Checking Sensitivity Signatures: Enumerating the Cell Lines Used

Keith A. Baggerly

September 24, 2009

Contents

1 Executive Summary

1.1 Introduction

In late 2006, Potti et al. [7] introduced a method for combining microarray profiles of cell lines with drug sensitivity data to derive "signatures" of sensitivity to specific drugs. These signatures could then be used to predict patient response. In theory, the approach is straightforward:

- Using drug sensitivity data for a panel of cell lines, select those that are most sensitive and most resistant to the drug of interest.
- Using array profiles of the identified cell lines, identify the most differentially expressed genes.
- Using the most differentially expressed genes, build a model that takes an array profile and returns a classification.

This report is part of a series in which we try to trace the specific steps involved in order to better understand the approach. In this report, we focus on identifying and tabulating the specific "sensitive" and "resistant" cell lines that have been used to assemble the signatures for various drugs.

1.2 Methods

We considered 12 sources of information about the signatures for 10 drugs. The drugs are docetaxel (D), paclitaxel (P), doxorubicin (adriamycin, A), fluorouracil (F), topotecan (T), etoposide (E), cyclophosphamide (C), pemetrexed (Pem), cisplatin (Cis), and temozolomide (Tem). The twelve sources are given below.

1. The heatmaps (D,P,A,F,T,E) presented in Potti et al. [7], November 2006.
2. The heatmaps (Cis, Pem) presented in Hsu et al. [6], October 2007.
3. The gene lists (D,P,A,F,T,E,C) presented on the Potti et al. [7] web site at the time of the article's first correction and the appearace of the initial Coombes et al. [4] critique, November 2007.*

4. The docetaxel (D) data posted on the Potti et al. [7] website at the time of first correction, November 2007.
5. The doxorubicin (A) data posted on the Potti et al. [7] website at the time of first correction, November 2007.
6. The "list of cell lines used" (D,P,A,F,T,E,C) and the "description of predictor generation" posted on the Potti et al. [7] website at the time of first correction, November 2007.*
7. The "list of cell lines used" (D,A,F,C) supplied as a webpanel supplement to Bonnefoi et al. [2], December 2007.
8. The numbers of sensitive and resistant lines (P,A,F,C) used by Salter et al. [10], April 2008.*
9. The "list of cell lines used" (D,P,A,F,T,E,C) posted on the Potti et al. [7] website at the time of second correction, August 2008.
10. The "list of cell lines used" (D,A,F,C) by Bonnefoi et al. [2] after correction, September 2008.
11. The numbers of sensitive and resistant lines (D,P,A,F,T,E,C) named by Riedel et al. [9], October 2008.*
12. The heatmap (Tem) presented in Augustine et al. [1], January 2009.

We draw inferences from heatmaps and gene lists when we are able to exactly match the reported results with the `binreg` software used by Potti et al. [7] This uniquely identifies the two groups being contrasted, but not the direction. Direction (which group is sensitive) is inferred from other statements in the relevant papers about what the figures represent. In some cases (indicated with asterisks above), either the direction or the identity of the cell lines is not precisely specified, so some information must be inferred from other sources.

1.3 Results

The NCI60 cell line lists and their sources are summarized in Figure 1. For every drug with measurements coming from more than one source, there is at least one inversion of the sensitive/resistant labeling. The sets of cell lines are different for most drugs, with two exceptions. First, the cell lines used for cyclophosphamide and pemetrexed have a stark overlap in terms of both cell line and direction. The list reported for cyclophosphamide is a subset of the set of lines used for pemetrexed, but the cell lines given for cyclophosphamide do not produce the gene list reported. The cell lines required to produce the cyclophosphamide gene list are a superset of the lines used for pemetrexed. Second, neither the set for cisplatin nor the set for temozolomide involve the NCI60 cell lines directly. The cisplatin signature is based on 30 cell lines assembled by Gyorffy et al. [5], and the heatmap reported for temozolomide is the same as the heatmap for cisplatin.

Both the gene lists initially supplied by Potti et al. [7] and Hsu et al. [6] are incorrect due to an "off-by-one" indexing error. Even after correcting for this error, gene lists from both sources contain "outliers" not derived from the cell line data; these are given at the end of the report in Table 1. The same gene outliers are present in the lists for several different drugs. In the case of the Hsu et al. [6] data for cisplatin, the outliers include probesets from a different array platform.

1.4 Conclusions

The cell lines used have changed over time, with the most common change being reversal of the sensitive and resistant labels. This inconsistency, which would reverse the classifications being produced, happens for all drugs measured more than once. A likely source of this problem is inconsistent use of 0/1 labels to represent sensitive and resistant. Indeed, the docetaxel and doxorubicin data examined (sources 4 and 5, respectively) use 0/1 labeling with the interpretations explicitly reversed, and the labeling given for docetaxel is wrong (or at any rate reversed relative to that in the description of predictor generation supplied at the same time). The mismatch between lists of cell lines supplied and what can be inferred from the gene lists and heatmaps shows that some of the lists now available use cell lines not used for Potti et al. [7]. The presence of outliers in the gene lists shows that either the annotation is faulty, or that the gene lists given as the signatures were not derived solely from the cell line data used to assemble the corresponding heatmaps.

The extreme overlap between the lists for cyclophosphamide and pemetrexed suggests that the list assembled for one drug may have been mistakenly used for the other as well, possibly with the directions reversed. The heatmap for cisplatin is indeed mistakenly used as the heatmap for temozolomide.

Either reversal of direction or use of cell lines chosen for a different drug should invalidate the predictions obtained.

2 Options and Libraries

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

3 Drugs, Cell Lines, and Quantifications

3.1 The Drugs of Interest

For each of the 10 drugs mentioned in at least one of Potti et al. [7], Hsu et al. [6], Bonnefoi et al. [2], Salter et al. [10] and Augustine et al. [1], we have collected (where possible) the Drug Name (Trade Name), NSC number, and mechanism of action. These are listed below. In the interests of completeness, we also list this information for daunorubicin and vincristine, which are used in some of the test data sets.

Drug	Trade Name	NSC #	Mechanism
Docetaxel	Taxotere	628503	Block Microtubule Disassembly
Paclitaxel	Taxol	125973	Block Microtubule Disassembly
Doxorubicin	Adriamycin	123127	Topoisomerase Inhibitor
Fluorouracil	5-FU	19893	Thymidine Antimetabolite
Topotecan	Hycamtin	609699	Topoisomerase 1 Inhibitor
Etoposide	Eposin	141540	Topoisomerase 2 Inhibitor
Cyclophosphamide	Cytoxan	26271	Crosslinking of DNA (Alkylating)
Pemetrexed	Alimta	698037	Folate Antimetabolite
Cisplatin		119875	Crosslinking of DNA (Alkylating-like)
Epirubicin	Ellence	256942	Topoisomerase Inhibitor
Daunorubicin		82151	Topoisomerase Inhibitor
Vincristine	Oncovin	67574	Block Microtubule Assembly
Temozolomide	Temodar	362856	Crosslinking of DNA (Triazene)

The NSC numbers are not given in the initial papers; they were found by querying for "(drug name) NSC number" on Google or by searching the NCI Drug Dictionary and NCI Thesaurus (for Pemetrexed). The mechanism information was acquired from Wikipedia. We store the drug name/NSC number pairs in a named vector below.

```
> nscNumbers <- c(628503, 125973, 123127, 19893, 609699, 141540,
+     26271, 698037, 119875, 256942, 82151, 67574, 362856)
```

```
> names(nscNumbers) <- c("docetaxel", "paclitaxel", "doxorubicin",
+     "fluorouracil", "topotecan", "etoposide", "cyclophosphamide",
+     "pemetrexed", "cisplatin", "epirubicin", "daunorubicin",
+     "vincristine", "temozolomide")
> nscNumbers

       docetaxel       paclitaxel      doxorubicin     fluorouracil
          628503           125973           123127            19893
       topotecan        etoposide cyclophosphamide       pemetrexed
          609699           141540            26271           698037
       cisplatin       epirubicin     daunorubicin      vincristine
          119875           256942            82151            67574
    temozolomide
          362856
```

