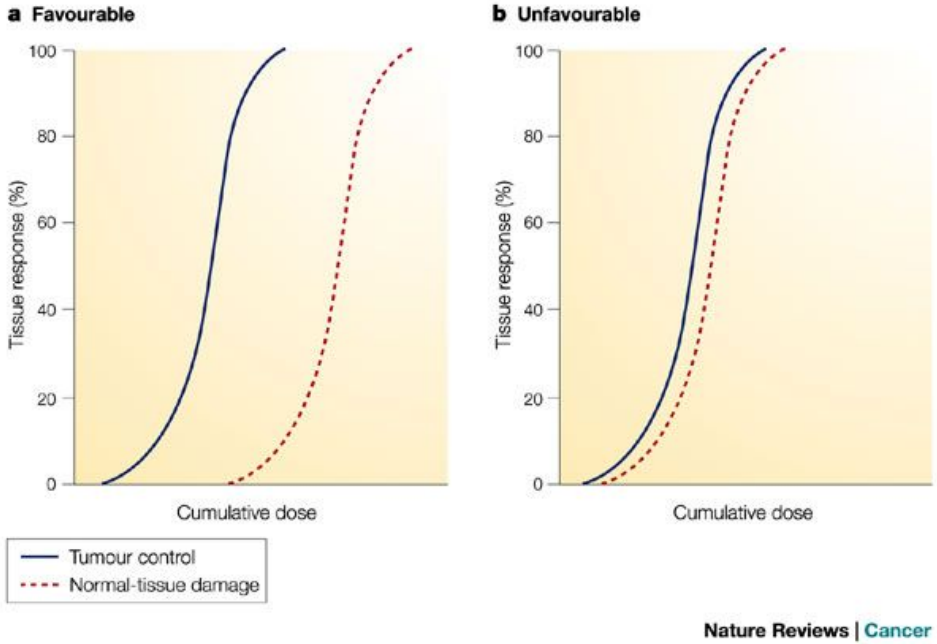
# Lymphoblastoid Cell Lines as a Model System for Chemotherapy Response in Breast Cancer Patients

Jeremy Ash and Alex Larsen

# What are LCLs?

- Lymphoblastoid Cell Lines (LCLs)
  - immortalized with Epstein-Barr virus
  - commonly used model system for pharmacogenomic studies
- Many studies showing that these immortalized cells are good surrogates for lymphocytes
  - review: Hussain and Mulherkar, 2012
  - close resemblance to parent lymphocytes
- May be a good model for in vivo drug exposure




Bernier et al. 2004

# Pros and Cons of LCL's

- Why use LCLs, or a model system at all?
  - Relatively few confounding variables compared to clinical trials genome mapping studies
  - Clinical trials have:
    - Changes in treatment regimen
    - Experimental design problems
  - Many well-characterized cell lines available from established cell line resources
    - 1000 Genomes Project, International HapMap Project, etc
- Disadvantages of LCLs
  - Cells from tissue where drug response is of interest are difficult and expensive to collect
  - "...arguably the biggest limitation of in vitro models is the difficulty in translating the relevance of in vitro phenotypes to in vivo outcomes. This is especially pertinent for highly toxic drugs, such as chemotherapies, where the therapeutic index is very narrow. Whether sensitivity in vitro relates to efficacy, toxicity, or both is an open, unanswered question..." (Jack et al, 2012)

# Taxane Chemotherapy for Breast Cancer Patients

- Chemotherapy treatments target cancer cells
  - Issued cyclically
  - Each treatment period is followed by recovery periods
- For this project, studying patient sensitivity to chemotherapy drug, Paclitaxel
- Paclitaxel is the chemical in Taxane (brand name)
  - Taxane commonly given as adjuvant therapy (post-surgery)
  - Interferes with a cell’s ability to divide

