Exploring Data Posted for Cisplatin and Pemetrexed

Keith A. Baggerly

November 9, 2009

Contents

1 Executive Summary **3**
1.1 Introduction . . . 3
1.2 Methods . . . 3
1.3 Results . . . 4
1.4 Conclusions . . . 5

2 Options and Libraries **6**

3 Earlier Rda Files **6**
3.1 Data from Baggerly and Coombes [2] . . . 6
3.2 Data from Baggerly et al. [1] . . . 7

4 Cell Lines in the Predictors **7**
4.1 Cisplatin . . . 7
4.2 Pemetrexed . . . 9
4.3 Summary . . . 12

5 Cisplatin Gene List **12**
5.1 Reading Data, Checking for Outliers . . . 12
5.2 Comparing this List with Others . . . 13
5.3 Summary . . . 14

6 Pemetrexed Gene List **14**
6.1 Reading Data . . . 14
6.2 Comparing this List with Others . . . 14
6.3 Summary . . . 15

7 Cisplatin Training Data **15**
7.1 Read Data . . . 15
7.2 Match to Györffy et al. [6] Cell Lines . . . 16
7.3 Attempted Reproduction with Binreg . . . 16
7.4 Summary . . . 18

8 Pemetrexed Training Data **18**
8.1 Read Data . . . 18
8.2 Match to NCI60 Cell Lines . . . 19
8.3 Attempted Reproduction with Binreg . . . 20
8.4 Summary . . . 20

9 Ovarian Testing Data **20**
9.1 Read Data . . . 20
9.2 Check Correlations . . . 21
9.3 Checking What Matches with What . . . 21
9.4 Checking Clinical Information for the Last 43 . . . 25
9.5 Checking Clinical Information (and Labels) for the First 16 . . . 26
9.6 Checking Missing Data . . . 27
9.7 Totaling Clinical Counts . . . 28
9.8 Matching the First 16 . . . 28
9.9 Summary . . . 32

10 Appendix **32**
10.1 File Location . . . 32
10.2 Saves . . . 32
10.3 SessionInfo . . . 32

List of Figures

1 Plot of the negative log10 concentrations required to achieve 50% growth inhibition (NLOGGGI50 values) for the NCI60 cell lines. The cell lines now reported are marked. TK-10, inferred as resistant by Baggerly and Coombes [2], and ACHN, labeled as resistant now instead of TK-10, are also shown. Shifting from TK-10 to ACHN does make the cell line groupings more coherent, but does not address the main problem with this data noted by Baggerly and Coombes [2]: cell lines sensitive to pemetrexed are labeled as resistant, and vice-versa. . . . 11
2 Heatmaps for cisplatin using the (a) the data now reported, and (b) the data inferred by Baggerly and Coombes [2]. The latter is a perfect match for Figure 1 of Hsu et al. [7]; the former is not. At some point, the cell lines identified by Baggerly and Coombes [2] were used. 17
3 Pairwise correlations between the ovarian sample quantifications now reported and RMA quantifications of the 146 CEL files examined by Bild et al. [3] The first 16 of the 59 samples now reported are clearly different. . . . 22
4 Pairwise correlations amongst the 59 samples now reported. The sample quantifications split into two qualitatively distinct blocks involving the first 16 and the last 43 samples, respectively, with high correlations within blocks and low correlations between. This shift is very extreme, suggesting an artifact, not biology, is the cause. In particular, this suggests that the probe labels were systematically misapplied for one of the two groups. . . . 23
5 Locations of very high correlations between the quantifications now reported and those derived from the Bild et al. [3] samples are shown. We can identify perfect matches for all of the last 43 samples. . . . 24
6 Pairwise correlations involving samples 1 (from the first block of 16) and 17 (from the second block of 43). The abrupt shift in values suggests misalignment of probeset values. . . . 29

7 Heatmap of the first 100 probeset intensities across all 59 samples now reported. The first 16 are qualitatively different from the rest, and since we can match the latter 43 the probeset ids are wrong for the first 16. 30
8 Side view of the first 100 probeset intensities for the 146 Bild samples (blue dots), with the corresponding values for the first 16 samples now supplied superimposed (red circles). This plot assumes that the probeset ordering given at GEO is what was used, in which case the first 68 values are from Affymetrix control probes that were nominally excluded. 31

List of Tables

1 Executive Summary

1.1 Introduction

Hsu et al. [7] constructed signatures for response to cisplatin and pemetrexed using cell lines. However, Baggerly and Coombes [2] have questioned the validity of these findings. Recently, a related supplementary data page for Hsu et al. [7], `http://data.genome.duke.edu/JCO`, was linked to the Duke IGSP main data page, `http://data.genome.duke.edu/`. We first noticed this link on Nov 6, 2009; it is not Google's snapshot of the main data page from Nov 1, 2009.

Here, we briefly explore the 6 data files posted to see if they can resolve the disagreement. The 6 files are:

- cell.lines.in.the.predictors.txt
- cis.pred.probes.txt (labeled "Cisplatin Predictor Gene List")
- pem.predictor.txt ("Pemetrexed Gene List")
- 6cmd1ngg.txt ("Cisplatin Predictor")
- pem.pred.txt ("Pemetrexed Predictor")
- Ovarian_cancer_n59_Validation.txt ("Ovarian Cancer Dataset")

1.2 Methods

We first loaded reference files previously assembled for Baggerly and Coombes [2] (cisplatinAll, pemetrexedAll, gyorffyAll, novartisA, and GI50_AUG08.bin) and Baggerly et al. [1] (ovcaRMAFromBild and clinicalInfo) to allow us to compare the results now presented with those reported earlier. These reference files are available at `http://bioinformatics.mdanderson.org/Supplements/ReproRsch-All`, and `http://bioinformatics.mdanderson.org/Supplements/ReproRsch-Ovary`, respectively.

The identities of the cell lines used to construct signatures for cisplatin and pemetrexed sensitivity were not initially presented in Hsu et al. [7]; these were inferred from the resulting heatmaps and gene lists by Baggerly and Coombes [2]. Since one point of disagreement was that "sensitive" and "resistant" labels might have been incorrectly applied, we compared the cell lines now reported for cisplatin and pemetrexed with those identified by Baggerly and Coombes [2].

Another point of disagreement involved the specific genes used in each signature. Baggerly and Coombes [2] claim that the gene lists initially reported by Hsu et al. [7] and currently available from the Journal of Clinical Oncology are incorrect: almost all of the probesets reported are "off-by-one" due to an indexing error,

and others reported as important (ERCC1, ERCC4, and FANCM) should not have been included at all and in some cases were not present on the U133A chips used. To see if this disagreement was now resolved, we compared the gene lists now reported for cisplatin and pemetrexed with those identified by Baggerly and Coombes [2] and those initially reported by Hsu et al. [7]

In attempting to reproduce these analysis results, it is critical that we start from the same raw data. Thus, we compared the training dataset values now reported for cisplatin and pemetrexed with expression values from Györffy et al. [6] and with expression values of the NCI-60 supplied by Novartis, respectively.

The validation data used by Hsu et al. [7] was not previously reported in detail; the only information supplied was that GEO dataset 3149 (containing 153 array profiles) was the source of the 59 samples examined. Since Baggerly and Coombes [2] encountered difficulties with other validation datasets examined (e.g., for doxorubicin), we compared the ovarian validation dataset quantifications, sample labels, and gene labels with those reported by Bild et al. [3] and those identified by Baggerly et al. [1].

1.3 Results

The cell lines now listed for cisplatin and pemetrexed are very similar to those reported by Baggerly and Coombes [2], with one difference for each drug. For cisplatin, they report using SNU182 as a resistant line when Baggerly and Coombes [2] inferred MeWo from the heatmap reported. For pemetrexed, they list ACHN as a resistant line instead of TK-10. However, as noted by Baggerly and Coombes [2], the drug senstivity data for pemetrexed available from the NCI (Figure 1) shows that the cell lines designated as "sensitive" are resistant to treatment, and vice-versa.

The new gene list for cisplatin does not include ERCC1, ERCC4, or FANCM. All of the genes are on the U133A. There are no genes in common with what was initially reported by Hsu et al. [7] There are 24/45 genes common to the new gene list and that reported by Baggerly and Coombes [2], suggesting that the off-by-one error has been addressed; the lack of perfect agreement is partially driven by the use of one different cell line in selecting genes.

The new gene list for pemetrexed is a perfect match for that reported by Baggerly and Coombes [2], and has no overlap with that initially reported by Hsu et al. [7] The agreement shows that the off-by-one error has been addressed. Perfect agreement is surprising, however, as changing even one of the cell lines used in selecting the genes tends to alter the list. As noted above, this partially affected the list for cisplatin (where SNU182 was reportedly used instead of MeWo), and here ACHN was reportedly used instead of TK-10.

The training data for cisplatin does involve numbers for SNU182 instead of MeWo. However, running this data through the binreg software posted by Potti and Nevins at `http://data.genome.duke.edu/NatureMedicine.php` does not produce the gene list now reported. Only 31/45 genes are matched (Figure 2). We have not been able to independently generate the list provided, and this should be possible when applying the software they used to the data they used. We have been able to perfectly reproduce the gene lists now reported for several other drugs (lists from Potti et al. [8]; reproductions are shown by Coombes et al. [4]).

The training data for pemetrexed does not involve numbers for ACHN. The numbers are those for TK-10, so the column was mislabeled. This explains why the overlap in gene lists is perfect, when we would expect the lists to be different if different cell lines were used to generate them.

The ovarian validation data appears to be log-transformed, and the values look like RMA values. Using correlation, we are able to identify perfect matches for the last 43 of the 59 data columns using the 146 ovarian CEL files posted by Bild et al. [3] at `http://data.genome.duke.edu/Oncogene.php` (Figure 5). Unfortunately, the sample names given are wrong for all 43. For the first 16 columns, the expression patterns are drastically different, with several normally intense probesets being light and vice-versa, suggesting that the probeset names are wrong (Figures 3, 4, 6, and 7). This interpretation is given more weight by the fact that the expression patterns for the first 68 reported probesets roughly correspond to those of the Affy

control probes that were nominally excluded (Figure 8). We haven't found a simple fix to correct the probeset mismatch, so the first 16 sample identities are currently unconfirmed. Even so, since 6 of the 16 reported labels match data in the last 43 columns, we know at least this many labels are incorrect (we assume that no single sample is present more than once). Mislabeling the samples might not be fatal for predicting response if the responder/nonresponder status was the same, which would point to simpler permutation within response groupings. However, checking the clinical data from Bild et al. [3] shows that responder/nonresponder status is different for 12/43 cases where we can identify the samples, so some sensitive samples are being treated as resistant when performance is being assessed, and vice-versa. Using this data with these labels would give incorrect results. The clinical data also shows that 40 of the reported samples are responders, 15 are nonresponders, and there is no information for 4. However, Figure 3 of Hsu et al. [7] shows values for 36 responders and 23 nonresponders, so the labels now being used (which were not supplied) differ from those used by Bild et al. [3] Either at least 4 cases labeled as nonresponders are now labeled as responders, or some different set of clinical labels entirely was used.

