Checking Drug Sensitivity of Cell Lines Used in Signatures

Keith A. Baggerly

Contents

1 Executive Summary

1.1 Introduction

In late 2006, Potti et al. [8] 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 take the cell lines identified previously (enumeratingCellLines.pdf) as those used in defining a drug sensitivity signature, and examine the corresponding drug sensitivity data from the NCI to clarify how the named lines were chosen.

1.2 Methods

We acquired text describing cell line selection from the supplementary methods for Potti et al. [8]; this text was repeated essentially verbatim by Hsu et al. [7] (methods), Bonnefoi et al. [2] (webappendix), and Augustine et al. [1] (supplementary information).

We acquired the August 2008 release of GI50, TGI, and LC50 summary sensitivity data from the NCI's Developmental Therapeutics Program. We extracted and plotted the data for 10 drugs analyzed directly by Potti et al. [8], Hsu et al. [7], Bonnefoi et al. [2], and Augustine et al. [1], and for 3 drugs to which an earlier signature was applied. The 10 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 3 other drugs are daunorubicin (Dauno), epirubicin (Epi), and vincristine (Vinc), all of which were modeled in part using the signature for doxorubicin. For the 10 initial drugs, we plotted the sensitivity values superimposing markers to indicate the cell lines actually used. For the 3 other drugs, we plotted the agreement between their sensitivity values and those of doxorubicin. Finally, noting a strong overlap between the cell lines used for cyclophosphamide and pemetrexed, we plotted the sensitivity results for pemetrexed superimposing the lines used for cyclophosphamide.

1.3 Results

For 6 of the 7 drugs examined by Potti et al. [8] (D, P, A, F, T, E), there was clear separation of the sensitive and resistant lines according to TGI and LC50 but not GI50, though there were some overlaps for 2 drugs (T, E). For all 6, the orientation supplied in August 2008 with the second correction to Potti et al. [8] appears correct. There was no separation of the lines used for the last (C) by any of the metrics available. For pemetrexed, there was clear separation of the groups using GI50 values (TGI and LC50 values were uninformative), but sensitive/resistant labels were reversed and one cell line was in the wrong group. For cisplatin, Hsu et al. [7] chose to use data from Győrffy et al. [5] to assemble the signature due to some problems with the NCI60 data; we do not see these problems. For temozolomide, we do not know what cell lines were used, but the sensitivity data is so uninformative that it is not clear how sets of cell lines can be chosen.

Plotting the cell lines used for doxorubicin against the sensitivity data for (Dauno, Epi, Vinc) shows good agreement for the first two and no agreement for the third. Plotting the cell lines used for cyclophosphamide against the sensitivity data for pemetrexed shows a clear separation by GI50 values, with the correct orientation and the same single cell line outlier as for the pemetrexed lines.

1.4 Conclusions

The initially reported method of selecting cell lines indicated a primary role for GI50 values, but the GI50 values for the sensitive and resistant groups overlap for 6 of the 7 drugs examined by Potti et al. [8]. The "Description of Predictor Generation" supplied for docetaxel with the first correction to Potti et al. [8] indicates that TGI and LC50 values were used, without using GI50 values. The "Description" also includes other rules for filtering cell lines that were not applied when cell lines were selected for other drugs. By contrast, selection of lines for pemetrexed was based on GI50 values. In short, the rules used for selecting cell lines are unclear and inconsistently applied.

The fact that the orientation supplied in August 2008 with the second correction to Potti et al. [8] is correct for 6 of 7 drugs means that the annotation for 5 drugs (P, A, F, T, E) was incorrect in Potti et al. [8], for docetaxel and for doxorubicin in the explicitly labeled datasets provided for these drugs in the November 2007 first correction to Potti et al. [8], for docetaxel in Bonnefoi et al. [2], for fluorouracil and cyclophosphamide in Salter et al. [11], and for the numbers given in Riedel et al. [10]. It is unclear that the correct set of orientations was actually employed in any of the papers listed. The orientation is simply incorrect for pemetrexed in Hsu et al. [7].

The fact that daunorubicin and epirubicin exhibit similar response patterns to doxorubicin but vincristine does not is not surprising, as the first (Dauno, Epi, D) share a common mechanism of action and the last (Vinc) does not. The lack of agreement between doxorubicin and vincristine is troubling because the August 2008 second correction to Potti et al. [8] indicates that some of the samples comprising the test set for doxorubicin were chosen on the basis of their sensitivity to vincristine.

All cell lines show almost complete resistance to cyclophosphamide and temozolomide. This is not surprising; both are prodrugs which need to be processed by the body to produce their active forms, and cell lines lack the processing apparatus. Thus, the NCI60 sensitivity data cannot directly identify lines as sensitive or resistant to these drugs. The overlap between the lines used for cyclophosphamide and pemetrexed suggests that the lines used for cyclophosphamide were chosen based on the sensitivity data for pemetrexed. The initial heatmap for temozolomide is actually the heatmap for cisplatin used by Hsu et al. [7], suggesting that a similar mixup happened there.

Selected points are illustrated in Figure 1.

2 Options and Libraries

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

3 Descriptions of Predictor Generation

The descriptions of how to identify the cell lines that are sensitive and resistant to a given drug are largely parallel across the papers considered.

The supplementary methods for Potti et al. [8] state

> NCI-60 predictors. The [-log10(M)] GI50/IC50, TGI (Total Growth Inhibition dose) and LC50 (50% cytotoxic dose) data was used to populate a matrix with MATLAB software, with the relevant expression data for the individual cell lines. Where multiple entries for a drug screen existed (by NCS number), the entry with the largest number of replicates was included. Incomplete data were assigned as Nan (not a number) for statistical purposes. To develop an in vitro gene expression based predictor of sensitivity/resistance from the pharmacologic data used in the NCI-60 drug screen studies, we chose cell lines within the NCI-60 panel that would represent the extremes of sensitivity to a given chemotherapeutic agent (mean GI50 +/- 1SD). Furthermore, since the TGI and LC50 dose also represent the cytostatic and cytotoxic levels of any given drug; the log transformed TGI and LC50 dose of the sensitive and resistant subsets was then correlated with the respective GI50 data to ascertain consistency between the TGI, LC50 and GI50 data. Cell lines with low GI50 ($<$ 1SD of mean) also needed to have a low LC50, and TGI concentration to be considered sensitive. Likewise, those with the highest GI50 ($>$ 1SD of mean), TGI and LC50 concentration for a given drug, were considered resistant. Our hypothesis was that such a rigorous selection would identify cell lines that represent the extremes of sensitivity to a given drug.

The methods section for Hsu et al. [7] states

> The [log10(M)] GI50/50% growth inhibitory dose (IC50), total growth inhibition (TGI), and 50% lethal concentrate (LC50) data on the NCI-60 cell line panel for pemetrexed was used to populate a matrix with MATLAB software (MathWorks, Natick,MA) with the relevant expression data for each individual cell line. When multiple entries for a drug screen existed (by National Service Center number), the entry with the largest number of replicates was included. To develop an in vitro gene expression-based predictor of pemetrexed sensitivity from the pharmacologic data used in the NCI-60 drug screen studies, we chose cell lines within the NCI-60 panel that would represent the extremes of sensitivity (`http://dtp.nci.nih.gov/docs/cancer/cancer_data.html`). Our hypothesis was that such a selection would identify cell lines that represent the extremes of sensitivity to a given drug. [8]

The webappendix for Bonnefoi et al. [2] states

> The negative log10 of the drug GI50 (50% growth inhibitory dose), TGI (Total Growth Inhibitory dose), and LC50 (50% cytotoxic dose) together with the relevant expression data for the individual cell lines were used to populate a matrix in MATLAB. For precise definitions of GI50, LC50, and TGI see the `http://dtp.nci.nih.gov` website. Where multiple NSC (National Service Center) entries for a drug screen existed, the entry with the largest number of replicates was included. Incomplete data were assigned Nan (not a number) for statistical purposes. Cell lines from the NCI-60 drug screen studies were chosen that would represent the extremes of sensitivity to a given chemotherapeutic agent (webpanel). The TGI and LC50 doses of the sensitive and

> resistant subsets were correlated with the respective GI50 data to ascertain consistency between the TGI, LC50, and GI50 data. Cell lines with low GI50/TGI needed to have a low LC50 to be considered sensitive. Likewise, those with the highest TGI and LC50 for a given drug were considered resistant. Our hypothesis was that such a rigorous selection would identify cell lines that represent the extremes of sensitivity to a given drug; in practice, the selected cell lines do indeed show significant differences in the means of the sensitive and resistant populations.

The methods section for Salter et al. [11] states

> The NCI-60 cell lines that were most resistant or sensitive to each chemotherapy agent were identified as previously described [8]

The supplementary methods for Augustine et al. [1] state

> Supplementary methods: Genomic predictors of sensitivity to Temozolomide. The [-log10(M)] GI50/IC50, TGI (Total Growth Inhibition dose) and LC50 (50% cytotoxic dose) data was used to populate a MATLAB software matrix with the relevant expression data for the individual cell lines. Where multiple entries for a drug screen existed (by NCS number), the entry with the largest number of replicates was included. Incomplete data were assigned as NAN (not a number) for statistical purposes. To develop an in vitro gene expression based predictor of sensitivity/resistance from the pharmacologic data used in the NCI-60 drug screen studies, we chose cell lines within the NCI-60 panel that would represent the extremes of sensitivity to temozolomide (mean GI50 +/- 1SD). Furthermore, since the TGI and LC50 dose also represent the cytostatic and cytotoxic levels of any given drug, the log transformed TGI and LC50 dose of the sensitive and resistant subsets was correlated with the respective GI50 data to ascertain consistency between the TGI, LC50 and GI50 data. For example, cell lines with low GI50 (< 1 SD of mean) also needed to have a low LC50 and TGI concentration to be considered sensitive. Likewise, those with the highest GI50 (> 1 SD of mean), TGI and LC50 concentrations for a given drug were considered resistant. Our hypothesis was that such a rigorous selection would identify cell lines that represent the extremes of sensitivity to temozolomide.

Starting from these descriptions, we assumed that the GI50 values for the most frequently tested concentration of a given drug should serve as the primary index for sorting the cell lines. In other words, we assumed that the GI50 values would define the most extreme lines, and the TGI and LC50 values would be used to confirm these selections. In particular, this implies that the GI50 values for the sensitive and resistant lines should be clearly separated.

4 Loading Data

4.1 Earlier Rda Files: The Cell Lines Used in Signatures

In enumeratingCellLines.pdf, we identified the NSC numbers (which the NCI uses for indexing) for drugs mentioned in Potti et al. [8], Hsu et al. [7], Bonnefoi et al. [2], and Augustine et al. [1] together with the lists of NCI60 cell lines used as sensitive or resistant for each drug and the sources from which such information was derived. We load that data here.

```
> rdaList <- c("cellLinesUsed")
> 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/cellLinesUsed.Rda  from cache

> 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

> nci60CellLines[1:5]

[1] "NCI-H23"   "NCI-H522"  "A549/ATCC" "EKVX"      "NCI-H226"

> names(cellLinesUsed[["docetaxel"]])

[1] "heatmapPottiNov06"        "geneListPotti06CorrNov07"
[3] "doceDataPotti06CorrNov07" "listPotti06CorrNov07"
[5] "listBonnefoiDec07"        "listPotti06CorrAug08"
[7] "listBonnefoi07CorrSep08"  "numbersRiedelOct08"

> cellLinesUsed[["docetaxel"]][["heatmapPottiNov06"]]

$Sensitive
[1] "HL-60(TB)" "SF-539"    "HT29"      "HOP-62"    "SK-MEL-2"  "SK-MEL-5"
[7] "NCI-H522"

$Resistant
[1] "EKVX"    "IGROV1"  "OVCAR-4" "786-0"   "CAKI-1"  "SN12C"   "TK-10"
```

4.2 The NCI Drug Sensitivity Summary Tables

We downloaded tables of drug sensitivity summary scores (50% growth inhibition, GI50, total growth inhibition, TGI, and 50% lethal concentration, LC50) for all drugs from the main data page for the NCI's Developmental Therapeutics Program (DTP), `http://dtp.nci.nih.gov/docs/cancer/cancer_data.html`.

We use the most recent data (the August 2008 release) here: GI50_AUG08.BIN, TGI_AUG08.BIN, and LC50_AUG08.BIN. As noted on the main data page, each of the BIN files is actually a gzipped compressed file; we used gunzip to uncompress them and then recompressed them after obtaining the relevant information. While some results may have changed slightly in the releases since the initial publication of Potti et al. [8], we are only looking for qualitative understanding of the process, for which we expect any differences from releases in 2005 onwards will be minor. We may incorporate data from other releases at some later date, so we set things up to allow for this extension.

We wrote a Perl script, "getKeyDrugs.pl", to extract the data for the drugs of interest from the larger tables. The intermediate results are stored in correspondingly named csv files; e.g., GI50_AUG08.bin.csv for GI50_AUG08.BIN. We now load all these concentration values into a list of drug sensitivities.

```
> tempGI50Files <- c("GI50_AUG08.BIN.csv")
> names(tempGI50Files) <- paste("gi50", c("aug08"), sep = ".")
> tempTGIFiles <- c("TGI_AUG08.BIN.csv")
> names(tempTGIFiles) <- paste("tgi", c("aug08"), sep = ".")
> tempLC50Files <- c("LC50_AUG08.BIN.csv")
> names(tempLC50Files) <- paste("lc50", c("aug08"), sep = ".")
> tempFiles <- c(tempGI50Files, tempTGIFiles, tempLC50Files)
> drugSensitivity <- vector("list", length(tempFiles))
> names(drugSensitivity) <- names(tempFiles)
> for (tempIndex in 1:length(tempFiles)) {
+     tempTable <- read.table(file.path("RawData", "NCI60", tempFiles[tempIndex]),
+         sep = ",", header = TRUE)
+     tempTable[, "CELL"] <- as.character(tempTable[, "CELL"])
+     tempTable[, "CELL"] <- gsub("[[:space:]]+$", "", tempTable[,
+         "CELL"])
+     drugSensitivity[[tempIndex]] <- tempTable
+ }
> rm(list = ls(pattern = "^temp"))
```

5 Exploring the Drug Sensitivity Data Format

To get a better feel for the data format, we skim the first few rows of one of the drug sensitivity files.

```
> drugSensitivity[["gi50.aug08"]][1:3, ]
```

```
    NSC CONCUNIT LCONC                PANEL      CELL PANELNBR CELLNBR NLOGGI50
1 19893       M     -4 Non-Small Cell Lung    NCI-H23        1       1    6.475
2 19893       M     -4 Non-Small Cell Lung   NCI-H522        1       3    5.138
3 19893       M     -4 Non-Small Cell Lung  A549/ATCC        1       4    6.723
  INDN TOTN           STDDEV
1    4    4 0.468
2    4    4  0.823
3    4    4   0.459
```

NSC gives the NSC number of the drug being tested; 19893 corresponds to fluorouracil.

CONCUNIT gives the units in which drug concentration is measured; M is molarity.

LCONC is the maximum log10 concentration of the drug applied in a set of dilution series measurements; 5-step dilution series are used, so a LCONC of -4 means that -8 was the weakest concentration tested. All dilution series assay results for the same drug/cell line/LCONC combination are collected in a single row. Typically, there are only a small number of distinct LCONC values, and one value will involve more assays than the others. This is because the first few series may involve finding the right "range" for seeing inhibition in the first place. Once the range has been established, further measurements use the same LCONC value.

PANEL is the textual name of the disease type to which this particular cell line belongs; e.g., "Non-Small Cell Lung", "Breast", or "Prostate".

CELL gives the name of the specific cell line being interrogated; e.g., "NCI-H23" or "NCI-H522".

PANELNBR gives the numerical encoding for the PANEL given earlier, and there is a one-to-one correspondence. The Non-Small Cell Lung PANEL has PANELNBR 1.

CELLNBR gives the numerical index of the specific cell line *within the panel*; NCI-H23 is CELLNBR 1 within the Non-Small Cell Lung PANEL. HT29 is CELLNBR 1 within the Colon PANEL.

NLOGGI50 is the average -log10 GI50 value, where the average is taken over all dilution series with the same LCONC. Because of the negative sign, a larger number indicates that the cell line is more sensitive.

INDN is the number of dilution series that have been successfully run with this drug/cell line/LCONC combination; NCI-H23 was assayed 4 times for fluourouracil as of August 2008.

TOTN is the number of times a given drug/LCONC combination was tested; for some cell lines which were not tested each time, INDN will be smaller than TOTN.

STDDEV is the standard deviation of the observed -log10 GI50 values.

