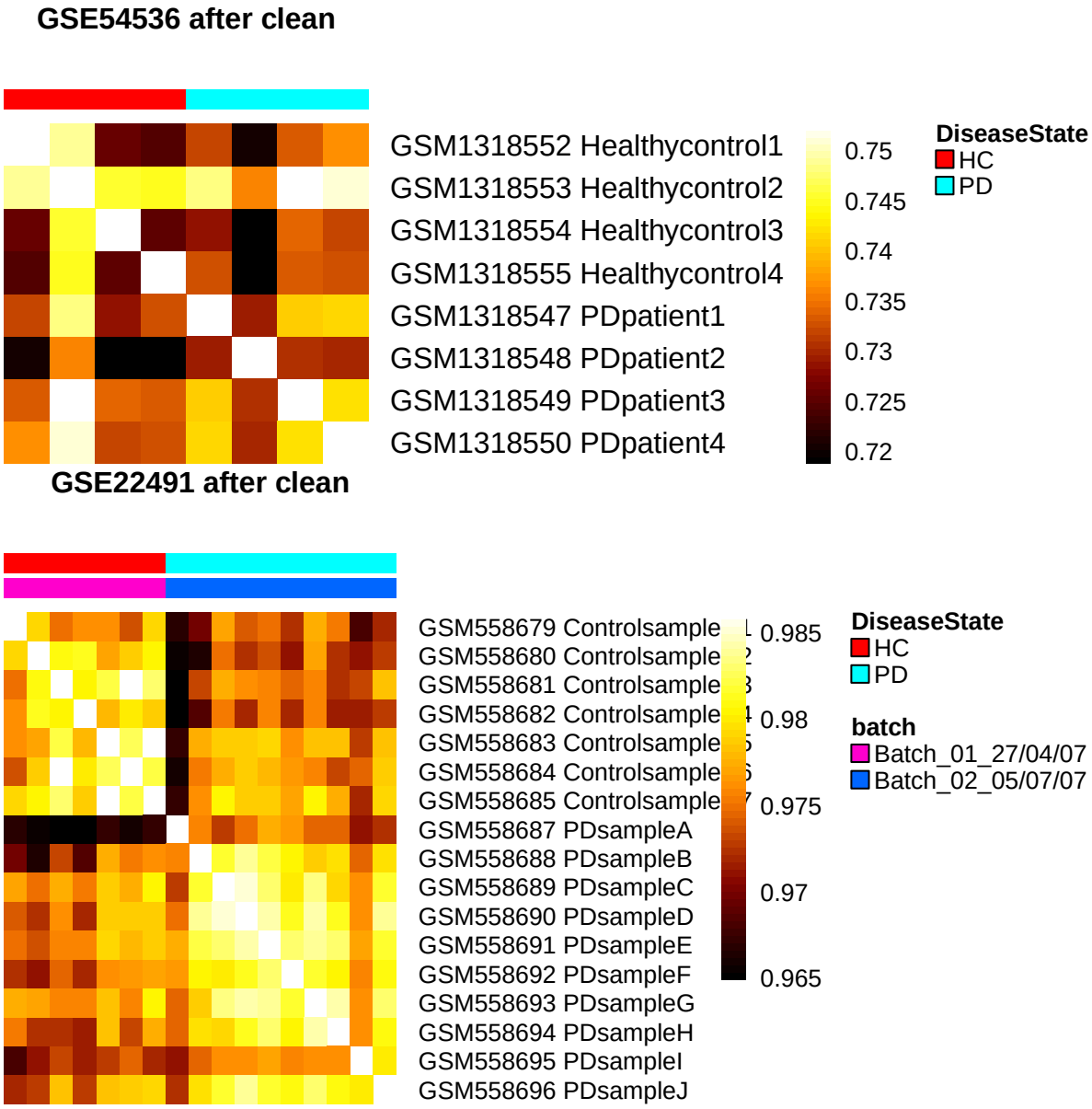
- GSE11838 batch effect still very evident.
- GSE6613 has a fairly clear outlier and some other samples of interest, as well as a likely batch effect (the band of samples that stick out, including the outlier). For the most part we are just ignoring this.
- GSE22491 shown for completeness; group difference looks less extreme, but the confound is still not fixable. The authors of GSE22491 also note an outlier sample, PD10 (GSM558687), likely due to contamination; it appears as an outlier here (lower correlations with other samples). The authors of GSE22491 didn't remove this sample, but "excluded genes differentially expressed in this subject". We retain the sample here as S&P did not remove it. In any case it doesn't affect the major problem of the confound.