1.4 Conclusions

There are several problems present in the data now posted. In our assessment, three of these are fatal flaws with respect to building a signature:

1. The sensitive and resistant labels for the pemetrexed signature are reversed. If the method works as advertised, and they use this signature to guide treatment, then patients will be actively guided to the wrong therapy.
2. At least 49/59, and possibly all, of the validation samples are mislabeled. All claims about how well these signatures work clinically are based on how well they predict outcomes for patient samples, and if you scramble the labels, you're predicting the wrong things.
3. For 16/59 validation samples, the genes are mislabeled, and not in a manner that immediately suggests a simple fix (like an off-by-one error). As with point 2, this means that for these samples, they're predicting the wrong thing. Further, this discrepancy makes these samples "look different", and to the extent that one group is overrepresented in these samples (i.e., if they think these are all "responders"), this can make the classification problem inappropriately easy, and potentially bias the results.

More concisely,

1. The sensitivity labels are wrong.
2. The sample labels are wrong.
3. The gene labels are wrong.

Until these "clerical" problems are fixed, results derived from these data should not be used to guide therapy.

There is another problem with the validation data.

- There is a discrepancy between the number of responders that can be linked to the reported labels and the number of responders initially reported by Hsu et al. [7] This mismatch shows that the labels now being used (which were not supplied) differ from those used by Bild et al. [3] Either at least 4 cases labeled as nonresponders are now labeled as responders, or some different set of clinical labels entirely was used. If the latter interpretation is correct, this is also a fatal flaw, since misapplying the clinical data means you're predicting the wrong thing.

More concisely,

- The clinical labels linked to the sample labels may be wrong.

This problem is "conditionally fatal".

We are also disturbed that

1. we can't reproduce the genes now reported for cisplatin using their data and their software, and that
2. a training data cell line was evidently misnamed in assembling the signature for pemetrexed.

While we would also like to see these problems resolved, we view them as (relatively) minor.

2 Options and Libraries

```
> options(width = 80)
```

3 Earlier Rda Files

3.1 Data from Baggerly and Coombes [2]

We begin by loading four Rda files available as part of the supplementary material for Baggerly and Coombes [2]: cisplatinAll (for the gene lists reported earlier and for those obtained from the binreg software), pemetrexedAll (same for pemetrexed), gyorffyAll (for the array quantifications used to assemble the cisplatin signature), and novartisA (for the array quantifications used to assemble the pemetrexed signature).

```
> rdaList <- c("cisplatinAll", "pemetrexedAll", "gyorffyAll", "novartisA")
> for (rdaFile in rdaList) {
+     rdaFullFile <- file.path("RawData", "BaggerlyAnnAppStat",
+         paste(rdaFile, "Rda", sep = "."))
+     cat("loading ", rdaFullFile, " from cache\n")
+     load(rdaFullFile)
+ }
```

```
loading  RawData/BaggerlyAnnAppStat/cisplatinAll.Rda  from cache
loading  RawData/BaggerlyAnnAppStat/pemetrexedAll.Rda  from cache
loading  RawData/BaggerlyAnnAppStat/gyorffyAll.Rda  from cache
loading  RawData/BaggerlyAnnAppStat/novartisA.Rda  from cache
```

We also load the drug sensitivity data for the NCI60 cell lines, and extract the entries for pemetrexed (NSC 698037); sensitivity information for cisplatin is contained in the gyorffyAll file already loaded.

```
> temp <- read.table(file.path("RawData", "BaggerlyAnnAppStat",
+     "GI50_AUG08.BIN.csv"), header = TRUE, sep = ",")
> temp <- temp[temp$NSC == 698037, ]
> pemetrexedNLOGGI50s <- temp[, "NLOGGI50"]
> names(pemetrexedNLOGGI50s) <- as.character(temp[, "CELL"])
> names(pemetrexedNLOGGI50s) <- gsub(" $", "", names(pemetrexedNLOGGI50s))
> pemetrexedNLOGGI50s <- sort(pemetrexedNLOGGI50s)
> pemetrexedNLOGGI50s[1:5]
```

```
  NCI-H23  NCI-H522      EKVX  NCI-H226 NCI-H322M
        4         4         4         4         4
```

3.2 Data from Baggerly et al. [1]

We also load two Rda files available as part of the supplementary material for Baggerly et al. [1]: ovcaRMAFromBild (RMA quantifications for the 146 ovarian cancer CEL files used by Bild et al. [3], which should be a superset of those used in the validation set here), and clinicalInfo (for the responder/nonresponder status for the 119 samples examined by Dressman et al. [5]).

```
> rdaList <- c("ovcaRMAFromBild", "clinicalInfo")
> for (rdaFile in rdaList) {
+     rdaFullFile <- file.path("RawData", "BaggerlyJCO", paste(rdaFile,
+         "Rda", sep = "."))
+     cat("loading ", rdaFullFile, " from cache\n")
+     load(rdaFullFile)
+ }
```

```
loading  RawData/BaggerlyJCO/ovcaRMAFromBild.Rda  from cache
loading  RawData/BaggerlyJCO/clinicalInfo.Rda  from cache
```

4 Cell Lines in the Predictors

4.1 Cisplatin

We begin by loading the cell lines reported for the cisplatin predictor.

```
> cisplatinReportedCellLines <- read.table(file.path("RawData",
+     "HsuJCO", "cell.lines.in.the.predictors.txt"), sep = "\t",
+     nrows = 2, colClasses = rep("character", 19))
> dim(cisplatinReportedCellLines)
```

```
[1]  2 19
```

```
> cisplatinReportedCellLines
```

```
                    V1        V2        V3        V4        V5        V6
1 Cisplatin Predictor       257P      A375     C8161       ES2      me43
2                     Resistant Resistant Resistant Resistant Resistant
         V7        V8        V9       V10       V11       V12       V13
1   SKMel19    SNU182    SNU423      Sw13      BT20      DV90     FUOV1
2 Resistant Resistant Resistant Resistant Sensitive Sensitive Sensitive
        V14       V15       V16 V17 V18 V19
1     OAW42     OVKAR      R103
2 Sensitive Sensitive Sensitive
```

```
> cisplatinReportedResistant <- as.character(cisplatinReportedCellLines[1,
+     which(cisplatinReportedCellLines[2, ] == "Resistant")])
> cisplatinReportedSensitive <- as.character(cisplatinReportedCellLines[1,
+     which(cisplatinReportedCellLines[2, ] == "Sensitive")])
```

Now we compare these cell lines with those reported by Baggerly and Coombes [2]. We check the resistant lines first.

```
> cbind(cisplatinReportedResistant, cisplatinResistantLines)
```

```
      cisplatinReportedResistant cisplatinResistantLines
 [1,] "257P"                     "257p"
 [2,] "A375"                     "A375"
 [3,] "C8161"                    "C8161"
 [4,] "ES2"                      "ES-2"
 [5,] "me43"                     "ME-43"
 [6,] "SKMel19"                  "MeWo"
 [7,] "SNU182"                   "SKMel19"
 [8,] "SNU423"                   "SNU423"
 [9,] "Sw13"                     "SW13"
```

Visual inspection shows a high overlap; 8/9 match. The distinction is that they report using SNU182 when we found MeWo.

We check the sensitive lines next.

```
> cbind(cisplatinReportedSensitive, cisplatinSensitiveLines)
```

```
     cisplatinReportedSensitive cisplatinSensitiveLines
[1,] "BT20"                     "BT20"
[2,] "DV90"                     "DV-90"
[3,] "FUOV1"                    "FU-OV-1"
[4,] "OAW42"                    "OAW42"
[5,] "OVKAR"                    "OVCAR3"
[6,] "R103"                     "R103"
```

Visual inspection shows that all 6 agree; we believe their "OVKAR" is the same as our "OVCAR3".

The sensitive/resistant assignments made by Györffy et al. [6] are listed below:

```
> gyorffyAllInfo[, c("origin", "Cisplatin")]
```

```
          origin Cisplatin
181/85p pancreas         R
257p     gastric         R
A375    melanoma         R
BT20      breast         S
C8161   melanoma         R
Colo699     lung         M
CX-2       colon         R
DU145   prostate         R
DV-90       lung         S
ES-2     ovarian         R
FU-OV-1  ovarian         S
Hep3B        HCC         R
HRT-18     colon         R
HT-29      colon         R
MDA-231   breast         M
ME-43   melanoma         R
MeWo    melanoma         R
```

```
OAW42    ovarian         S
OVCAR3   ovarian         S
R103      breast         S
R193      breast         S
SKBR3     breast         R
SKMel13 melanoma         M
SKMel19 melanoma         R
SKOV-3   ovarian         R
SNU182       HCC         R
SNU423       HCC         R
SNU449       HCC         R
SNU475       HCC         R
SW13    prostate         R
```

Visual checking confirms these are consistent with the labels now assigned.

4.2 Pemetrexed

We now load the cell lines reported for the pemetrexed predictor.

```
> pemetrexedReportedCellLines <- read.table(file.path("RawData",
+     "HsuJCO", "cell.lines.in.the.predictors.txt"), sep = "\t",
+     nrows = 2, colClasses = rep("character", 19), skip = 3)
> dim(pemetrexedReportedCellLines)

[1]  2 19

> pemetrexedReportedCellLines

                    V1        V2        V3        V4        V5        V6
1 Pemetrexed Predictor     K-562    MOLT-4 HL-60(TB)      MCF7  HCC-2998
2                      Resistant Resistant Resistant Resistant Resistant
         V7        V8        V9       V10       V11             V12        V13
1   HCT-116  NCI-H460      ACHN    SNB-19   HS 578T MDA-MB-231/ATCC MDA-MB-435
2 Resistant Resistant Resistant Sensitive Sensitive       Sensitive  Sensitive
        V14       V15       V16       V17       V18       V19
1  NCI-H226       M14  MALME-3M  SK-MEL-2 SK-MEL-28     SN12C
2 Sensitive Sensitive Sensitive Sensitive Sensitive Sensitive

> pemetrexedReportedResistant <- as.character(pemetrexedReportedCellLines[1,
+     which(pemetrexedReportedCellLines[2, ] == "Resistant")])
> pemetrexedReportedSensitive <- as.character(pemetrexedReportedCellLines[1,
+     which(pemetrexedReportedCellLines[2, ] == "Sensitive")])
```

Now we compare these cell lines with those reported by Baggerly and Coombes [2]. We check the resistant lines first.

```
> cbind(pemetrexedReportedResistant, pemetrexedResistantLines)
```

```
     pemetrexedReportedResistant pemetrexedResistantLines
[1,] "K-562"                     "K-562"
[2,] "MOLT-4"                    "MOLT-4"
[3,] "HL-60(TB)"                 "HL-60(TB)"
[4,] "MCF7"                      "MCF7"
[5,] "HCC-2998"                  "HCC-2998"
[6,] "HCT-116"                   "HCT-116"
[7,] "NCI-H460"                  "NCI-H460"
[8,] "ACHN"                      "TK-10"
```

Visual inspection shows good agreement; 7/8 resistant lines overlap. The distinction is that they report using ACHN when we found TK-10.