6 Functions for Working with the Data

We first construct a generic function for extracting data of a given type. We then construct generic tools for working with the lists that the extraction tool provides (e.g., plotting functions).

```
> getSensitivity <- function(drugName, lconc, metric = "gi50",
+     version = "aug08", drugSensitivityList = drugSensitivity,
+     nci60CellLineNames = nci60CellLines, drugNscNumbers = nscNumbers) {
+     dataToGet <- paste(metric, version, sep = ".")
+     dataToGetIndex <- match(dataToGet, names(drugSensitivityList))
+     if (is.na(dataToGetIndex)) {
+         print(paste("Info for", metric, version, "not found"))
+         print(paste("Data is available for:", names(drugSensitivityList)))
+         return()
+     }
+     sensitivityData <- drugSensitivityList[[dataToGetIndex]]
+     drugNSC <- drugNscNumbers[drugName]
+     if (is.na(drugNSC)) {
+         print(paste("NSC number for", drugName, "not found."))
+         print(paste("NSC numbers are available for:", names(drugNscNumbers)))
```

```
+         return()
+     }
+     sensitivityData <- sensitivityData[sensitivityData[, "NSC"] ==
+         drugNSC, ]
+     validConcentrations <- unique(sensitivityData[, c("LCONC",
+         "TOTN")])
+     concentrationIndex <- match(lconc, validConcentrations[,
+         "LCONC"])
+     if (is.na(concentrationIndex)) {
+         print(paste("No data found for an LCONC of", lconc))
+         print("Valid concentrations (and numbers of tests) are:")
+         print(validConcentrations)
+         return()
+     }
+     sensitivityData <- sensitivityData[sensitivityData[, "LCONC"] ==
+         lconc, ]
+     keyRows <- match(nci60CellLineNames, sensitivityData[, "CELL"])
+     valueColumn <- grep("^NLOG", names(sensitivityData))
+     sensitivityValues <- sensitivityData[keyRows, valueColumn]
+     names(sensitivityValues) <- nci60CellLineNames
+     list(sensitivityValues = sort(sensitivityValues, na.last = TRUE),
+         drug = drugName, nsc = drugNscNumbers[drugName], metric = metric,
+         version = version, lconc = lconc, totn = validConcentrations[validConcentrations[,
+             "LCONC"] == lconc, "TOTN"])
+ }
```

Before constructing plotting functions, we first load generic "capitalization" function, supplied in the description of the "toupper" function in R.

```
> capwords <- function(s, strict = FALSE) {
+     cap <- function(s) paste(toupper(substring(s, 1, 1)), {
+         s <- substring(s, 2)
+         if (strict)
+             tolower(s)
+         else s
+     }, sep = "", collapse = " ")
+     sapply(strsplit(s, split = " "), cap, USE.NAMES = !is.null(names(s)))
+ }
```

We now construct a plotting function that can work with a single set of sensitivity data.

```
> plotSensitivity <- function(sensitivityList, sensitiveLines = NULL,
+     resistantLines = NULL, xlab = NULL, ylab = NULL, main = NULL,
+     sensitiveColor = NULL, resistantColor = NULL) {
+     grabWarning <- capture.output(attach(sensitivityList))
+     if (is.null(xlab)) {
+         xlab = "Cell Lines"
+     }
+     drawMtext <- TRUE
```

```
+     if (!(is.null(ylab))) {
+         drawMtext <- FALSE
+     }
+     if (is.null(ylab)) {
+         ylab = paste("-log10(", toupper(metric), ")", sep = "")
+     }
+     if (is.null(main)) {
+         main = paste(capwords(drug), toupper(metric), sep = " ")
+     }
+     if (is.null(sensitiveColor)) {
+         sensitiveColor = "gray"
+     }
+     if (is.null(resistantColor)) {
+         resistantColor = "red"
+     }
+     plot(sensitivityValues, xlab = xlab, ylab = ylab, main = main)
+     abline(v = which(is.na(sensitivityValues)), lty = "dashed",
+         col = "gray")
+     if (drawMtext) {
+         mtext(expression("" %->% Sensitive), side = 2, line = 3,
+             adj = 1)
+         mtext(expression(Resistant %<-% ""), side = 2, line = 3,
+             adj = 0)
+     }
+     if (!is.null(sensitiveLines)) {
+         sensitiveIndices <- match(sensitiveLines, names(sensitivityValues))
+         points(sensitiveIndices, sensitivityValues[sensitiveIndices],
+             pch = 21, cex = 2, bg = "gray")
+     }
+     if (!is.null(resistantLines)) {
+         resistantIndices <- match(resistantLines, names(sensitivityValues))
+         points(resistantIndices, sensitivityValues[resistantIndices],
+             pch = 24, cex = 2, bg = "red")
+     }
+     detach(sensitivityList)
+ }
```

We also add one more function for plotting two sets of sensitivity data against each other.

```
> plotTwoSensitivities <-
+   function(xSensitivityList,
+            ySensitivityList,
+            xlab=NULL, ylab=NULL,
+            sensitiveLines=NULL, resistantLines=NULL){
+
+     drawXMtext <- TRUE
+     if(!(is.null(xlab))){
+       drawXMtext <- FALSE
+     }
```

```
+     if(is.null(xlab)){
+       xlab=paste(capwords(xSensitivityList$drug), ", ",
+         paste("-log10(",toupper(xSensitivityList$metric),")",sep=""),
+         sep="")
+     }
+     drawYMtext <- TRUE
+     if(!(is.null(ylab))){
+       drawYMtext <- FALSE
+     }
+     if(is.null(ylab)){
+       ylab=paste(capwords(ySensitivityList$drug), ", ",
+         paste("-log10(",toupper(ySensitivityList$metric),")",sep=""),
+         sep="")
+     }
+     plot(xSensitivityList$sensitivityValues,
+          ySensitivityList$sensitivityValues[
+              names(xSensitivityList$sensitivityValues)],
+          xlab=xlab,
+          ylab=ylab)
+     if(drawXMtext){
+       mtext(expression("" %->% Sensitive), side=1, line=3, adj=1)
+       mtext(expression(Resistant %<-% ""), side=1, line=3, adj=0)
+     }
+     if(drawYMtext){
+       mtext(expression("" %->% Sensitive), side=2, line=3, adj=1)
+       mtext(expression(Resistant %<-% ""), side=2, line=3, adj=0)
+     }
+     if(!is.null(sensitiveLines)){
+       points(xSensitivityList$sensitivityValues[sensitiveLines],
+              ySensitivityList$sensitivityValues[sensitiveLines],
+              pch=21,cex=2,bg="gray")
+     }
+     if(!is.null(resistantLines)){
+       points(xSensitivityList$sensitivityValues[resistantLines],
+              ySensitivityList$sensitivityValues[resistantLines],
+              pch=24,cex=2,bg="red")
+     }
+   }
```

7 Docetaxel

We begin by examining the sensitivity data for docetaxel. Checking the list of information sources for this drug,

```
> names(cellLinesUsed[["docetaxel"]])

[1] "heatmapPottiNov06"        "geneListPotti06CorrNov07"
[3] "doceDataPotti06CorrNov07" "listPotti06CorrNov07"
```

```
[5] "listBonnefoiDec07"        "listPotti06CorrAug08"
[7] "listBonnefoi07CorrSep08"  "numbersRiedelOct08"
```

we see that there were 8 sources. However, looking at the plot in enumeratingCellLines.pdf shows that the two groups of cell lines being contrasted were always the same; it was merely the sensitive/resistant direction that would occasionally change. We will work with the lists oriented per the list supplied by Potti et al. [8] for their second correction in August of 2008. This agrees with what what was given in the detailed "Description of Predictor Generation" they supplied for this drug.

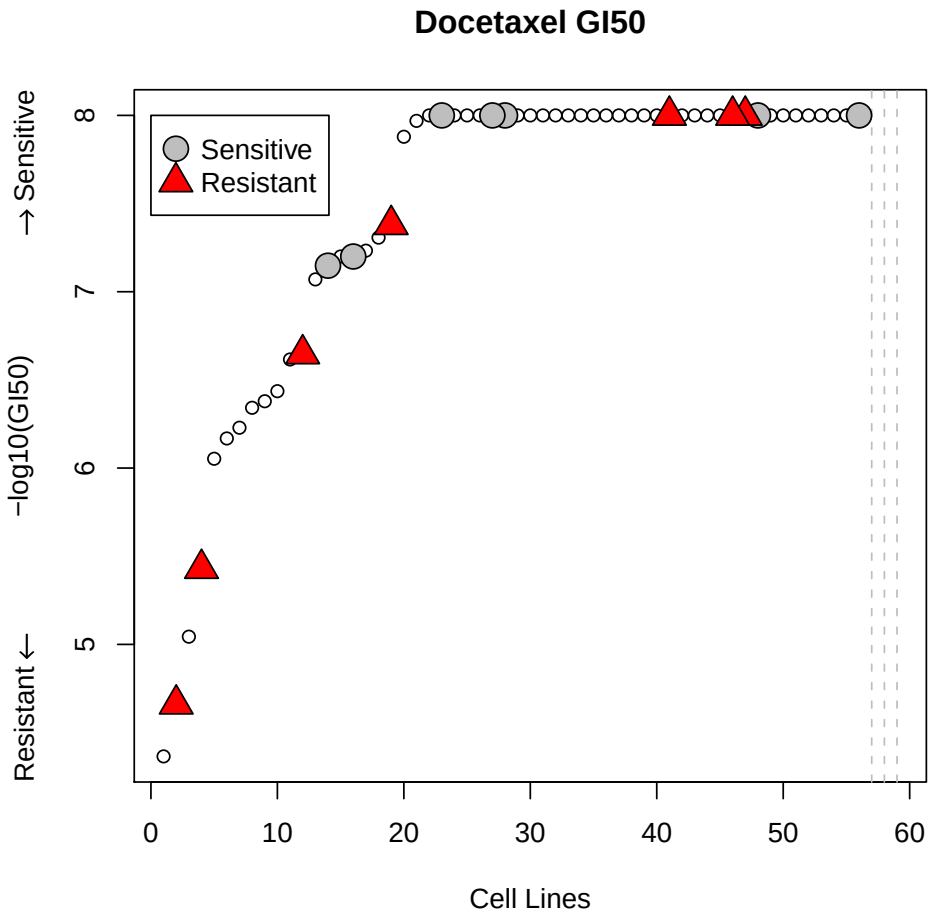
Before extracting the data, we need to decide which set of concentration values to work with. We can check the options available by simply trying to extract results for an unused concentration, e.g. 0.

```
> getSensitivity("docetaxel", 0, "gi50")
```

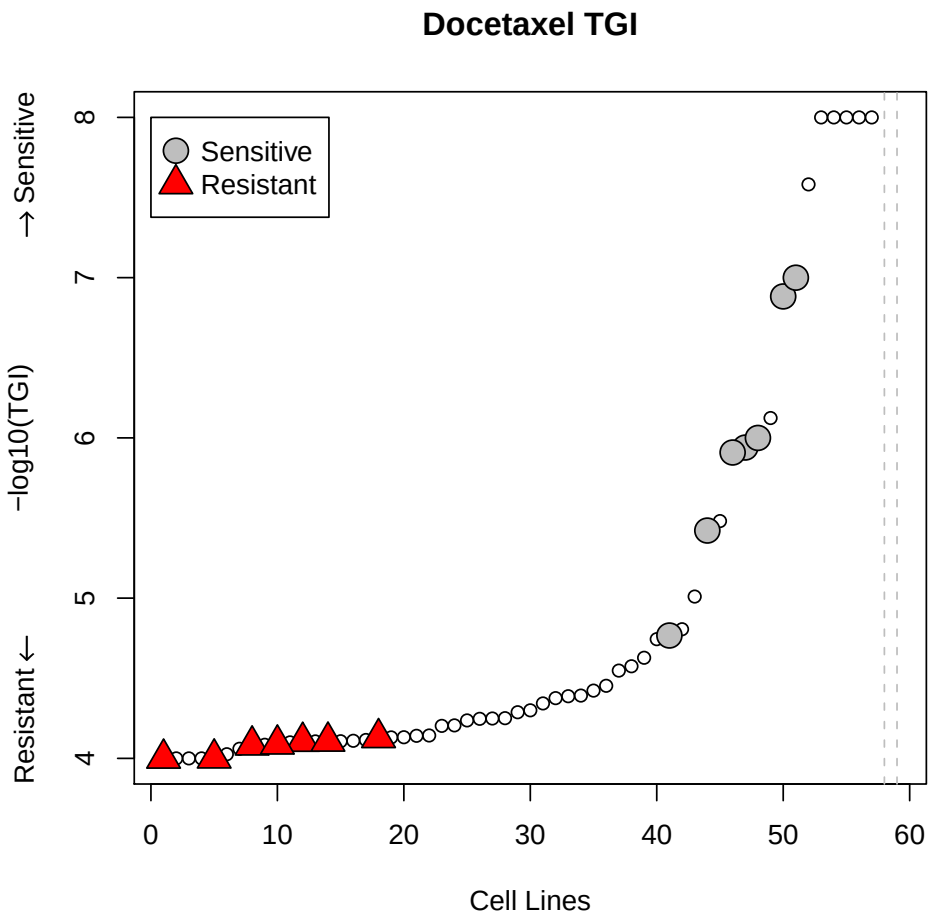
```
[1] "No data found for an LCONC of 0"
[1] "Valid concentrations (and numbers of tests) are:"
     LCONC TOTN
2143    -7    1
2200    -6    2
2268    -4    5
NULL
```

For docetaxel, we want to work with the data associated with the -4 LCONC value, since that has the largest number of associated assays (5 as of August 2008). We begin by plotting the GI50 values.

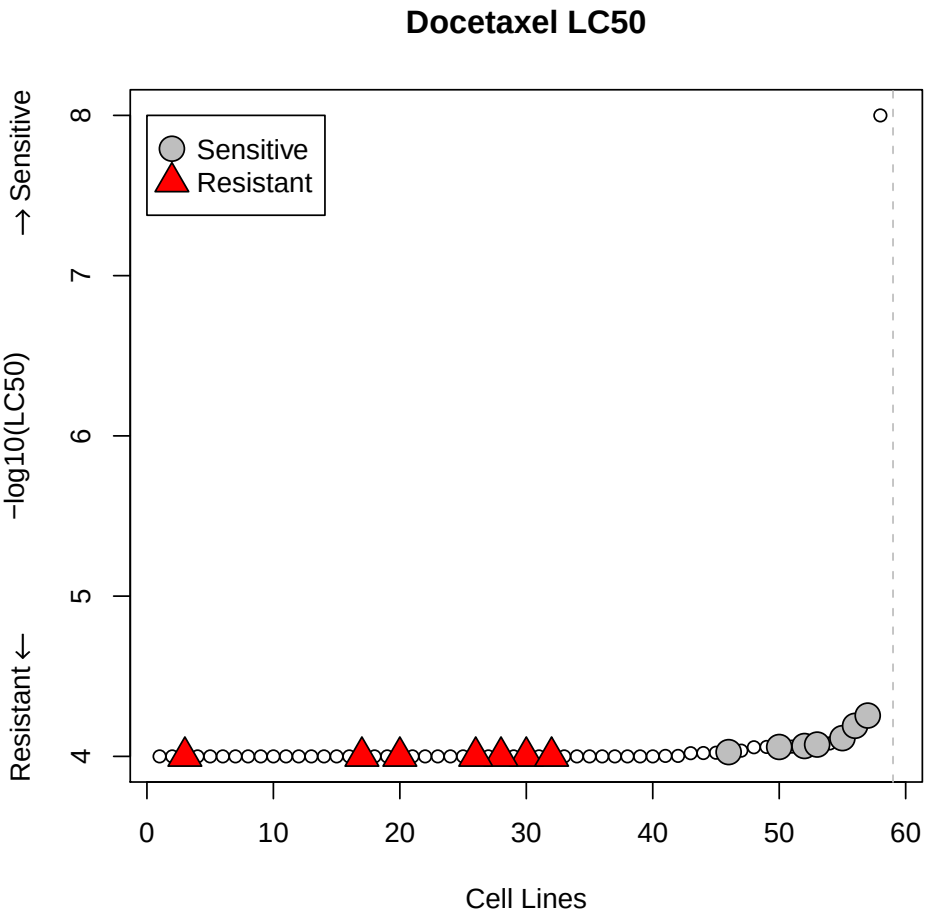
```
> temp <- getSensitivity("docetaxel",-4,"gi50")
> plotSensitivity(temp,
+     sensitiveLines=
+         cellLinesUsed[["docetaxel"]][["listPotti06CorrAug08"]][["Sensitive"]],
+     resistantLines=
+         cellLinesUsed[["docetaxel"]][["listPotti06CorrAug08"]][["Resistant"]])
> legend(x=0,y=8,legend=c("Sensitive","Resistant"),
+        pch=c(21,24),pt.bg=c("gray","red"),pt.cex=2)
```



While the orientation appears correct (the most resistant of the lines labeled "Resistant" are more resistant than the "Sensitive" lines), the GI50 plot still shows a problem with the description provided in the supplementary methods for Potti et al. [8]. The GI50 values for the sensitive and resistant lines overlap, which makes no sense if these values are used as primary selectors. The "Description of Predictor Generation", by contrast, ignores the GI50 values in favor of the TGI and LC50 values, so we plot those next.

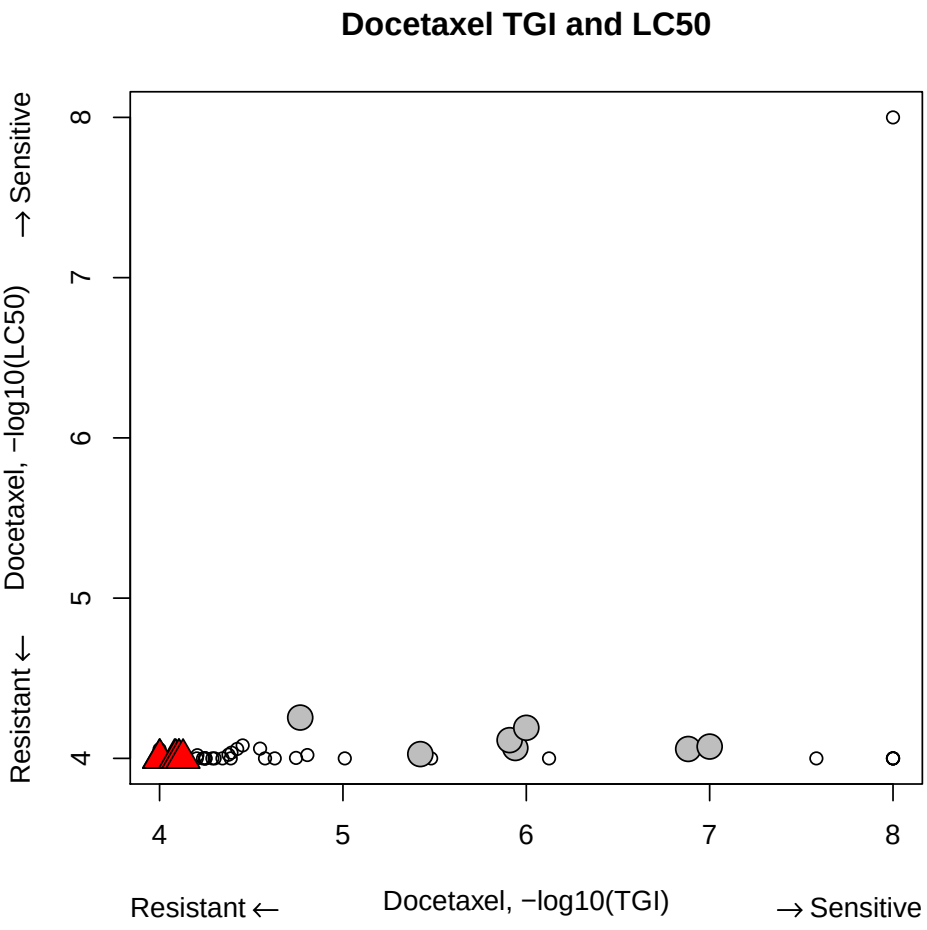


Using TGI, the story is much cleaner. The sensitive and resistant lines clearly separate, and the orientation is correct (resistant lines are at the bottom). However, the "most sensitive" lines by this metric (those at the top of the plot) are not included.