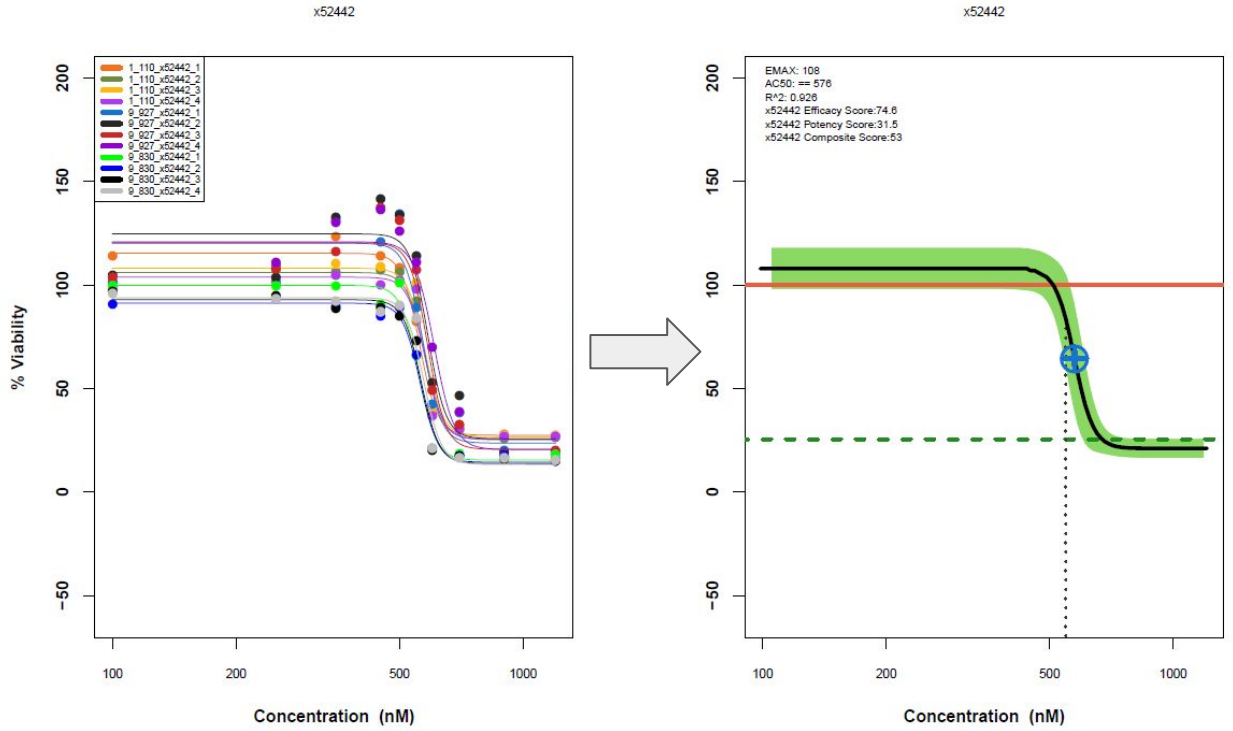
# Project Objective:

Is in-vitro exposure of lymphoblastoid cell lines to Paclitaxel a good predictor of cancer patient responses to chemotherapy?

- Are there relationships between breast cancer patient outcomes and LCL model behavior?
- Which of our set of clinical variables best explain variation in patient dose response profiles?

# LCL Data

- Dose response measurements for 93 c lines collected from bre cancer patients
  - Percent viability measured at 10 different doses
  - Extensive QC pipeline
- Curve Fits
  - IC50 values extracted



# LCL Data QC Pipeline

- Data normalized to positive and negative controls (100% viability at negative control, 0% viability at positive control)
- Coefficient of variation smoothing
- Removed replicates with abnormally low cell counts
  - Obviously sick
- Monotonicity filter
  - Would ideally do regression smoothing but could not get to work
- **Final Data**: mean of remaining replicates for each dose response