We check the sensitive lines next.

```
> cbind(pemetrexedReportedSensitive, pemetrexedSensitiveLines)
```

```
      pemetrexedReportedSensitive pemetrexedSensitiveLines
 [1,] "SNB-19"                    "SNB-19"
 [2,] "HS 578T"                   "HS 578T"
 [3,] "MDA-MB-231/ATCC"           "MDA-MB-231/ATCC"
 [4,] "MDA-MB-435"                "MDA-MB-435"
 [5,] "NCI-H226"                  "NCI-H226"
 [6,] "M14"                       "M14"
 [7,] "MALME-3M"                  "MALME-3M"
 [8,] "SK-MEL-2"                  "SK-MEL-2"
 [9,] "SK-MEL-28"                 "SK-MEL-28"
[10,] "SN12C"                     "SN12C"
```

Visual inspection shows that all 10 agree.

We now plot the drug sensitivity data for pemetrexed. This is shown in Figure 1. The cell lines now reported are marked. TK-10, inferred as resistant by Baggerly and Coombes [2], and ACHN, labeled as resistant now instead of TK-10, are also shown. Shifting from TK-10 to ACHN does make the cell line groupings more coherent, but does not address the main problem with this data noted by Baggerly and Coombes [2]: cell lines sensitive to pemetrexed are labeled as resistant, and vice-versa.

We list the numerical values below to allow for more explicit checking if desired (we note in passing that two cell lines in the standard panel, CCRF-CEM and UO-31, were not successfully evaluated for pemetrexed sensitivity, so no values are reported).

```
> pemetrexedNLOGGI50s
```

```
   NCI-H23    NCI-H522        EKVX    NCI-H226   NCI-H322M
     4.000       4.000       4.000       4.000       4.000
    HOP-62     HOP-92        HT29      SW-620    COLO 205
     4.000       4.000       4.000       4.000       4.000
    HCT-15        KM12     HS 578T  MDA-MB-435       MDA-N
     4.000       4.000       4.000       4.000       4.000
    BT-549      T-47D     OVCAR-3     OVCAR-4     OVCAR-5
     4.000       4.000       4.000       4.000       4.000
   IGROV1     SK-OV-3          SR      SN12C      CAKI-1
     4.000       4.000       4.000       4.000       4.000
```

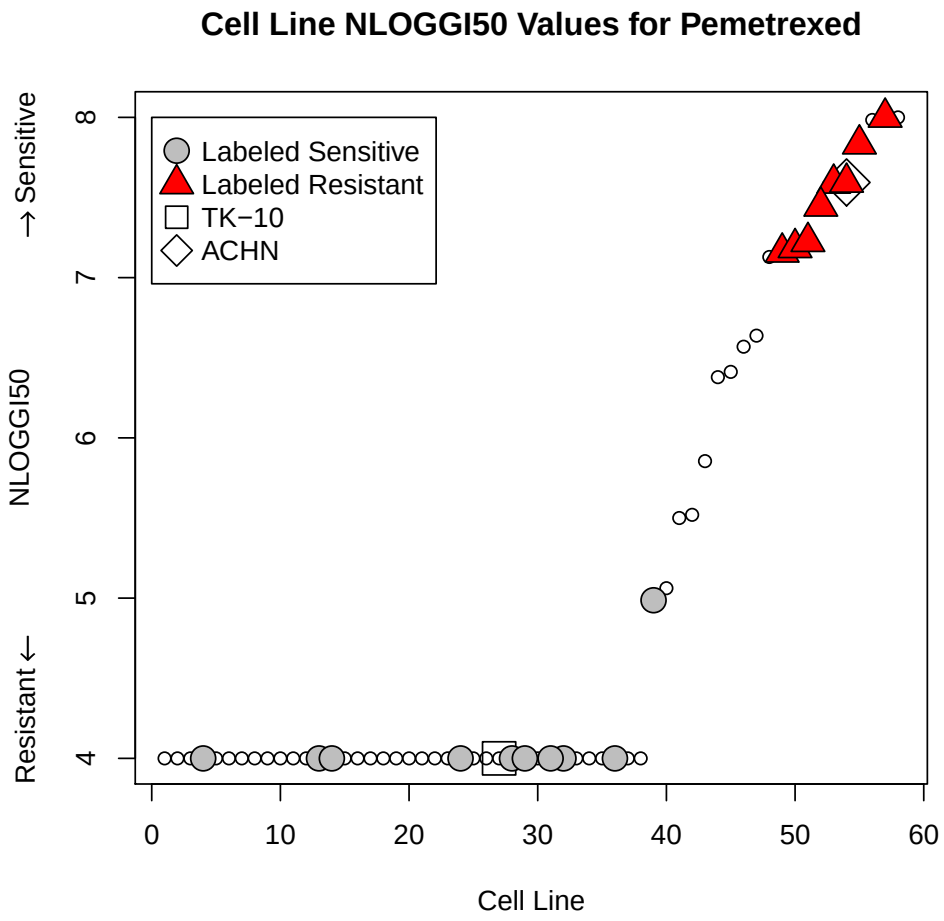

Figure 1: Plot of the negative log10 concentrations required to achieve 50% growth inhibition (NLOGGGI50 values) for the NCI60 cell lines. The cell lines now reported are marked. TK-10, inferred as resistant by Baggerly and Coombes [2], and ACHN, labeled as resistant now instead of TK-10, are also shown. Shifting from TK-10 to ACHN does make the cell line groupings more coherent, but does not address the main problem with this data noted by Baggerly and Coombes [2]: cell lines sensitive to pemetrexed are labeled as resistant, and vice-versa.

```
      RXF 393          TK-10       MALME-3M       SK-MEL-2       SK-MEL-5
        4.000          4.000          4.000          4.000          4.000
   SK-MEL-28            M14       UACC-257           PC-3         DU-145
        4.000          4.000          4.000          4.000          4.000
      SNB-19         SNB-75           U251 MDA-MB-231/ATCC       RPMI-8226
        4.000          4.000          4.000          4.987          5.062
     OVCAR-8     NCI/ADR-RES      A549/ATCC         SF-295         SF-268
        5.500          5.520          5.855          6.379          6.412
        A498        UACC-62       LOX IMVI         MOLT-4       HL-60(TB)
        6.569          6.638          7.129          7.158          7.186
     HCT-116        NCI-H460          K-562           ACHN           MCF7
        7.224          7.445          7.588          7.594          7.833
        786-0       HCC-2998         SF-539
        7.984          8.000          8.000
```

4.3 Summary

Agreement between the cell lines now listed and those inferred by Baggerly and Coombes [2] is very good, with only one disagreement per drug. However, the drug sensitivity data for pemetrexed shows that the sensitive/resistant labels are reversed for that drug. Assuming the method works as described, this is a fatal flaw, as it could actively assign patients to the wrong therapy. This problem was noted by Baggerly and Coombes [2].

5 Cisplatin Gene List

5.1 Reading Data, Checking for Outliers

We now turn to the cisplatin gene list.

```
> cisplatinNewGeneTable <- read.table(file.path("RawData", "HsuJCO",
+     "cis.pred.probes.txt"), sep = "\t", header = TRUE)
> dim(cisplatinNewGeneTable)

[1] 45  3

> cisplatinNewGeneTable[1:3, ]

  Probe.Set.ID Gene.Symbol                          Gene.Title
1  200076_s_at    C19orf50 chromosome 19 open reading frame 50
2  200711_s_at        SKP1 S-phase kinase-associated protein 1
3    200719_at        SKP1 S-phase kinase-associated protein 1

> sort(as.character(cisplatinNewGeneTable[, "Gene.Symbol"]))

 [1] "---"      "ANKRD5"   "ASS1"     "BTN2A1"   "C19orf50" "CCDC86"
 [7] "CD24"     "CD24"     "CEBPD"    "CLU"      "CREG1"    "CSK"
[13] "DPY19L1"  "EMP3"     "EP400"    "EXPH5"    "EXPH5"    "FGFR2"
[19] "FGFR2"    "FGFR2"    "FGFR2"    "FOLR1"    "GOLSYN"   "HDGF"
[25] "HMGCS1"   "HMGCS1"   "IFI30"    "IMPA2"    "LIMCH1"   "LIMCH1"
```

```
[31] "LIMCH1"   "MAP7"     "MEST"     "METTL7A"  "OBFC1"    "POLQ"
[37] "PPP2CB"   "PTPRA"    "RABL4"    "RREB1"    "SKP1"     "SKP1"
[43] "SLPI"     "ST14"     "TRMT1"
```

```
> cisplatinNewGeneList <- as.character(cisplatinNewGeneTable[,
+     "Probe.Set.ID"])
> match(cisplatinNewGeneList, rownames(gyorffyAll))
```

```
 [1]    97   239   247   728   903   950  1533  1544  1857  2418  2550  2648
[13]  2654  3165  3166  3228  3256  3500  3964  4564  5349  6602  7267  7282
[25]  7733  8286  9259 10689 10816 11710 11712 11713 11760 12177 13175 13308
[37] 14110 14993 15750 15855 18056 18464 19508 21110 21637
```

The new gene list does not include ERCC1, ERCC4, or FANCM, which Hsu et al. [7] had initially named as important, and which Baggerly and Coombes [2] had identified as "outliers" that should not be present. All of the probesets now listed are on the U133A platform; Baggerly and Coombes [2] noted that the initially reported list contained two probesets from the U133B platform.

5.2 Comparing this List with Others

We now compare this list of probesets with that identified by Baggerly and Coombes [2] and with that reported by Hsu et al. [7]

```
> intersect(cisplatinNewGeneList, softwareCisplatinProbesets)
```

```
 [1] "200076_s_at" "200719_at"   "201200_at"   "202016_at"   "202329_at"
 [6] "203021_at"   "203119_at"   "203638_s_at" "203639_s_at" "203729_at"
[11] "204437_s_at" "205822_s_at" "207076_s_at" "207761_s_at" "208228_s_at"
[16] "209771_x_at" "211401_s_at" "212327_at"   "213929_at"   "214734_at"
[21] "216379_x_at" "218692_at"   "220144_s_at" "221750_at"
```

```
> setdiff(cisplatinNewGeneList, softwareCisplatinProbesets)
```

```
 [1] "200711_s_at" "201375_s_at" "201422_at"   "202005_at"   "202889_x_at"
 [6] "203126_at"   "203701_s_at" "203973_s_at" "205037_at"   "207746_at"
[11] "208791_at"   "211256_x_at" "212325_at"   "212328_at"   "212375_at"
[16] "212792_at"   "213795_s_at" "215620_at"   "216484_x_at" "219100_at"
[21] "222278_at"
```

```
> setdiff(softwareCisplatinProbesets, cisplatinNewGeneList)
```

```
 [1] "213546_at"   "201924_at"   "203063_at"   "203258_at"   "206687_s_at"
 [6] "202167_s_at" "201015_s_at" "218966_at"   "218929_at"   "212727_at"
[11] "215631_s_at" "213540_at"   "220165_at"   "209794_at"   "209772_s_at"
[16] "212729_at"   "208651_x_at" "206747_at"   "266_s_at"    "202769_at"
[21] "208650_s_at"
```

```
> intersect(cisplatinNewGeneList, cisplatinReportedProbesets)
```

```
data frame with 0 columns and 0 rows
```

Of the probesets now reported, 24/45 match those reported by Baggerly and Coombes [2], and none match those reported by Hsu et al. [7].