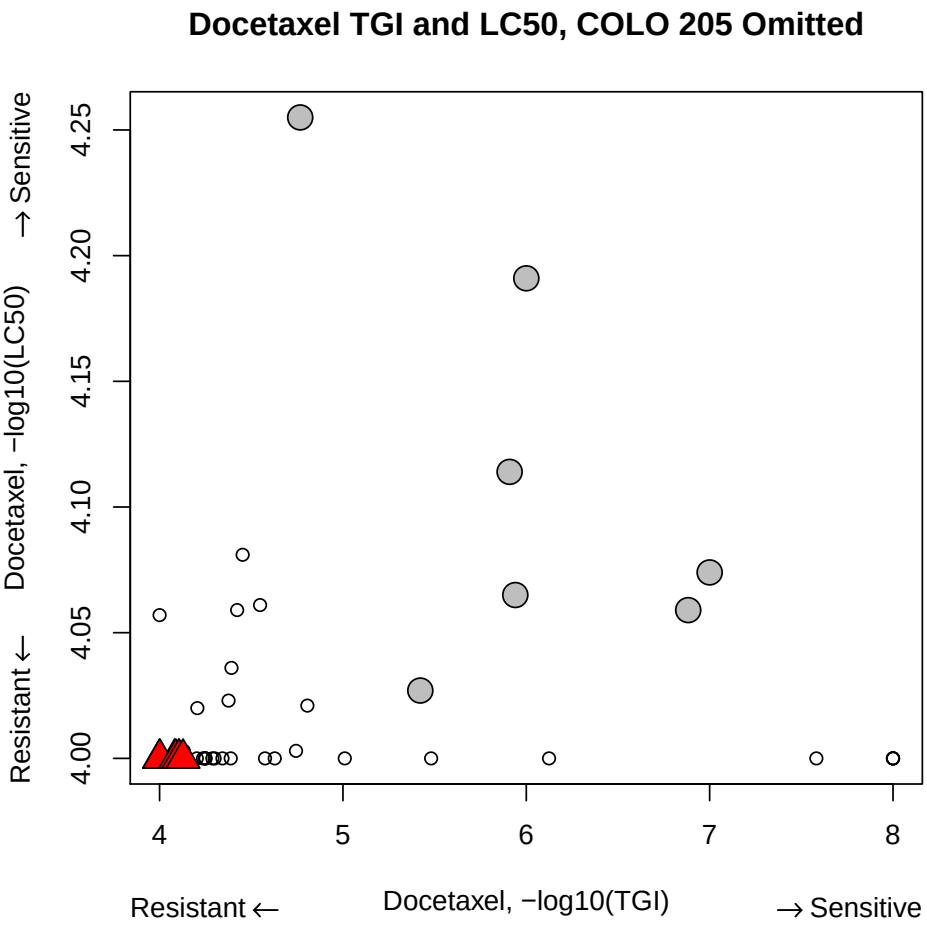
Using LC50, the lines still separate, but one dramatic feature of the plot is the extremely sensitive outlier (COLO 205) which is not part of the sensitive list. As their "Description" notes, this cell line was left off of their list because it was too sensitive. However, no algorithmic guidance for exclusion of other "tail" cases is given. Since it is not immediately clear from this plot alone how the specific sensitive and resistant lines were chosen from among the set available, we next look at a joint plot of the TGI and LC50 values.

```
> tempTGI  <- getSensitivity("docetaxel",-4,"tgi")
> tempLC50 <- getSensitivity("docetaxel",-4,"lc50")
> plotTwoSensitivities(tempTGI, tempLC50,
+   sensitiveLines=
+     cellLinesUsed[["docetaxel"]][["listPotti06CorrAug08"]][["Sensitive"]],
+   resistantLines=
+     cellLinesUsed[["docetaxel"]][["listPotti06CorrAug08"]][["Resistant"]])
> title(main="Docetaxel TGI and LC50")
```



While the division into groups is a bit clearer, the COLO 205 outlier is even more stark, and the decision to exclude it seems even odder. To focus our attention more, let's look at the same plot after excluding COLO 205.

```
> tempLC50[["sensitivityValues"]][["COLO 205"]] <- NA
```



Here, the sensitive and resistant lines separate in the way that we would wish, though the amount of separation along the LC50 axis is quite small. There is also some overplotting at the most extreme resistance levels in the lower left corner, so it's worth looking at the specific cell lines involved.

```
> tempDoce <-
+   data.frame(tgi=tempTGI$sensitivityValues,
+              lc50=tempLC50$sensitivityValues[
+                       names(tempTGI$sensitivityValues)],
+              status=sensitivityMatrix[
+                       names(tempTGI$sensitivityValues),6])
> rownames(tempDoce) <- names(tempTGI$sensitivityValues)
> tempDoce[1:10,]

            tgi  lc50    status
EKVX      4.000 4.000 Resistant
HCT-116   4.000 4.057  Not Used
```

```
PC-3      4.000 4.000  Not Used
SNB-19    4.000 4.000  Not Used
786-0     4.001 4.000 Resistant
NCI-H322M 4.026 4.000  Not Used
OVCAR-8   4.060 4.000  Not Used
IGROV1    4.082 4.000 Resistant
A549/ATCC 4.085 4.000  Not Used
CAKI-1    4.088 4.000 Resistant
```

What this shows is that while the resistant lines do indeed appear resistant, several lines that were apparently just as resistant on these scales were not used. To understand this exclusion, we looked more carefully at the "Description of Predictor Generation". They note that

> In addition, the NCI provides the above concentrations as a calculated -log data for convenience and cursory evaluations. This form of data is however truncated at times, if the concentrations are large. The NCI also provides the raw data on the various replicate experiments performed (sometimes at different time points, three different experiments for docetaxel). We felt that only those experiments that were reproducible should be used in determining drug sensitivity, by cell line. By using this generic criteria, and matching up exactly with cell lines that have corresponding Affymetrix expression data (Novartis A dataset, not all cell lines valuated in cell proliferation studies have expression data), 49 cell lines were then identified to be used for further analyses.

In passing, we note that there were actually eight, not three, experiments run for docetaxel; there were three different maximum concentrations at which these assays were performed. Unfortunately, without further clarification as to how "those experiments that were reproducible" were identified, we do not know how to apply this filtration step ourselves.

They further note that

> We then go on to exclude "duplicates" of the same parent cell line to avoid over fitting by including the same cell lines in to a resistant or sensitive category, and especially if they had different inhibitory concentrations (in the case of docetaxel, it was only two cell lines - OVCAR-3 and SK-OV-3 - this was also confirmed recently by John Weinstein when he gave us the U133A data).

However, examination of the cell lines used in other predictor signatures shows that both OVCAR-3 and SK-OV-3 are included in the "resistant" group for doxorubicin, so this rule was not always followed.

In short, there does appear to be some separation of the lines used using TGI and LC50, but the methods for selecting the precise lines used appear to be idiosyncratic and inconsistently applied. Further, the orientation used here (which is correct) has not always been applied. In particular, an incorrect orientation was explicitly supplied with the dataset posted for docetaxel in November 2007 to accompany the first correction of Potti et al. [8]

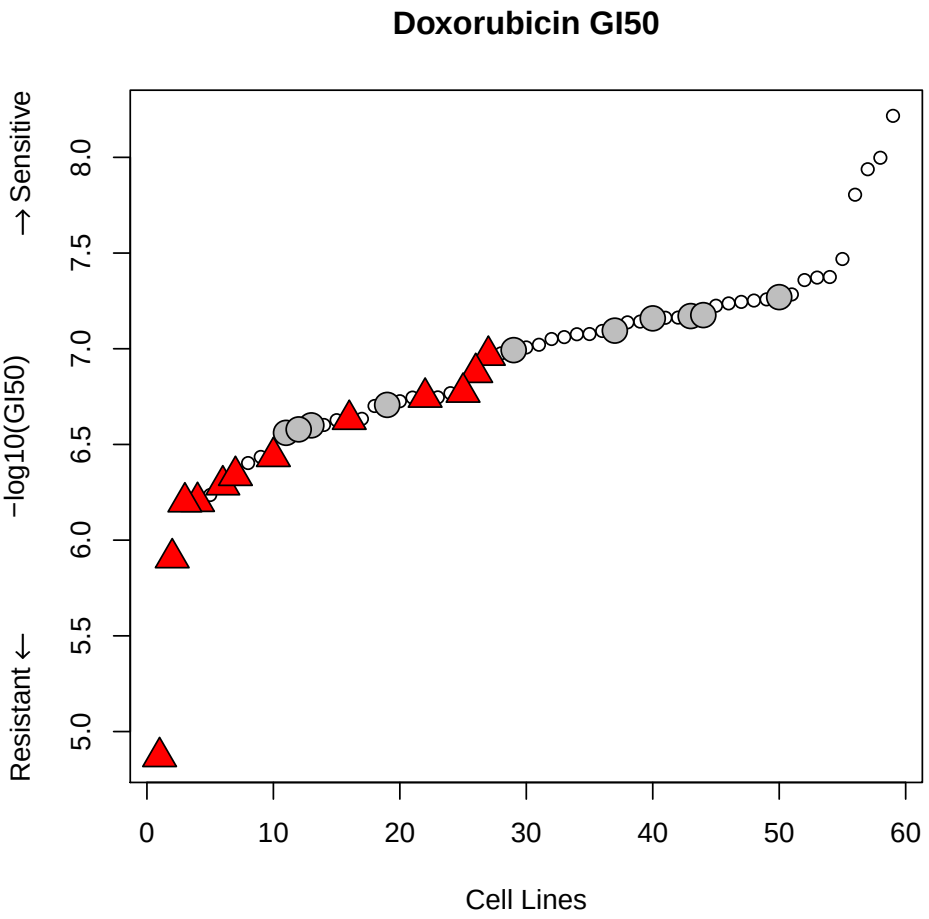
8 Doxorubicin

We next shift to examination of the sensitivity values for doxorubicin. As before, we look for the most common concentration before extracting values.

```
> temp <- getSensitivity("doxorubicin", 0, "tgi")
```

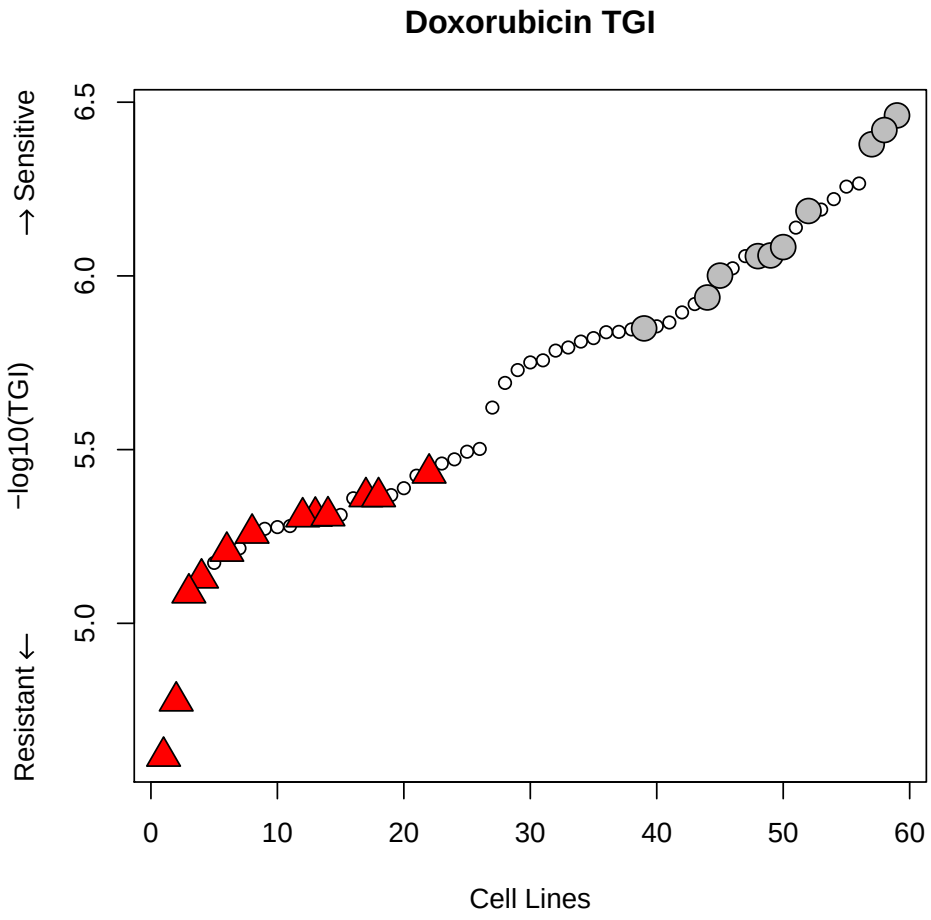
```
[1] "No data found for an LCONC of 0"
[1] "Valid concentrations (and numbers of tests) are:"
     LCONC TOTN
1042  -6.6    1
1097  -4.6 2008
1191  -4.0    2
```

The -4.6 LCONC value should be used. We now look at the GI50 values, working with the lists oriented per the list supplied by Potti et al. [8] for their second correction in August of 2008 as above. As with docetaxel, the cell lines used for doxorubicin have remained consistent; only the sensitive/resistant orientation has changed.

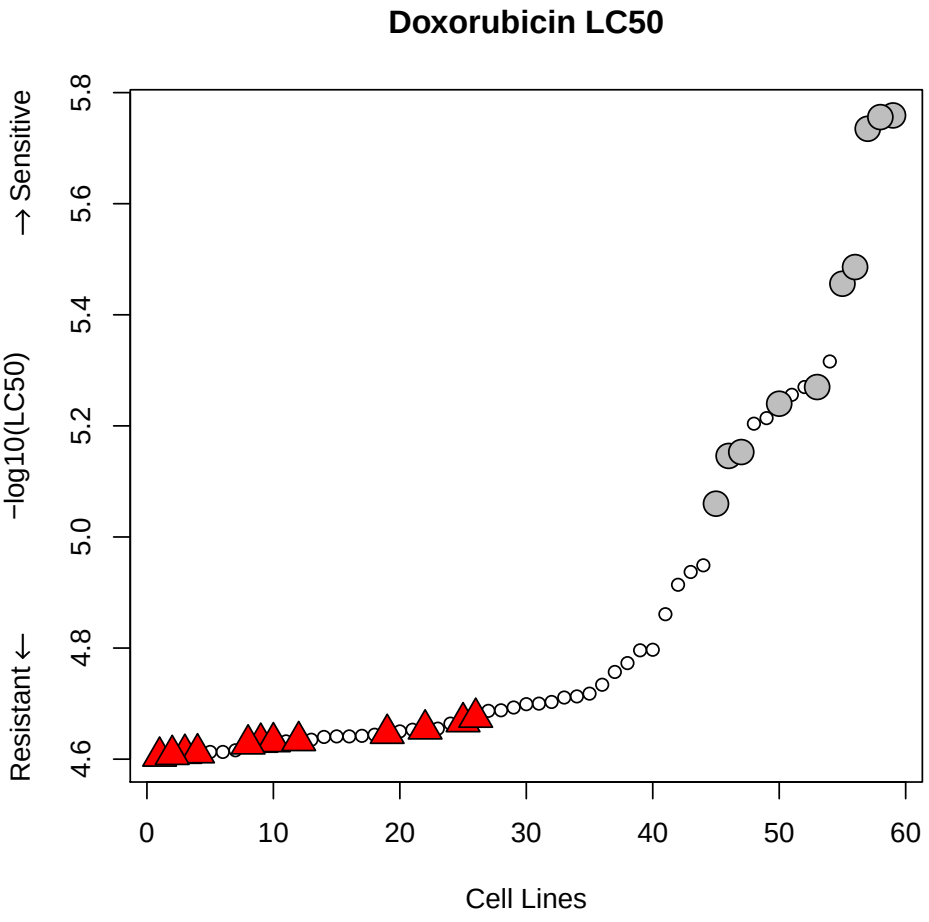


While the most resistant cell line is indeed in the "resistant" group (suggesting that this orientation is correct), the GI50 values for the sensitive and resistant groups still overlap.