GSE18838 after clean



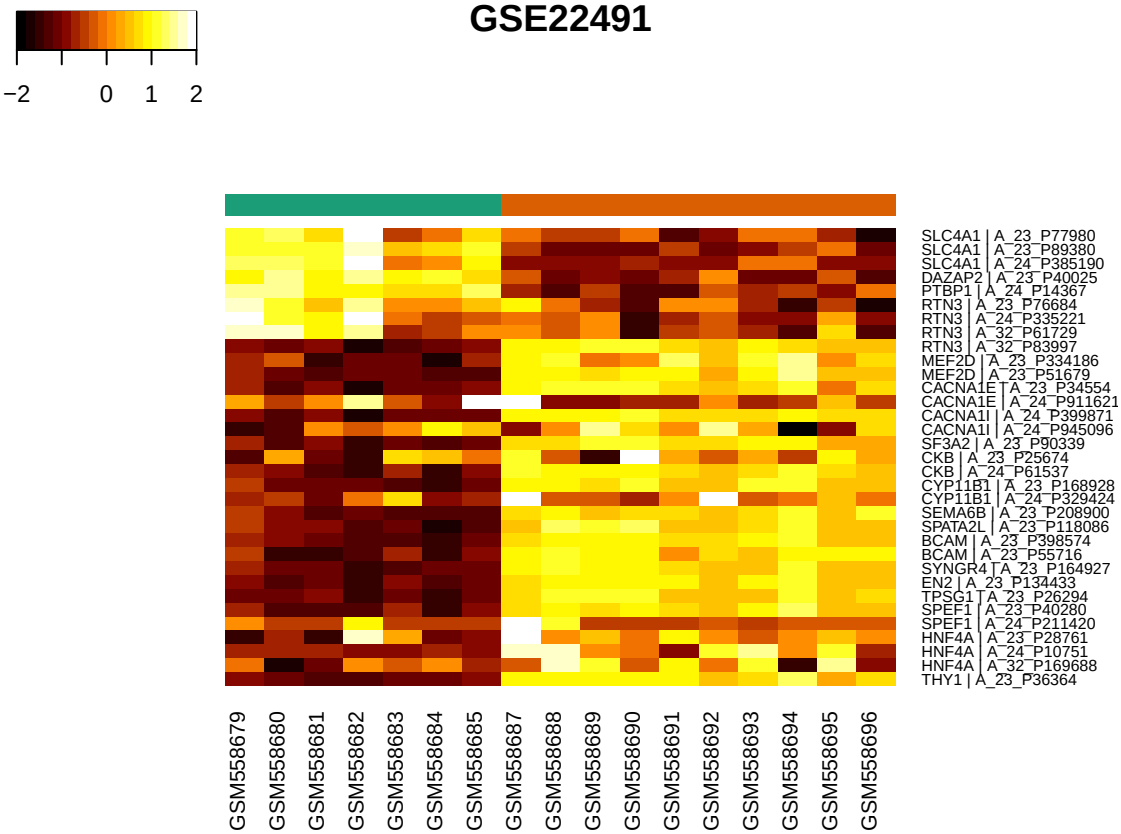
GSE6613 after clean

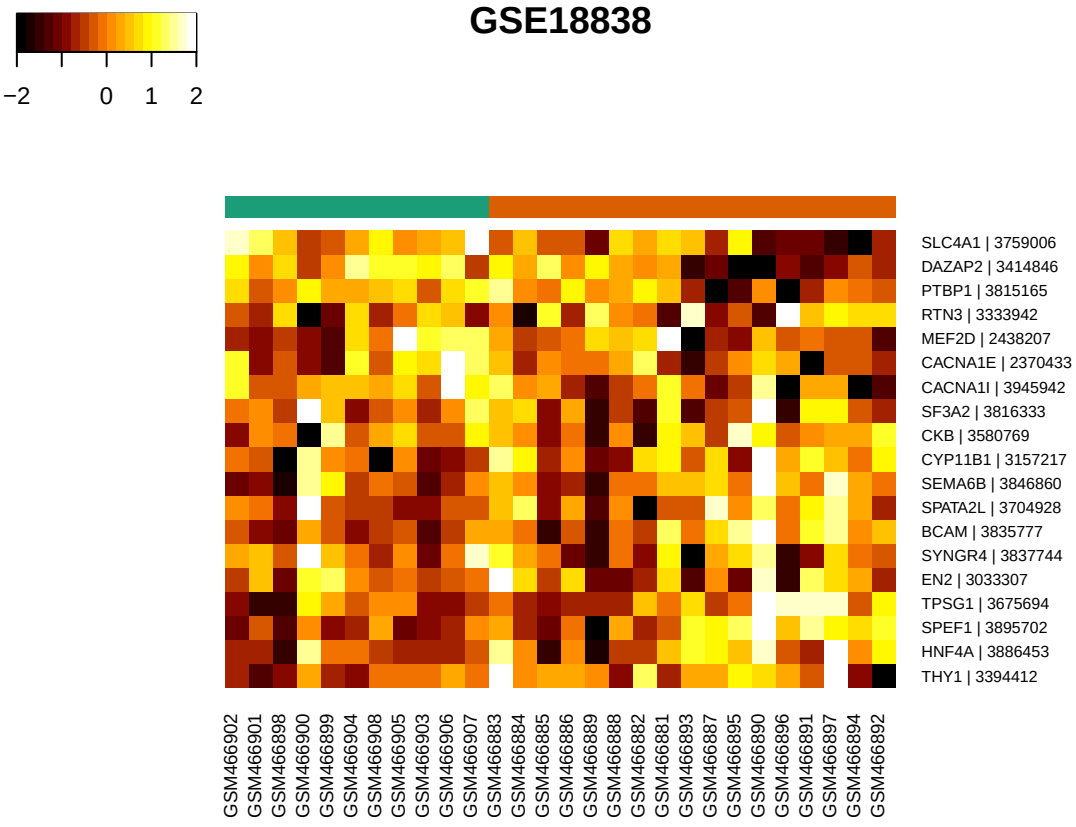


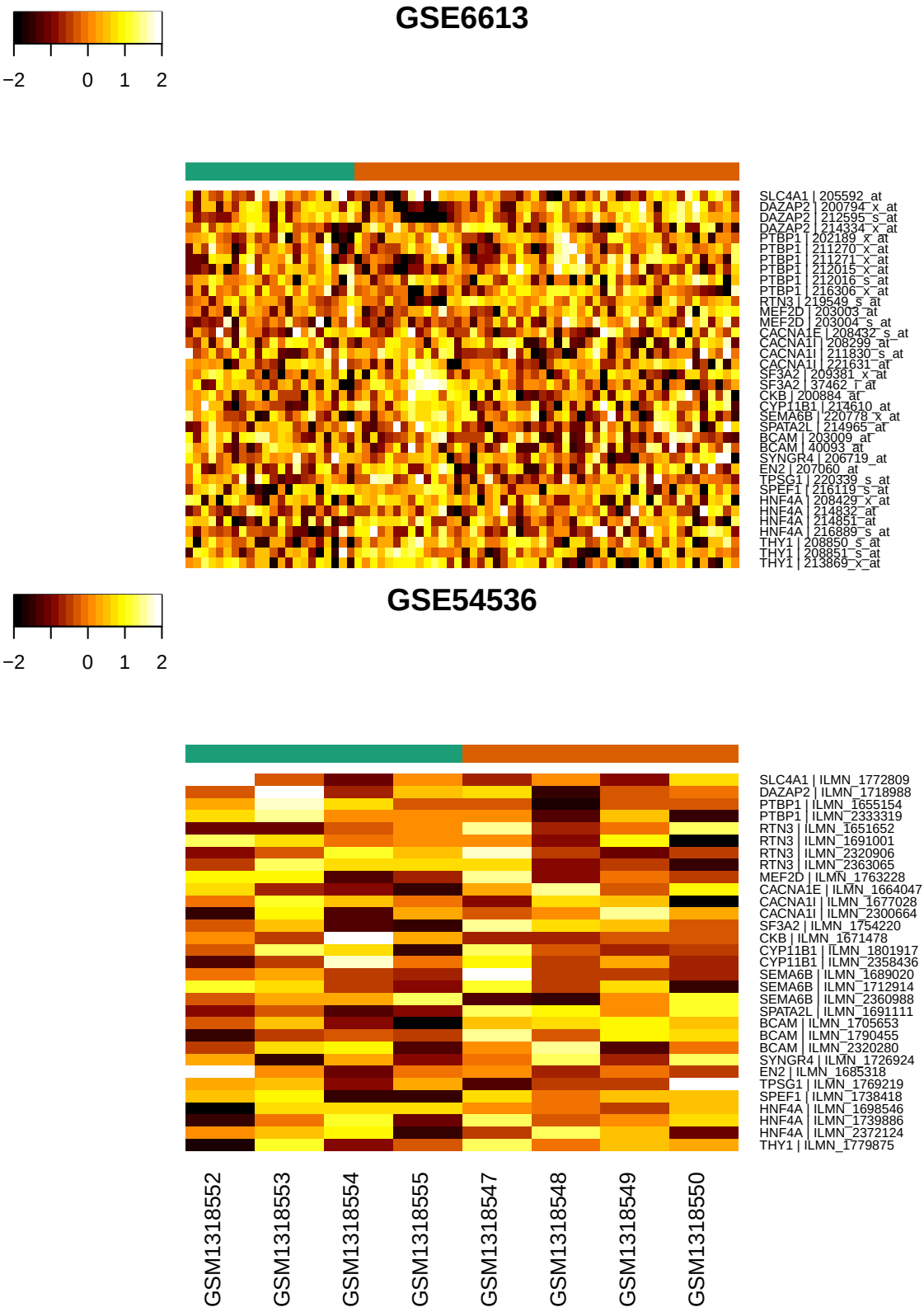


3.5 Heatmaps for the top 19 genes reported

Figure 1 in S&P was taken from the INMEX interface (apparently edited slightly to move the scale bar to the top), and shows the top few genes identified. Here we attempt to approximate the images. See the Appendix for a direct attempt to replicate Figure 1 using INMEX. We omit the sample names for GSE6613 due to readability. Note that there are only 19 genes in Figure 1 of S&P, not 20 as stated in the caption.


GSE22491
−2
0
1
2
SLC4A1 | A_23_P77980
SLC4A1 | A_23_P89380
SLC4A1 | A_24_P385190
DAZAP2 | A_23_P40025
PTBP1 | A_24_P14367
RTN3 | A_23_P76684
RTN3 | A_24_P335221
RTN3 | A_32_P61729
RTN3 | A_32_P83997
MEF2D | A_23_P334186
MEF2D | A_23_P51679
CACNA1E | A_23_P34554
CACNA1E | A_24_P911621
CACNA1I | A_24_P399871
CACNA1I | A_24_P945096
SF3A2 | A_23_P90339
CKB | A_23_P25674
CKB | A_24_P61537
CYP11B1 | A_23_P168928
CYP11B1 | A_24_P329424
SEMA6B | A_23_P208900
SPATA2L | A_23_P118086
BCAM | A_23_P398574
BCAM | A_23_P55716
SYNGR4 | A_23_P164927
EN2 | A_23_P134433
TPSG1 | A_23_P26294
SPEF1 | A_23_P40280
SPEF1 | A_24_P211420
HNF4A | A_23_P28761
HNF4A | A_24_P10751
HNF4A | A_32_P169688
THY1 | A_23_P36364
GSM558679
GSM558680
GSM558681
GSM558682
GSM558683
GSM558684
GSM558685
GSM558687
GSM558688
GSM558689
GSM558690
GSM558691
GSM558692
GSM558693
GSM558694
GSM558695
GSM558696


GSE18838
−2
0
1
2
SLC4A1 | 3759006
DAZAP2 | 3414846
PTBP1 | 3815165
RTN3 | 3333942
MEF2D | 2438207
CACNA1E | 2370433
CACNA1I | 3945942
SF3A2 | 3816333
CKB | 3580769
CYP11B1 | 3157217
SEMA6B | 3846860
SPATA2L | 3704928
BCAM | 3835777
SYNGR4 | 3837744
EN2 | 3033307
TPSG1 | 3675694
SPEF1 | 3895702
HNF4A | 3886453
THY1 | 3394412
GSM466902
GSM466901
GSM466898
GSM466900
GSM466899
GSM466904
GSM466908
GSM466905
GSM466903
GSM466906
GSM466907
GSM466883
GSM466884
GSM466885
GSM466886
GSM466889
GSM466888
GSM466882
GSM466881
GSM466893
GSM466887
GSM466895
GSM466890
GSM466896
GSM466891
GSM466897
GSM466894
GSM466892