5.3 Summary

The lack of outliers and the presence of partial agreement with the list obtained by Baggerly and Coombes [2] suggests that the major errors noted previously have been fixed. The lack of perfect agreement now could well be driven by the use of a single different cell line in the training data.

6 Pemetrexed Gene List

6.1 Reading Data

We now turn to the pemetrexed gene list.

```
> pemetrexedNewGeneTable <- read.table(file.path("RawData", "HsuJCO",
+     "pem.predictor.txt"), sep = "\t", header = TRUE)
> dim(pemetrexedNewGeneTable)

[1] 85  3

> pemetrexedNewGeneTable[1:3, ]

  Probe.Set.ID Gene.Symbol
1      1101_at       APBB1
2    1228_s_at      CTAGE5
3      1319_at        DDR2
                                                              Gene.Title
1 amyloid beta (A4) precursor protein-binding, family B, member 1 (Fe65)
2                                                CTAGE family, member 5
3                           discoidin domain receptor tyrosine kinase 2

> pemetrexedNewGeneList <- as.character(pemetrexedNewGeneTable[,
+     "Probe.Set.ID"])
```

6.2 Comparing this List with Others

We compare this list of probesets with that identified by Baggerly and Coombes [2] and with that reported by Hsu et al. [7]

```
> intersect(pemetrexedNewGeneList, softwarePemetrexedProbesets)

 [1] "1101_at"    "1228_s_at"  "1319_at"    "1356_at"    "242_at"
 [6] "243_g_at"   "31463_s_at" "31511_at"   "31538_at"   "31546_at"
[11] "32226_at"   "32252_at"   "32260_at"   "32318_s_at" "32434_at"
[16] "32574_at"   "32749_s_at" "32836_at"   "32893_s_at" "33145_at"
[21] "33362_at"   "33378_at"   "33452_at"   "33614_at"   "33855_at"
[26] "33919_at"   "34246_at"   "34319_at"   "34859_at"   "34860_g_at"
[31] "35352_at"   "35435_s_at" "356_at"     "35748_at"   "35763_at"
[36] "36119_at"   "36192_at"   "36536_at"   "36585_at"   "36989_at"
[41] "37345_at"   "37375_at"   "37485_at"   "37745_s_at" "37747_at"
[46] "38120_at"   "38288_at"   "38405_at"   "38479_at"   "38546_at"
[51] "38909_at"   "39019_at"   "39150_at"   "39170_at"   "39248_at"
```

```
[56] "39329_at"   "39330_s_at" "39351_at"   "39544_at"   "39750_at"
[61] "39798_at"   "39800_s_at" "40213_at"   "40328_at"   "40394_at"
[66] "40493_at"   "40684_at"   "40822_at"   "40855_at"   "40865_at"
[71] "40953_at"   "41128_at"   "41235_at"   "41403_at"   "41436_at"
[76] "41443_at"   "41449_at"   "41460_at"   "41644_at"   "41739_s_at"
[81] "41758_at"   "41834_g_at" "41854_at"   "591_s_at"   "798_at"
```

```
> setdiff(pemetrexedNewGeneList, softwarePemetrexedProbesets)
```

```
character(0)
```

```
> setdiff(softwarePemetrexedProbesets, pemetrexedNewGeneList)
```

```
character(0)
```

```
> intersect(pemetrexedNewGeneList, pemetrexedReportedProbesets)
```

```
data frame with 0 columns and 0 rows
```

The probesets now reported perfectly match those reported by Baggerly and Coombes [2]; none match those reported by Hsu et al. [7].

6.3 Summary

The agreement with the list obtained by Baggerly and Coombes [2] suggests that the off-by-one error noted previously has been fixed. The perfect agreement seen, however, is surprising, given that the cell lines named for pemetrexed above do not agree perfectly with those inferred by Baggerly and Coombes [2], and we would expect different cell lines to yield different genes.

7 Cisplatin Training Data

7.1 Read Data

We now examine the cisplatin training data.

```
> cisplatinTrainingData <- read.table(file.path("RawData", "HsuJCO",
+     "6cmd1ngg.txt"), sep = "\t", header = TRUE)
> dim(cisplatinTrainingData)
```

```
[1] 22215    15
```

```
> cisplatinTrainingData[1:3, ]
```

```
        X0     X0.1     X0.2     X0.3     X0.4     X0.5     X0.6     X0.7
1 8.830360 9.232587 8.837760 9.227460 9.647339 9.771864 8.669036 9.658593
2 7.975630 8.601956 8.068175 8.032434 8.293609 8.216027 9.879587 8.430591
3 7.674556 7.571382 7.456545 7.639730 7.903592 7.562891 7.355005 7.380951
      X0.8        X1     X1.1     X1.2     X1.3      X1.4     X1.5
1 9.007352 10.538850 9.480232 9.208135 9.832805 10.686688 9.224756
2 8.767707  8.512155 8.541081 7.187853 8.511809  8.252671 8.791642
3 7.434239  7.463574 7.457873 7.512885 7.558033  7.482426 7.488156
```

The first 9 column headers are “0”, and the last 6 are “1”. Simple counting suggests that 0=Resistant, and 1=Sensitive.

7.2 Match to Györffy et al. [6] Cell Lines

We now compare the numbers reported with those given by Györffy et al. [6].

```
> gyorffyAll[1, ]
```

```
           181/85p    257p     A375     BT20   C8161 Colo699    CX-2    DU145
1007_s_at 9.977162 8.83036 9.232587 10.53885 8.83776 9.80518 9.65463 10.04867
            DV-90    ES-2  FU-OV-1    Hep3B   HRT-18    HT-29  MDA-231   ME-43
1007_s_at 9.480232 9.22746 9.208135 9.525418 9.859844 9.402371 9.501546 9.64734
             MeWo    OAW42   OVCAR3     R103     R193    SKBR3 SKMel13  SKMel19
1007_s_at 9.48662 9.832805 10.68669 9.224756 9.214594 9.384983 9.52101 9.771864
            SKOV-3   SNU182   SNU423  SNU449   SNU475     SW13
1007_s_at 9.518115 8.669036 9.658593 9.18172 9.793796 9.007352
```

```
> cisplatinMatchingLines <- c("257p", "A375", "C8161", "ES-2",
+     "ME-43", "SKMel19", "SNU182", "SNU423", "SW13", "BT20", "DV-90",
+     "FU-OV-1", "OAW42", "OVCAR3", "R103")
> all(cisplatinTrainingData == gyorffyAll[1:22215, cisplatinMatchingLines])
```

```
[1] TRUE
```

```
> c(cisplatinReportedResistant, cisplatinReportedSensitive)
```

```
 [1] "257P"    "A375"    "C8161"   "ES2"     "me43"    "SKMel19" "SNU182"
 [8] "SNU423"  "Sw13"    "BT20"    "DV90"    "FUOV1"   "OAW42"   "OVKAR"
[15] "R103"
```

All of the numbers match columns from the first 22215 rows of the Györffy data. Visual inspection suggests a mapping, which exact comparison then confirms. The numbers correspond to the cell lines now reported.

7.3 Attempted Reproduction with Binreg

Baggerly and Coombes [2] identified the cell lines they reported by perfectly matching the cisplatin heatmap reported by Hsu et al. [7]. Using the data now reported will not produce the earlier heatmap, but should give rise to the genes now reported.

To test this, we ran the supplied data through binreg, using scripts in MatlabFiles/Cisplatin, producing both a heatmap and an associated gene list. The heatmap is shown in Figure 2, together with the heatmap constructed by Baggerly and Coombes [2]. The latter is a perfect match for Figure 1 of Hsu et al. [7]; the former is not. At some point, the cell lines identified by Baggerly and Coombes [2] were used.

We also checked the gene list to see whether it matches the list now reported.

```
> newCis <- read.table(file.path("MatlabFiles", "Cisplatin", "topCisplatinGenesInHeatmapOrder.txt"))
> newCisNames <- as.character(newCis[, 1])
> intersect(newCisNames, cisplatinNewGeneList)
```

```
 [1] "200719_at"   "214734_at"   "212327_at"   "213929_at"   "200076_s_at"
 [6] "203639_s_at" "203729_at"   "211256_x_at" "207746_at"   "202889_x_at"
[11] "204437_s_at" "201200_at"   "202329_at"   "212375_at"   "205822_s_at"
[16] "203126_at"   "202016_at"   "220144_s_at" "203021_at"   "208228_s_at"
```

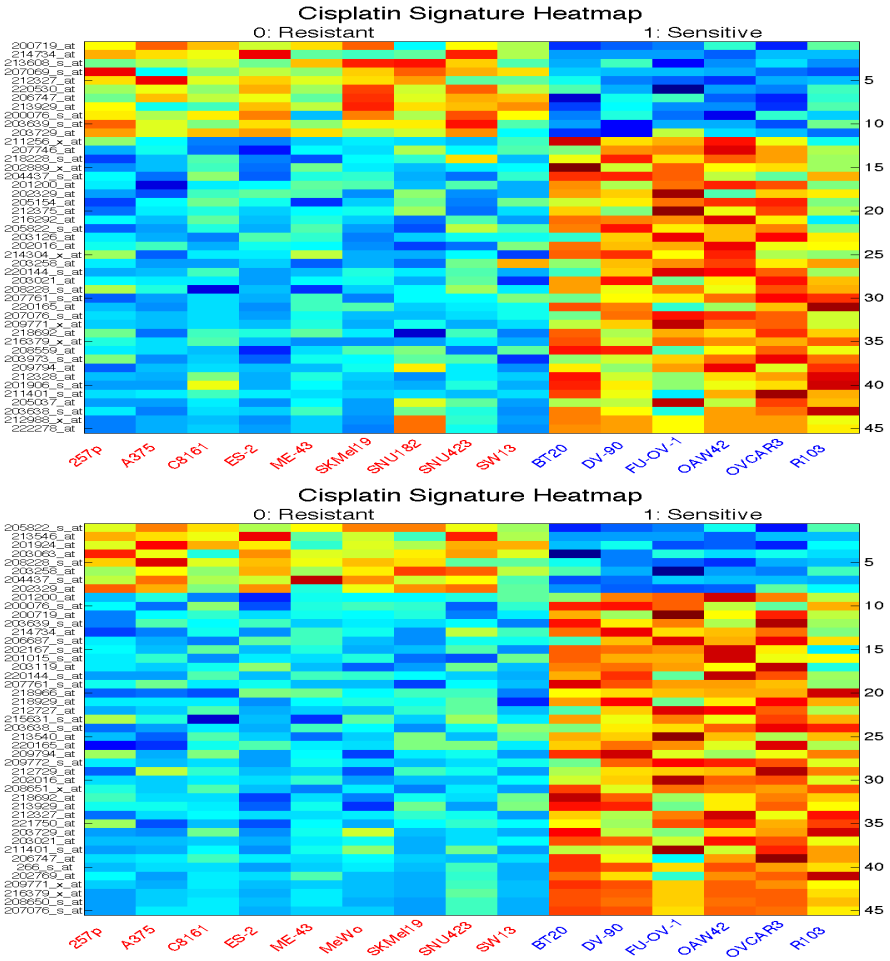

Figure 2: Heatmaps for cisplatin using the (a) the data now reported, and (b) the data inferred by Baggerly and Coombes [2]. The latter is a perfect match for Figure 1 of Hsu et al. [7]; the former is not. At some point, the cell lines identified by Baggerly and Coombes [2] were used.

```
[21] "207761_s_at" "207076_s_at" "209771_x_at" "218692_at"   "216379_x_at"
[26] "203973_s_at" "212328_at"   "211401_s_at" "205037_at"   "203638_s_at"
[31] "222278_at"
```

```
> setdiff(cisplatinNewGeneList, newCisNames)
```

```
 [1] "200711_s_at" "201375_s_at" "201422_at"   "202005_at"   "203119_at"
 [6] "203701_s_at" "208791_at"   "212325_at"   "212792_at"   "213795_s_at"
[11] "215620_at"   "216484_x_at" "219100_at"   "221750_at"
```

Only 31/45 genes now reported match those produced by binreg; we have not yet been able to match the other 14.