3.2 The NCI60 Cell Lines

For all of the drugs save cisplatin, the cell lines come from the NCI-60 panel. A listing of these cell lines can be obtained from a number of sources, including http://dtp.nci.nih.gov/docs/misc/common_files/cell_list.html and http://discover.nci.nih.gov/cellminer/celllinequery.do. There is no array data for MDA-N, so we need consider only 59 cell lines. These are listed below using formatting chosen to match that in the drug sensitivity tables available from the NCI.

```
> nci60CellLines <- c("NCI-H23", "NCI-H522", "A549/ATCC", "EKVX",
+     "NCI-H226", "NCI-H322M", "NCI-H460", "HOP-62", "HOP-92",
+     "HT29", "HCC-2998", "HCT-116", "SW-620", "COLO 205", "HCT-15",
+     "KM12", "MCF7", "NCI/ADR-RES", "MDA-MB-231/ATCC", "HS 578T",
+     "MDA-MB-435", "BT-549", "T-47D", "OVCAR-3", "OVCAR-4", "OVCAR-5",
+     "OVCAR-8", "IGROV1", "SK-OV-3", "CCRF-CEM", "K-562", "MOLT-4",
+     "HL-60(TB)", "RPMI-8226", "SR", "UO-31", "SN12C", "A498",
+     "CAKI-1", "RXF 393", "786-0", "ACHN", "TK-10", "LOX IMVI",
+     "MALME-3M", "SK-MEL-2", "SK-MEL-5", "SK-MEL-28", "M14", "UACC-62",
+     "UACC-257", "PC-3", "DU-145", "SNB-19", "SNB-75", "U251",
+     "SF-268", "SF-295", "SF-539")
```

3.3 Earlier Rda Files: The NCI-60 Array Quantifications

As noted by Coombes et al. [4], the NCI-60 array quantifications used are from arrays run in triplicate by Novartis, of which only the first set (the A set) was used. We extracted the A set array quantifications and modified the formatting of the cell line names to match those given above.

```
> rdaList <- c("novartisA")
> for (rdaFile in rdaList) {
+     rdaFullFile <- file.path("RDataObjects", paste(rdaFile, "Rda",
+         sep = "."))
+     if (file.exists(rdaFullFile)) {
+         cat("loading ", rdaFullFile, " from cache\n")
+         load(rdaFullFile)
+     }
+     else {
+         cat("building ", rdaFullFile, " from raw data\n")
+         Stangle(file.path("RNowebSource", paste("buildRda", rdaFile,
+             "Rnw", sep = ".")))
+         source(paste("buildRda", rdaFile, "R", sep = "."))
+     }
+ }
```

```
loading  RDataObjects/novartisA.Rda  from cache
```

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

```
            CCRF-CEM     K-562    MOLT-4
36460_at  41.16584  95.75866  68.12313
36461_at  80.50482  98.03310  97.47627
36462_at 113.68166 200.20106 248.60211
```

```
> all(sort(colnames(novartisA)) == sort(nci60CellLines))
```

```
[1] TRUE
```

3.4 The Györffy et al. [5] Cell Lines

For cisplatin, the signature was assembled using the set of 30 cell lines profiled by Györffy et al. [5]. We list those 30 lines here, in alphabetical order.

```
> gyorffyCellLines <- c("181P", "257P", "A375", "BT20", "C8161",
+     "colo699", "Cx2", "Du145", "DV90", "ES2", "FUOV1", "Hep3B",
+     "HRT18", "HT29", "mda231", "me43", "MeWo", "OAW42", "OVKAR",
+     "R103", "R193", "SKBR3", "SKMel13", "SKMel19", "Skov3", "SNU182",
+     "SNU423", "SNU449", "SNU475", "Sw13")
```

4 Defining the Storage Structure

Here, we simply define and allocate space for a hierarchichal structure of lists for containing all of the results to be compiled below. The hierarchy involves three levels:

1. Drug, e.g., "docetaxel",
2. Source of information about the cell lines, e.g., "matching heatmaps from Potti et al. [7]", and
3. Direction, e.g., "Sensitive" or "Resistant".

The initial structure allocates space for all possible drug/source combinations; unused pairs will be pruned near the end of the report.

```
> cellLinesUsed <- vector("list", length(nscNumbers))
> names(cellLinesUsed) <- names(nscNumbers)
> informationSources <- c("heatmapPottiNov06", "heatmapHsuOct07",
+     "geneListPotti06CorrNov07", "doceDataPotti06CorrNov07", "doxoDataPotti06CorrNov07",
+     "listPotti06CorrNov07", "listBonnefoiDec07", "numbersSalterApr08",
+     "listPotti06CorrAug08", "listBonnefoi07CorrSep08", "numbersRiedelOct08",
+     "heatmapAugustineJan09")
> tempLabels <- c("Sensitive", "Resistant")
> for (tempIndex1 in 1:length(cellLinesUsed)) {
+     cellLinesUsed[[tempIndex1]] <- vector("list", length(informationSources))
+     names(cellLinesUsed[[tempIndex1]]) <- informationSources
+     for (tempIndex2 in 1:length(informationSources)) {
+         cellLinesUsed[[tempIndex1]][[tempIndex2]] <- vector("list",
+             length(tempLabels))
+         names(cellLinesUsed[[tempIndex1]][[tempIndex2]]) <- tempLabels
+     }
+ }
> rm(list = ls(pattern = "^temp"))
```

5 Heatmaps and Gene Lists from Potti et al. [7], Nov 06

Our first sources of information about the cell lines used for each drug are the heatmaps and gene lists given in Potti et al. [7]. These are produced using the `binreg` software, which is comprised of a set of Matlab scripts. These scripts are available from the web site for Potti et al. [7], `http://data.genome.duke.edu/NatureMedicine.php`. More details about how `binreg` works are given in supplementary report SR9 of Coombes et al. [4], available at `http://bioinformatics.mdanderson.org/Supplements/Repro-Rsch-Chemo`.

5.1 The Heatmaps

In a sense, the heatmaps are the "best defined" source of information, in that reproducing the heatmap means that we are indeed using the same cell lines that were in use when the graph was produced. As noted in SR9 from Coombes et al. [4], we can precisely match the heatmaps for 6 of the 7 drugs examined by Potti et al. [7] (cyclophosphamide is the exception), so we can infer the two groups of cell lines for each. We can get direction information from two sources. Figure 1 of Potti et al. [7] shows the heatmap for docetaxel, and labels the left side "Resistant" and the right side "Sensitive". This gives us the direction for docetaxel. Figure 2 and Supplementary Figure 2 of Potti et al. [7] show the heatmaps for the other 6 drugs (with the map for paclitaxel mislabeled as coming from cyclophosphamide), and the caption for Supplementary Figure 2 notes that "Panel A shows the gene expression models selected for predicting response to the indicated drugs, with resistant lines on the left, sensitive on the right for each predictor." This gives us the directions for all of the other drugs we can match. As noted, we are unable to match the heatmap for cyclophosphamide, so we have no information for this drug from this source. The lists are as follows.

```
> cellLinesUsed[["docetaxel"]][["heatmapPottiNov06"]][["Resistant"]] <- c("EKVX",
+     "IGROV1", "OVCAR-4", "786-0", "CAKI-1", "SN12C", "TK-10")
> cellLinesUsed[["docetaxel"]][["heatmapPottiNov06"]][["Sensitive"]] <- c("HL-60(TB)",
+     "SF-539", "HT29", "HOP-62", "SK-MEL-2", "SK-MEL-5", "NCI-H522")
> cellLinesUsed[["paclitaxel"]][["heatmapPottiNov06"]][["Resistant"]] <- c("SF-295",
+     "SF-539", "HS 578T", "MDA-MB-435", "COLO 205", "HCC-2998",
+     "HT29", "OVCAR-3", "DU-145")
> cellLinesUsed[["paclitaxel"]][["heatmapPottiNov06"]][["Sensitive"]] <- c("CCRF-CEM",
+     "SW-620", "A549/ATCC", "EKVX", "MALME-3M", "SK-MEL-28", "OVCAR-8",
+     "786-0")
> cellLinesUsed[["doxorubicin"]][["heatmapPottiNov06"]][["Resistant"]] <- c("SF-539",
+     "SNB-75", "MDA-MB-435", "NCI-H23", "M14", "MALME-3M", "SK-MEL-2",
+     "SK-MEL-28", "SK-MEL-5", "UACC-62")
> cellLinesUsed[["doxorubicin"]][["heatmapPottiNov06"]][["Sensitive"]] <- c("NCI/ADR-RES",
+     "HCT-15", "HT29", "EKVX", "NCI-H322M", "IGROV1", "OVCAR-3",
+     "OVCAR-4", "OVCAR-5", "OVCAR-8", "SK-OV-3", "CAKI-1")
> cellLinesUsed[["fluorouracil"]][["heatmapPottiNov06"]][["Resistant"]] <- c("MCF7",
+     "COLO 205", "HCT-116", "NCI-H460", "LOX IMVI", "SK-MEL-5",
+     "A498", "UO-31")
> cellLinesUsed[["fluorouracil"]][["heatmapPottiNov06"]][["Sensitive"]] <- c("NCI/ADR-RES",
+     "MDA-MB-435", "SW-620", "EKVX", "M14", "SN12C", "OVCAR-8")
> cellLinesUsed[["topotecan"]][["heatmapPottiNov06"]][["Resistant"]] <- c("SF-539",
+     "SNB-75", "U251", "HS 578T", "HOP-62", "NCI-H226", "NCI-H23",
+     "LOX IMVI", "OVCAR-8", "A498", "ACHN", "CAKI-1", "UO-31")
> cellLinesUsed[["topotecan"]][["heatmapPottiNov06"]][["Sensitive"]] <- c("K-562",
+     "RPMI-8226", "MDA-MB-435", "SK-MEL-5", "HCC-2998", "HCT-116",
+     "HCT-15", "NCI-H322M", "SK-MEL-28", "COLO 205")
> cellLinesUsed[["etoposide"]][["heatmapPottiNov06"]][["Resistant"]] <- c("SF-539",
+     "BT-549", "MDA-MB-231/ATCC", "NCI/ADR-RES", "HOP-62", "NCI-H226",
+     "SK-MEL-28", "UACC-257", "786-0")
> cellLinesUsed[["etoposide"]][["heatmapPottiNov06"]][["Sensitive"]] <- c("MCF7",
+     "HCC-2998", "HCT-15", "SW-620", "NCI-H322M", "PC-3", "OVCAR-4",
+     "OVCAR-5")
```