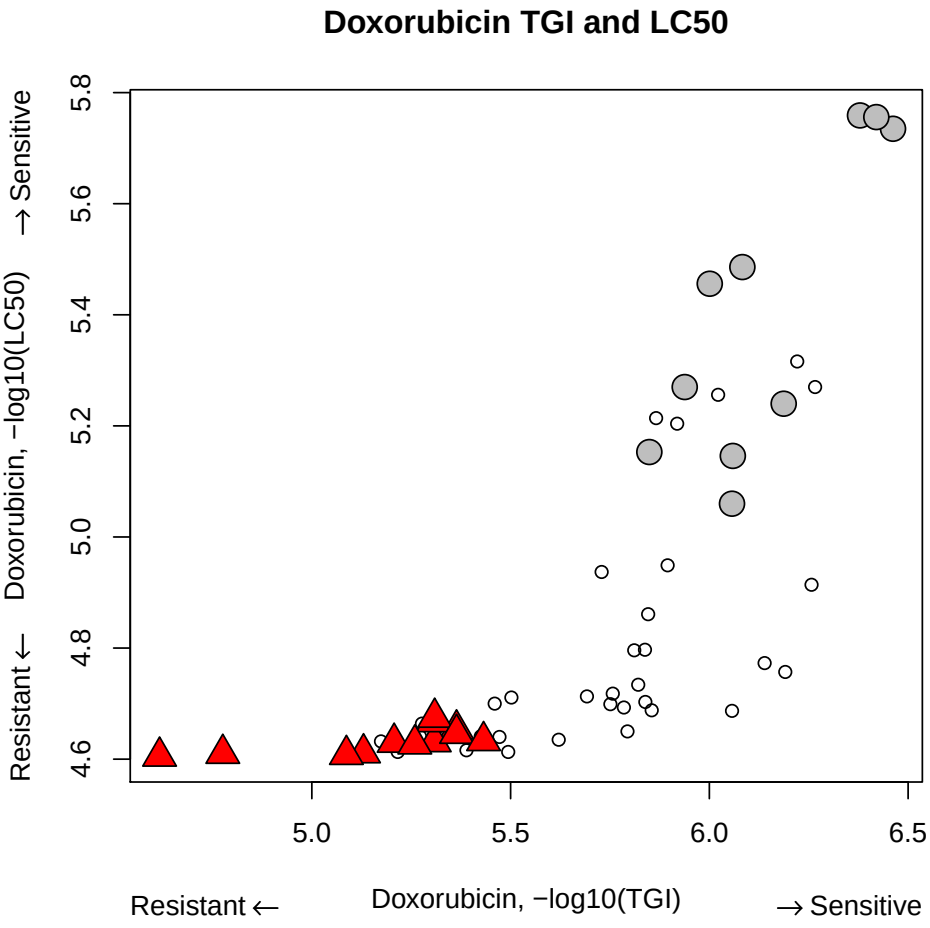
We next look at TGI.



Using this metric, the separation is much clearer, and the orientation is clearly correct. We next look at LC50.



Again, the separation is stark. Finally, we look at TGI and LC50 jointly.



Separation appears pretty good using a bivariate TGI/LC50 plot. However, as with docetaxel, the lines selected do not appear to include all of the most extreme ones overall.

While the orientation is correct, we note that this conflicts with the orientation initially asserted for the Potti et al. [8] heatmaps and with the orientation supplied with the doxorubicin dataset posted in November of 2007 as part of the first correction to Potti et al. [8]

9 Paclitaxel

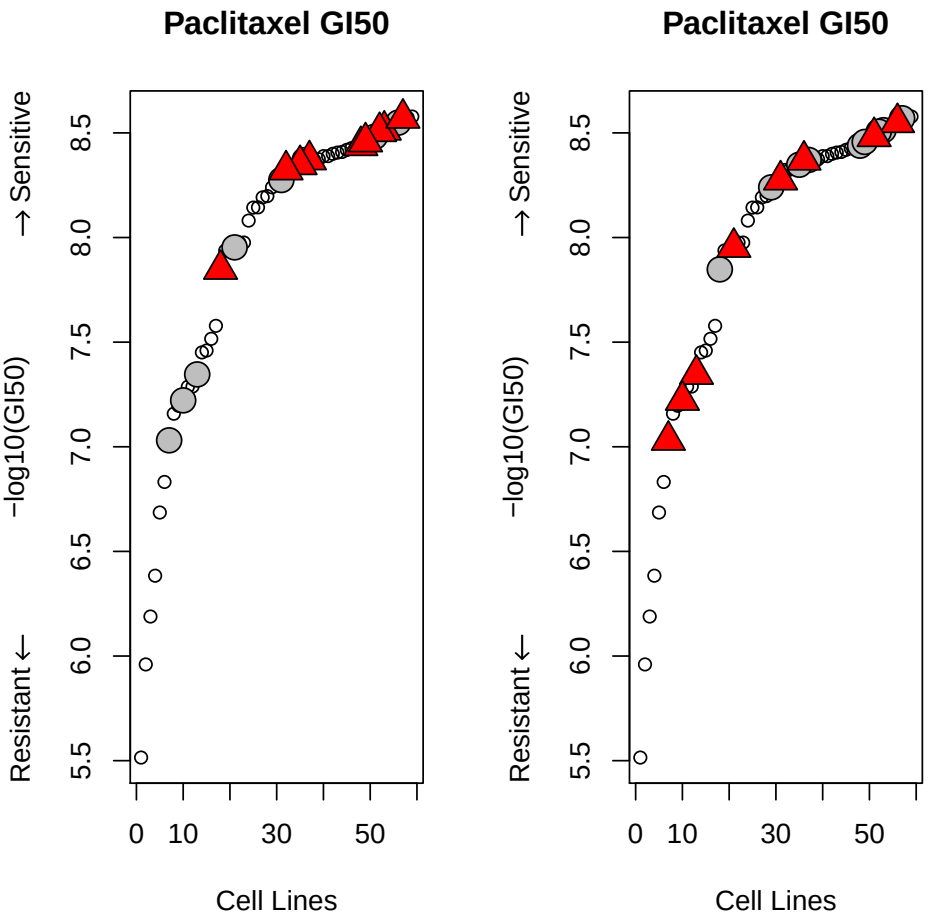
We turn next to paclitaxel. Here, there was a shift between the cell lines used for the initial Potti et al. [8] heatmap and the cell line lists reported in the 2007 and 2008 corrections, so we will look at both sets. Again, we first identify the LCONC value to focus on.

```
> temp <- getSensitivity("paclitaxel", 0, "gi50")
```

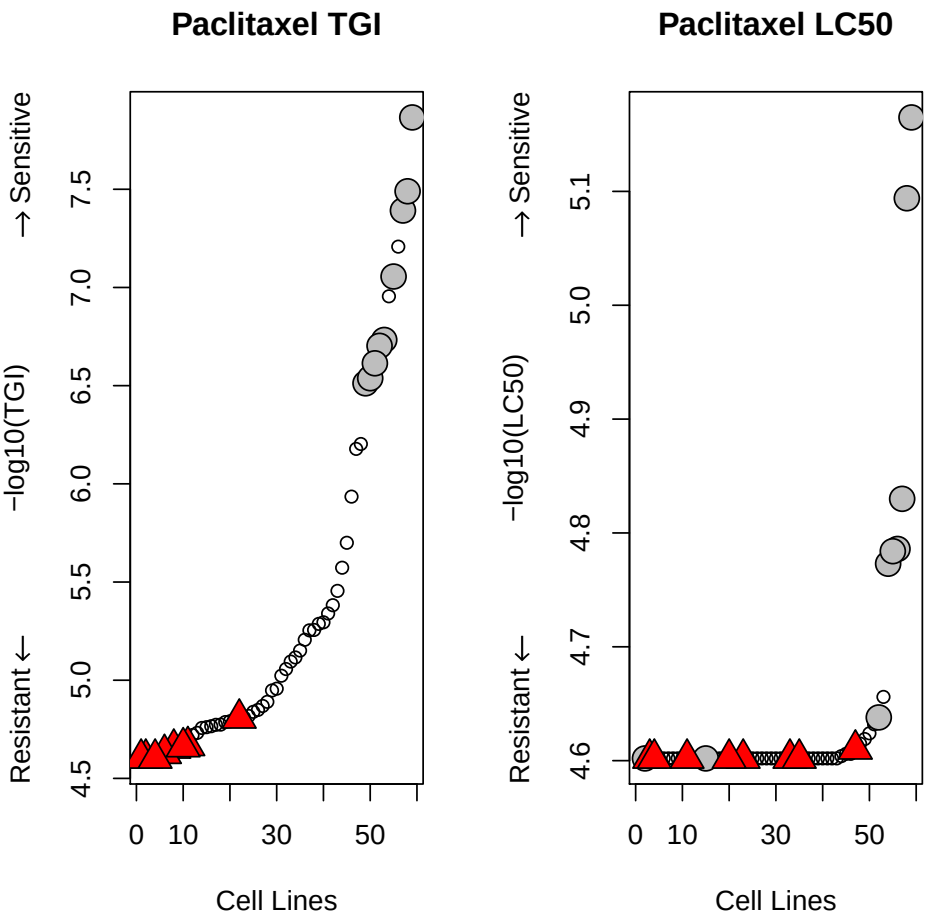
```
[1] "No data found for an LCONC of 0"
[1] "Valid concentrations (and numbers of tests) are:"
```

```
     LCONC TOTN
1272  -6.0   22
1359  -5.0    1
1413  -4.6   33
1489  -4.0    9
```

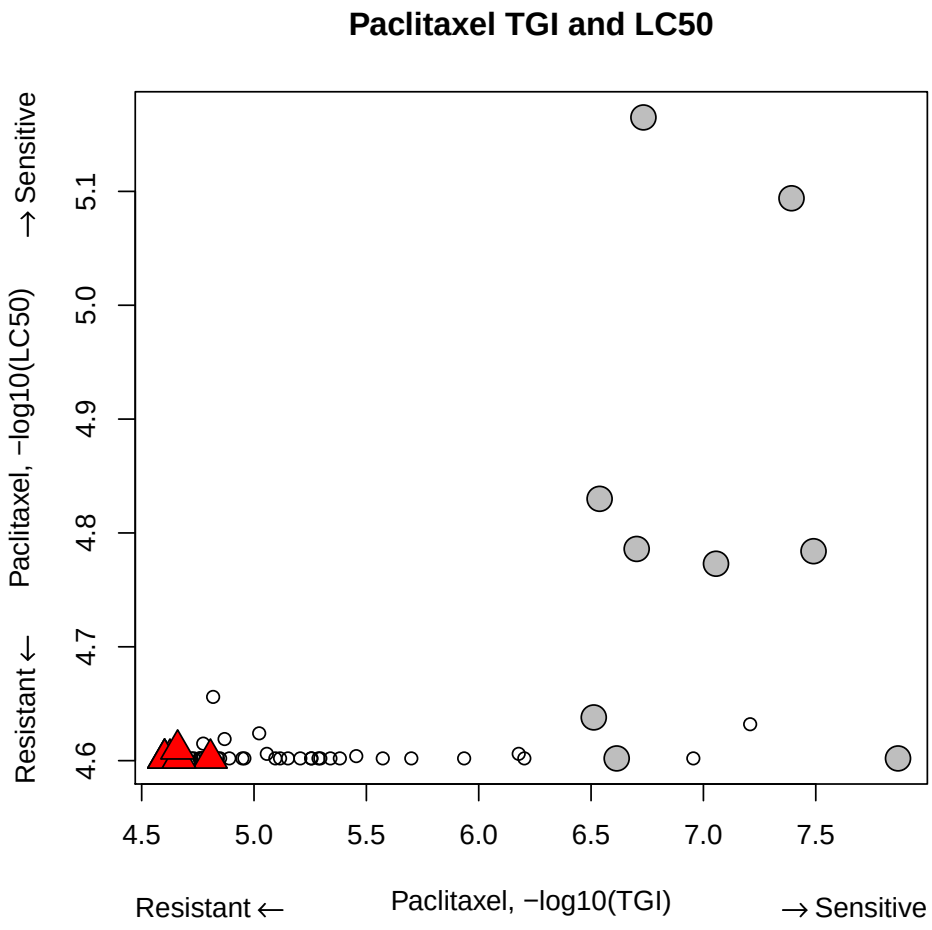
The -4.6 concentration is the one we use. We check the GI50 values first, using both the ordering and cell lines from the initial heatmaps of Potti et al. [8] (at left) and those from the second correction as of August 2008 (at right).



The latter orientation appears more correct, and the shift of one cell line does not appear to alter the story. The values for the sensitive and resistant cell lines still overlap, and the most extreme cell lines do not appear to have been chosen. We look at the TGI and LC50 values next, using just the August 2008 ordering of lines.



The separation is decidedly cleanest using TGI. We now plot TGI against LC50.



Separation appears pretty good using a bivariate TGI/LC50 plot, though again some of the most extreme lines do not appear to have been included.

In sum, the August 2008 orientation of cell lines appears correct, though this conflicts with the annotation reported for the heatmaps of Potti et al. [8]. The separation uses TGI and LC50 values, and GI50 values have apparently not been used.

10 Fluorouracil

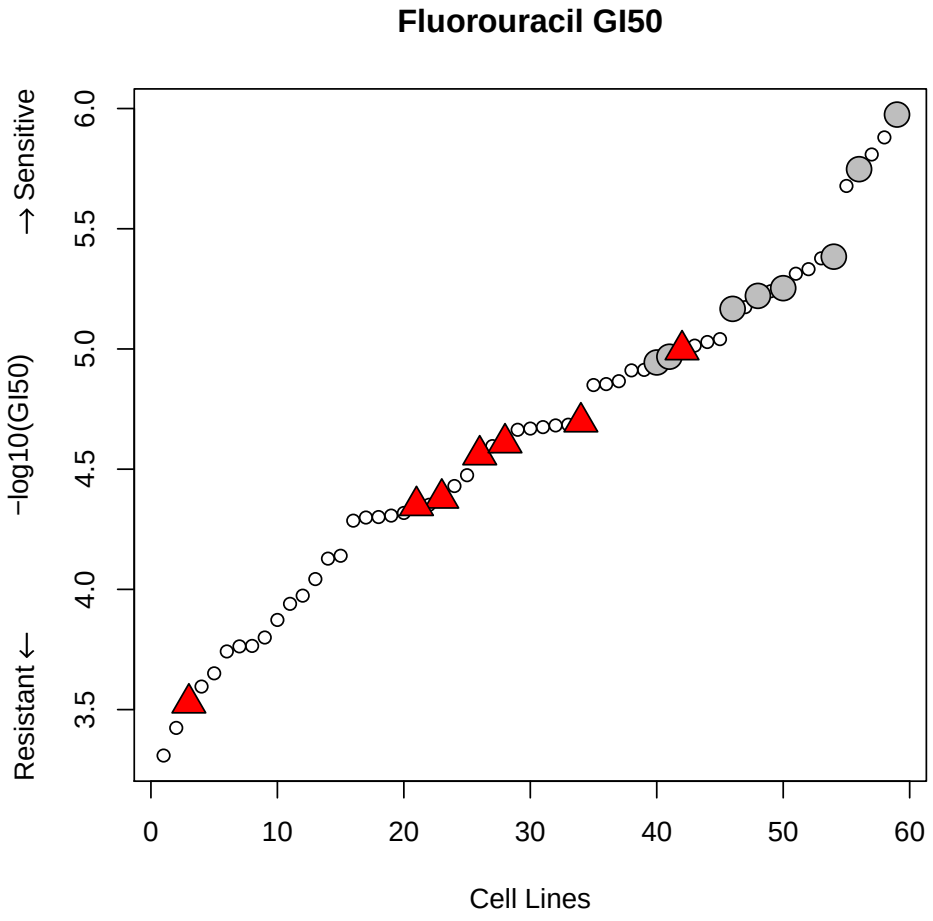
We turn next to fluorouracil. The sets of cell lines used are consistent across the sources of information we have surveyed; only the sensitive/resistant orientation has changed. Thus, we use the August 2008 groupings as above. We begin by checking the LCONC values available.

```
> temp <- getSensitivity("fluorouracil", 0, "gi50")
```

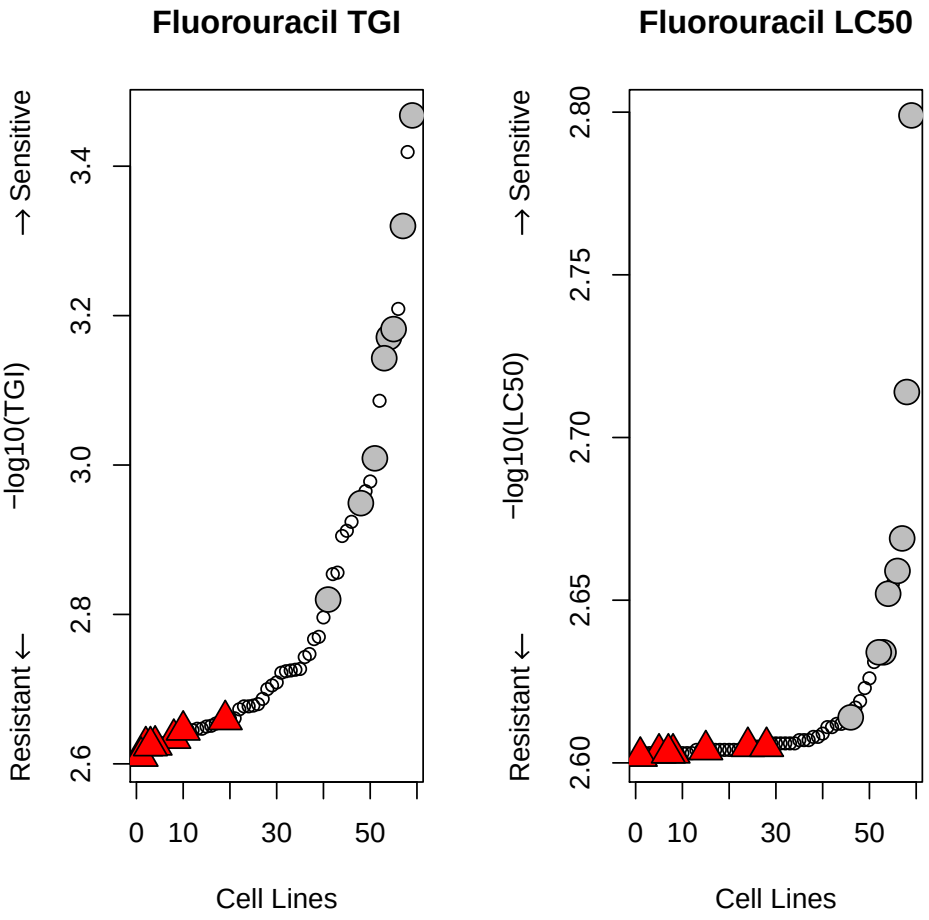
```
[1] "No data found for an LCONC of 0"
[1] "Valid concentrations (and numbers of tests) are:"
```

```
   LCONC TOTN
1   -4.0    4
60  -2.6 1889
```