7.4 Summary

The data supplied matches that given by Györffy et al. [6] for the cell lines now named. However, running this data through the binreg software does not produce the heatmap initially reported by Hsu et al. [7] or the gene list reported now. The former shows that the cell lines inferred by Baggerly and Coombes [2] were used at one point, though this is a comparatively minor problem if the new list of cell lines represents a fix. The fact that we cannot reproduce the gene list they now report when applying their software to their data is more troubling, but not necessarily fatal.

8 Pemetrexed Training Data

8.1 Read Data

We now examine the pemetrexed training data.

```
> pemetrexedTrainingData <- read.table(file.path("RawData", "HsuJCO",
+     "pem.pred.txt"), sep = "\t", header = TRUE)
> dim(pemetrexedTrainingData)
```

```
[1] 12558    18
```

```
> pemetrexedTrainingData[1:3, ]
```

```
         X0      X0.1     X0.2     X0.3     X0.4      X0.5     X0.6      X0.7
1  95.75866  68.12313 113.8921 110.5558 116.0983  77.84743  63.7117  78.51872
2  98.03310  97.47627 112.8283 202.5027 110.1074 118.65627  60.2855  81.23810
3 200.20106 248.60211 208.6047 283.7809 165.3601 207.78110 202.9525 208.93970
         X1      X1.1      X1.2     X1.3      X1.4      X1.5      X1.6     X1.7
1  42.05887  74.29423  63.72237 125.7948 110.08205  79.06321  97.71843 119.4534
2  90.93796  77.06199  91.43547 218.3590  85.21201  77.45312  62.93141 126.5399
3 205.60739 159.19189 232.45746 211.1068 267.43652 216.54279 261.73642 204.2937
       X1.8      X1.9
1  66.41752  90.80546
2 108.48688 124.90649
3 232.38895 190.62155
```

The first 8 column headers are "0", and the last 10 are "1". Simple counting suggests that 0=Resistant, and 1=Sensitive.

8.2 Match to NCI60 Cell Lines

We now compare the numbers reported with those given in the A set of replicates from the triplicate set of U95Av2 quantifications of the NCI60 run by Novartis.

```
> novartisA[1, ]
```

```
      CCRF-CEM          K-562         MOLT-4      HL-60(TB)       RPMI-8226
     41.165840       95.758659       68.123131      113.892136       83.760979
            SR         SF-268         SF-295         SF-539         SNB-19
      7.254722       56.562138       82.410301       41.671947       42.058872
        SNB-75           U251         BT-549        HS 578T           MCF7
     21.820335       23.525270      105.377037       74.294228      110.555771
MDA-MB-231/ATCC    NCI/ADR-RES    MDA-MB-435          T-47D       COLO 205
     63.722366       66.192032      125.794838       78.400185      135.164444
      HCC-2998         HCT-116         HCT-15           HT29           KM12
    116.098328       77.847427      102.882088      101.626663      144.640198
         SW-620       A549/ATCC            EKVX         HOP-62         HOP-92
    135.111832       51.387577       63.307228       27.854231       65.249519
      NCI-H226         NCI-H23       NCI-H322M        NCI-H460        NCI-H522
    110.082054       93.459251       61.847382       63.711700      156.090027
       LOX IMVI             M14        MALME-3M        SK-MEL-2       SK-MEL-28
    153.429443       79.063210       97.718430      119.453430       66.417519
       SK-MEL-5        UACC-257         UACC-62         DU-145           PC-3
     88.215401      104.077370       62.631435       48.746502      262.895172
         IGROV1         OVCAR-3         OVCAR-4         OVCAR-5         OVCAR-8
    120.009315      138.865372       65.809509       21.589413      101.995895
        SK-OV-3            786-0            A498            ACHN         CAKI-1
    104.582458       93.567451       77.914803       80.786690      110.608986
        RXF 393            SN12C           TK-10           UO-31
     59.924706       90.805458       78.518715        8.625963
```

```
> all(pemetrexedTrainingData == novartisA[-grep("^AFFX", rownames(novartisA)),
+     c(pemetrexedResistantLines, pemetrexedSensitiveLines)])
```

```
[1] TRUE
```

```
> c(pemetrexedResistantLines, pemetrexedSensitiveLines)
```

```
 [1] "K-562"           "MOLT-4"          "HL-60(TB)"       "MCF7"
 [5] "HCC-2998"        "HCT-116"         "NCI-H460"        "TK-10"
 [9] "SNB-19"          "HS 578T"         "MDA-MB-231/ATCC" "MDA-MB-435"
[13] "NCI-H226"        "M14"             "MALME-3M"        "SK-MEL-2"
[17] "SK-MEL-28"       "SN12C"
```

Again, visual inspection suggests a mapping, which exact comparison then confirms. The numbers correspond to the cell lines identified by Baggerly and Coombes [2], not those now reported. In particular, the numbers in column X0.7 in the pemetrexed training data match TK-10, not ACHN.

8.3 Attempted Reproduction with Binreg

Since the numbers exactly match those used by Baggerly and Coombes [2], their report matchPemetrexed-Heatmap.pdf shows how running these numbers through binreg produces both the heatmap initially reported in Hsu et al. [7] and the gene list reported now.

8.4 Summary

The data supplied matches the NCI60 data for the cell lines named by Baggerly and Coombes [2], not the cell lines reported here. Running this data through the binreg software does not produce the heatmap initially reported by Hsu et al. [7] or the gene list reported now. The column of numbers for TK-10 was apparently labeled as coming from ACHN. This is disturbing, but possibly not fatal if mislabeling is a very rare event.

9 Ovarian Testing Data

9.1 Read Data

We now turn to the ovarian cancer testing data.

```
> ovarianTestingData <- read.table(file.path("RawData", "HsuJCO",
+     "Ovarian_cancer_n59_Validation.txt"), sep = "\t", header = TRUE,
+     row.names = "Probes")
> dim(ovarianTestingData)

[1] 22215    59

> colnames(ovarianTestingData) <- gsub("^X", "", colnames(ovarianTestingData))
> ovarianTestingData[1, ]

              M337     M485     M503     M810    M1055    M1241    M1390
1007_s_at 7.159988 7.159235 7.361196 7.134395 7.086862 6.608204 6.573475
             M1503    M2515    M2729    M2807    M3514    M3627   M4171
1007_s_at 6.599638 6.681299 6.644108 6.932458 6.520604 6.672259 6.51526
             M5775     1451    D2247     2324    D2332    D2421     2422
1007_s_at 6.545945 6.456417 11.01506 11.47455 10.90974 10.96553 10.41815
             D2432    D2433     2479     2542    D2560    D2572    D2575
1007_s_at 9.611253 11.27089 11.02993 10.93970 11.20753 11.03937 10.87672
             D2668     2673    D2689    D2691   D2700    D2733    D2749
1007_s_at 10.72335 10.86992 10.77384 10.96321 10.2842 10.68992 10.25562
             M1891     2465      M17     M359     M444    M1054    M1572
1007_s_at 10.70913 11.20442 10.88532 10.98397 11.21423 11.05373 11.06826
             M2070    M3142    M4161     6488     2476    D2480    D2557
1007_s_at 10.75442 11.12686 10.72843 11.22538 10.68763 10.92112 11.55423
              2573    D2576     2581   D2611    D2629    D2640    D2648
1007_s_at 11.37468 10.81826 10.54384 9.99899 10.98939 10.09090 10.85199
             D2727    D2738   D2792
1007_s_at 10.92684 11.93534 9.53838
```

9.2 Check Correlations

This set of 59 arrays is nominally a subset of those examined earlier by Bild et al. [3] The numerical values suggest some type of log-scale quantification, most likely RMA. RMA quantifications for the Bild et al. [3] CEL files were computed by Baggerly et al. [1] and were loaded above. Now we check the sample by sample correlations.

```
> corHsuWBild <- cor(ovarianTestingData, ovcaRMAFromBild[rownames(ovarianTestingData),
+     ])
> corHsuWHsu <- cor(ovarianTestingData)
```

We first look at the pattern of correlations overall, to see if any structure is evident. These correlations are depicted in Figure 3.

Red indicates low values in Figure 3, so there is clearly something very different about the first 16 samples as opposed to the last 43. As a check, we also examine the correlations between pairs of the Hsu et al. [7] quantification columns. These are shown in Figure 4. This shows that the first 16 of the Hsu et al. [7] samples behave similarly, even if they're not like the others. There are no tied columns in the data, or the count of high correlation pairs would be more than 59.

Now we look for very high correlations (above 0.99), between the Hsu et al. [7] quantifications and the RMA values from the Bild et al. [3] CEL files. Such high correlations suggest sample matches. These are shown in Figure 5. We have perfect matches for the last 43 of the Hsu et al. [7] quantifications, which we take as confirmation that we are indeed dealing with RMA values and correct probeset labels for these samples.