# Clinical Trial Data

- Recovery status (none, partial, full)
- Age
- Menopause
- Estrogen receptor
- Treatment regimen
- Number of cycles
- Total weeks
- Pre-experiment cancer stage
- Race
- Smoking status
- Her2 (human epidermal growth factor receptor 2)

# Methods

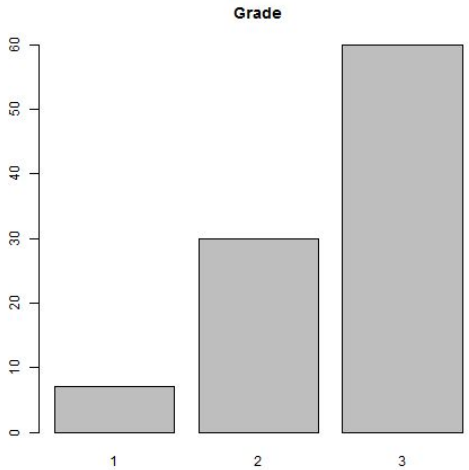
The goal is to look for associations between the LCL data and the clinical trial data. We employ the following:

- Correlations
- Significance Testing
  - Dose-Response Curve vs. Clinical Trial variables
  - IC50 vs. Clinical Trial variables
- Model Fitting
  - Ordinal regression
- Variable selection
  - Regression Trees, Random Forests

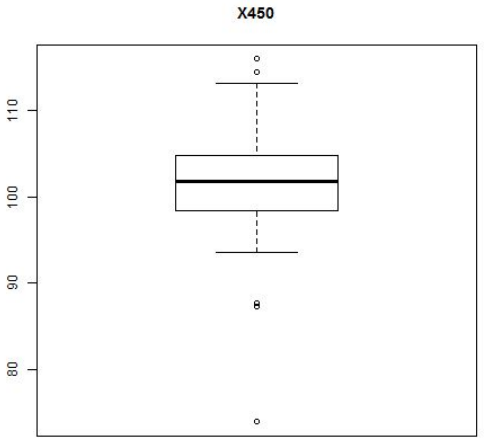
# Initial Data Exploration

- Continuous Variables
  - Boxplots, Histograms, Summaries of distributions
  - Identification of outliers
- Categorical Variables
  - Barplots, Contingency tables