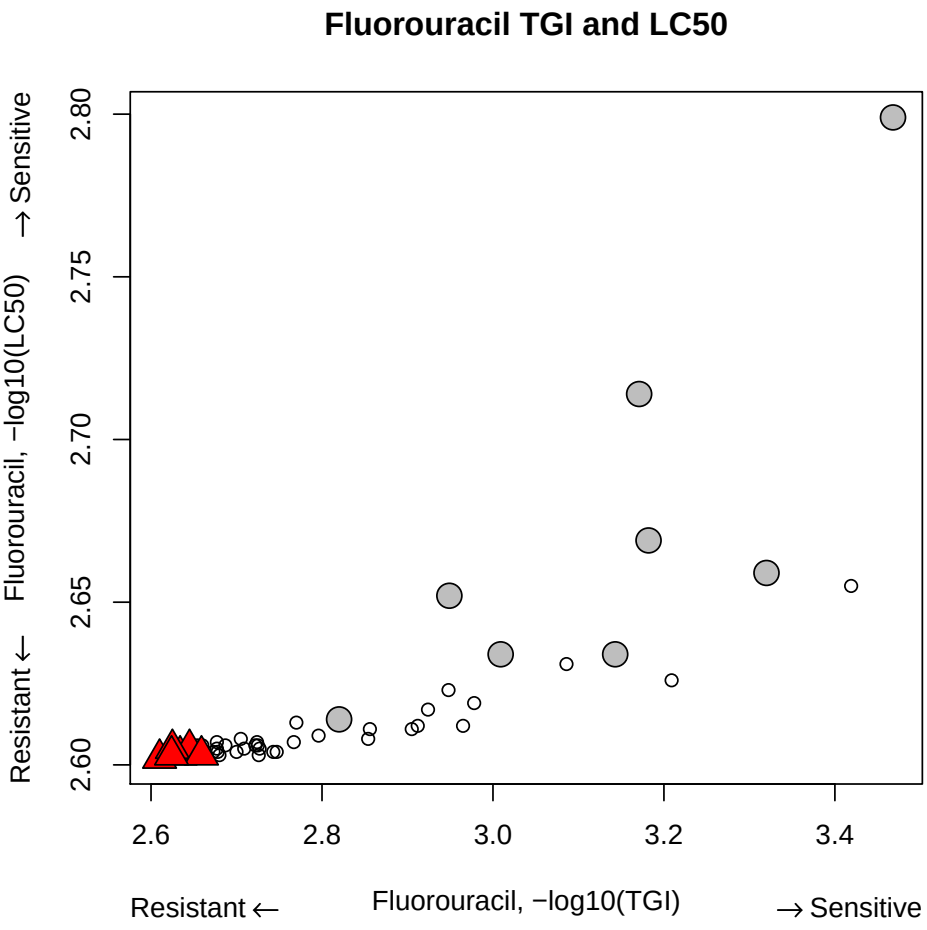
The -2.6 concentration is the clear winner, so that's what we'll use. We check the GI50 values first.



The orientation appears correct, but again the sensitive and resistant values overlap. We check the TGI and LC50 values next.



The orientation appears correct, though the range of LC50 values is very narrow. We next plot the TGI and LC50 values jointly.



Separation appears pretty good using a bivariate TGI/LC50 plot.

In sum, the August 2008 orientation of cell lines appears correct, though this contradicts the heatmaps from Potti et al. [8] and the orientation used in Salter et al. [11]. Again, GI50 values have apparently not been used.

11 Topotecan

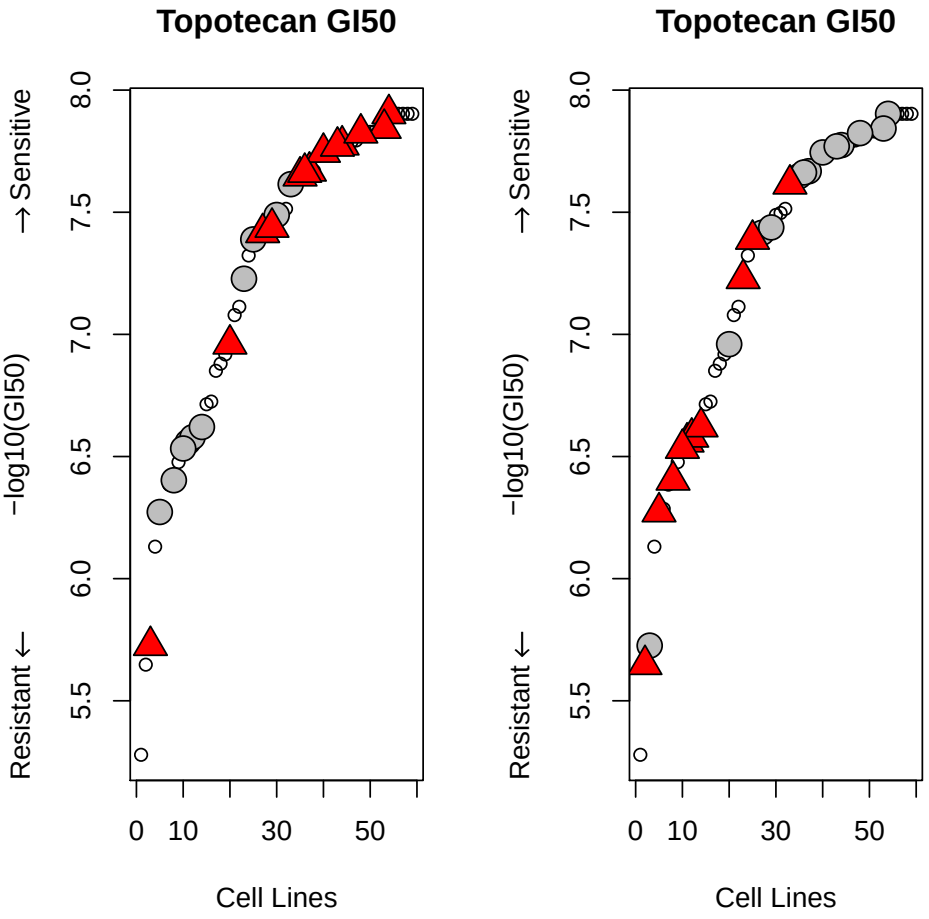
We turn next to topotecan. Here, there was a shift between the cell lines used for the initial Potti et al. [8] heatmap and the cell line lists reported in the 2007 and 2008 corrections, so we will look at both sets of lines. Again, we first identify the LCONC value to focus on.

```
> temp <- getSensitivity("topotecan", 0, "gi50")
```

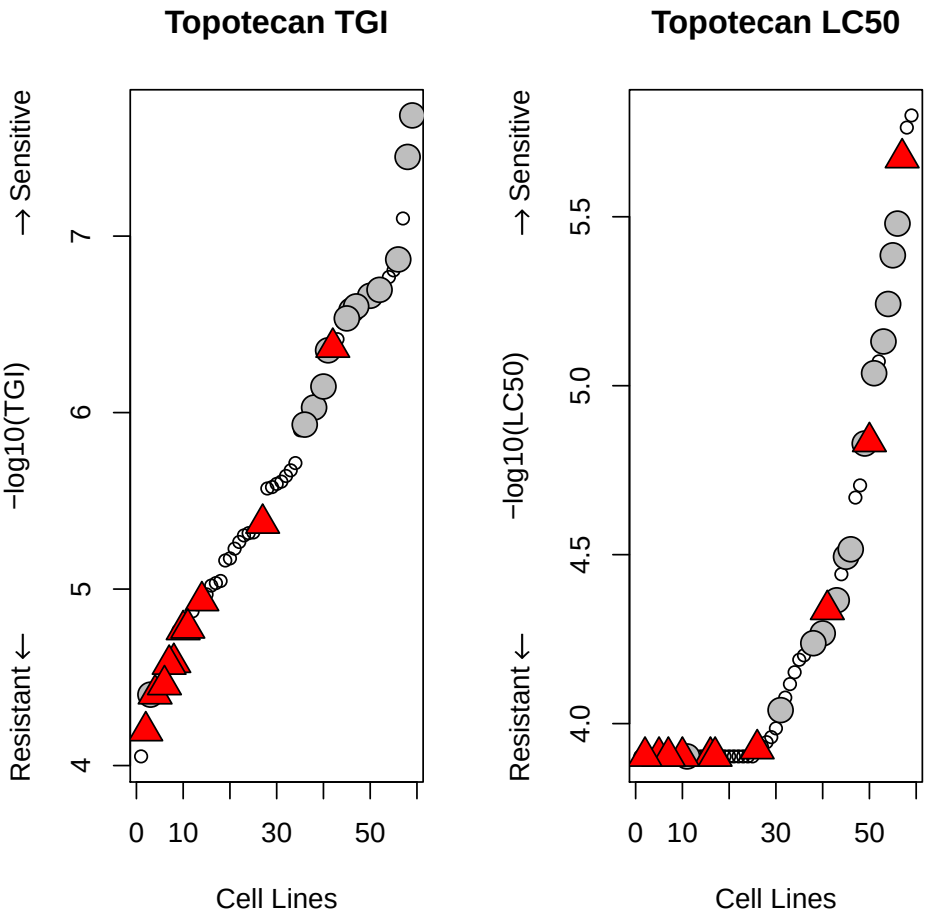
```
[1] "No data found for an LCONC of 0"
[1] "Valid concentrations (and numbers of tests) are:"
     LCONC TOTN
```

```
1942  -4.6    4
2002  -4.0    5
2072  -3.9    8
```

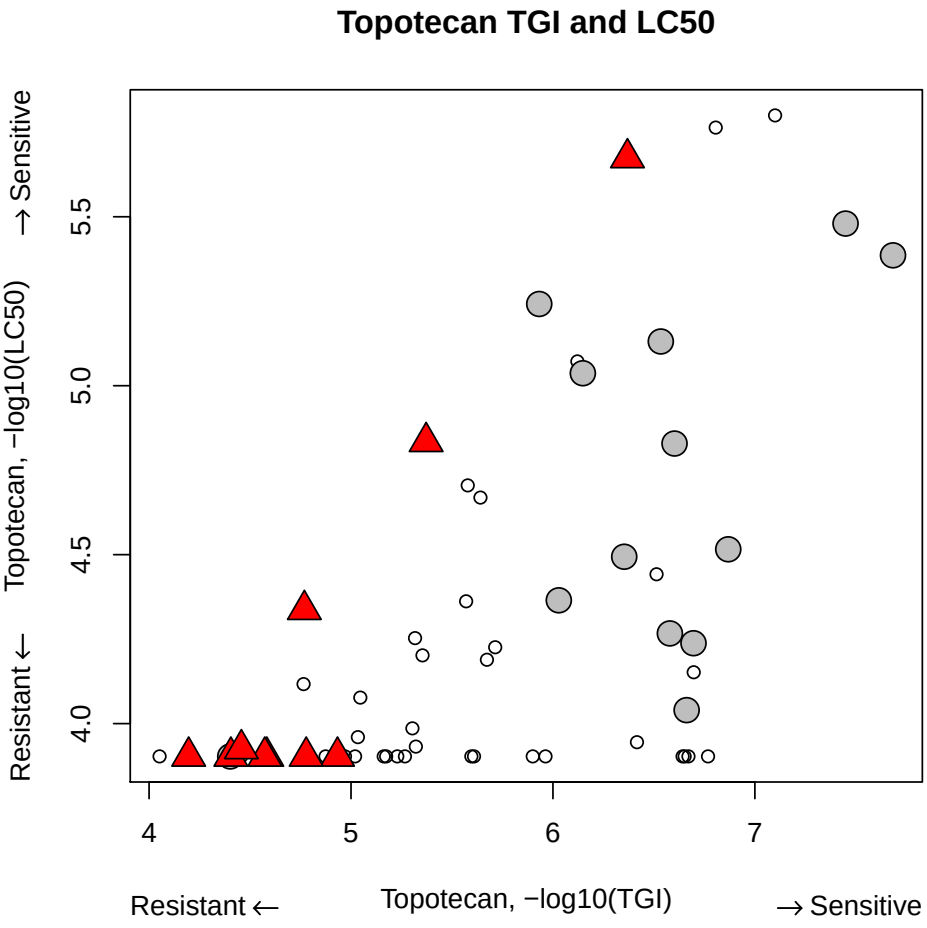
The -3.9 concentration is the one we use. We check the GI50 values first, using both the ordering and cell lines from the initial heatmaps of Potti et al. [8] (at left) and those from the second correction as of August 2008 (at right).



The latter orientation appears correct, and the shift of cell lines does not appear to alter the story. The values for the sensitive and resistant cell lines still overlap, and the most extreme cell lines do not appear to have been chosen. We look at the TGI and LC50 values next, using just the August 2008 ordering of lines.



Here, while the plots show we do have the correct orientation, the sensitive and resistant values continue to overlap; shifting the metric used does not remove this ambiguity. We next try plotting TGI and LC50 jointly to see if this resolves the issue.



Unfortunately, the ambiguity is still present. At least one resistant line is more sensitive than some sensitive lines along both measurement axes. Further, the lines chosen do not include some of the most extreme cases.

In sum, the August 2008 orientation appears correct, though this conflicts with the annotation for the Potti et al. [8] heatmaps. Values for the sensitive and resistant lines overlap along all three axes considered, so it is not clear how these lines were chosen.

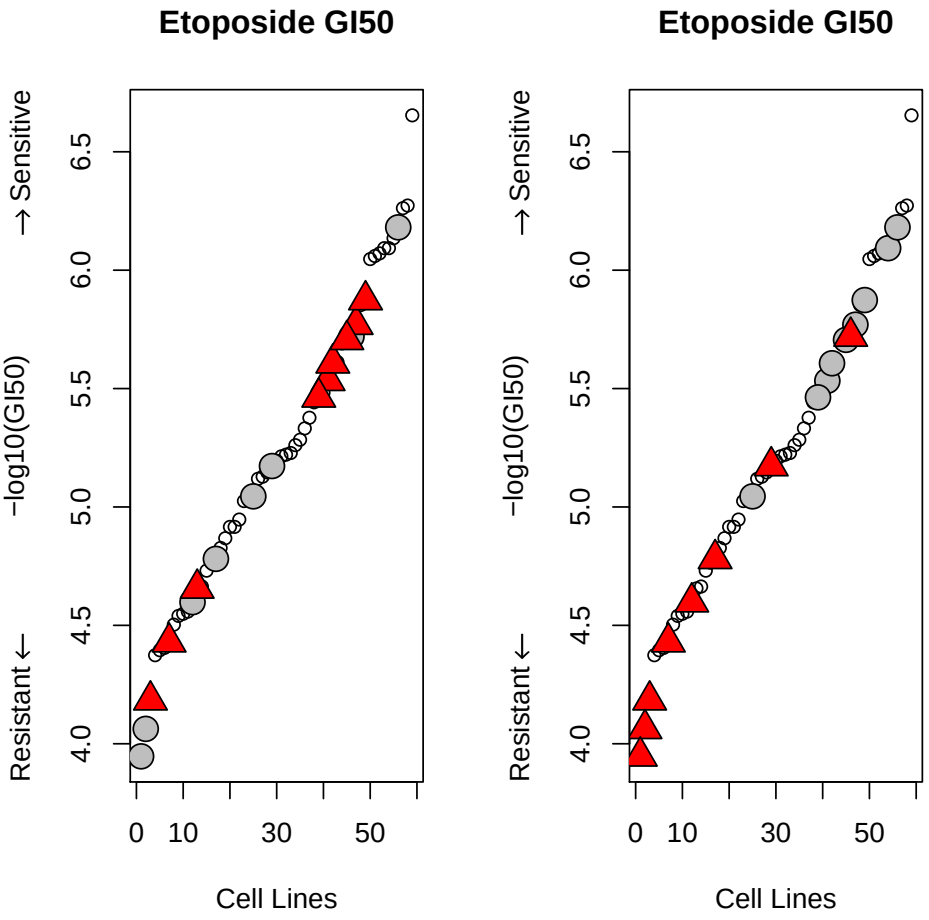
12 Etoposide

We turn next to etoposide. Here, there was a shift between the cell lines used for the initial Potti et al. [8] heatmap and the cell line lists reported in the 2007 and 2008 corrections, so we will look at both sets. Again, we first identify the LCONC value to focus on.