For all of the drugs except docetaxel, simply counting the numbers of sensitive and resistant lines provides a sanity check, since the boundary between the two groups is starkly visible in each heatmap. The numbers in each group are as follows.

Drug	# Resistant	# Sensitive
Docetaxel	7	7
Paclitaxel	9	8
Doxorubicin	10	12
Fluorouracil	8	7
Topotecan	13	10
Etoposide	9	8

A quick check shows that these counts hold for the lists above.

5.2 The Original Gene Lists, Nov 06

The original supplementary table 1 for Potti et al. [7] gave the lists of probesets associated with each predictor. In theory, matching the gene lists when using the `binreg` software should provide another source of information about the groups of cell lines being contrasted. Unfortunately, as noted by Coombes et al. [4], there was an indexing error that caused the reported probesets to be “off by one” from those intended, relative to the probeset order supplied to `binreg`. The ordering used was that of the raw Novartis data, which was largely alphabetic. After adjusting for this offset, (a) we still have no agreement for cyclophosphamide, (b) the gene lists matched the `binreg` output exactly for three drugs (fluorouracil, topotecan, and etoposide), and (c) the gene lists partially match for the remaining three (docetaxel, paclitaxel, and doxorubicin). In the case of docetaxel, 14 of the 19 unmatched genes appear as a contiguous block in the list of “important” genes identified by Chang et al. [3] from the data being used as the docetaxel test set. For the three partially matching cases, the unmatched “outliers” in the gene lists are as follows:

```
docetaxel:
1258_s_at  1751_g_at  1802_s_at  1878_g_at  1997_s_at
  "ERCC4"   "FARSLA"    "ERBB2"    "ERCC1"      "BAX"
2085_s_at   31431_at 31432_g_at   31638_at   32099_at
 "CTNNA1"    "FCGRT"    "FCGRT"         NA    "SAFB2"
 32331_at   32523_at 32843_s_at   33047_at   33133_at
  "AK3L1"     "CLTB"      "FBL"  "BCL2L11"     "FLII"
 33214_at 33285_i_at 33371_s_at   40567_at
 "MRPS12"     "SIKE"    "RAB31" "K-ALPHA-1"

As noted, the docetaxel list breaks down into groups of 14 and 5:

docetaxel in Chang (14):
1751_g_at  1997_s_at
 "FARSLA"      "BAX"
2085_s_at   31431_at  31432_g_at  31638_at  32099_at
 "CTNNA1"    "FCGRT"     "FCGRT"        NA   "SAFB2"
 32331_at   32523_at  32843_s_at  33133_at
  "AK3L1"     "CLTB"       "FBL"    "FLII"
 33214_at 33285_i_at 33371_s_at
"MRPS12"      "SIKE"    "RAB31"

docetaxel not in Chang (5):
1258_s_at   1802_s_at   1878_g_at   33047_at   40567_at
  "ERCC4"     "ERBB2"     "ERCC1"  "BCL2L11" "K-ALPHA-1"

paclitaxel:
1258_s_at   1802_s_at   1878_g_at   33047_at   36519_at
  "ERCC4"     "ERBB2"     "ERCC1"  "BCL2L11"    "ERCC1"
 40567_at    114_r_at
"K-ALPHA-1"    "MAPT"

doxorubicin:
1258_s_at   1847_s_at     1909_at  1910_s_at  2034_s_at
  "ERCC4"      "BCL2"      "BCL2"     "BCL2"   "CDKN1B"
```

All 5 of the "non-Chang" docetaxel outliers are also in the list of 7 outliers for paclitaxel. The two others in the paclitaxel list are 36519_at, ERCC1, and 114_r_at, MAPT. The first of these, ERCC1, is also interrogated by 1878_g_at, which is a common outlier for both paclitaxel and docetaxel. There is one probeset, 1258_s_at or ERCC4, which is an outlier for docetaxel, paclitaxel, and doxorubicin. The ERCC family is overrepresented in the outliers. Many of the outlier genes are explicitly named in the main text of Potti et al. [7] For docetaxel, "both predictors were linked to expected targets for docetaxel, including BCL2, WDR7 (also known as TRAG), ERBB2 and tubulin genes". For paclitaxel, "An examination of the genes that constituted the paclitaxel predictor identified microtubule-associated protein tau (MAPT), described previously as a determinant of paclitaxel sensitivity." For doxorubicin, "excision repair genes (for example, ERCC4), retinoblastoma pathway genes and BCL2 constituted the adriamycin predictor." While we are unable to explain how they were chosen, the outliers are evidently important.

6 Heatmaps and Gene Lists from Hsu et al. [6], Oct 07

In October of 2007, Hsu et al. [6] used the approach of Potti et al. [7] to develop sensitivity signatures for cisplatin and pemetrexed. The NCI-60 data was used to generate the pemetrexed signature, but the cisplatin signature was derived from cell line data supplied by Györffy et al. [5] who used the Affymetrix U133A platform as opposed to the U95Av2 for the NCI-60. The associated heatmaps for the two signatures are shown in Figure 1 of Hsu et al. [6]; the corresponding gene lists were supplied by Hsu et al. [6] as Supplementary tables 1 and 2 (for cisplatin and pemetrexed, respectively). As with Potti et al. [7], we matched the heatmaps in order to infer the cell lines and genes involved, and checked the gene lists to see if these told a consistent story. Details of this matching, which involves both intelligent guessing and brute force, are given in a separate report, matchingHsuHeatmaps.pdf; only the results are summarized here.

6.1 The Cisplatin Heatmap

Based on the heatmap, the Györffy et al. [5] cell lines used in the cisplatin signature are as follows:

```
> cellLinesUsed[["cisplatin"]][["heatmapHsuOct07"]][["Resistant"]] <- c("257P",
+     "A375", "C8161", "ES2", "me43", "MeWo", "SKMel19", "SNU423",
+     "Sw13")
> cellLinesUsed[["cisplatin"]][["heatmapHsuOct07"]][["Sensitive"]] <- c("BT20",
+     "DV90", "FUOV1", "OAW42", "OVKAR", "R103")
```

The heatmap sides are explicitly labeled sensitive and resistant, so the plot supplies direction as well as grouping. All of the lines used are indeed listed by Györffy et al. [5] as sensitive or resistant, respectively, so that is consistent. (The cell line labeled OVCAR3 in Figure 2 of Györffy et al. [5] is labeled OVKAR in the table of quantifications.) However, other cell lines in the set of 30 examined were also labeled as sensitive or resistant; only 3 of the 30 were labeled as indeterminate. We have no idea how or why this subset of just 15 samples was selected.

6.2 The Pemetrexed Heatmap

Based on the heatmap, the NCI-60 cell lines used in the pemetrexed signature are as follows:

```
> cellLinesUsed[["pemetrexed"]][["heatmapHsuOct07"]][["Resistant"]] <- c("K-562",
+     "MOLT-4", "HL-60(TB)", "MCF7", "HCC-2998", "HCT-116", "NCI-H460",
+     "TK-10")
> cellLinesUsed[["pemetrexed"]][["heatmapHsuOct07"]][["Sensitive"]] <- c("SNB-19",
+     "HS 578T", "MDA-MB-231/ATCC", "MDA-MB-435", "NCI-H226", "M14",
+     "MALME-3M", "SK-MEL-2", "SK-MEL-28", "SN12C")
```

As with cisplatin, the heatmap sides are explicitly labeled sensitive and resistant, so the plot supplies direction as well as grouping.