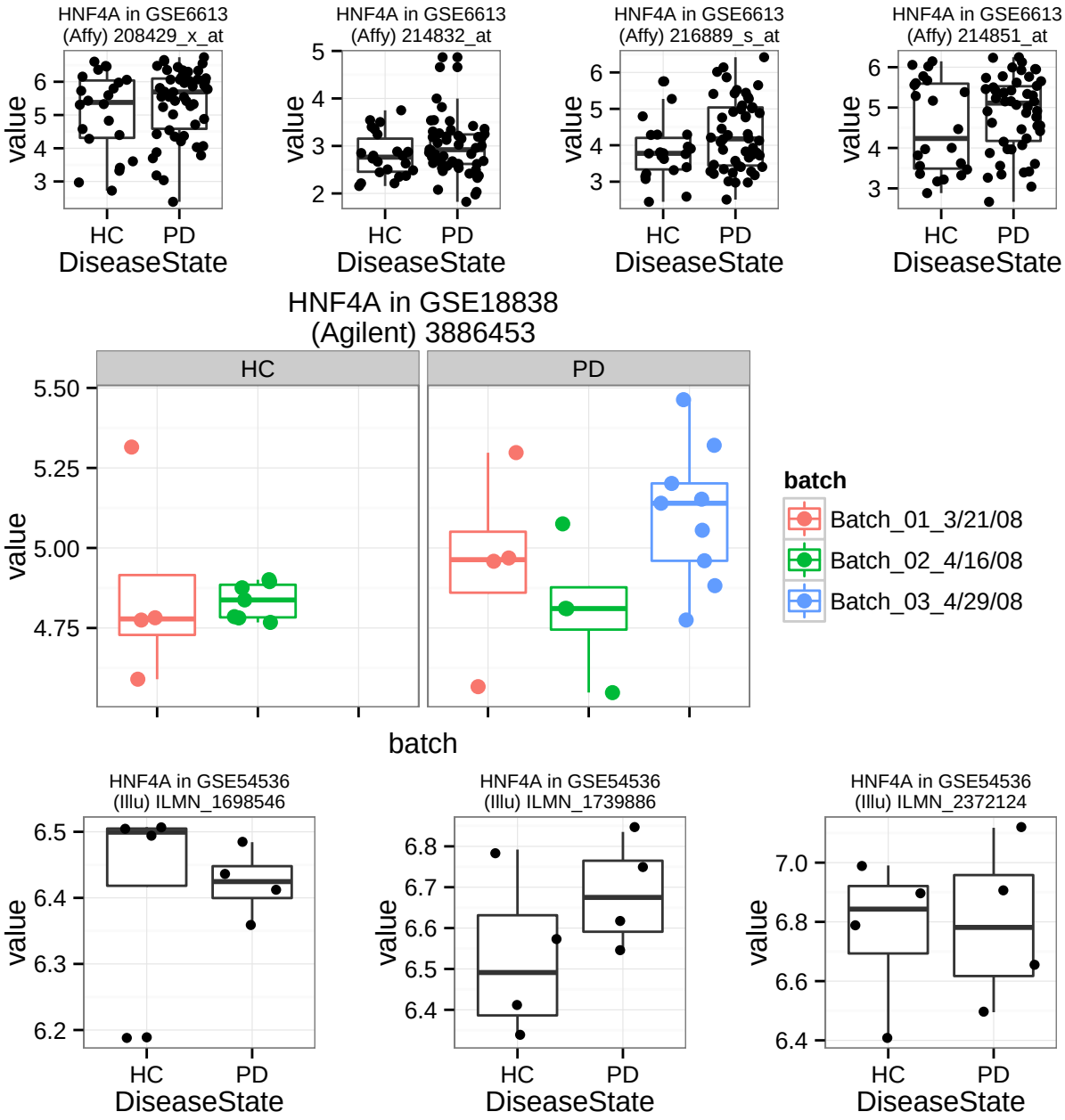


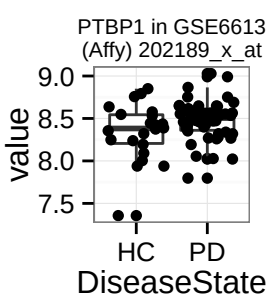
Our observation at this point are that except for GSE22491 (which has a confounding batch effect) differences between HC and PD are very hard to see, in agreement with Figure 1 of S&P.

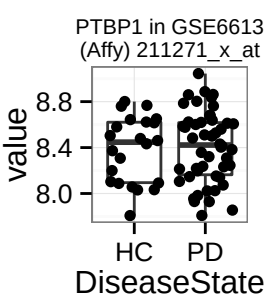
3.6 Inspection of data for the key genes.

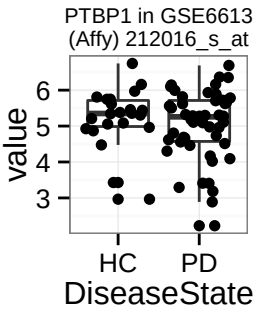
Each plot is for a probe, and the plots are grouped by experiment. Be sure to note the y-axis scales and the differences in the mean expression.

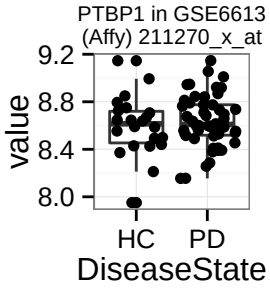
Because GSE18838 has a batch artifact that is partly confounded with the PD status we have shown the data for that data set faceted by batch. This visualization suggests any changes in GSE18838 in HNF4A and PTBP1 seem largely attributable to the third confounded batch.