9.3 Checking What Matches with What

Now let's check what matches with what.

```
> bestMatches <- which(corHsuWBild > 0.99, arr.ind = TRUE)
> nameMatches <- cbind(colnames(corHsuWBild)[bestMatches[, "col"]],
+     rownames(corHsuWBild)[bestMatches[, "row"]])
> nameMatches
```

```
      [,1]                          [,2]
 [1,] "0074_02394_h133a_2476.cel"  "M1891"
 [2,] "0074_02400_h133a_2895.cel"  "M1572"
 [3,] "0074_02403_h133a_2981.cel"  "M4161"
 [4,] "0074_02484_h133a_3250.cel"  "6488"
 [5,] "0074_2030_h133a_1024.cel"   "D2691"
 [6,] "0074_2031_h133a_2739.cel"   "D2700"
 [7,] "0074_2032_h133a_2673.cel"   "D2733"
 [8,] "0074_2033_h133a_2505.cel"   "D2749"
 [9,] "0074_2395_h133a_1447.cel"   "2465"
[10,] "0074_2396_h133a_1913.cel"   "M17"
[11,] "0074_2397_h133a_1552.cel"   "M359"
[12,] "0074_2398_h133a_1578.cel"   "M444"
[13,] "0074_2399_h133a_3107.cel"   "M1054"
[14,] "0074_2401_h133a_3018.cel"   "M2070"
[15,] "0074_2402_h133a_3090.cel"   "M3142"
[16,] "0193_00000_h133a_D1805.cel" "2476"
[17,] "0193_00000_h133a_D1859.cel" "D2557"
```

```
> image(1:59, 1:146, corHsuWBild, xlab = "Hsu et al. Quantifications",
+     ylab = "RMA from Bild et al. CEL Files", main = "Pairwise Correlations")
> box()
```

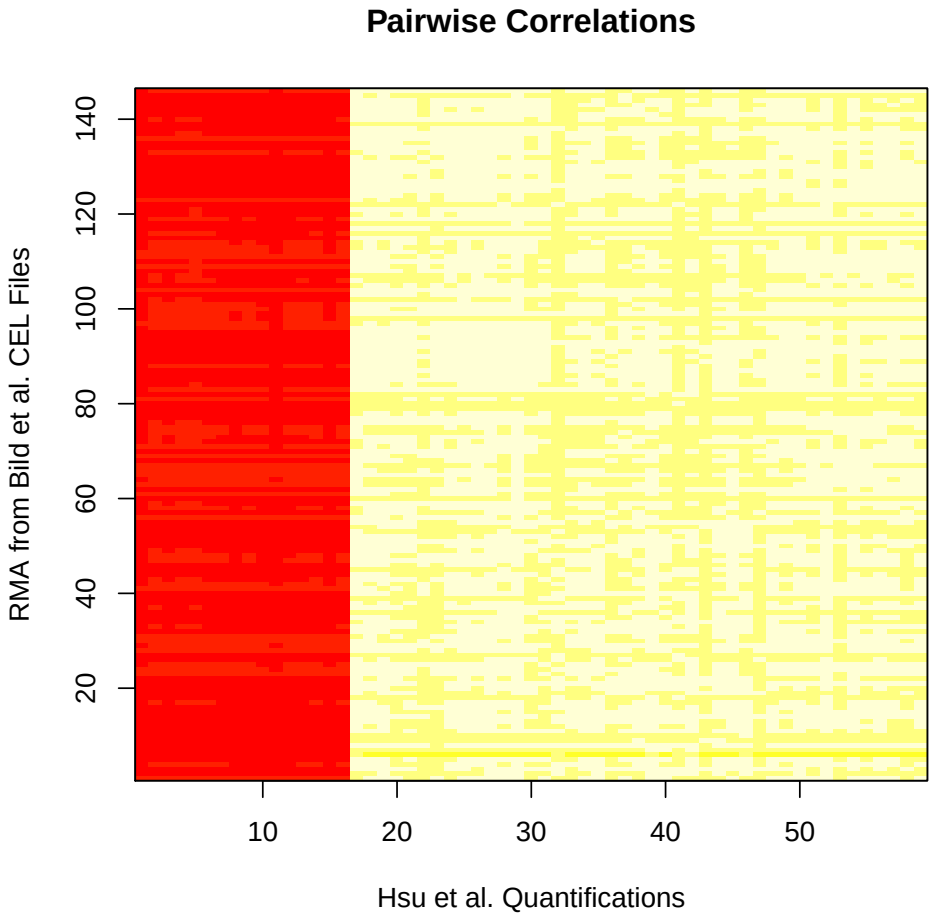

Figure 3: Pairwise correlations between the ovarian sample quantifications now reported and RMA quantifications of the 146 CEL files examined by Bild et al. [3] The first 16 of the 59 samples now reported are clearly different.

```
> image(1:59, 1:59, corHsuWHsu, xlab = "Hsu et al. Quantifications",
+     ylab = "Hsu et al. Quantifications", main = "Pairwise Correlations")
> box()
> sum(corHsuWHsu > 0.99)

[1] 59
```

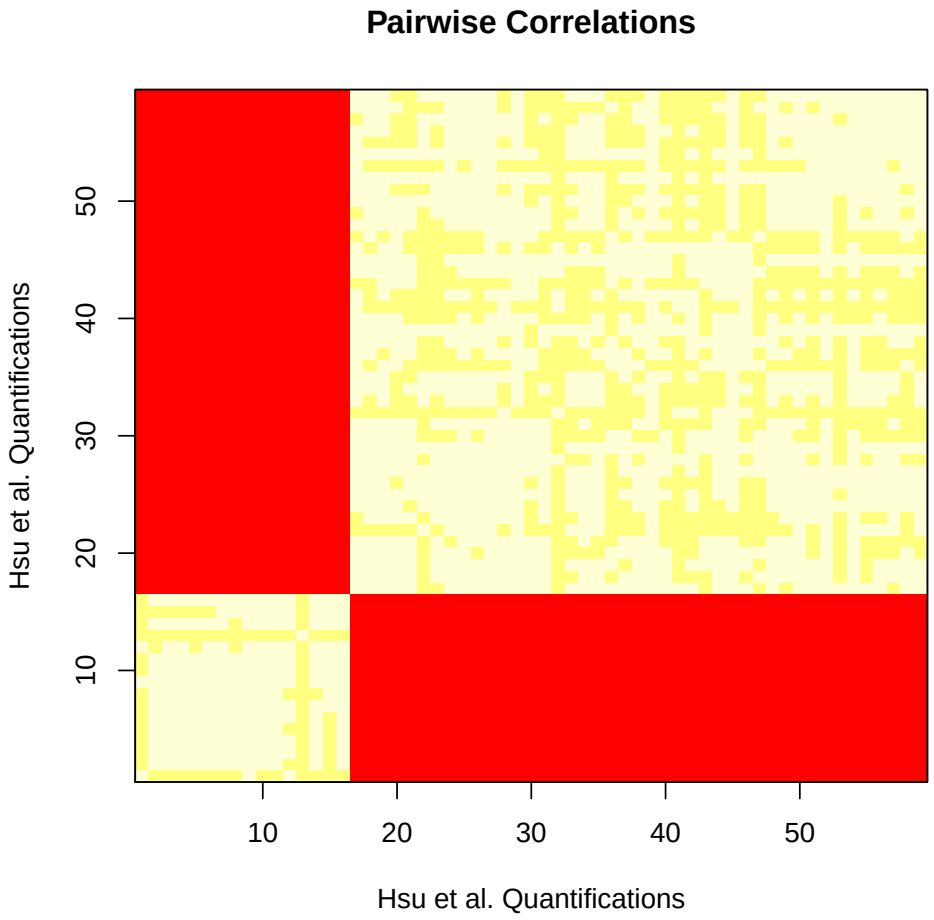

Figure 4: Pairwise correlations amongst the 59 samples now reported. The sample quantifications split into two qualitatively distinct blocks involving the first 16 and the last 43 samples, respectively, with high correlations within blocks and low correlations between. This shift is very extreme, suggesting an artifact, not biology, is the cause. In particular, this suggests that the probe labels were systematically misapplied for one of the two groups.

```
> image(1:59, 1:146, corHsuWBild < 0.99, xlab = "Hsu et al. Quantifications",
+     ylab = "RMA from Bild et al. CEL Files", main = "Pairwise Correlations > 0.99 (Sample Matches)")
> box()
> sum(corHsuWBild > 0.99)

[1] 43
```

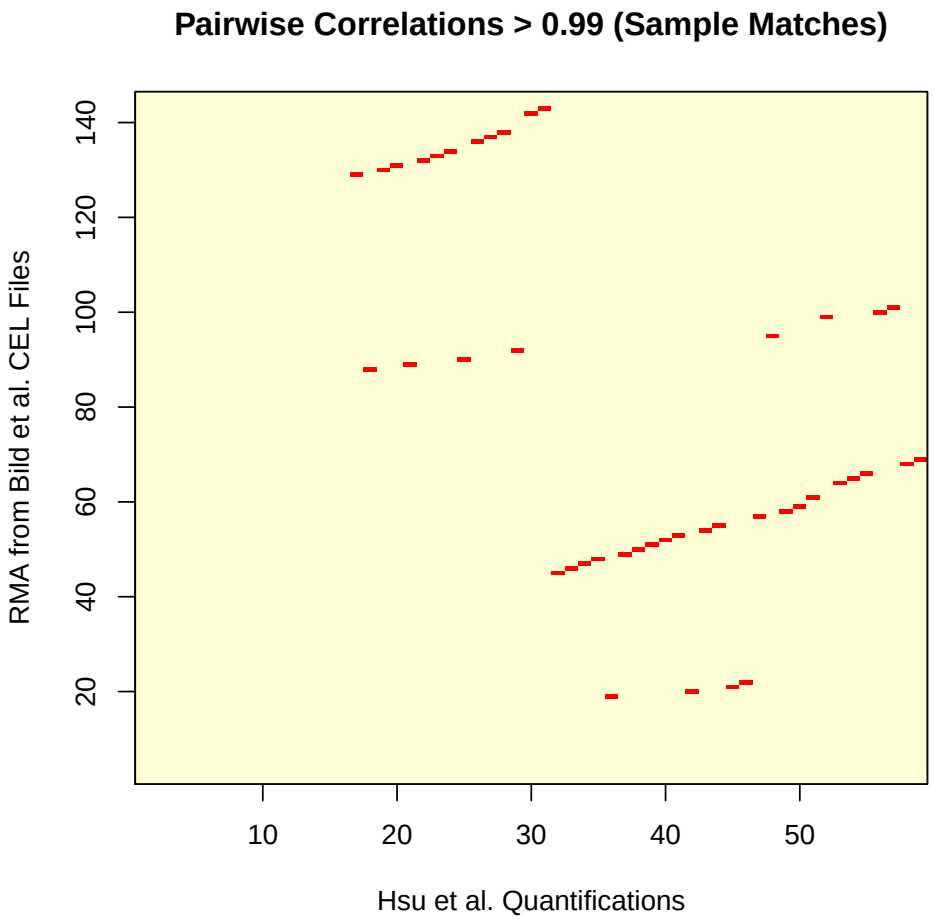

Figure 5: Locations of very high correlations between the quantifications now reported and those derived from the Bild et al. [3] samples are shown. We can identify perfect matches for all of the last 43 samples.

```
[18,] "0193_00000_h133a_D2098.cel" "2573"
[19,] "0193_00000_h133a_D2208.cel" "D2576"
[20,] "0193_00000_h133a_D2342.cel" "D2611"
[21,] "0193_00000_h133a_D2358.cel" "D2629"
[22,] "0193_00000_h133a_D2421.cel" "D2640"
[23,] "0193_00000_h133a_D2480.cel" "D2738"
[24,] "0193_00000_h133a_D2557.cel" "D2792"
[25,] "0193_00000_h133a_M2070.cel" "2324"
[26,] "0193_00000_h133a_M2437.cel" "2422"
[27,] "0193_00000_h133a_M3142.cel" "2542"
[28,] "0193_00000_h133a_M4161.cel" "D2668"
[29,] "0193_10000_h133a_D1837.cel" "D2480"
[30,] "0193_10000_h133a_D2332.cel" "2581"
[31,] "0193_10000_h133a_D2432.cel" "D2648"
[32,] "0193_10000_h133a_D2433.cel" "D2727"
[33,] "0193_10000_h133a_M1891.cel" "D2247"
[34,] "0193_10000_h133a_M2097.cel" "D2332"
[35,] "0193_10000_h133a_M2184.cel" "D2421"
[36,] "0193_10000_h133a_M2515.cel" "D2432"
[37,] "0193_10000_h133a_M2729.cel" "D2433"
[38,] "0193_10000_h133a_M2807.cel" "2479"
[39,] "0193_10000_h133a_M3484.cel" "D2560"
[40,] "0193_10000_h133a_M3514.cel" "D2572"
[41,] "0193_10000_h133a_M3627.cel" "D2575"
[42,] "0193_10000_h133a_M5668.cel" "2673"
[43,] "0193_10000_h133a_M5775.cel" "D2689"
```

Of the 43 quantifications that we can qualitatively match, *none* of the names match the CEL file from which the numbers were derived. This is disturbing, since all claims about how well these signatures work clinically are based on how well they predict outcomes for patient samples, and if you scramble the labels, you're predicting the wrong things.