6.3 Gene Lists from Hsu et al. [6]

The two supplementary tables for Hsu et al. [6] give the lists of probesets associated with the cisplatin and pemetrexed predictors, respectively. Unfortunately, as with Potti et al. [7], there is an indexing error that causes the reported probesets to be “off by one” from the one intended, relative to the probeset order supplied to `binreg`. The ordering used for pemetrexed was that of the raw Novartis data, which was largely alphabetic, and the ordering used for cisplatin was that of the quantification table supplied by Györffy et al. [5]. After adjusting for this offset, the gene list matches the `binreg` output exactly for pemetrexed. The gene list matches the `binreg` output in 41/45 cases for cisplatin. For cisplatin, the unmatched "outliers" in the gene list are as follows:

```
cisplatin:
203719_at  210158_at  228131_at  231971_at
  "ERCC1"    "ERCC4"    "ERCC1"*   "FANCM"*

* probeset on the U133B chip, not the U133A.
```

Hsu et al. [6] note that

> The cisplatin sensitivity predictor includes DNA repair genes such as ERCC1 and ERCC4, among others, that had altered expression in the list of cisplatin sensitivity predictor genes. Interestingly, one previously described mechanism of resistance to cisplatin therapy results from the increased capacity of cancer cells to repair DNA damage incurred, by activation of DNA repair genes.

As with the outliers in the initial gene lists from Potti et al. [7], while we are unable to explain how they were chosen (particularly in the case of probesets from a different chip platform), the outliers are evidently important.

7 Corrected Gene Lists from Potti et al. [7], Nov 07

The gene lists for Potti et al. [7] were corrected twice, first in May of 2007 and again in October of 2007. The October 2007 lists were also provided on the Potti et al. [7] web site at the time of the first official correction of the article, in November 2007. The May 2007 corrections addressed the off-by-one issue, but using a different initial ordering and without excluding the outliers. There was no list given for cyclophosphamide. The October 2007 corrections addressed both offsets and outliers, and included a new gene list for cyclophosphamide. Using the latter set of gene lists we identified sets of cell lines that would produce them. For the 6 drugs where we matched heatmaps, the sets of cell lines are the same as those inferred above. Identifying the cell lines required to produce the gene list reported for cyclophosphamide, is the subject of a separate report, identifyingCyclophosphamideLines.pdf; only the results are given here. The groups of cell lines that yield the reported gene list are given below.

```
> tempDrugs <- c("docetaxel", "paclitaxel", "doxorubicin",
+                "fluorouracil", "topotecan", "etoposide")
> for(tempIndex in 1:length(tempDrugs)){
+   cellLinesUsed[[tempDrugs[tempIndex]]][["geneListPotti06CorrNov07"]] <-
+     cellLinesUsed[[tempDrugs[tempIndex]]][["heatmapPottiNov06"]]
+ }
> cellLinesUsed[["cyclophosphamide"]][["geneListPotti06CorrNov07"]][["Sensitive"]] <-
+   c("K-562","MOLT-4","HL-60(TB)","MCF7","HCC-2998","HCT-116",
+     "NCI-H460","TK-10")
> cellLinesUsed[["cyclophosphamide"]][["geneListPotti06CorrNov07"]][["Resistant"]] <-
+   c("SNB-19","HS 578T","MDA-MB-231/ATCC","MDA-MB-435",
+     "SW-620", "EKVX", "NCI-H226","M14",
+     "MALME-3M","SK-MEL-2","SK-MEL-28","SN12C")
```

In this case, we do not know the directions explicitly, so we infer them by looking for maximal concordance with the lists of cell lines posted on the Potti et al. [7] website in November of 2007, the time of the Coombes et al. [4] and the reply by Potti et al. [8]. These lists are given in the next section.

8 Cell Line Lists from Potti 06, Correction Nov 07

In November of 2007, the web site for Potti et al. [7] contained at least 5 files that supplied information about the cell lines used:

- GeneLists.zip
- CellLinesInEachChemoPredictor.xls
- DescriptorOfPredictorGeneration.doc
- BreastData.txt
- Adria_ALL.txt

We discussed the gene lists in the previous section. We cover the next two files in this section, and the last two in the sections following.

The most directly relevant file is **CellLinesInEachChemoPredictor.xls**. As its name suggests, it has one row per drug, and each row has names of cell lines in two colors: yellow first, then blue. For example, the first row is “Taxotere” (docetaxel), followed by EKVX, IGROV1, OVCAR-4, 786-0, CAKI-1, SN12C and TK-10 in yellow and by HL-60(TB), SF-539, HT29, HOP-62, SK-MEL-2, SK-MEL-5, and NCI-H522 in blue. These are precisely the resistant and sensitive lines identified for docetaxel above, suggesting that *Yellow = resistant*, and *Blue = sensitive*. We apply this direction throughout. We note that there is no further information in this table, just drug names, cell line names, and colors.

These lists are *almost*, but not quite, the same as those inferred above. We give these lists below.

```
> cellLinesUsed[["docetaxel"]][["listPotti06CorrNov07"]][["Resistant"]] <- c("EKVX",
+     "IGROV1", "OVCAR-4", "786-0", "CAKI-1", "SN12C", "TK-10")
> cellLinesUsed[["docetaxel"]][["listPotti06CorrNov07"]][["Sensitive"]] <- c("HL-60(TB)",
+     "SF-539", "HT29", "HOP-62", "SK-MEL-2", "SK-MEL-5", "NCI-H522")
> cellLinesUsed[["paclitaxel"]][["listPotti06CorrNov07"]][["Resistant"]] <- c("SF-295",
+     "SF-539", "HS 578T", "MDA-MB-435", "COLO 205", "HCC-2998",
+     "HT29", "OVCAR-3", "NCI-H522")
> cellLinesUsed[["paclitaxel"]][["listPotti06CorrNov07"]][["Sensitive"]] <- c("CCRF-CEM",
+     "SW-620", "A549/ATCC", "EKVX", "MALME-3M", "SK-MEL-28", "OVCAR-8",
+     "786-0")
> cellLinesUsed[["doxorubicin"]][["listPotti06CorrNov07"]][["Resistant"]] <- c("SF-539",
+     "SNB-75", "MDA-MB-435", "NCI-H23", "M14", "MALME-3M", "SK-MEL-2",
+     "SK-MEL-28", "SK-MEL-5", "UACC-62")
> cellLinesUsed[["doxorubicin"]][["listPotti06CorrNov07"]][["Sensitive"]] <- c("NCI/ADR-RES",
+     "HCT-15", "HT29", "EKVX", "NCI-H322M", "IGROV1", "OVCAR-3",
+     "OVCAR-4", "OVCAR-5", "OVCAR-8", "SK-OV-3", "CAKI-1")
> cellLinesUsed[["fluorouracil"]][["listPotti06CorrNov07"]][["Resistant"]] <- c("MCF7",
+     "COLO 205", "HCT-116", "NCI-H460", "LOX IMVI", "SK-MEL-5",
+     "A498", "UO-31")
> cellLinesUsed[["fluorouracil"]][["listPotti06CorrNov07"]][["Sensitive"]] <- c("NCI/ADR-RES",
+     "MDA-MB-435", "SW-620", "EKVX", "M14", "SN12C", "OVCAR-8")
> cellLinesUsed[["topotecan"]][["listPotti06CorrNov07"]][["Resistant"]] <- c("SF-539",
+     "SNB-75", "U251", "HS 578T", "HOP-62", "NCI-H226", "NCI-H23",
+     "LOX IMVI", "OVCAR-8", "A498", "ACHN", "CAKI-1", "UO-31")
> cellLinesUsed[["topotecan"]][["listPotti06CorrNov07"]][["Sensitive"]] <- c("K-562",
+     "RPMI-8226", "MDA-MB-435", "MDA-MB-231/ATCC", "HCC-2998",
+     "HCT-116", "HCT-15", "NCI-H322M", "SK-MEL-28", "COLO 205")
> cellLinesUsed[["etoposide"]][["listPotti06CorrNov07"]][["Resistant"]] <- c("SF-539",
+     "BT-549", "MDA-MB-231/ATCC", "HCC-2998", "HOP-62", "NCI-H226",
+     "M14", "PC-3", "786-0")
> cellLinesUsed[["etoposide"]][["listPotti06CorrNov07"]][["Sensitive"]] <- c("MCF7",
+     "NCI/ADR-RES", "HCT-15", "SW-620", "NCI-H322M", "UACC-257",
+     "OVCAR-4", "OVCAR-5")
> cellLinesUsed[["cyclophosphamide"]][["listPotti06CorrNov07"]][["Resistant"]] <- c("K-562",
+     "MOLT-4", "HL-60(TB)", "MCF7", "HCC-2998", "HCT-116", "NCI-H460",
+     "TK-10")
> cellLinesUsed[["cyclophosphamide"]][["listPotti06CorrNov07"]][["Sensitive"]] <- c("SNB-19",
+     "HS 578T", "MDA-MB-231/ATCC", "MDA-MB-435", "NCI-H226", "M14",
+     "MALME-3M", "SK-MEL-2")
```

The specific differences are as follows:

Docetaxel : No difference.