Imbalanced Classes

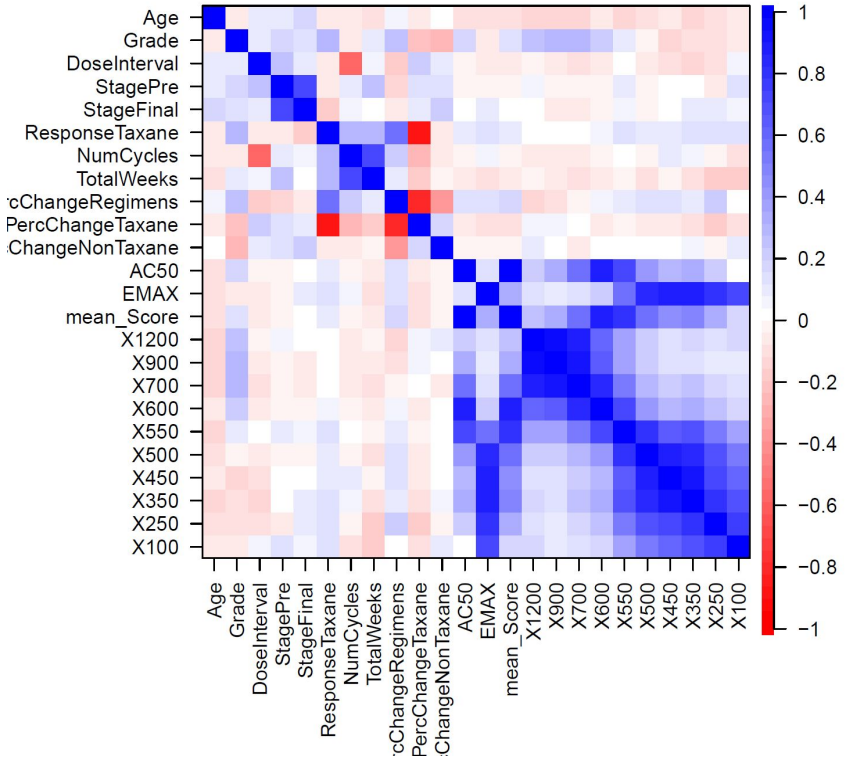



Large Outliers

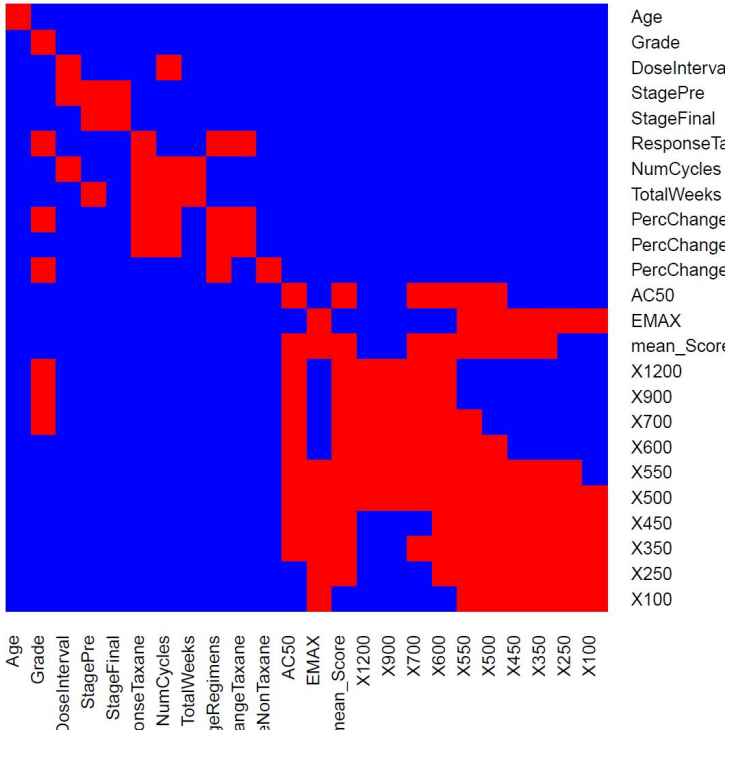


# Continuous Variable Associations