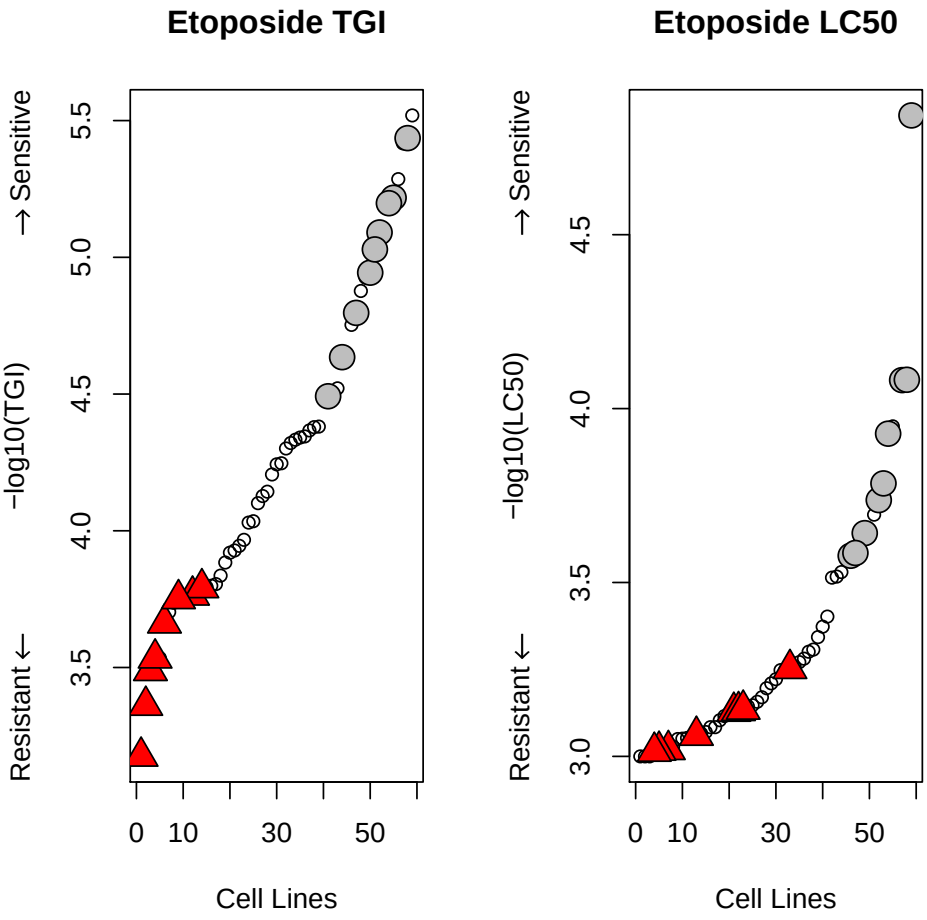
```
> temp <- getSensitivity("etoposide", 0, "tgi")
```

```
[1] "No data found for an LCONC of 0"
[1] "Valid concentrations (and numbers of tests) are:"
     LCONC TOTN
1520  -6.0    3
1580  -4.0    6
1649  -3.6    1
1694  -3.0   43
```

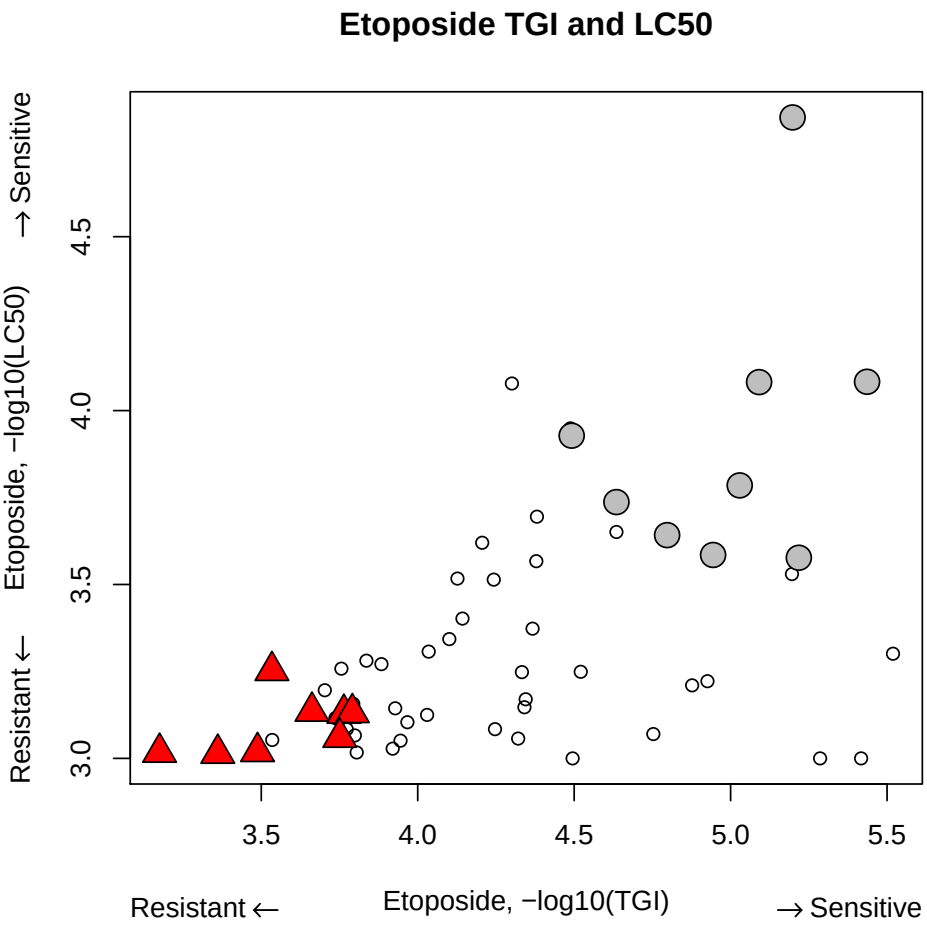
The -3.0 concentration is the one we use. We check the GI50 values first, using both the ordering and cell lines from the initial heatmaps of Potti et al. [8] (at left) and those from the second correction as of August 2008 (at right).



The latter orientation appears correct, and the shift of cell lines appears to make the story somewhat clearer at the resistant end. The values for the sensitive and resistant cell lines, however, still overlap and the most extremely sensitive cell lines do not appear to have been chosen. We look at the TGI and LC50 values next, using just the August 2008 ordering of lines.



Here, the plots show we have the correct orientation, with values for the sensitive and resistant lines clearly separated. We next try plotting TGI and LC50 jointly.



The story is pretty clear. There continues to be stark separation between the sensitive and resistant lines, though there is at least one extra cell line on the resistant side whose omission appears curious.

In sum, the August 2008 orientation appears correct, though this conflicts with the annotation for the Potti et al. [8] heatmaps.

13 Cyclophosphamide

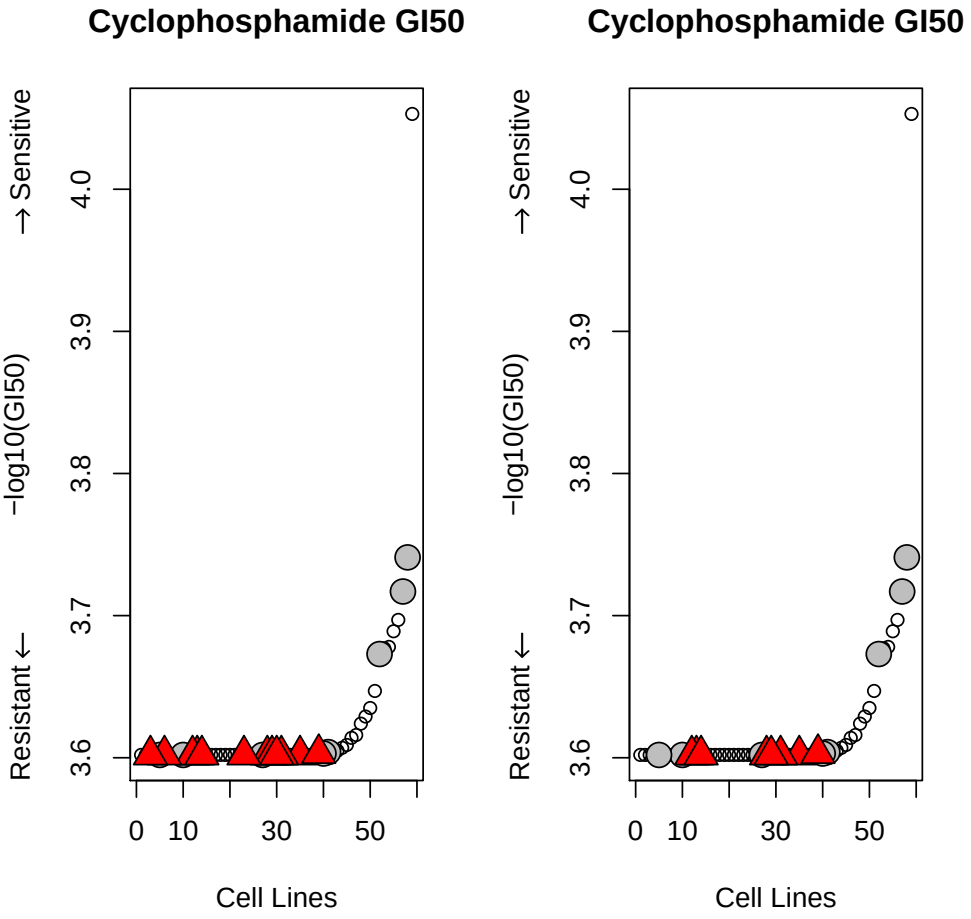
We turn next to cyclophosphamide. Here, there is a shift between the cell lines we can infer from the gene list reported with the November 2007 correction to Potti et al. [8] and the cell line lists reported in the 2007 and 2008 corrections, so we will look at both sets of lines. Again, we first identify the LCONC value to focus on.

```
> temp <- getSensitivity("cyclophosphamide", 0, "tgi")
```

```
[1] "No data found for an LCONC of 0"
[1] "Valid concentrations (and numbers of tests) are:"
```

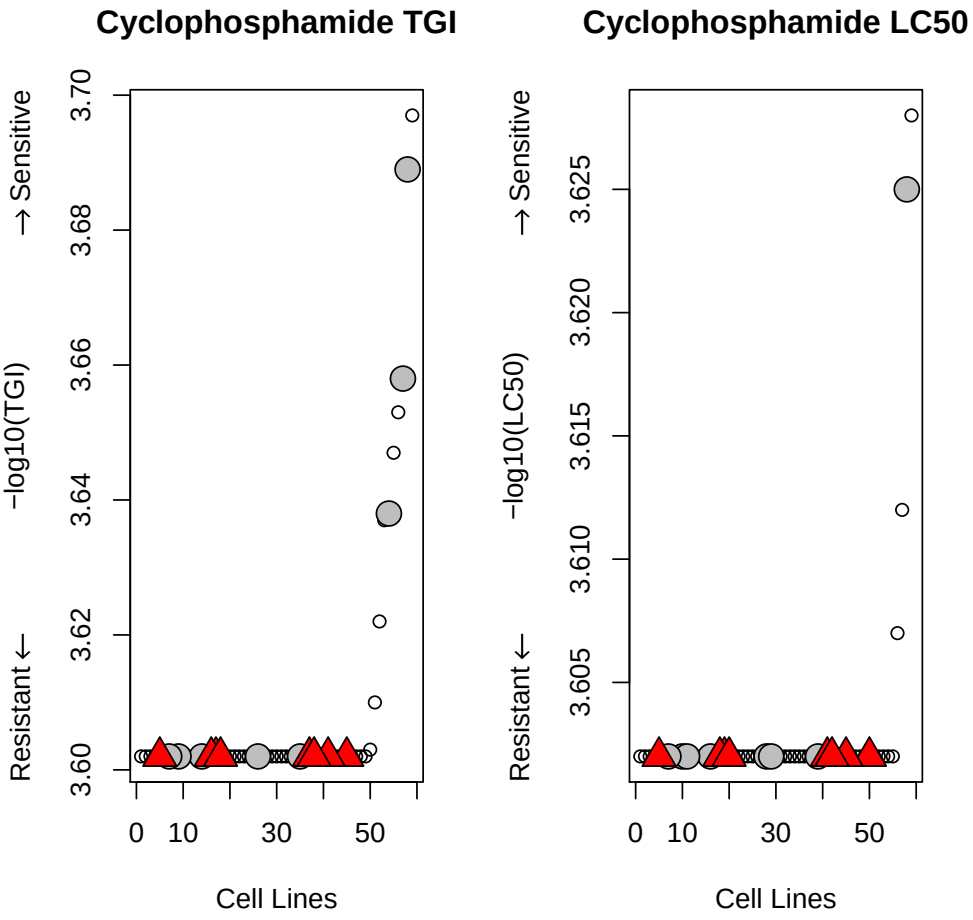
```
    LCONC TOTN
152  -3.6   59
```

Here, the -3.6 concentration is the only one available. We'll check the GI50 values first, using both the ordering and cell lines from the gene list (at left) and those from the cell line list given with the second correction as of August 2008 (at right).



The difference in cell lines does not affect the story appreciably; there is still clear overlap between the values for the sensitive and resistant lines. Another notable feature is the very narrow range of values seen. Almost all of the cell lines are at the resistant limit; deviations may well be due to experimental error.

We now look at the TGI and LC50 values, using just the more recent list.



Here, the story does not become clearer with a shift of metric. There is still effectively no separation to be seen, and the vertical axes are even narrower are before. The cell lines are all effectively "resistant" to cyclophosphamide. This is not surprising – if we don't hit 50% growth inhibition for a cell line at the maximum concentration measured, we certainly can't hit total growth inhibition or 50% lethal concentration.

The problem is that cyclophosphamide is a prodrug, and needs to be processed by the body into an active form. Cell lines, lacking the processing apparatus, do not respond. It is not clear how this NCI60 sensitivity data can be used to select lines for cyclophosphamide.

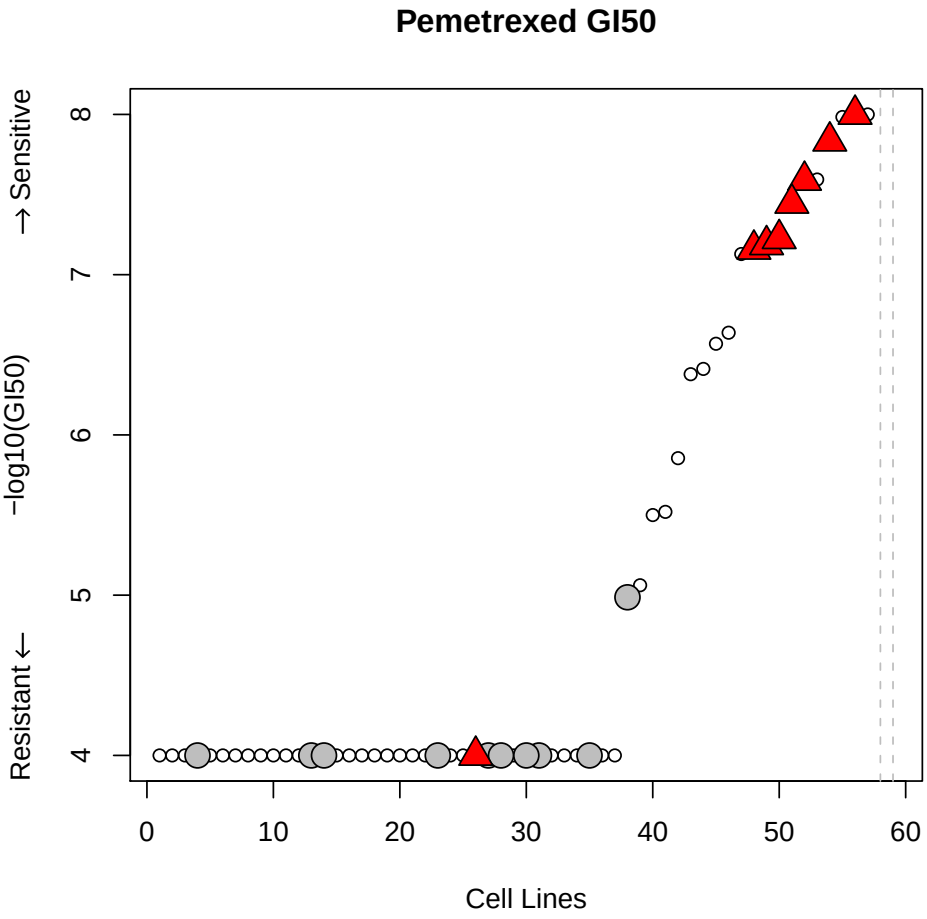
14 Pemetrexed

We turn next to pemetrexed. Here, our only source of information involves inference from the heatmap published in Hsu et al. [7]. Full details of the identification of the cell lines used are given in matchingHsuHeatmaps.pdf. Again, we first identify the LCONC value to focus on.

```
> temp <- getSensitivity("pemetrexed", 0, "gi50")
```

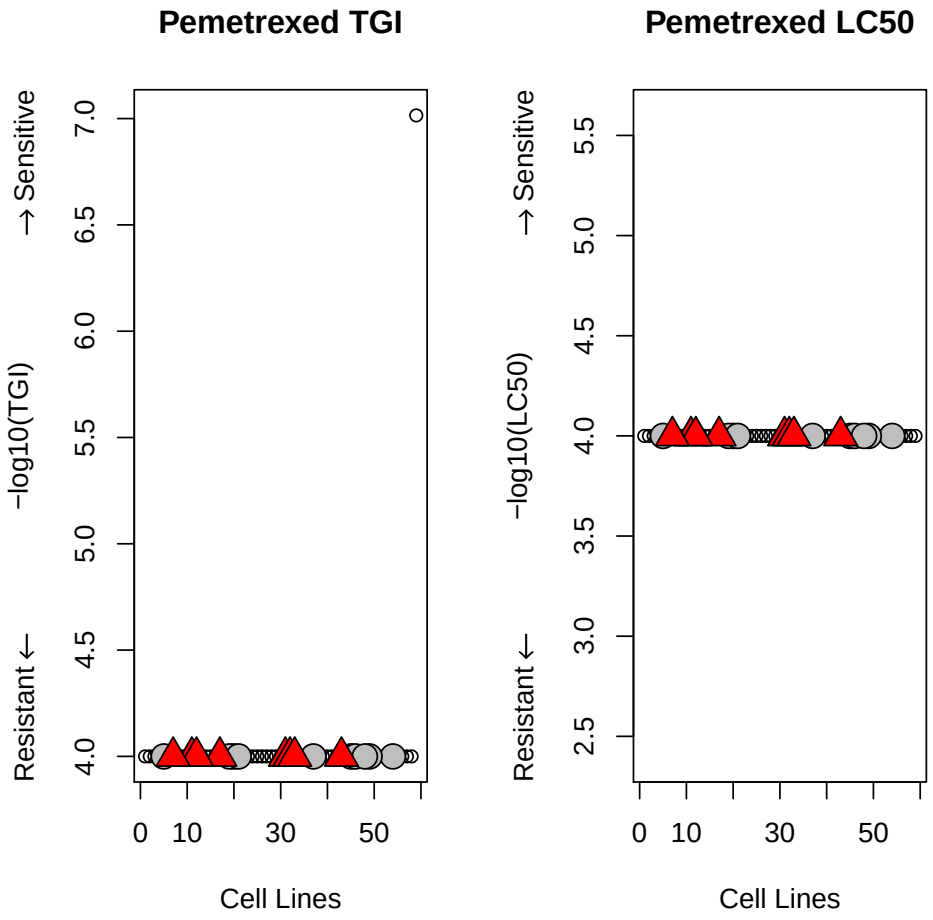
```
[1] "No data found for an LCONC of 0"
[1] "Valid concentrations (and numbers of tests) are:"
     LCONC TOTN
2333    -4    2
```

Here, -4 is the only concentration available. We check the GI50 values first.