Paclitaxel : Resistant line DU-145 from the heatmap is replaced with NCI-H522.

Doxorubicin : No difference.

Fluorouracil : No difference.

Topotecan : Sensitive line SK-MEL-5 from the heatmap is replaced with MDA-MB-231/ATCC.

Etoposide : Sensitive lines HCC-2998 and PC-3 from the heatmap are now listed as resistant. Resistant lines NCI/ADR-RES and UACC-257 from the heatmap are now listed as sensitive. Resistant line SK-MEL-28 from the heatmap has been replaced with M14.

Cyclophosphamide : Resistant lines SW-620, EKVX, SK-MEL-28, and SN12C from the gene list set are not in the list supplied.

In the cases where there are differences, using the lists of cell lines supplied does not produce the corresponding gene lists. As noted, the above discussion assumes that the *Yellow = resistant*, *Blue = sensitive* mapping holds. The other files provide further evidence as to direction, but this evidence (discussed below) is somewhat muddled.

The **DescriptorOfPredictorGeneration.doc** file contains a description of how the sensitive and resistant lists for docetaxel were assembled. The cell lines named and their direction match the mapping used above.

9 Docetaxel Data from Potti et al. [7], Correction Nov 07

The **BreastData.txt** file posted on the Potti et al. [7] web site in November of 2007 contains quantifications and some limited annotation for 38 arrays: 14 cell lines and 24 test data samples. The file is set up to use the cell line data for docetaxel to predict the response status of patients treated with single agent docetaxel, with the latter data coming from Chang et al. [3]. The first 3 rows and 14 columns of this table are as follows:

```
Columns 1-7:
Docetaxel 0         0           0           0           0           0           0
Sensitive   Sensitive   Sensitive   Sensitive   Sensitive   Sensitive   Sensitive
     0.54        2.21        0.59        1.28      1.85         1.2        0.87

Columns 8-14:
        1           1           1           1           1           1  Docetaxel1
Resistant   Resistant   Resistant   Resistant   Resistant   Resistant   Resistant
     1.97        0.22        1.53        0.09        2.18        1.12        3.92
```

Here, the numerical encoding of sample status is apparently *0=Sensitive, 1=Resistant.* In order to figure out where these numbers came from, we returned to the raw array data. As noted by Coombes et al. [4], this is from a set of arrays run in triplicate by Novartis, of which only the first set (the A set) was used. The first row in the Novartis dataset corresponds to probeset 36460_at. We give this row and a processed version of it below.

```
> temp <- novartisA["36460_at", c("EKVX", "IGROV1", "OVCAR-4",
+     "786-0", "CAKI-1", "SN12C", "TK-10", "HL-60(TB)", "SF-539",
+     "HT29", "HOP-62", "SK-MEL-2", "SK-MEL-5", "NCI-H522")]
> temp

     EKVX    IGROV1   OVCAR-4     786-0    CAKI-1     SN12C     TK-10 HL-60(TB)
 63.30723 120.00932  65.80951  93.56745 110.60899  90.80546  78.51872 113.89214
   SF-539      HT29    HOP-62  SK-MEL-2  SK-MEL-5  NCI-H522
 41.67195 101.62666  27.85423 119.45343  88.21540 156.09003

> t(round(exp(scale(log(temp))), 2))

     EKVX IGROV1 OVCAR-4 786-0 CAKI-1 SN12C TK-10 HL-60(TB) SF-539 HT29 HOP-62
[1,] 0.54   2.21    0.59  1.28   1.85   1.2  0.87      1.97   0.22 1.53   0.09
     SK-MEL-2 SK-MEL-5 NCI-H522
[1,]     2.18     1.12     3.92
attr(,"scaled:center")
[1] 4.426113
attr(,"scaled:scale")
[1] 0.4568785

> rm(list = ls(pattern = "^temp"))
```

We get the values reported by taking the Novartis data for the docetaxel cell lines, log transforming it, centering and scaling the result, exponentiating to undo the log transform, and then rounding to two decimal places. However, the order should be noted: the first 7 columns, which are labeled “Sensitive”, correspond to the cell lines identified previously as being Resistant, and so on. *One set of labels is reversed.* Given the detail presented in the description of predictor generation, we’re fairly certain that the data labeling examined here is incorrect.

```
> cellLinesUsed[["docetaxel"]][["doceDataPotti06CorrNov07"]][["Sensitive"]] <-
+   cellLinesUsed[["docetaxel"]][["listPotti06CorrNov07"]][["Resistant"]]
> cellLinesUsed[["docetaxel"]][["doceDataPotti06CorrNov07"]][["Resistant"]] <-
+   cellLinesUsed[["docetaxel"]][["listPotti06CorrNov07"]][["Sensitive"]]
```

10 Doxorubicin Data from Potti et al. [7], Correction Nov 07

The **Adria_ALL.txt** file is likewise set up to use the cell line data for doxorubicin (adriamycin) to predict the response status of patients. The file contains quantification information for 144 arrays: 22 cell lines and 122 test data samples. The first 3 rows and 22 columns of this file are as follows:

```
Columns 1-5:
Adria0             0          0           0          0
Resistant   Resistant   Resistant   Resistant   Resistant
     1.18        1.12        3.46        0.65        3.07

Columns 6-10:
        0           0           0           0           0
Resistant   Resistant   Resistant   Resistant   Resistant
     1.57        0.13        1.05        2.38        1.53

Columns 11-16:
Adria1    1      1      1      1      1
Sens   Sens   Sens   Sens   Sens   Sens
6.58   1.34   0.31   1.38   0.47   0.85

Columns 17-22:
   1      1      1      1      1      1
Sens   Sens   Sens   Sens   Sens   Sens
0.91   0.41   0.31   3.78   0.18   1.22
```

Here, the numerical encoding of sample status is apparently *0=Resistant, 1=Sensitive*, which is the reverse of that used for docetaxel. Simply counting the number of sensitive and resistant lines suggests concordance with those identified above. In order to figure out where these numbers came from, we returned to the raw array data. However, there is a new complexity, in that there are only 8958 rows of data here, not the 12625 from the full Novartis data or the 12558 we get by excluding the Affymetrix control probesets. This is because the test data was run on the U133A platform, and “chip comparer” (`http://tenero.duhs.duke.edu/genearray/perl/chip/chipcomparer.pl`) was used to match probes across platforms. After some experimentation, we find that probeset 35753_at is interesting.

```
> temp <- novartisA["35753_at", c("SF-539", "SNB-75", "MDA-MB-435",
+     "NCI-H23", "M14", "MALME-3M", "SK-MEL-2", "SK-MEL-28", "SK-MEL-5",
+     "UACC-62", "NCI/ADR-RES", "HCT-15", "HT29", "EKVX", "NCI-H322M",
+     "IGROV1", "OVCAR-3", "OVCAR-4", "OVCAR-5", "OVCAR-8", "SK-OV-3",
+     "CAKI-1")]
> temp
```

```
    SF-539      SNB-75  MDA-MB-435     NCI-H23         M14    MALME-3M
  901.5941    884.5792   1351.3036    722.8375   1291.4535   1005.0535
```

```
   SK-MEL-2   SK-MEL-28    SK-MEL-5     UACC-62 NCI/ADR-RES      HCT-15
   399.0167    864.2867   1173.3961    995.1872   1717.5609    945.9998
       HT29        EKVX   NCI-H322M      IGROV1     OVCAR-3     OVCAR-4
   546.7088    958.4876    638.6286    797.8848    819.8338    605.1154
    OVCAR-5     OVCAR-8     SK-OV-3      CAKI-1
   547.6945   1396.6946    441.9525    915.3911
```

```
> t(round(exp(scale(log(temp))), 2))
```

```
     SF-539 SNB-75 MDA-MB-435 NCI-H23  M14 MALME-3M SK-MEL-2 SK-MEL-28 SK-MEL-5
[1,]   1.18   1.12       3.46    0.65 3.07     1.57     0.13      1.05     2.38
     UACC-62 NCI/ADR-RES HCT-15 HT29 EKVX NCI-H322M IGROV1 OVCAR-3 OVCAR-4
[1,]    1.53        6.58   1.34 0.31 1.38      0.47   0.85    0.91     0.4
     OVCAR-5 OVCAR-8 SK-OV-3 CAKI-1
[1,]    0.31    3.78    0.17   1.22
attr(,"scaled:center")
[1] 6.743783
attr(,"scaled:scale")
[1] 0.3742465
```

```
> t(round(exp(scale(log(temp))), 3))[1, c("OVCAR-4", "SK-OV-3")]
```

```
OVCAR-4 SK-OV-3
  0.405   0.175
```

```
> rm(list = ls(pattern = "^temp"))
```