9.4 Checking Clinical Information for the Last 43

To the extent that the samples are only being used to predict response, the results will not be affected by sample relabeling if the clinical response status of the underlying samples is the same. To check this, we shorten the filenames in nameMatches above to focus on the sample identifiers.

```
> shortNameMatches <- nameMatches
> colnames(shortNameMatches) <- c("BildLabel", "HsuLabel")
> shortNameMatches[, "BildLabel"] <- gsub("\\.cel$", "", shortNameMatches[,
+     "BildLabel"])
> temp <- unlist(strsplit(shortNameMatches[, "BildLabel"], "h133a_"))
> temp <- temp[seq(2, length(temp), 2)]
> shortNameMatches[, "BildLabel"] <- temp
```

Now we do some quick checks to assess the degree of overlap.

```
> sum(shortNameMatches[, "BildLabel"] == shortNameMatches[, "HsuLabel"])
```

```
[1] 0
```

```
> intersect(shortNameMatches[, "BildLabel"], shortNameMatches[,
+     "HsuLabel"])
```

```
 [1] "2476"  "2673"  "D2421" "D2480" "D2557" "M2070" "M3142" "M4161" "D2332"
[10] "D2432" "D2433" "M1891"
```

```
> table(clinicalInfo[shortNameMatches[, "BildLabel"], "Response"])
```

```
CR NR
35  7
```

```
> table(clinicalInfo[shortNameMatches[, "HsuLabel"], "Response"])
```

```
CR NR
25 15
```

```
> table(clinicalInfo[shortNameMatches[, "BildLabel"], "Response"],
+     clinicalInfo[shortNameMatches[, "HsuLabel"], "Response"])
```

```
    CR NR
  CR 23 10
  NR  2  5
```

```
> fisher.test(table(clinicalInfo[shortNameMatches[, "BildLabel"],
+     "Response"], clinicalInfo[shortNameMatches[, "HsuLabel"],
+     "Response"]))
```

```
        Fisher's Exact Test for Count Data

data:
p-value = 0.08116
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.7442456 66.8254981
sample estimates:
odds ratio
  5.476575
```

Only 12 of the samples named are even present, albeit in other locations. Looking at the marginal distributions of response associated with response shows that the two labelings can't be equivalent, as the total numbers (35 and 7, and 25 and 15) don't match. Checking the cross-tabulation shows that the association is not strongly significant.

9.5 Checking Clinical Information (and Labels) for the First 16

While we can't yet assess the Bild labels (and hence the response status) for the first 16 samples, we can do this for the Hsu labels.

```
> shortHsuNames16 <- colnames(ovarianTestingData)[1:16]
> table(clinicalInfo[shortHsuNames16, "Response"])
```

```
CR NR
15  0
```

Almost all of these are complete responders; we have one case for which we don't know the outcome.

We can also use the short names for the last 43 columns to check whether any of the labels given to the first 16 columns match the data used in the latter 43. Any overlap would suggest that either those samples were mislabeled as well, or that the same sample had been included twice.

```
> intersect(shortNameMatches[, "BildLabel"], colnames(ovarianTestingData)[1:16])
```

```
[1] "M2515" "M2729" "M2807" "M3514" "M3627" "M5775"
```

At least 6 of the first 16 samples were also mislabeled.

9.6 Checking Missing Data

We can look at the cases of missing data more closely to see if there's something else to be gleaned.

```
> shortHsuNames16[which(is.na(clinicalInfo[shortHsuNames16, "Response"]))]
```

```
[1] "M4171"
```

```
> which(is.na(clinicalInfo[shortNameMatches[, "BildLabel"], "Response"]))
```

```
[1] 4
```

```
> which(is.na(clinicalInfo[shortNameMatches[, "HsuLabel"], "Response"]))
```

```
[1]  4 30 33
```

```
> shortNameMatches[c(4, 30, 33), ]
```

```
     BildLabel HsuLabel
[1,] "3250"    "6488"
[2,] "D2332"   "2581"
[3,] "M1891"   "D2247"
```

```
> rownames(clinicalInfo)
```

```
  [1] "0.08"  "860"   "872"   "922"   "1024"  "1447"  "1451"  "1504"  "1526"
 [10] "1552"  "1578"  "1590"  "1615"  "1623"  "1665"  "1674"  "1675"  "1774"
 [19] "1784"  "1834"  "1846"  "1877"  "1913"  "1929"  "2046"  "2063"  "2064"
 [28] "2075"  "2198"  "2204"  "2324"  "2419"  "2422"  "2424"  "2465"  "2476"
 [37] "2479"  "2505"  "2542"  "2573"  "2673"  "2739"  "2802"  "2849"  "2895"
 [46] "2967"  "2981"  "2999"  "3018"  "3090"  "3102"  "3107"  "3142"  "3249"
 [55] "D1805" "D1837" "D1859" "D2098" "D2208" "D2332" "D2342" "D2358" "D2421"
 [64] "D2432" "D2433" "D2480" "D2557" "D2559" "D2560" "D2572" "D2575" "D2576"
 [73] "D2581" "D2603" "D2611" "D2629" "D2640" "D2648" "D2668" "D2689" "D2691"
 [82] "D2700" "D2726" "D2727" "D2733" "D2738" "D2749" "D2776" "D2792" "M1054"
 [91] "M1055" "M120"  "M1241" "M1390" "M1503" "M1572" "M17"   "M1891" "M2070"
[100] "M2097" "M2184" "M2437" "M2515" "M2729" "M2807" "M3142" "M337"  "M3484"
[109] "M3514" "M359"  "M3627" "M4161" "M444"  "M485"  "M503"  "M5668" "M5775"
[118] "M6199" "M810"
```

```
> clinicalInfo["3249", ]

     SurvMonths Censoring CensoringBild Response Stage Grade Debulk CA125 POST
3249         15      Dead          Dead       NR     4     2      O can't find
```

Looking at the names, we see that one of the cases that fails to match is sample 3250 (Bild labeling). As noted in Baggerly et al. [1], there is no CEL file for 3249, so we suspect these two may be the same sample (hence the "can't find" entry for CA125 level). If that is the case, there would be one more NR using the Bild labels.

9.7 Totaling Clinical Counts

Using the labels reported, we have 40 CR, 15 NR, and 4 unknown. Using the Bild labels, we have 35 CR, 8 NR, and 16 unknown. However, Figure 3 of Hsu et al. [7] shows 36 responders and 23 nonresponders. The counts reported can't match this (there are too many CRs seen). The counts associated with the Bild labels could match this, but only by introducing far more instances where the CR/NR labels for a given column conflict between the two sets of labels. This mismatch shows that the labels now being used (which were not supplied) differ from those used by Bild et al. [3] Either at least 4 cases labeled as nonresponders are now labeled as responders, or some different set of clinical labels entirely was used.

9.8 Matching the First 16

We now look at the first 16 columns in more detail, to see if we can figure out part of the matching problem. We begin by looking at the pairwise correlations associated with samples 1 (from the first 16) and 17 (from the remainder) in more quantitative detail. These are shown in Figure 6.

The correlations shift from very high to very low as we shift between blocks. The most frequent cause of this problem in our experience is a mixup in probe labeling.

We can check this interpretation further by looking at an image map of the first 100 probesets across all 59 samples, as shown in Figure 7.

There is clearly a qualitative difference, so the probeset labels are incorrect for the first 16 columns.

Given that the supplied probeset ordering doesn't work, our next question is whether there are some natural other orderings to try. Since the paper mentions that this data comes from GEO dataset GSE3149, we'll try the probeset ordering used there. In particular, we use the ordering extracted from the table for sample M337, GSM70578. We begin by loading the table of reported MAS5 values.

```
> m337FromGEO <- read.table(file.path("RawData", "m337FromGEO"),
+     header = TRUE, skip = 2, nrow = 22283, row.names = 1)
```

Next, we take a look at the first 100 values, as shown in Figure 8.

For the most part, these track pretty well; certainly better than the reported ordering. However, there are still some probesets where the alignment is poor, so this is likely not the final word. Even so, we note a problem, namely that the first 68 probesets in the GEO ordering are the Affy control probes that were nominally excluded.

At present, we do not have a fix for this problem, so we cannot map the first 16 samples. All we can say (based mostly on the discordant expression patterns across the 59 samples but partially on the apparent agreement with 68 probes that shouldn't be there) is that the assigned probeset labels are wrong.

```
> plot(corHsuWHsu[1, ], xlab = "Column in Ovarian Testing Set",
+     ylab = "Correlation with Column 1 (o) or 17 (*)", main = "Pairwise Correlations Involving Ovarian
> points(corHsuWHsu[17, ], col = "red", pch = "*")
```