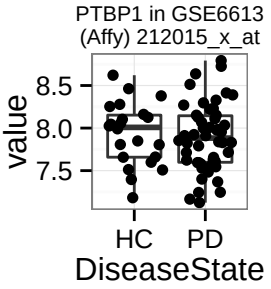


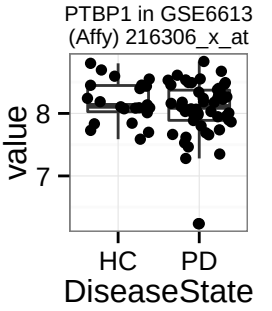

PTBP1 in GSE6613
(Affy) 202189_x_at
value
9.0
8.5
8.0
7.5
HC
PD
DiseaseState

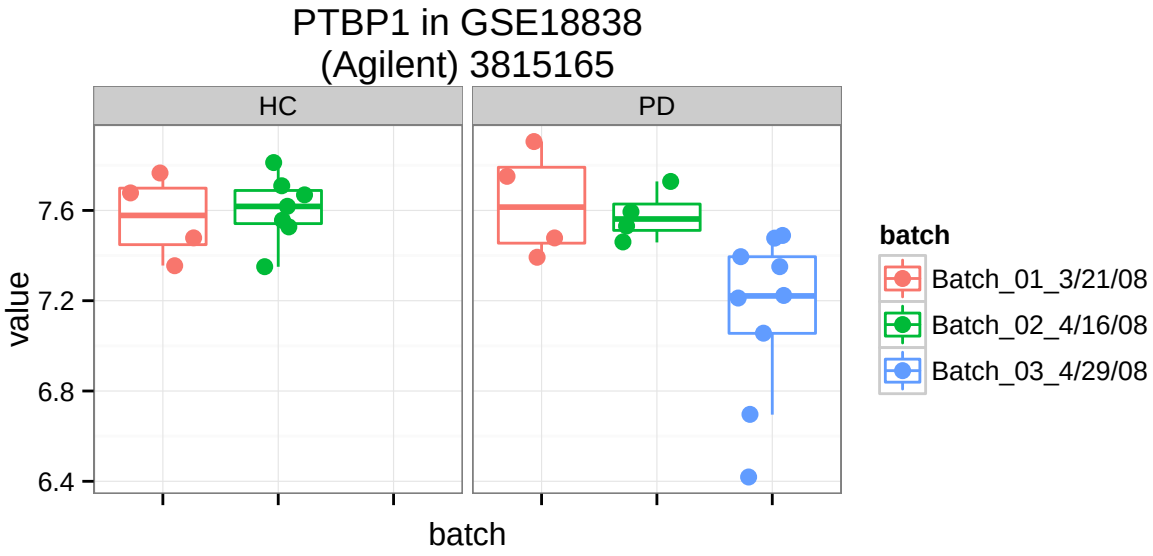

PTBP1 in GSE6613
(Affy) 211271_x_at
value
8.8
8.4
8.0
HC
PD
DiseaseState

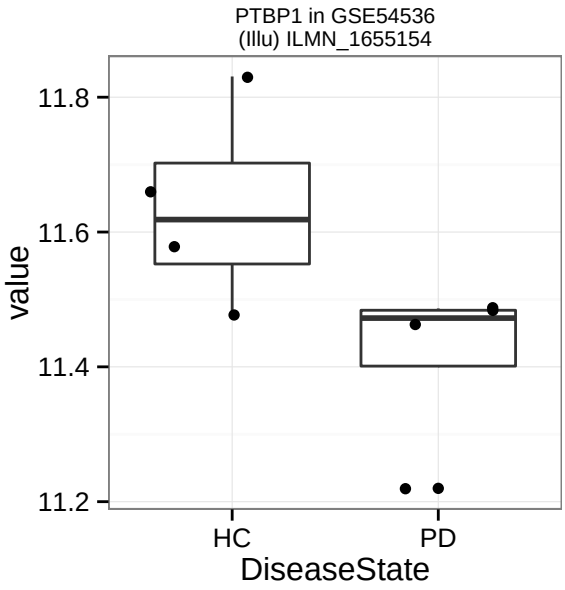

PTBP1 in GSE6613
(Affy) 212016_s_at
value
6
5
4
3
HC
PD
DiseaseState

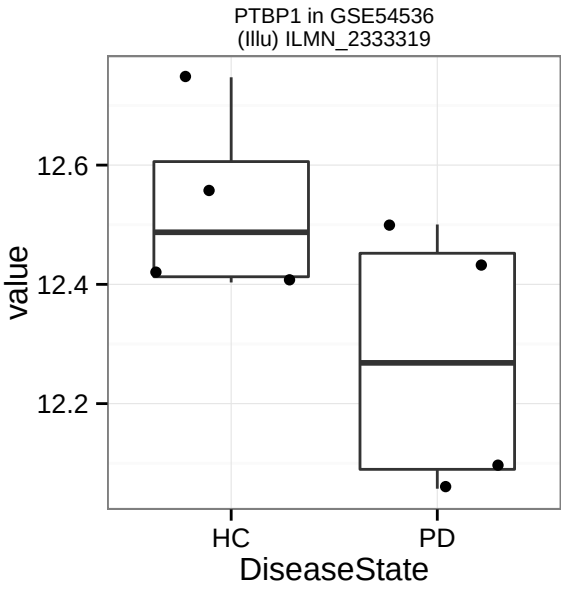

PTBP1 in GSE6613
(Affy) 211270_x_at
value
9.2
8.8
8.4
8.0
HC
PD
DiseaseState


PTBP1 in GSE6613
(Affy) 212015_x_at
value
8.5
8.0
7.5
HC
PD
DiseaseState


PTBP1 in GSE6613
(Affy) 216306_x_at
value
8
7
HC
PD
DiseaseState


PTBP1 in GSE18838
(Agilent) 3815165
HC
PD
value
7.6
7.2
6.8
6.4
batch
batch
Batch_01_3/21/08
Batch_02_4/16/08
Batch_03_4/29/08





4 Differential expression analysis

S&P used limma, via INMEX. For genes with more than one feature, INMEX reports "an average for combined probes". Checking with the authors of INMEX, we found that they take the average of the expression profiles as an initial step in data preprocessing. We do not favour this approach for a number of reasons, and for clarity at this stage it is better to keep the probes separate.

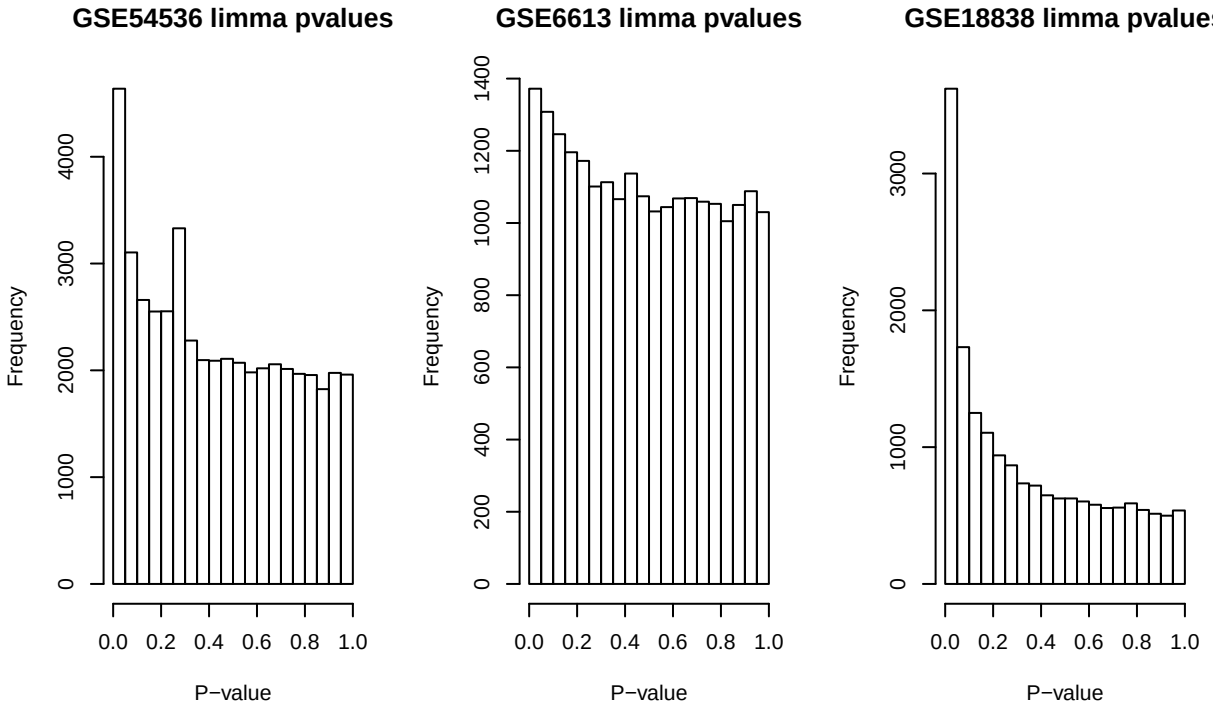
Note that we aren't doing anything special with the direction of change if there are probes showing conflicting directions; we're just looking at the data probe-by-probe, and GSE18838 has not been batch-corrected.

We have omitted GSE22491 from this section because of its problematic batch confound. For an analysis of that data see the Appendix.

4.1 P-value histograms

There is some apparent differential expression in all three data sets, as evidenced by the skewed p-value distributions. Overall the signal is weak, except in GSE18838 (which has a batch artifact). The biggest data set, GSE6613, has a weak signal despite having the most power.

Bear in mind that these are two-tailed p-values.



The weakness of the signals is further evident in using "qvalue" as an alternative FDR control and gene selection method. These tables show how many results are selected at different thresholds.

```
Call:
qvalue(p = gse54536limmar[, "P.Value"][!is.na(gse54536limmar[,     "P.Value"])])

pi0:    0.8180837

Cumulative number of significant calls:

          <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1    <1
p-value       30    267  1420   2776  4637 7741 47231
q-value        0      0     0      0     0    0 47231
local FDR      0      0     0      0     0    0 40115


Call:
qvalue(p = gse6613limmar[, "P.Value"])

pi0:    0.9362576

Cumulative number of significant calls:

          <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1    <1
p-value        1     30   274    710  1372 2680 22283
q-value        0      0     0      0     0    0 22283
local FDR      0      0     0      0     0    0 12343


Call:
```

```
qvalue(p = gse18838limmar[, "P.Value"])

pi0:    0.5781609

Cumulative number of significant calls:

          <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1    <1
p-value       15    194  1300   2397  3620 5351 17839
q-value        0      0     0      0     0 2105 17839
local FDR      0      0     0      0   167 1099 17839
```

4.2 P-values for the genes important to the study.

The main observation is that for the most part, the differences are not even nominally significant (almost all p>0.05) much less interesting at a reasonable FDR (e.g. 0.1). In all cases the data are characterized by very small fold changes, nearly all much less than 2-fold (recall that log2(2)=1), as little as 10%.

For PTBP1 in GSE6613, the direction of change varies from probe to probe (and the differences are miniscule); most likely this is just noise though other interpretations are possible.

It is important not to forget that GSE18838 has a batch artifact as well that is contributing; this is a "best-case" scenario for S&P. Also, these are also two-tailed p-values.