We get the values reported using the same processing applied with docetaxel; in the two cases with slight disagreements the mismatches are likely driven by rounding errors. Again, the order should be noted: the first 10 columns, which are labeled "Resistant", correspond to the Resistant cell lines identified previously.

```
> cellLinesUsed[["doxorubicin"]][["doxoDataPotti06CorrNov07"]] <-
+   cellLinesUsed[["doxorubicin"]][["listPotti06CorrNov07"]]
```

11 Cell Line Lists from Bonnefoi et al. [2], Dec 07

The initial publication of Bonnefoi et al. [2] included (as a webpanel) lists of the cell lines designated as sensitive and resistant for docetaxel, doxorubicin, fluorouracil, and cyclophosphamide. These lists were identical to those given in the **CellLinesInEachChemoPredictor.xls** file on the Potti et al. [7] website, but with the sensitive and resistant labels reversed from those inferred above for all four drugs.

```
> tempDrugs <- c("docetaxel", "doxorubicin", "fluorouracil",
+                "cyclophosphamide")
> for(tempIndex in 1:length(tempDrugs)){
+   cellLinesUsed[[tempDrugs[tempIndex]]][["listBonnefoiDec07"]][["Sensitive"]] <-
+     cellLinesUsed[[tempDrugs[tempIndex]]][["listPotti06CorrNov07"]][["Resistant"]]
+   cellLinesUsed[[tempDrugs[tempIndex]]][["listBonnefoiDec07"]][["Resistant"]] <-
+     cellLinesUsed[[tempDrugs[tempIndex]]][["listPotti06CorrNov07"]][["Sensitive"]]
+ }
> rm(list=ls(pattern="^temp"))
```

12 Numbers from Salter et al. [10], Apr 08

Salter et al. [10] revisit paclitaxel, doxorubicin, fluorouracil and cyclophosphamide using NCI-60 data from the U133A platform, noting that "The NCI-60 cell lines that were most resistant or sensitive to each chemotherapy agent were identified as previously described [7]." We have not attempted to fully match the heatmaps here, but have simply counted the numbers of sensitive and resistant lines (direction can be inferred from Supplementary Figure S1). These are tabulated below.

Drug	# Resistant	# Sensitive
Paclitaxel	8	9
Doxorubicin	12	10
Fluorouracil	8	7
Cyclophosphamide	8	8

These numbers disagree with those from the postings in 2007 for paclitaxel and doxorubicin, but agree in the case of fluorouracil. These numbers alone do not resolve the direction issue in the case of cyclophosphamide. Since the heatmap for cyclophosphamide shows most genes to be higher in the left (resistant) lines, we loaded the MAS5.0 quantifications for these data and checked which direction would work using the cell lines named in the Potti et al. [7] Nov 2007 posting; details are given in the report checkingSalterNumbers.pdf. This check shows that cyclophosphamide has the same orientation as fluorouracil, so two drugs agree and two drugs disagree with the directions inferred previously. Related comments about microRNAs higher in the "resistant" groups show that all four directions used here were treated as valid.

```
> tempDrugs <- c("paclitaxel", "doxorubicin")
> for(tempIndex in 1:length(tempDrugs)){
+   cellLinesUsed[[tempDrugs[tempIndex]]][["numbersSalterApr08"]][["Sensitive"]] <-
+     cellLinesUsed[[tempDrugs[tempIndex]]][["listPotti06CorrNov07"]][["Resistant"]]
+   cellLinesUsed[[tempDrugs[tempIndex]]][["numbersSalterApr08"]][["Resistant"]] <-
+     cellLinesUsed[[tempDrugs[tempIndex]]][["listPotti06CorrNov07"]][["Sensitive"]]
+ }
> tempDrugs <- c("fluorouracil", "cyclophosphamide")
> for(tempIndex in 1:length(tempDrugs)){
+   cellLinesUsed[[tempDrugs[tempIndex]]][["numbersSalterApr08"]] <-
+     cellLinesUsed[[tempDrugs[tempIndex]]][["listPotti06CorrNov07"]]
+ }
> rm(list=ls(pattern="^temp"))
```

13 Cell Line Lists from Potti et al. [7], Second Correction Aug 08

In August of 2008, a second correction was provided for Potti et al. [7], and the posted data were revised again. The **BreastData.txt** and **Adria_ALL.txt** files were removed. The latter file was replaced with a file naming the test samples used, but saying nothing about the cell lines used to construct the signature. The **GeneLists.zip** and **DescriptorOfPredictorGeneration.doc** files remained unchanged. The **CellLinesInEachChemoPredictor.xls** file was changed to **Celllines_in_each_predictor1.xls**, with the preface

> Please note that the prediction of taxotere (docetaxel) as reported in the Nature Medicine paper (Figure 1) was for resistance rather than sensitivity. The remaining signatures predicted sensitivity to the drugs. The identity of the cell lines with respect to resistant or sensitive is labeled below.

The sets of cell lines used are the same as in the previously provided Excel file, but the Sensitive and Resistant labels have been made explicit, and are reversed from those inferred above for all drugs save docetaxel.

```
> tempDrugs <- c("paclitaxel", "doxorubicin", "fluorouracil",
+                "topotecan", "etoposide", "cyclophosphamide")
> for(tempIndex in 1:length(tempDrugs)){
+   cellLinesUsed[[tempDrugs[tempIndex]]][["listPotti06CorrAug08"]][["Sensitive"]] <-
+     cellLinesUsed[[tempDrugs[tempIndex]]][["listPotti06CorrNov07"]][["Resistant"]]
+   cellLinesUsed[[tempDrugs[tempIndex]]][["listPotti06CorrAug08"]][["Resistant"]] <-
+     cellLinesUsed[[tempDrugs[tempIndex]]][["listPotti06CorrNov07"]][["Sensitive"]]
+ }
> tempDrugs <- c("docetaxel")
> for(tempIndex in 1:length(tempDrugs)){
+   cellLinesUsed[[tempDrugs[tempIndex]]][["listPotti06CorrAug08"]] <-
+     cellLinesUsed[[tempDrugs[tempIndex]]][["listPotti06CorrNov07"]]
+ }
> rm(list=ls(pattern="^temp"))
```

14 Cell Line Lists from Bonnefoi et al. [2], Correction Sep 08

In September of 2008, a correction was posted for Bonnefoi et al. [2], noting simply that “In the web-panel of this article, the headings for docetaxel should have been ‘Resistant cell lines’ (first) and ‘Sensitive cell lines’ (second).” This brings the lists for these four drugs into alignment with those given in **Celllines_in_each_predictor1.xls**.

```
> tempDrugs <- c("docetaxel", "doxorubicin", "fluorouracil",
+                "cyclophosphamide")
> for(tempIndex in 1:length(tempDrugs)){
+   cellLinesUsed[[tempDrugs[tempIndex]]][["listBonnefoi07CorrSep08"]] <-
+     cellLinesUsed[[tempDrugs[tempIndex]]][["listPotti06CorrAug08"]]
+ }
> rm(list=ls(pattern="^temp"))
```

15 Numbers from Riedel et al. [9], Oct 08

Riedel et al. [9] pursue a related problem, citing the chemotherapy signatures in the context of assembling pathway signatures. They do not list the cell lines used in the drug signatures by name, but they do list the numbers of resistant and sensitive lines for each drug. These are tabulated below.

Drug	# Resistant	# Sensitive
Docetaxel	7	7
Paclitaxel	9	8
Doxorubicin	10	12
Fluorouracil	8	7
Topotecan	13	10
Etoposide	9	8
Cyclophosphamide	8	8

These numbers agree with those from the postings in 2007, which are reversed from the list posted in August 2008 on the Potti et al. [7] web site in the cases where we can distinguish them. We have taken this to indicate that the directions were reversed for all drugs; we cannot explicitly check this in the cases of docetaxel and cyclophosphamide.

```
> tempDrugs <-
+   c("docetaxel", "paclitaxel","doxorubicin", "fluorouracil",
+     "topotecan", "etoposide", "cyclophosphamide")
> for(tempIndex in 1:length(tempDrugs)){
+   cellLinesUsed[[tempDrugs[tempIndex]]][["numbersRiedelOct08"]] <-
+     cellLinesUsed[[tempDrugs[tempIndex]]][["listPotti06CorrNov07"]]
+ }
> rm(list=ls(pattern="^temp"))
```

16 Heatmap from Augustine et al. [1], Jan 09

In January of 2009, Augustine et al. [1] used the same approach to construct a signature of sensitivity to temozolomide in the context of melanoma. Figure 4A of that paper shows the associated heatmap, nominally derived from the NCI-60. Supplementary Figure 3 of that paper shows the associated metagene scores.