Here, the orientation is clearly reversed; all of the "sensitive" lines are already at the resistant limit. Indeed, more than half of the cell lines profiled are at this limit, and it is not clear how only a subset was chosen. Further, the sensitive and resistant values overlap due to an outlying "resistant" value.

We check TGI and LC50 next.



Here the story is quite different than for the previous drugs – these two metrics are uninformative as all of the lines specified are at the resistant limit. Indeed, all of the cell lines, period, are at the limit for LC50 and only one (SF-539) misses it for TGI.

In sum, with one exception, the lines used for pemetrexed separate according to their GI50 values, with the sensitive and resistant ordering reversed.

15 Cisplatin

For cisplatin, Hsu et al. [7] used the 30 cell lines from Györffy et al. [5] to define a sensitivity signature instead of using the NCI60. They note that

> The collection of data in the NCI-60 data occasionally does not represent a significant diversity in resistant and sensitive cell lines to any given drug. Thus, if a drug screening experiment did not result in widely variable GI50/IC50 and/or LC50 data, the generation of a genomic predictor is not possible using our methods, as in the case of cisplatin. Thus, we used data published by Györffy et al. [5] where they had determined definitive resistance and sensitivity to cisplatin in

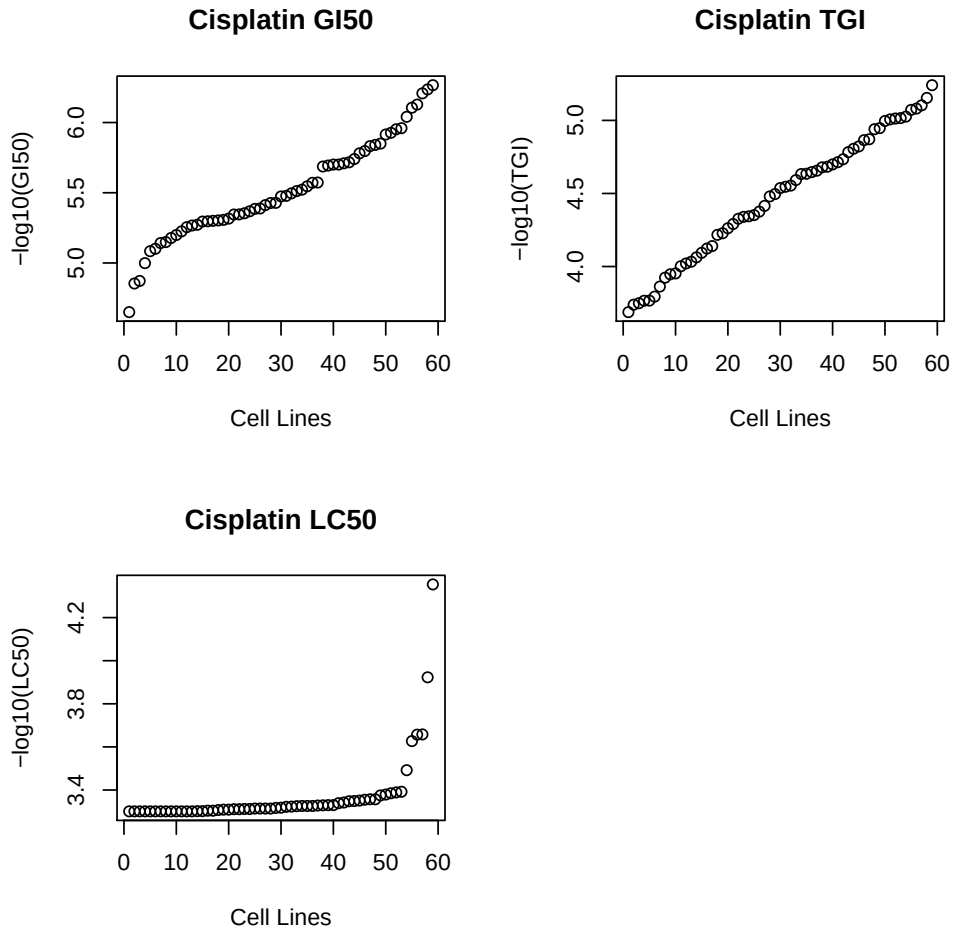
30 cancer cell lines. Importantly, we also had access to corresponding gene expression data to facilitate the generation of a model that would predict sensitivity to cisplatin.

Here, we simply check the quality of the NCI60 drug sensitivity data available for cisplatin. We begin by checking the concentrations available.

```
> temp <- getSensitivity("cisplatin", 0, "gi50")
```

```
[1] "No data found for an LCONC of 0"
[1] "Valid concentrations (and numbers of tests) are:"
    LCONC TOTN
713  -5.6    2
771  -4.3   16
864  -4.0    7
944  -3.3  143
```

The -3.3 concentration is the one we use.



It is not clear to us that the data for cisplatin are markedly worse than that for several of the other drugs for which signatures have been assembled.

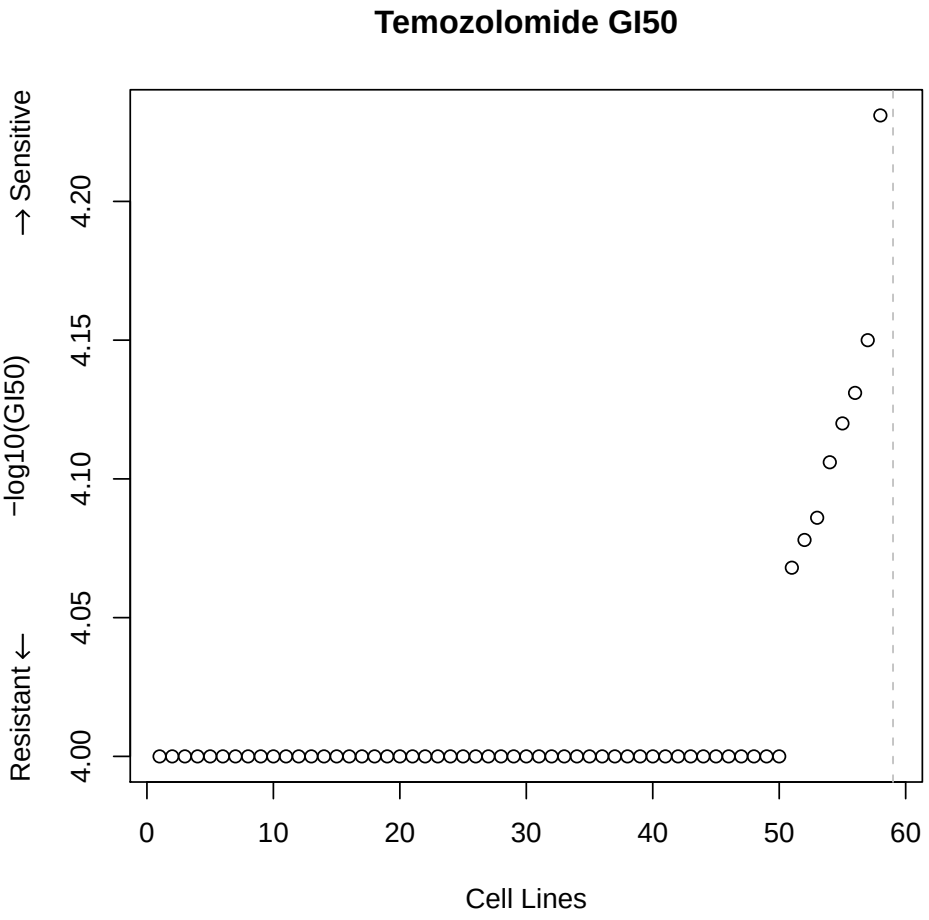
16 Temozolomide

We turn next to temozolomide. Here, our only source of information involves inference from the heatmap published in Augustine et al. [1], which was in fact noninformative because the heatmap shown was actually the one for cisplatin from Hsu et al. [7]. Thus, we are simply exploring the range of values available here, without focusing on the specific cell lines used. Again, we first identify the LCONC value to focus on.

```
> temp <- getSensitivity("temozolomide", 0, "gi50")
```

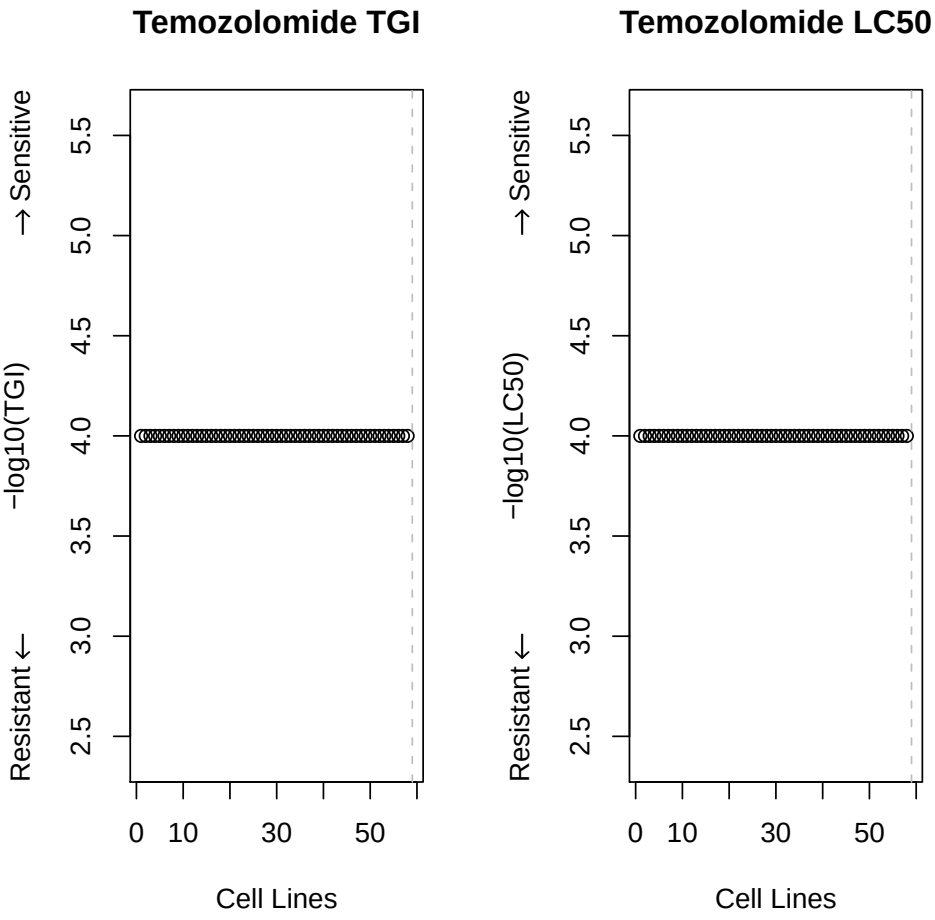
```
[1] "No data found for an LCONC of 0"
[1] "Valid concentrations (and numbers of tests) are:"
     LCONC TOTN
1875    -4    2
```

Here, the -4 concentration is the only one available. We check the GI50 values first.



The GI50 plot here shows a different problem than that of reusing a heatmap – there is effectively no difference between the cell lines in terms of response to the drug. The cell lines are all resistant (note the very narrow range of GI50 values).

We check the TGI and LC50 values next.



The TGI and LC50 values are completely noninformative.

This resistance is similar to what we saw with cyclophosphamide, and for the same reason: temozolomide is a prodrug, and needs to be processed by the body into an active form. Cell lines, lacking the processing apparatus, do not respond. As with cyclophosphamide, it is not clear how the NCI60 data can be used to select lines for temozolomide.

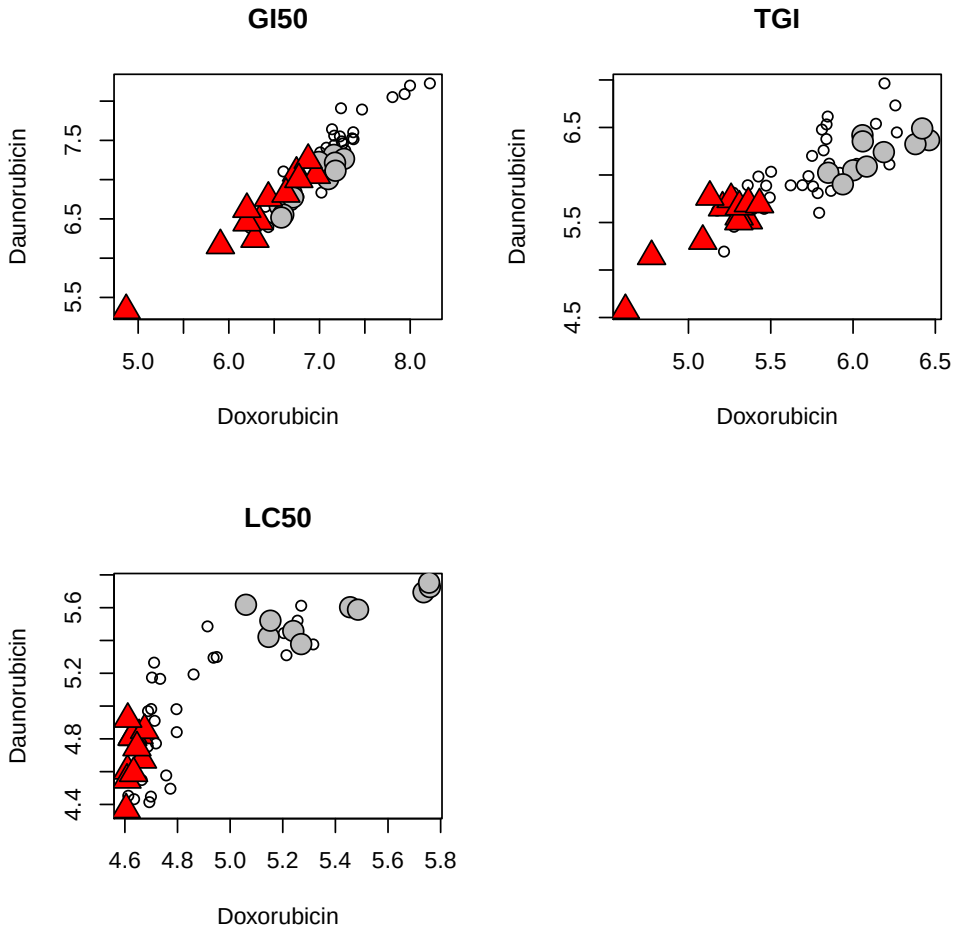
17 Doxorubicin and Daunorubicin

We turn now to daunorubicin. This drug is not explicitly modeled by Potti et al. [8], but it was the drug used in the Holleman et al. [6] study that Potti et al. [8] use as a test set for doxorubicin. Here, we simply confirm that daunorubicin and doxorubicin appear to have similar sensitivity profiles. We begin by checking concentrations for daunorubicin.

```
> temp <- getSensitivity("daunorubicin", 0, "gi50")
```

```
[1] "No data found for an LCONC of 0"
[1] "Valid concentrations (and numbers of tests) are:"
    LCONC TOTN
395  -5.0    2
461  -4.3   97
577  -4.0    2
637  -3.3   55
```

We use the -4.3 LCONC value.



The agreement between drugs is pretty good, so using a doxorubicin signature to predict daunorubicin response seems reasonable.