Table 1: HNF4A in GSE54536

	NCBIid	GeneSymbol	logFC	AveExpr	t	P.Value	adj.P.Val	B
ILMN_1698546	3172	HNF4A	0.00	6.42	-0.01	0.99	1.00	-6.06
ILMN_1739886	3172	HNF4A	0.15	6.60	1.32	0.22	0.74	-5.23
ILMN_2372124	3172	HNF4A	0.02	6.78	0.13	0.90	0.98	-6.05

Table 2: HNF4A in GSE6613

	NCBIid	GeneSymbol	logFC	AveExpr	t	P.Value	adj.P.Val	B
208429_x_at	3172	HNF4A	0.29	5.27	1.05	0.29	0.90	-4.78
214832_at	3172	HNF4A	0.16	2.92	1.10	0.27	0.89	-4.75
216889_s_at	3172	HNF4A	0.48	4.16	2.08	0.04	0.80	-3.73
214851_at	3172	HNF4A	0.34	4.76	1.35	0.18	0.86	-4.55

Table 3: HNF4A in GSE18838

	GeneSymbol	NCBIid	logFC	AveExpr	t	P.Value	adj.P.Val	B
3886453	HNF4A	3172	0.15	4.94	1.71	0.1	0.33	-4.54

Table 4: PTBP1 in GSE54536

	NCBIid	GeneSymbol	logFC	AveExpr	t	P.Value	adj.P.Val	B
ILMN_1655154	5725	PTBP1	-0.22	11.52	-2.24	0.05	0.52	-4.02
ILMN_2333319	5725	PTBP1	-0.26	12.40	-1.98	0.08	0.57	-4.38

Table 5: PTBP1 in GSE6613

	NCBIid	GeneSymbol	logFC	AveExpr	t	P.Value	adj.P.Val	B
202189_x_at	5725	PTBP1	0.14	8.44	1.91	0.06	0.81	-3.95
211270_x_at	5725	PTBP1	0.04	8.62	0.61	0.54	0.95	-5.03
211271_x_at	5725	PTBP1	0.02	8.40	0.27	0.79	0.98	-5.13
212015_x_at	5725	PTBP1	-0.04	7.90	-0.40	0.69	0.97	-5.10
212016_s_at	5725	PTBP1	-0.19	5.09	-0.83	0.41	0.93	-4.93
216306_x_at	5725	PTBP1	-0.11	8.11	-1.06	0.29	0.90	-4.78

Table 6: PTBP1 in GSE18838

	GeneSymbol	NCBIid	logFC	AveExpr	t	P.Value	adj.P.Val	B
3815165	PTBP1	5725	-0.23	7.45	-1.98	0.06	0.26	-4.11

4.3 Meta-analysis of HNF4A and PTBP1

The data thus far look at the data one dataset at a time, and one point of a meta-analysis is to identify potentially small but reasonably consistent changes across studies.

We performed a meta-analysis of the results for the two genes in question, using the same method as S&P (Fisher's combined probability test). We exclude GSE22491 and for GSE11838 we include batch as a blocking factor in the call to lmFit (alternatively, pretreating the data with ComBat gave similar results).

If a gene has multiple probes in a data set, we take the geometric mean of the p-values to represent it in the meta-analysis (this is different from what INMEX does, where the data are combined up-front at the profile level). A anti-conservative approach which takes the best pvalue would nudge the p-values lower, but this does not affect the conclusion. We consider one-tailed p-values here, doing one meta-analysis test for "up" and one for "down" regulation with respect to HC (HNF4A was reported to be upregulated by S&P, PTBP1 downregulated, so we only show those tests).

As shown below, the p-values are unremarkable for both genes. A caveat is that we are unsure of how close this is to the data used by S&P.

```
[1] "Meta-P for HNF4A upregulation = 0.09 (Chi-squared 10.81 df=6) rank = 1484"
```

```
[1] "Meta-P for PTBP1 downregulation = 0.07 (Chi-squared 11.84 df=6) rank = 1396"
```

5 Closing remarks

Our analysis shows that without the flawed data set GSE22491, the microarray data does not support PTBP1 and HNF4A being biomarkers of PD.

There are other questions about the design of the meta-analysis. Despite reporting adherence to meta-analysis standards S&P do not mention which data sets were considered for inclusion. It is thus unclear why S&P did not use a fifth microarray dataset, GSE34287, which they had been involved in generating using the PROBE cohort, despite it appearing to meet their inclusion criterion of "10 samples or more". The removal of the pooled samples from GSE53536 drops the study sample size to eight, implying it should not have met the criteria set by the authors.

The positive results with the qPCR on independent cohorts ("PROBE" and "HBS") and a protein microarray (for PTBP1; HNF4A protein was "not identified" in blood) might be offered as the data that excuses the

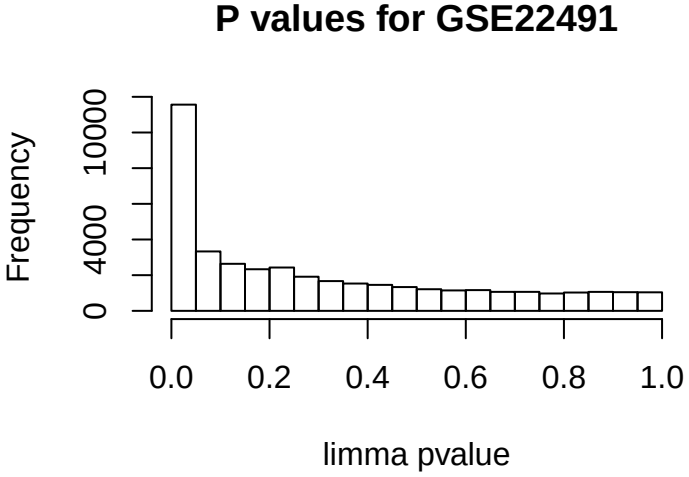
problems we identified with the meta-analysis. However, the other data are not compelling in isolation. Perhaps most importantly, the prediction performance cited (AU-ROC >0.9) for the qPCR data was based on using the same data for testing and for training the classifier (that is, there was no cross-validation or independent validation, a point that Dr. Potashkin confirmed in an email). This procedure generally leads to overestimated accuracy. The failure of the biomarkers to discriminate cases from controls at a later time point (Figure 5 of S&P) is interpreted by the authors as an interesting and potentially useful phenomenon of "longitudinal dynamics". An alternative interpretation is regression to the mean. Finally, the use of a Student's t-test for data distributed as illustrated in Figure 3 is questionable, as is the use of linear regression in Figure 4. A caveat here is that our investigation has been limited by a lack of availability of the raw qPCR data to us at this time, but we at least suggest these data deserve further scrutiny before acceptance as a stand-alone demonstration of the power of these RNAs to predict PD status.

6 Appendices

6.1 Differential expression in GSE22491

We are showing GSE22491 separately to make it clear we consider any differential expression in this study is likely to reflect a technical artifact. Limma was run on this data just as for the other data sets.

As can be seen in the p-value distribution and table of significant calls for different FDR thresholds, this data set is an outlier even compared to GSE11838 (which also has a batch effect).




```
Call:
qvalue(p = gse22491limmar[, "P.Value"])

pi0:    0.4954147

Cumulative number of significant calls:

          <1e-04 <0.001 <0.01 <0.025 <0.05  <0.1    <1
p-value     2257   3846  7026   9219 11563 14895 41000
q-value     1227   2329  4724   6625  8817 12322 41000
local FDR    827   1510  2991   4036  5185  7023 40992
```

Here are the pvalues for the key genes in GSE22491, which we repeat is likely attributable to a technical artifact. According to one probe, HNF4A has a 6.2-fold change in PD vs HC. PTBP1 is "down-regulated" nearly 2-fold.

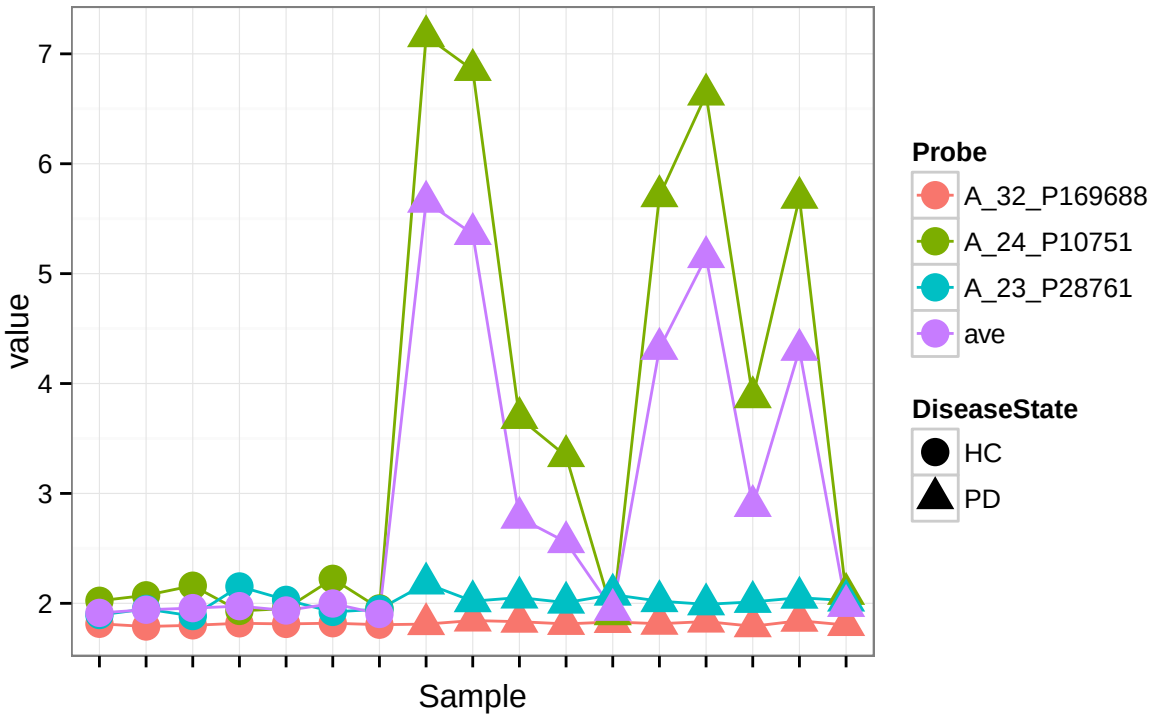
Table 7: HNF4A in GSE22491

	NCBIid	GeneSymbol	logFC	AveExpr	t	P.Value	adj.P.Val	B
A_32_P169688	3172	HNF4A	0.01	1.82	0.42	0.68	0.81	-7.02
A_24_P10751	3172	HNF4A	2.65	3.60	3.69	0.00	0.02	-1.89
A_23_P28761	3172	HNF4A	0.08	2.01	1.63	0.12	0.31	-5.82