We use the phrase “nominally derived from the NCI-60” because, as it happens, the heatmap is not derived from these cell lines. This heatmap and the associated metagene score plot are the same as Figure 1A of Hsu et al. [6], which corresponds to the signature for cisplatin derived from the cell lines of Györffy et al. [5], not from the NCI-60. The cell lines required to generate this figure are identified in the section on the Hsu et al. [6] heatmap above.

```
> cellLinesUsed[["temozolomide"]][["heatmapAugustineJan09"]] <-
+   cellLinesUsed[["cisplatin"]][["heatmapHsuOct07"]]
```

17 Consolidating the Data

At this point, we have collected all of the cell line membership information; we now need to process and display the information. This section focuses on the first of these tasks. We begin by pruning the list structure down to retain only the drug/source of information combinations that were actually seen.

```
> for(tempDrug in names(nscNumbers)){
+   for(tempSource in informationSources){
+     for(tempStatus in c("Sensitive","Resistant")){
+       if(is.null(cellLinesUsed[[tempDrug]][[tempSource]][[tempStatus]])){
+         cellLinesUsed[[tempDrug]][[tempSource]][[tempStatus]] <- NULL
+       }
+     }
+     if(length(cellLinesUsed[[tempDrug]][[tempSource]]) == 0){
+       cellLinesUsed[[tempDrug]][[tempSource]] <- NULL
+     }
+   }
+   if(length(cellLinesUsed[[tempDrug]]) == 0){
+     cellLinesUsed[[tempDrug]] <- NULL
+   }
+ }
> rm(list=ls(pattern="^temp"))
```

Next, we run through the remaining structure to assemble a "sensitivity matrix" for the NCI-60 cell lines, where the rows are cell lines, the columns are drug/information source combinations, and the entries are "Sensitive", "Resistant", or "Not Used".

```
> tempColumnCount <- 0
> for (tempIndex in 1:length(cellLinesUsed)) {
+     tempColumnCount <- tempColumnCount + length(cellLinesUsed[[tempIndex]])
+ }
> sensitivityMatrix <- matrix("Not Used", nrow = length(nci60CellLines),
+     ncol = tempColumnCount)
> rownames(sensitivityMatrix) <- nci60CellLines
> tempColumn <- 1
> for (tempDrug in 1:length(cellLinesUsed)) {
+     for (tempSource in 1:length(cellLinesUsed[[tempDrug]])) {
+         if (all(!is.na(match(cellLinesUsed[[tempDrug]][[tempSource]][["Sensitive"]],
+             nci60CellLines)))) {
+             sensitivityMatrix[cellLinesUsed[[tempDrug]][[tempSource]][["Sensitive"]],
+                 tempColumn] <- "Sensitive"
+             sensitivityMatrix[cellLinesUsed[[tempDrug]][[tempSource]][["Resistant"]],
+                 tempColumn] <- "Resistant"
+         }
+         tempColumn <- tempColumn + 1
+     }
+ }
> rm(list = ls(pattern = "^temp"))
```

Finally, we assemble a binary "source matrix" indicating the source used for every column in the sensitivity matrix above.

```
> tempColumn <- 1
> sourceMatrix <- matrix(FALSE, nrow = length(informationSources),
+     ncol = dim(sensitivityMatrix)[2])
> rownames(sourceMatrix) <- informationSources
> for (tempDrug in 1:length(cellLinesUsed)) {
+     for (tempSource in 1:length(cellLinesUsed[[tempDrug]])) {
+         tempSourceRow <- match(names(cellLinesUsed[[tempDrug]])[tempSource],
+             informationSources)
+         sourceMatrix[tempSourceRow, tempColumn] <- TRUE
+         tempColumn <- tempColumn + 1
+     }
+ }
> rm(list = ls(pattern = "^temp"))
```

18 Plotting the Results

We now assemble a figure summarizing the directions and identities of the cell line lists identified in the previous sections. This figure uses 3 different panels for display:

- Panel 1 is the main display of cell line against drug/information source, using color coding to indicate direction.
- Panel 2 is a cross-linking display below the first, indicating what information source was used to infer the cell line identities.
- Panel 3 is a simple legend indicating which colors are “Sensitive” and “Resistant”.

We begin by collecting some useful index numbers (number of drugs, etc) and defining an overall ”shrink factor” to scale the figure down to plottable size.

```
> tempNDrugs <- length(cellLinesUsed)
> tempDrugCounts <- rep(0, tempNDrugs)
> for (tempDrug in 1:tempNDrugs) {
+     tempDrugCounts[tempDrug] <- length(cellLinesUsed[[tempDrug]])
+ }
> names(tempDrugCounts) <- names(cellLinesUsed)
> tempDrugLastIndices <- cumsum(tempDrugCounts)
> tempDrugCenters <- (c(1, tempDrugLastIndices[1:(tempNDrugs -
+     1)] + 1) + tempDrugLastIndices)/2
> tempShrink <- 1
```

Next, we partition the figure into panels and define the sensitive and resistant plotting colors.

```
> tempMainImageMapPanel <- c(10/46, 45/46, 11.5/60, 55.5/60)
> tempSourceMapPanel <- c(10/46, 45/46, 1.5/60, 10.5/60)
> tempLegendPanel <- c(2/46, 7.5/46, 56.5/60, 59/60)
> tempSensitiveColor <- "blue"
> tempResistantColor <- "red"
```

Now we draw in the main panel. We first add the blocks indicating the sensitive lines, and then superimpose a plot indicating the resistant lines. Text strings for the cell line and drug names are added, with character sizes scaled down a bit further in order to fit a single row of the image drawn. Some of the drug names need to be vertically displaced so as not to overlap; we add vertical tick marks to connect the names to the corresponding image columns.

```
> par(plt = tempMainImageMapPanel)
> image(1:max(tempDrugLastIndices), 1:length(nci60CellLines), t(sensitivityMatrix ==
+     "Sensitive"), xlab = "", ylab = "", xaxt = "n", yaxt = "n",
+     ylim = c(length(nci60CellLines) + 0.5, 0.5), col = c("transparent",
+         tempSensitiveColor))
> par(plt = tempMainImageMapPanel, new = TRUE)
> image(1:max(tempDrugLastIndices), 1:length(nci60CellLines), t(sensitivityMatrix ==
+     "Resistant"), xlab = "", ylab = "", xaxt = "n", yaxt = "n",
+     ylim = c(length(nci60CellLines) + 0.5, 0.5), col = c("transparent",
+         tempResistantColor))
> box()
> abline(v = (tempDrugLastIndices[1:(tempNDrugs - 1)] + 0.5))
> tempScale = 0.75 * tempShrink
> mtext(paste(nci60CellLines, c(" "), sep = ""), side = 2, at = 1:length(nci60CellLines),
+     las = 2, cex = tempScale)
```

```
> tempDrugNames <- names(cellLinesUsed)
> tempCap <- toupper(substr(tempDrugNames, 1, 1))
> substr(tempDrugNames, 1, 1) <- tempCap
> tempNameLocs <- tempDrugCenters
> tempNameHeights <- c(0.3, 0.3, 0.3, 0.3, 0.3, 1.1, 0.3, 1.1,
+     1.9, 2.7) * tempShrink
> tempNameAdj <- c(rep(0.5, 7), rep(1, 3))
> mtext(tempDrugNames, side = 3, at = tempNameLocs, line = tempNameHeights,
+     adj = tempNameAdj, cex = tempScale)
> axis(side = 3, at = tempDrugCenters[8] + 0.2, labels = "", tcl = -1.6 *
+     tempShrink)
> axis(side = 3, at = tempDrugCenters[9] + 0.2, labels = "", tcl = -2.4 *
+     tempShrink)
> axis(side = 3, at = tempDrugCenters[10] + 0.2, labels = "", tcl = -3.2 *
+     tempShrink)
> axis(side = 2, at = (length(nci60CellLines) + 1)/2, labels = "NCI-60 Cell Lines",
+     line = 8, tick = FALSE, cex = tempShrink)
> axis(side = 3, at = (max(tempDrugLastIndices) + 1)/2, labels = "Drug",
+     line = 1, tick = FALSE, cex = tempShrink)
```