18 Doxorubicin and Epirubicin

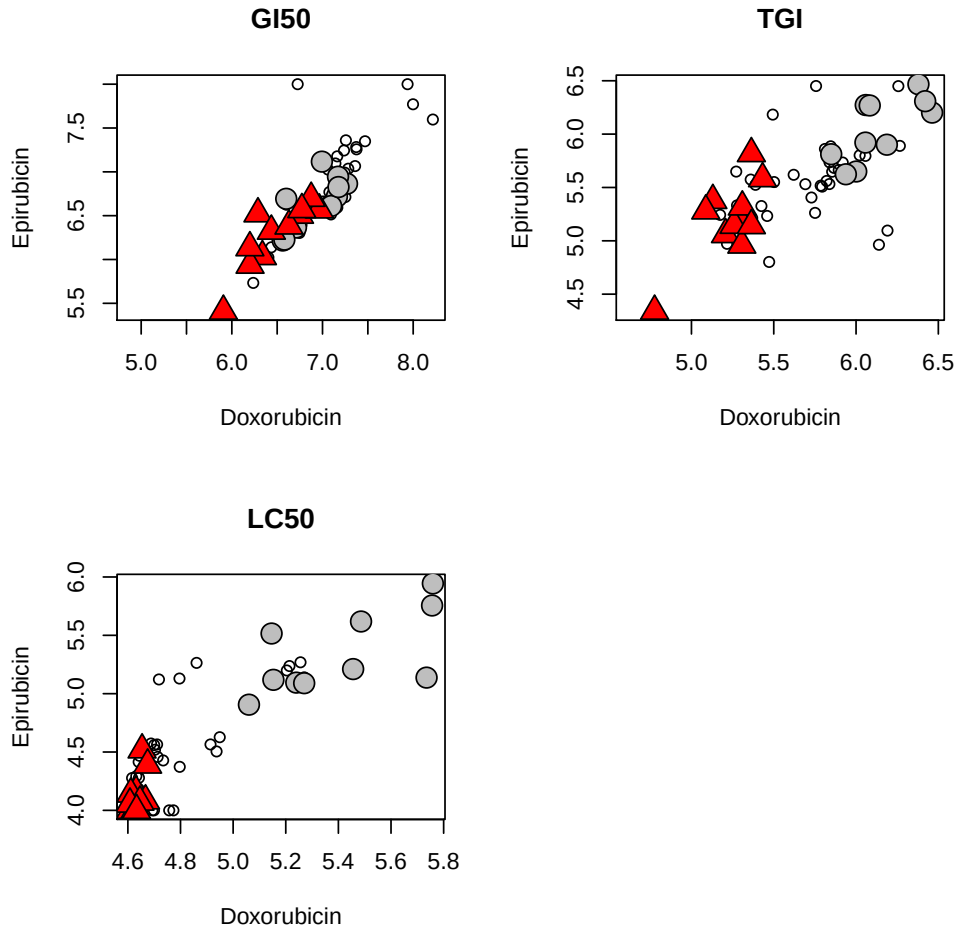
We turn now to epirubicin. This drug is not explicitly modeled by Potti et al. [8], but it is modeled by Bonnefoi et al. [2] using the doxorubicin signature. Here, we simply confirm that epirubicin and doxorubicin

appear to have similar sensitivity profiles.

```
> temp <- getSensitivity("epirubicin", 0, "gi50")
```

```
[1] "No data found for an LCONC of 0"
[1] "Valid concentrations (and numbers of tests) are:"
     LCONC TOTN
1809    -4    3
```

Here, -4 is the only concentration available.



The agreement between drugs is pretty good, so using a doxorubicin signature to predict epirubicin response seems reasonable.

19 Doxorubicin and Vincristine

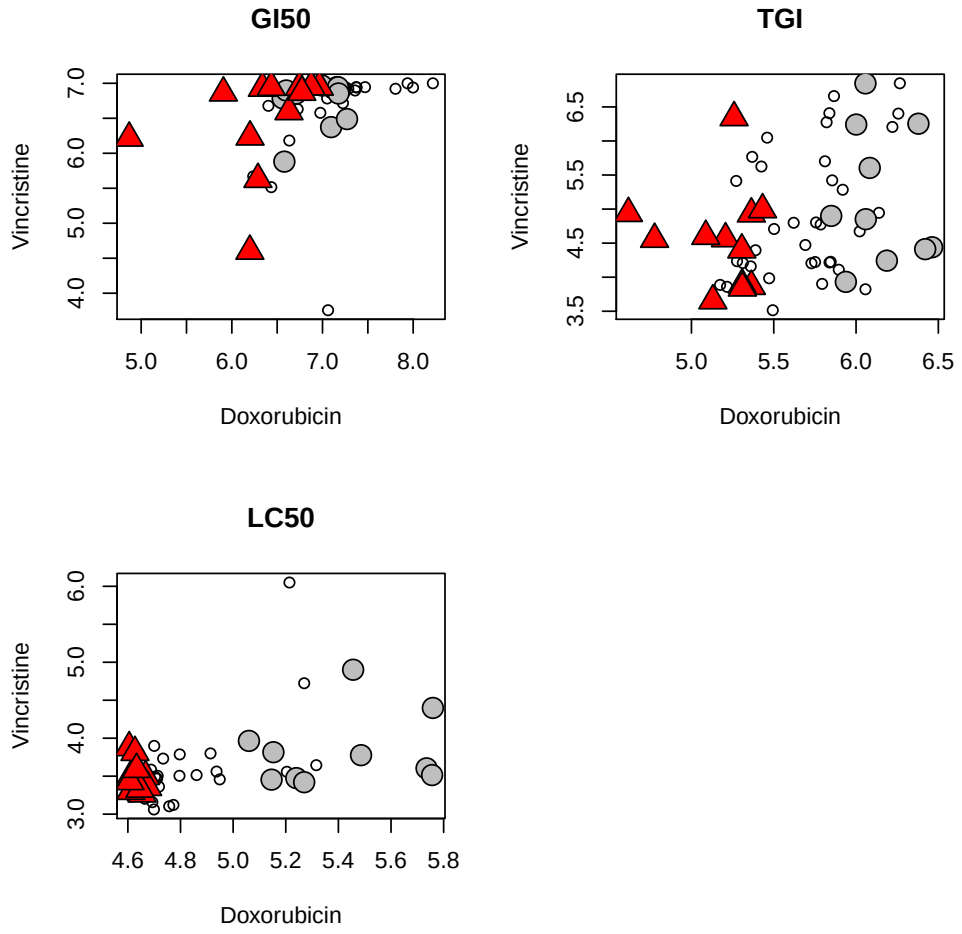
We turn now to vincristine. This drug is not explicitly modeled by Potti et al. [8], but the second correction to Potti et al. [8] indicates that the signature for doxorubicin was tested in part on patients chosen according

to their resistance to vincristine.

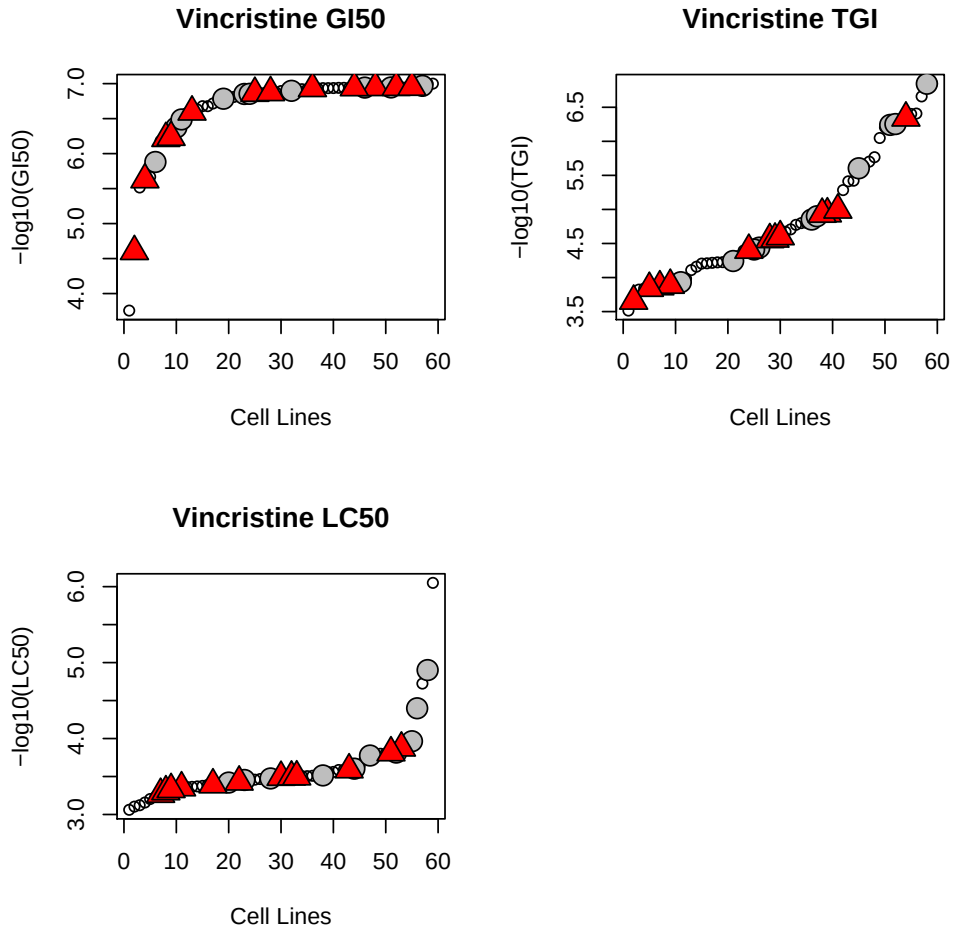
```
> temp <- getSensitivity("vincristine", 0, "gi50")
```

```
[1] "No data found for an LCONC of 0"
[1] "Valid concentrations (and numbers of tests) are:"
    LCONC TOTN
235    -5    9
295    -3   63
```

We use the -3 concentration.



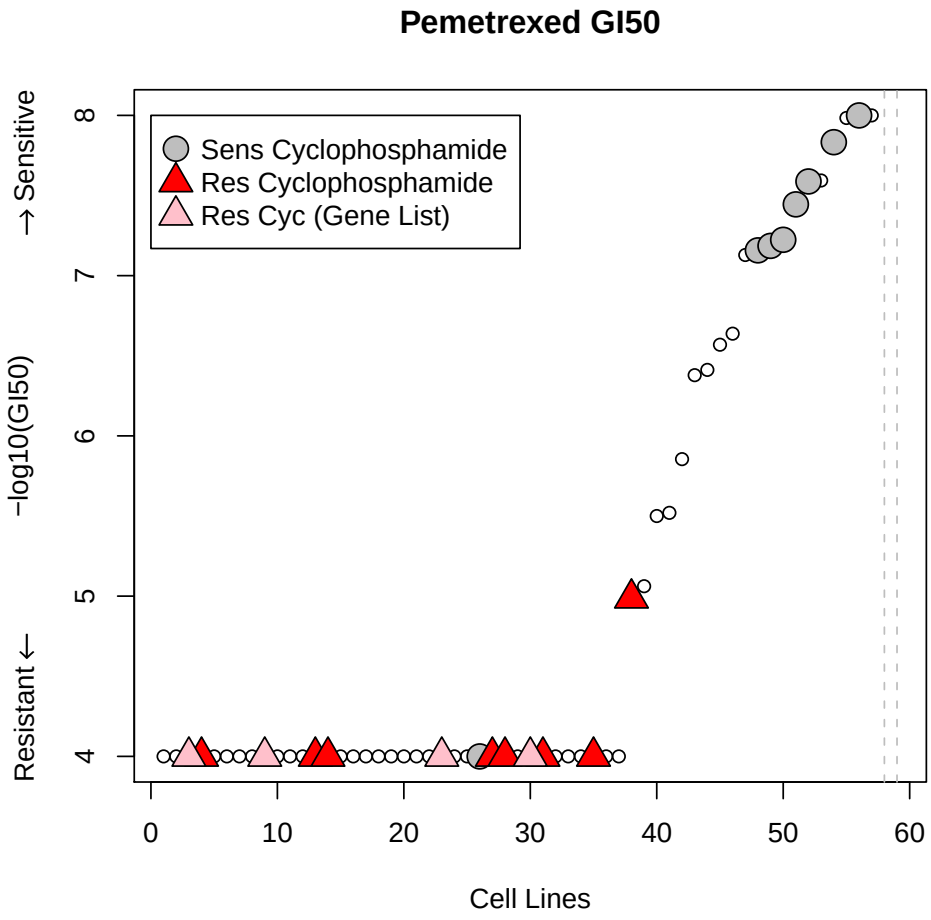
Unfortunately, the agreement here is very poor. Lines that separate according to doxorubicin response do not separate according to vincristine response. This may be more evident if we simply plot the vincristine values, only indicating the doxorubicin involvement through symbols.



The sensitive and resistant groups are clearly intermingled. Using a doxorubicin signature to predict response to vincristine is a bad idea.

20 Pemetrexed and Cyclophosphamide

We turn last to a consideration of pemetrexed and cyclophosphamide, motivated by the similarity of the cell lines involved. Thus, we examine the drug sensitivity values for pemetrexed, labeling the lines identified as sensitive and resistant for cyclophosphamide. There are 4 cell lines not in the list of cell lines reported for cyclophosphamide that inference from the gene list suggests should be in the resistant group: EKVX, SW-620, SN12C, and SK-MEL-28. We add these using a slightly different shading. (SN12C and SK-MEL-28 were used for pemetrexed.)



But for one outlier, the separation is as stark as we have seen for the other drugs. This suggests to us that the drug sensitivity data for pemetrexed was used to select the cell lines used for cyclophosphamide.

21 Summary Figure

Here we collect some figures illustrating specific points.

```
quartz
     2
```

22 Appendix

22.1 File Location

```
> getwd()
```

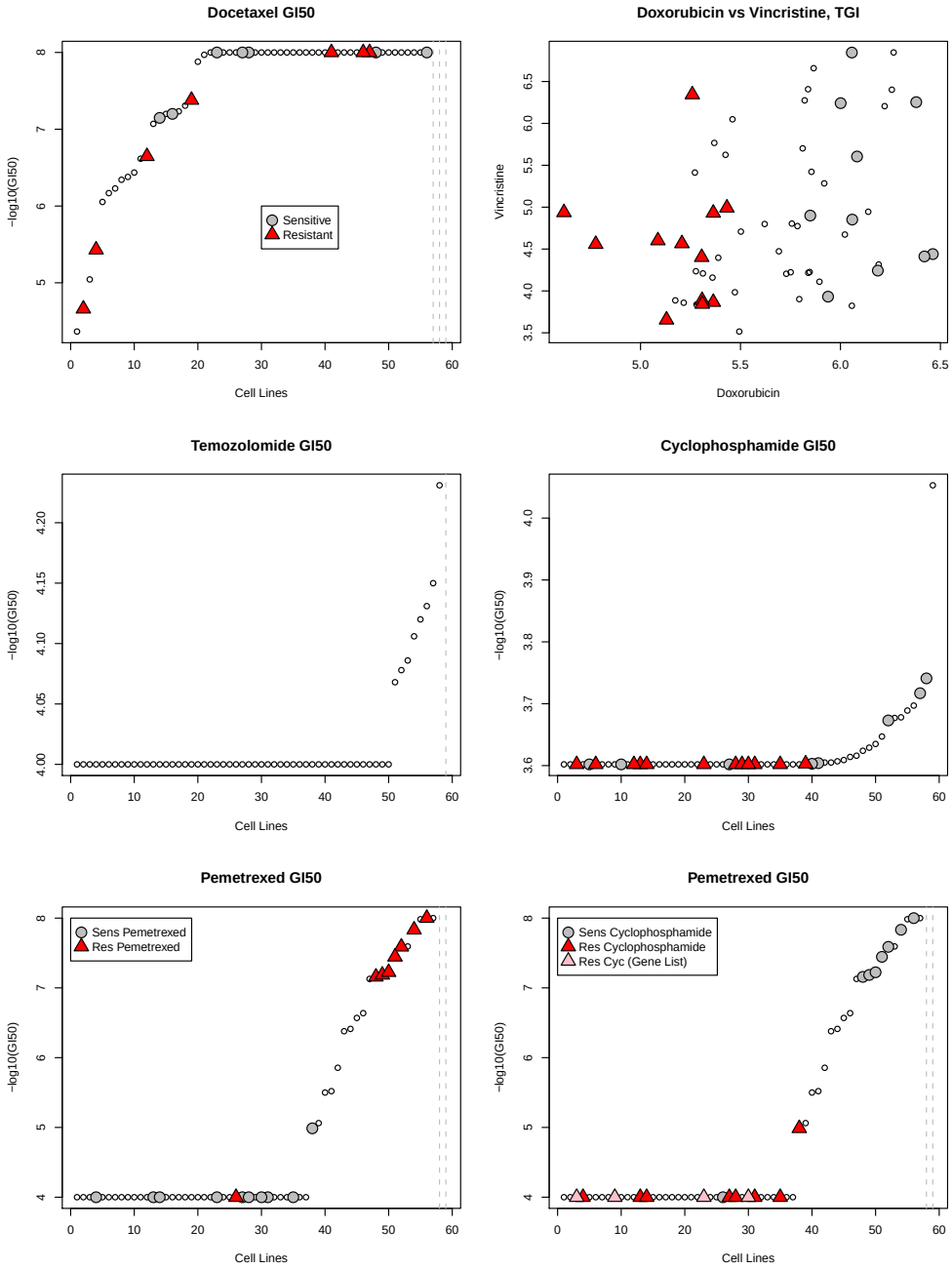


Figure 1: A. GI50 values for docetaxel. GI50 values for the sensitive and resistant lines overlap, which shouldn't happen. B. TGI values for doxorubicin and vincristine. Some test samples for doxorubicin signature were chosen on the basis of their sensitivity to vincristine, but the lack of agreement means using doxorubicin sensitivity to predict vincristine response is a bad idea. C. GI50 values for temozolomide. All cell lines have GI50 values very close to the resistant limit. Temozolomide is a prodrug that must be processed by the body into an active form. Cell lines lack the processing machinery. D. GI50 values to cyclophosphamide. Cyclophosphamide, like temozolomide, is a prodrug. It is not clear how sensitivity data for a prodrug can be used to choose cell lines. E. GI50 values for pemetrexed. The sensitive and resistant labels have been reversed. F. GI50 values for pemetrexed, with cyclophosphamide cell lines indicated. The pattern suggests that cell lines for cyclophosphamide were chosen based on sensitivity data for pemetrexed. The heatmap used by Augustine et al. [1] for temozolomide likewise belongs to cisplatin ([7]).

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

22.2 Saves

22.3 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] Holleman A, Cheok MH, den Boer ML, et al.: Gene-Expression patterns in drug-resistant acute lymphoblastic leukemia cells and response to treatment. *NEJM*, **351**:533-42, 2004.

[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

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

[10] 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.

[11] 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.