Table 8: PTBP1 in GSE22491

	NCBIid	GeneSymbol	logFC	AveExpr	t	P.Value	adj.P.Val	B
A_24_P14367	5725	PTBP1	-0.94	6.51	-8.84	0	0	7.87

For HNF4A, these p-values do not reflect what S&P might have seen, because in INMEX the data for each gene is first averaged, creating an expression profile per-gene. This does not affect PTBP1 because there is only one probe for that gene (at least, in the annotations we are using).The data for the three probes for HNF4A was shown in the heatmap above, but that obscures the underlying values. Here is the data plotted more conventionally, along with the average.



In the HC samples (left-hand of plots), expression of all three probes is very low and invariant. But one probe dramatically changes its behaviour in the PD. The average shows very strong differential expression with respect to PD status with very low variance in the HC group. We caution that we are not exactly sure if this is reproducing what S&P used. For this plot the data were averaged before log-transforming, which we believe is what INMEX did. Obviously taking the arithmetic mean of the data after log transformation would cause the average to look less like the "outlier" probe but the impression is the same.

6.2 Replication of INMEX analysis

To bolster our confidence that we were using similar data to what S&P used, we attempted to run INMEX on the data. There were some challenges; at this writing not everything has been sorted out so this is preliminary data. Some of the aspects where we didn't know how exactly S&P handled INMEX:

- INMEX does not have annotations for the platform used in GSE22491 (the platform is GEO ID GPL6480, Agilent array), GSE18838 (GPL5175, Affy exon array) or GSE54536 (GPL10558, Illumina). While S&P do not state it clearly, they must have used annotations from elsewhere. For our analysis we use the annotations from Gemma.
- In our hands, entering data containing negative values caused the meta-analysis to fail in INMEX. As a work-around we added 100 to the GSE54536 expression values prior to upload, as we did for our independent analysis above.
- Inspection of Figure 1 of S&P shows that in GSE6613 there are only 21 control samples and not 22 as we found in the GEO data. Our guess is that S&P may have removed the outlier sample we noted above, as it is also obvious in the PCA offered by INMEX. We removed this sample for the INMEX analysis. This has only a minor effect as GSE6613 is a large data set.
- Minor: Figure 1 of S&P only has 19 genes, though the legend says it is the top 20.
- We found that repeated runs of INMEX could yield different results if the data sets were entered in a different order. The heatmaps generated by INMEX often have the colored bars labeling the conditions (along the top) mislocated. We are working with the authors of INMEX to isolate these problems.

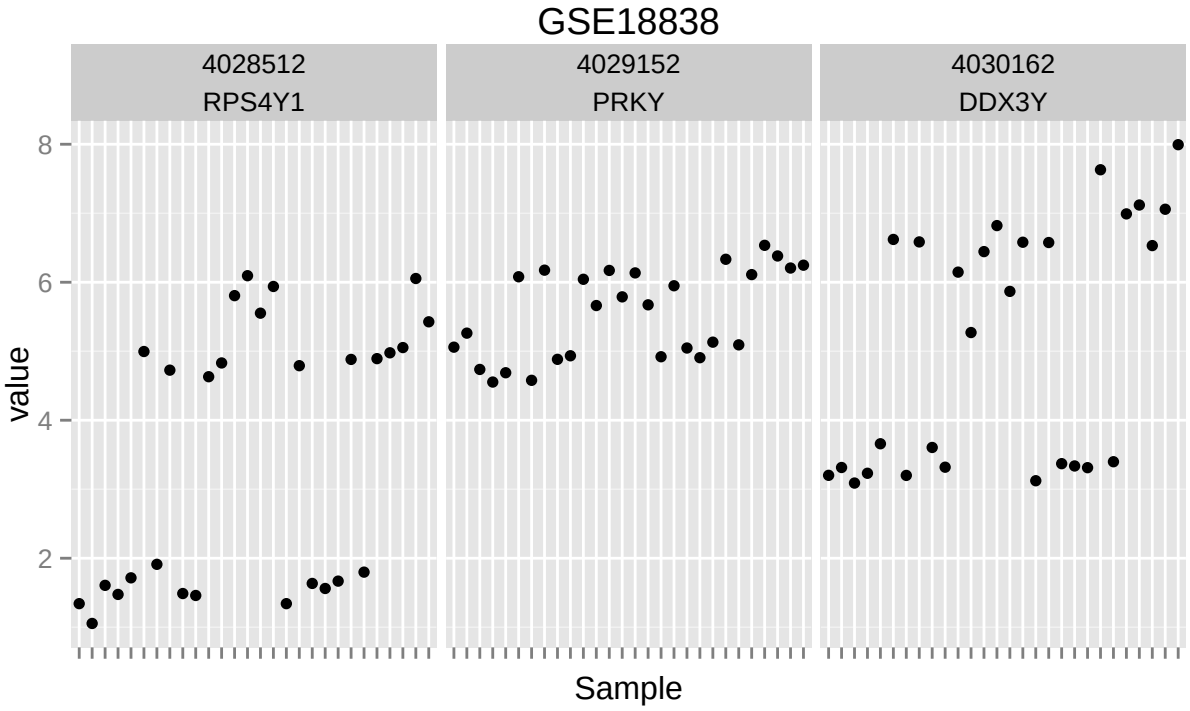
In our closest replication to date, HNF4A is ranked 797 and PTBP1 is ranked 19; 13 of the top 20 genes are the same as those reported as the top 19 by S&P (our top 20 is actually 24 due to tied p-values reported by INMEX). Any difference might be attributable to differences in annotations and the data itself, but because INMEX does not give completely consistent results in repeated runs, we are not sure.

6.3 Inspection of gender markers

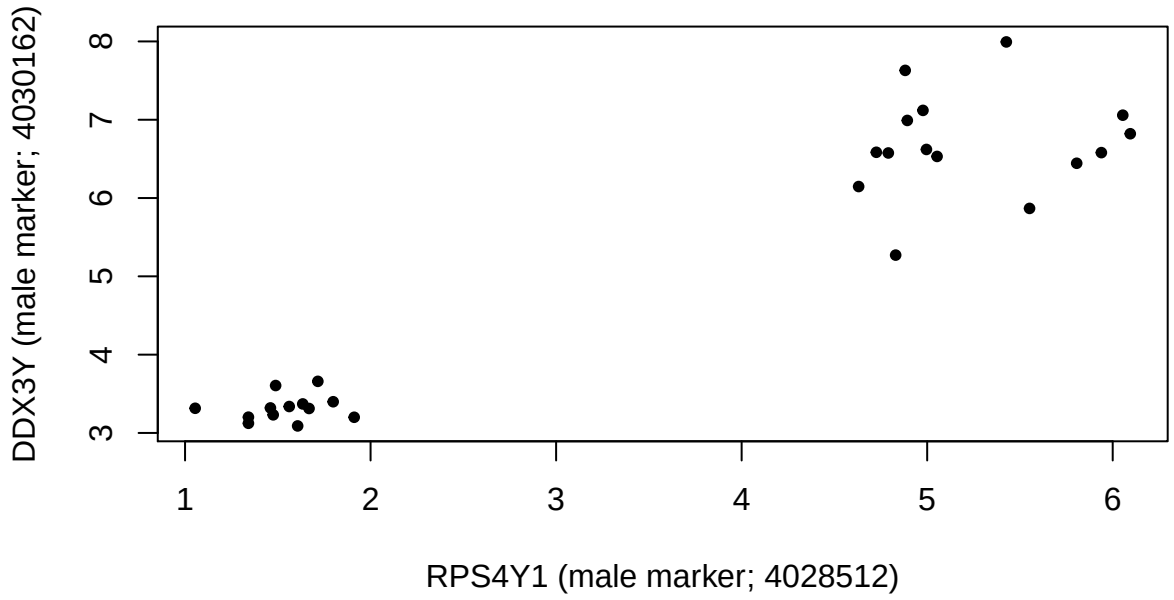
We have gotten into the habit of predicting gender in data sets using known markers, and comparing it to available annotations. This is a good idea because it is very common for samples to have been mixed up by experimenters, or for them to be misannotated: the annotated gender does not match the gender predicted based on markers. Samples that seem mistaken are candidates for removal. Here we are not removing such cases (only two data sets provided gender information, GSE22491 and GSE18838) but just noting them.