Now we shift to the panel showing the information sources, which is a simple indicator matrix – every column should have exactly one information source.

```
> par(plt = tempSourceMapPanel, new = TRUE)
> tempSourceNames <- c("Potti Heatmap, Nov 06", "Hsu Heatmap, Oct 07",
+     "Potti Corr 1 Gene List, Nov 07", "Potti Corr 1 Doce Data, Nov 07",
+     "Potti Corr 1 Doxo Data, Nov 07", "Potti Corr 1 List, Nov 07",
+     "Bonnefoi List, Dec 07", "Salter Numbers, Apr 08", "Potti Corr 2 List, Aug 08",
+     "Bonnefoi Corr List, Sep 08", "Riedel Numbers, Oct 08", "Augustine Heatmap, Jan 09")
> image(1:max(tempDrugLastIndices), 1:length(tempSourceNames),
+     ylim = c(length(tempSourceNames) + 0.5, 0.5), t(sourceMatrix),
+     xlab = "", ylab = "", xaxt = "n", yaxt = "n", col = c("white",
+         "black"))
> box()
> abline(v = (tempDrugLastIndices[1:(tempNDrugs - 1)] + 0.5))
> abline(h = c(1:length(tempSourceNames)), col = "gray", lty = "dashed")
> mtext(paste(tempSourceNames, c(" "), sep = ""), side = 2, at = 1:length(tempSourceNames),
+     las = 2, cex = tempScale)
> axis(side = 2, at = (length(tempSourceNames) + 1)/2, labels = "Information Source",
+     line = 8, tick = FALSE, cex = tempShrink)
```

Finally, we add the image panel and clean up the interim variables.

```
> par(plt = tempLegendPanel, new = TRUE)
> plot(c(0, 1), c(0, 1), type = "n", xlab = "", ylab = "", xaxt = "n",
+     yaxt = "n")
> legend(x = "center", legend = c("Sensitive", "Resistant"), fill = c(tempSensitiveColor,
+     tempResistantColor), cex = tempShrink)
> rm(list = ls(pattern = "^temp"))
```

```
quartz
     2
```

```
quartz
     2
```

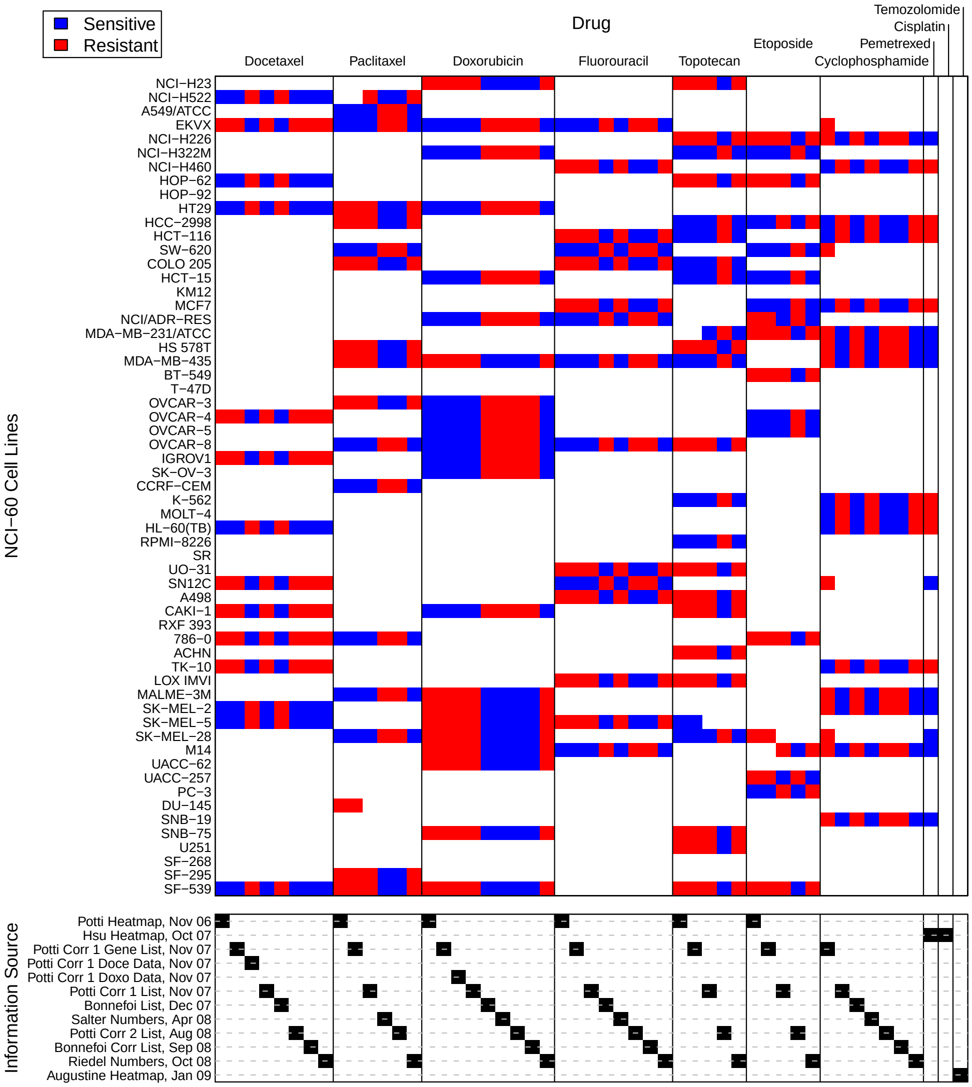

Figure 1: Mapping illustrating which NCI-60 cell lines were used as sensitive or resistant to define sensitivity signatures for which drugs, and the source of information used to make the inference. For example, there are 8 sources of information (columns) for docetaxel; in the third column, corresponding to the data table on docetaxel treated patients posted with the first correction to Potti et al. [7], cell line NCI-H522 was listed as resistant. All drug groupings with more than one column show color flips, corresponding to reversal of the sensitive/resistant labeling. The groups of cell lines are different for most drugs save cyclophosphamide and pemetrexed. For this pair the strength of the overlap suggests that the same set of cell lines was used for both drugs. No cell lines are indicated for cisplatin or temozolomide; in the first case because the signature was derived from a different set of cell lines, and in the second because the heatmap reported for temozolomide is actually the heatmap for cisplatin.

U95Av2 Probeset	Docetaxel	Paclitaxel	Doxorubicin	U133A Probeset	Cisplatin
114_r_at		MAPT		203719_at	ERCC1
1258_s_at	ERCC4	ERCC4	ERCC4	210158_at	ERCC4
1802_s_at	ERBB2	ERBB2		228131_at	ERCC1*
1847_s_at			BCL2	231971_at	FANCM*
1878_g_at	ERCC1	ERCC1			
1909_at			BCL2		
1910_s_at			BCL2		
2034_s_at			CDKN1B		
33047_at	BCL2L11	BCL2L11			
36519_at		ERCC1			
40567_at	K-ALPHA-1	K-ALPHA-1			

Table 1: "Outlier" probesets in the initially reported gene lists that are not derived from the cell line data alone for Potti et al. [7] (docetaxel, paclitaxel, doxorubicin) and Hsu et al. [6] (cisplatin). ERCC1 and ERCC4 are common. Probesets for cisplatin marked with an asterisk come from the U133B array platform, not the U133As used. The table does not include 14 outliers in the initial signature for docetaxel that are also found in the list reported by Chang et al. [3] as being effective separators in the docetaxel test set.

19 Gene List Outliers

As noted in the sections dealing with the Potti et al. [7] and Hsu et al. [6] heatmaps, the initially reported gene lists contained outliers not derivable from the cell line data alone.

20 Appendix

20.1 File Location

```
> getwd()
```

```
[1] "/Users/kabagg/ReproRsch/AnnAppStat"
```

20.2 Saves

```
> save(cellLinesUsed, nscNumbers, nci60CellLines, gyorffyCellLines,
+     informationSources, sensitivityMatrix, sourceMatrix, file = file.path("RDataObjects",
+         "cellLinesUsed.Rda"))
```

20.3 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

other attached packages:
[1] XML_2.6-0
```

References

[1] Augustine CK, Yoo JS, Potti A, et al.: Genomic and molecular profiling predicts response to temozolomide in melanoma. *Clin Cancer Res*, **15**:502-10, 2009.

[2] Bonnefoi H, Potti A, Delorenzi M, et al.: Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the EORTC 10994/BIG 00-01 clinical trial. *Lancet Oncology*, **8**:1071-8, 2007.

[3] Chang JC, Wooten EC, Tsimelzon A, et al.: Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet*, **362**:362-369 (2003).

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

[5] 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**:1699-1712, 2006

[6] 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

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

[8] Potti A, Nevins JR: Reply to Microarrays: retracing steps. *Nat Med*, **13**:1277-8, 2007.

[9] Riedel RF, Porrello A, Pontzer E, et al.: A genomic approach to identify molecular pathways associated with chemotherapy resistance. *Mol Cancer Ther*, **7(10)**:3141-9.

[10] Salter KH, Acharya CR, Walters KS, et al.: An integrated approach to the prediction of chemotherapeutic response in patients with breast cancer. *PLoS ONE*, **3**:e1908, 2008.