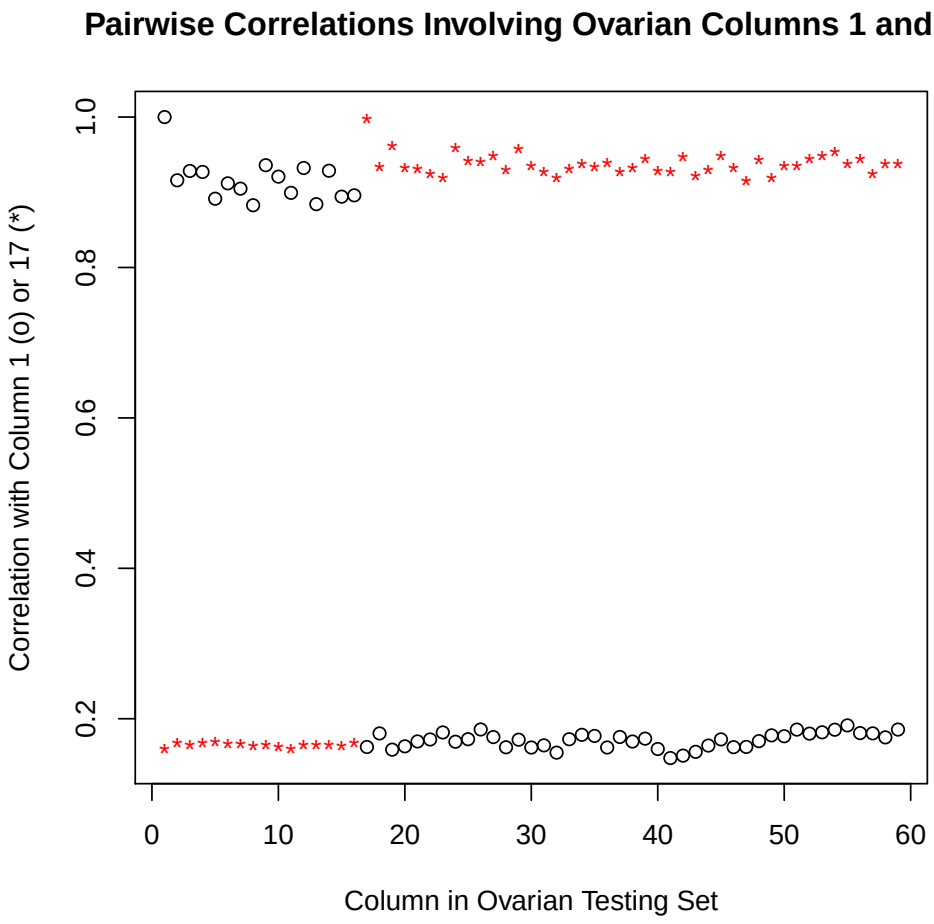

Figure 6: Pairwise correlations involving samples 1 (from the first block of 16) and 17 (from the second block of 43). The abrupt shift in values suggests misalignment of probeset values.

```
> image(1:100, 1:59, as.matrix(ovarianTestingData[1:100, ]), xlab = "Probeset Row",
+     ylab = "Sample Column", main = "First 100 Probeset Intensities, By Sample (Red is Low)")
```

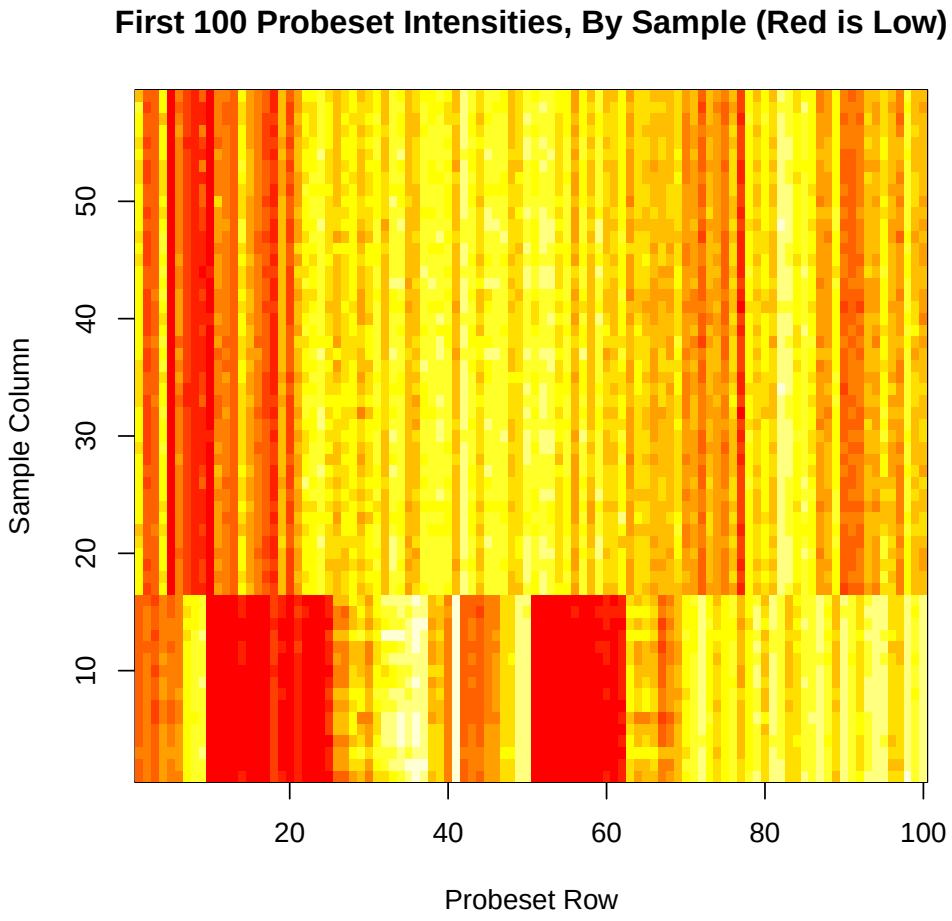

Figure 7: Heatmap of the first 100 probeset intensities across all 59 samples now reported. The first 16 are qualitatively different from the rest, and since we can match the latter 43 the probeset ids are wrong for the first 16.

```
> matplot(ovcaRMAFromBild[rownames(m337FromGEO)[1:100], ], col = "blue",
+     pch = ".", cex = 2, xlab = "Probeset Row Index", ylab = "Probeset Intensity",
+     main = "Bild (blue) and Hsu (red) Probeset Intensities, Using GEO Order")
> matpoints(ovarianTestingData[1:100, 1:16], col = "red", pch = "o",
+     cex = 0.5)
```

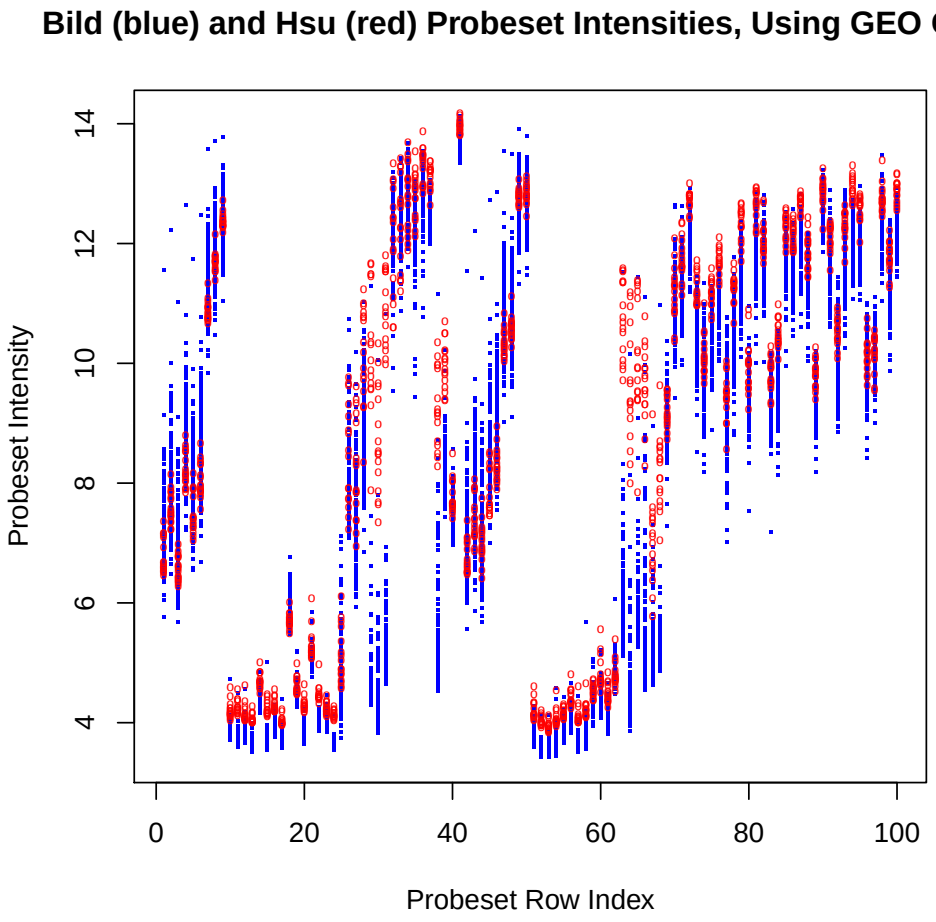

Figure 8: Side view of the first 100 probeset intensities for the 146 Bild samples (blue dots), with the corresponding values for the first 16 samples now supplied superimposed (red circles). This plot assumes that the probeset ordering given at GEO is what was used, in which case the first 68 values are from Affymetrix control probes that were nominally excluded.

9.9 Summary

There are major problems with the ovarian validation data supplied.

All of the last 43 and at least 6 of the first 16 columns are mislabeled. Extensive sample mislabeling is a fatal flaw, since all claims about how well these signatures work clinically are based on how well they predict outcomes for patient samples, and if you scramble the labels, you're predicting the wrong things.

The genes are mislabeled for the first 16 samples. Extensive gene mislabeling is a fatal flaw, since (as above) for these samples, they're predicting the wrong thing. Further, this discrepancy makes these samples "look different", and to the extent that one group is overrepresented in these samples (i.e., if they think these are all "responders"), this can make the classification problem inappropriately easy, and potentially bias the results.

Finally, there is a discrepancy between the number of responders that can be linked to the reported labels and the number of responders initially reported by Hsu et al. [7] This mismatch shows that the labels now being used (which were not supplied) differ from those used by Bild et al. [3] Either at least 4 cases labeled as nonresponders are now labeled as responders, or some different set of clinical labels entirely was used. If the latter interpretation is correct, this is also a fatal flaw, since misapplying the clinical data means you're predicting the wrong thing.

10 Appendix

10.1 File Location

```
> getwd()

[1] "/Users/kabagg/MicroarrayCourse/Nevins/JCO-LungCa09"
```

10.2 Saves

10.3 SessionInfo

```
> sessionInfo()

R version 2.9.1 (2009-06-26)
i386-apple-darwin8.11.1

locale:
en_US.UTF-8/en_US.UTF-8/C/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base
```

References

[1] Baggerly KA, Coombes KR, Neeley ES: Run batch effects potentially compromise the usefulness of genomic signatures for ovarian cancer. *J Clin Oncol*, **26**:1186-7, author reply 1187-8, 2008.

[2] Baggerly KA, Coombes KR: Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Ann App Statist*, *to appear*, 2009. Preprint available at `http://www.imstat.org/aoas/next_issue.html`.

[3] Bild AH, Yao G, Chang JT, et al.: Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, **439**:353-7, 2006.

[4] Coombes KR, Wang J, Baggerly KA: Microarrays: retracing steps. *Nat Med*, **13**:1276-7, 2007. Author reply, 1277-8.

[5] Dressman HK, Berchuck A, Chan G, et al.: An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer. *J Clin Oncol*, **25**:517-25, 2007.

[6] Györffy B, Surowiak P, Kiesslich O, et al.: Gene expression profiling of 30 cancer cell lines predicts resistance towards 11 anticancer drugs at clinically achieved concentrations. *Int J Cancer*, **118**:1699-712, 2006.

[7] Hsu DS, Balakumaran BS, Acharya CR, et al.: Pharmacogenomic strategies provide a rational approach to the treatment of cisplatin-resistant patients with advanced cancer. *J Clin Oncol*, **25**:4350-4357, 2007

[8] Potti A, Dressman HK, Bild A, et al: Genomic signatures to guide the use of chemotherapeutics. *Nat Med*, **12**:1294-1300, 2006.