We are also using these markers to provide one rough estimate of background noise as mentioned under Methods. For GSE18838, background is ~3-4. For GSE22491, ~2-3; for GSE6613, around 6-7; for GSE54536, around 6-7. Comparing this analysis to the expression levels in the differential expression section (above) suggests that HNF4A is expressed at very low levels (near background in most cases).

We notice that in GSE18838, the number of males and females predicted based on the markers does not match that reported by the authors (Shehadeh et al.).



The plot above suggests there are 13 females, 15 males, which is made clearer by a direct comparison of two markers (there's no good female marker on this platform)

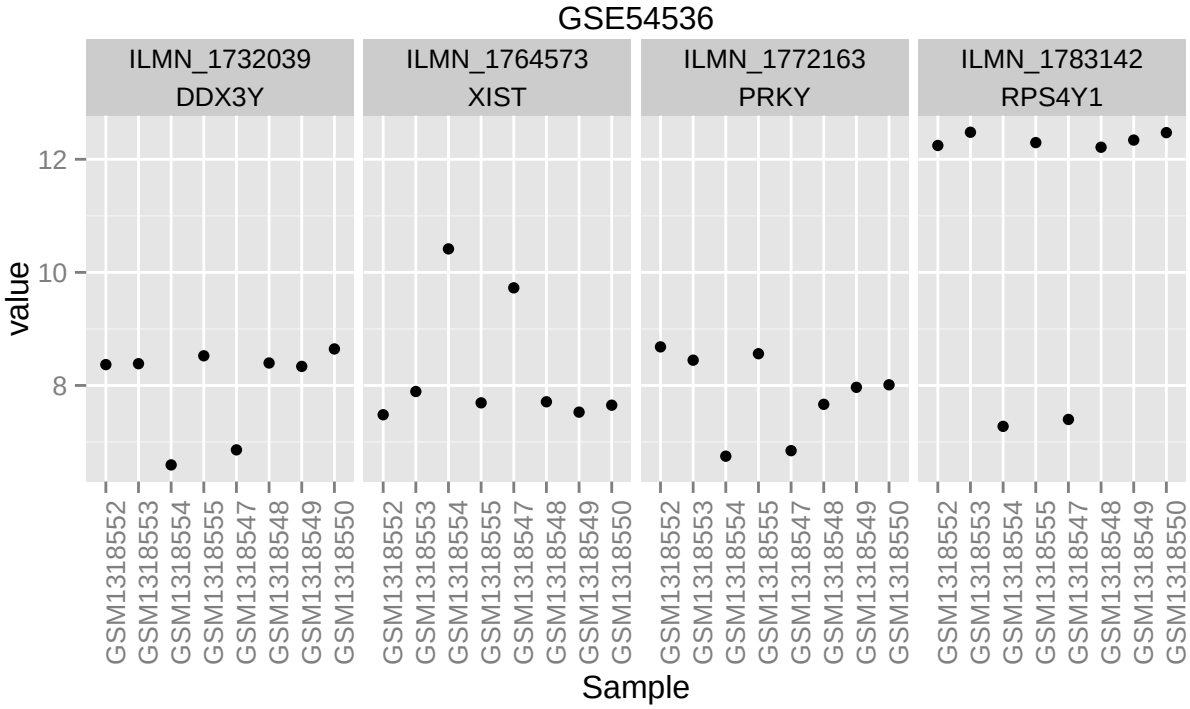


The reported gender counts from supplementary table 6 of Shehadeh et al. differ from this:

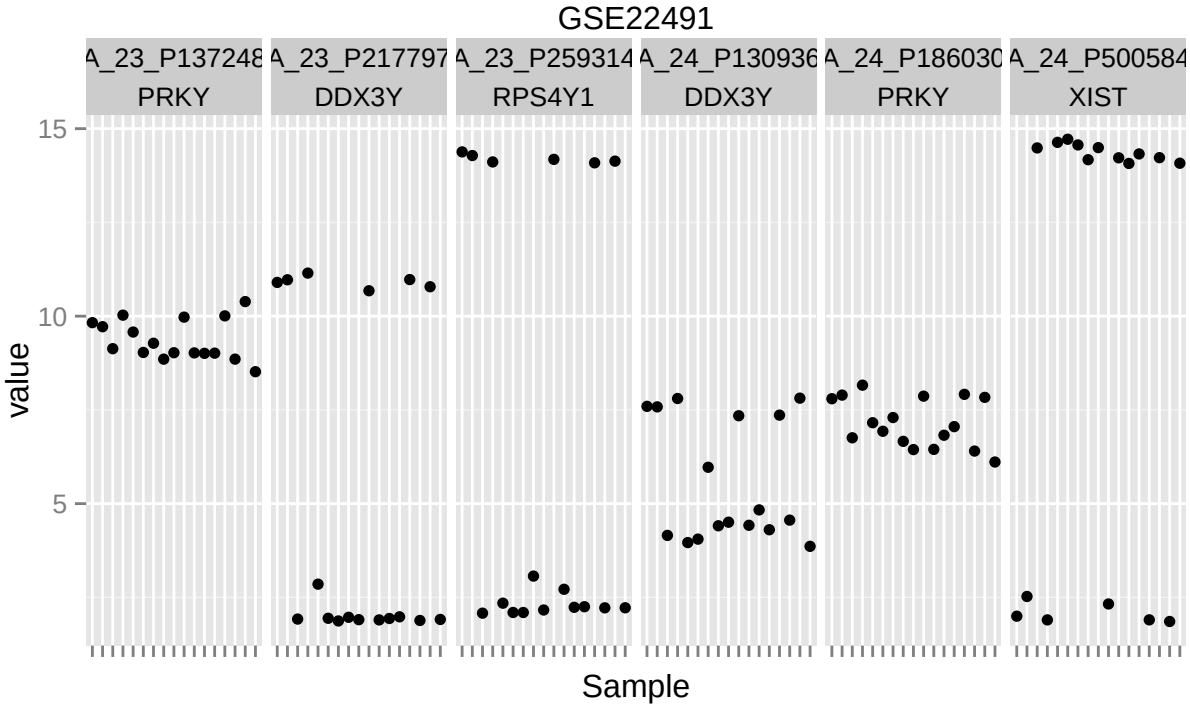
```
 F  M
10 18
```

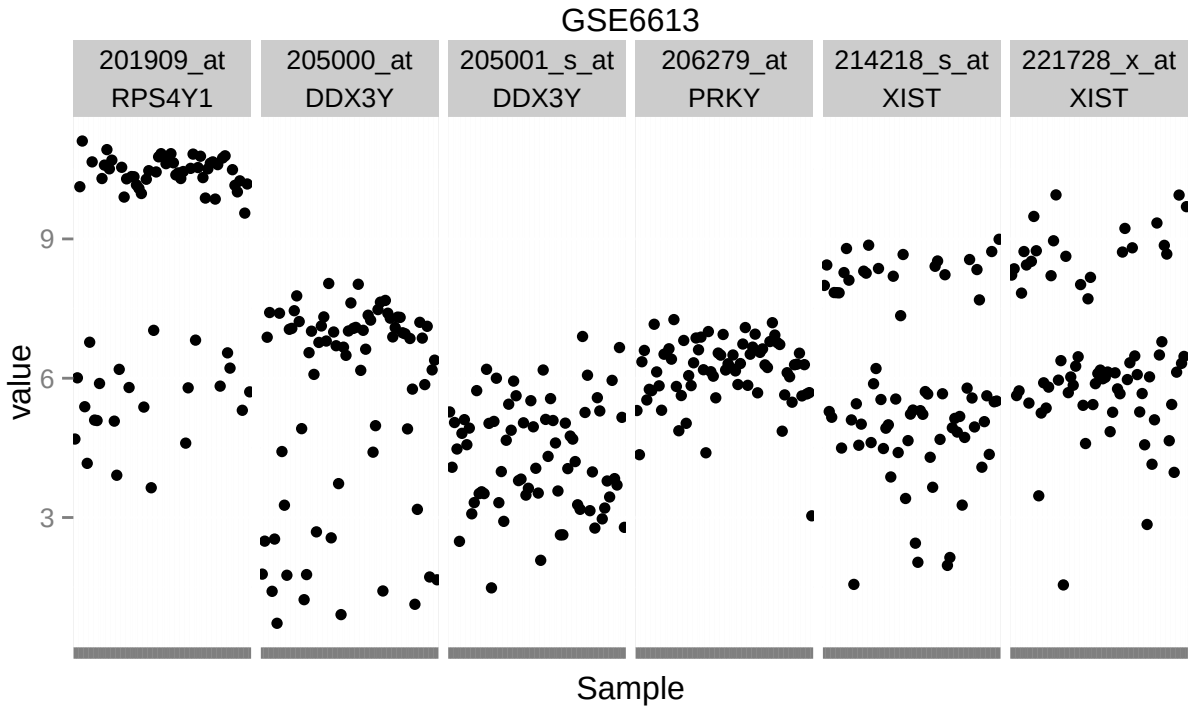
This is surprising because Shehadeh et al. report confirming gender and provide a figure that purports do document this, Figure S2. But Figure S2 of Shehadeh et al. has sample labels that are not present in the supplementary table, such as PD8_26R. We conclude tentatively that the meta-data is somehow erroneous. We have brought this to their attention.

For GSE54536, it apears there are 6 males and 2 females, which disagrees with the gender matching reported by Alieva et al. (2014):



For completeness we show similar plots for the other two data sets, to judge background expression levels. For GSE22491 the predicted genders match the count reported in the paper; for GSE6613 we don't have any additional information.

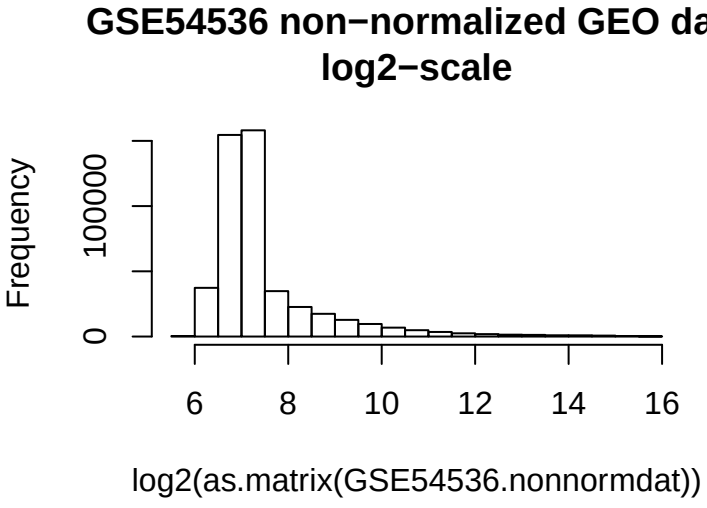


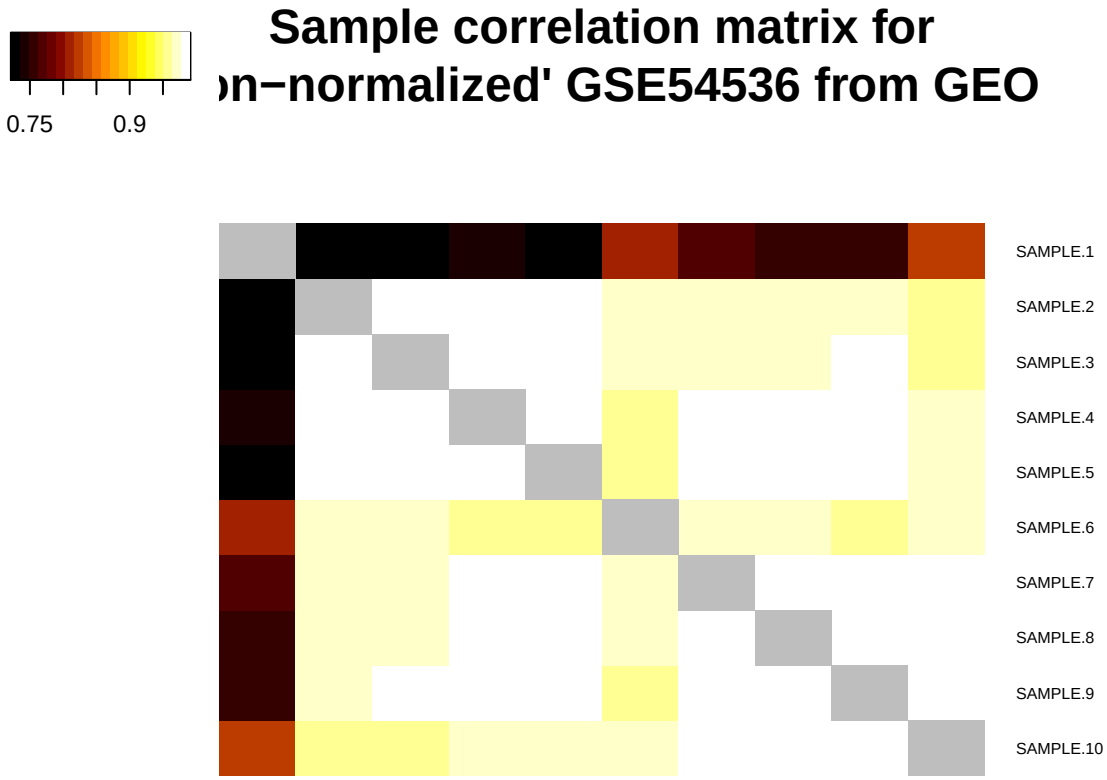


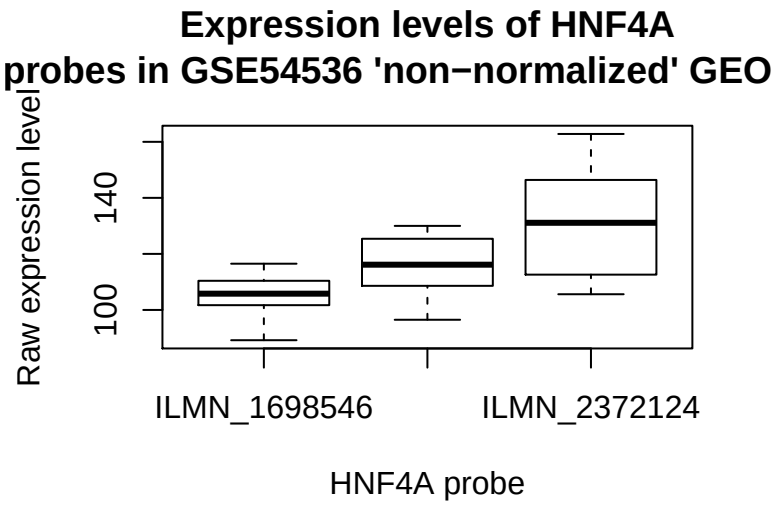
6.4 Investigating the GSE54536 'non-normalized' data

The GSE54536 data in GEO are quite oddly distributed and the paper (Alieva et al. 2014) is unclear as to how it was treated. We looked at the 'non-normalized' data provided in the supplementary files in GEO. Unfortunately those data do not have any sample labels, so we could not determine which were HC and which were PD, thus it is not possible to use that data for analysis. We have emailed the corresponding author of the source paper (March 30 2015) for clarification without response to date.

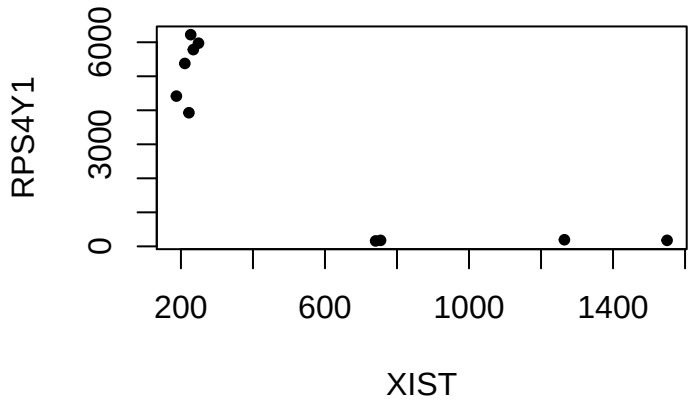
This 'nonnormalized' data looks more or less okay (non-logged) but SAMPLE.1 is an outlier. No outlier is evident in the data we are using in the main section, in any case.







Alieva et al. say they "matched" cases and controls for gender but they don't say what that gender was for each sample.



Based on this it appears there might be four females and six males. Based on the gender markers in the data we used for the main analysis, after removing the "pool" samples there are 2 females and six males. On this plot it looks possible that the two points near x=750 are the pooled samples as they have the lowest expression of XIST. In any case, what Alieva et al. may have meant is that they balanced gender, with an equal number of males in each group.

The males have expression of XIST in the range 200: higher than HNF4A.

Thus HNF4A is expressed at background levels in the GSE54536 'non-normalized' data, supporting our hypothesis that HNF4A is difficult to detect in blood.