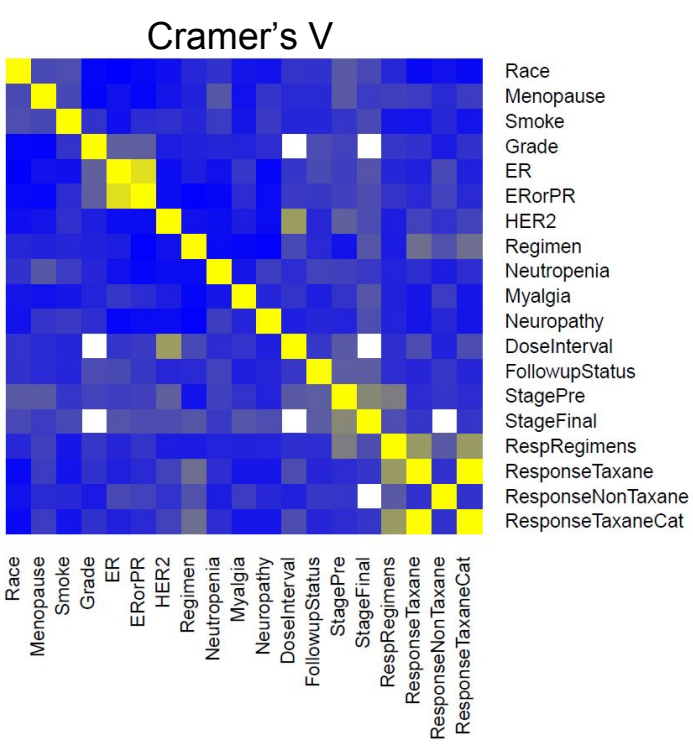
Pearson's Correlations

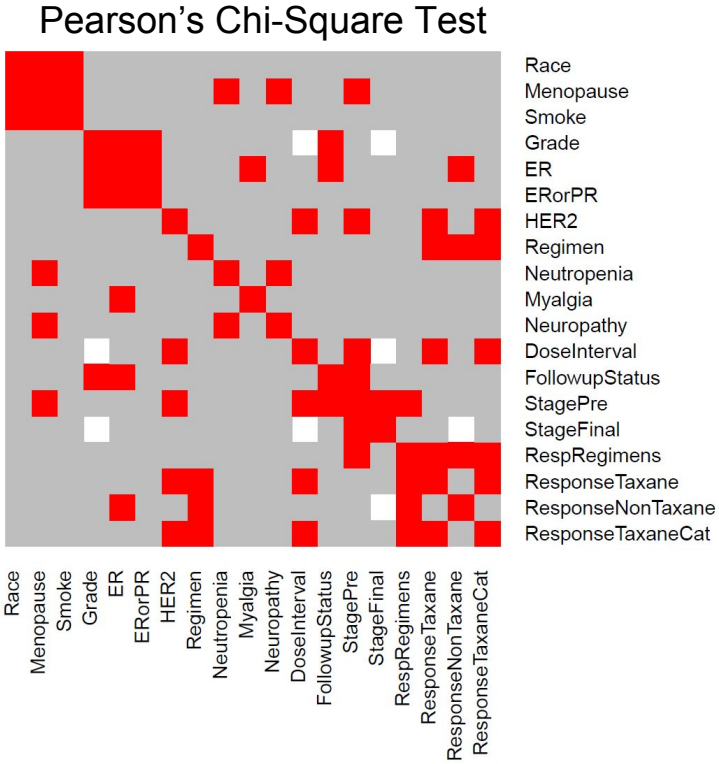



SLR p-values

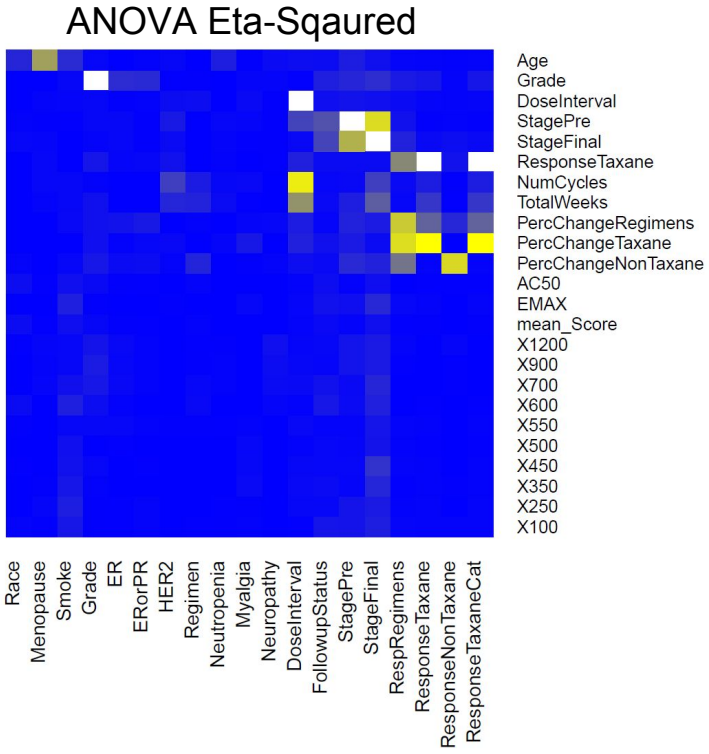


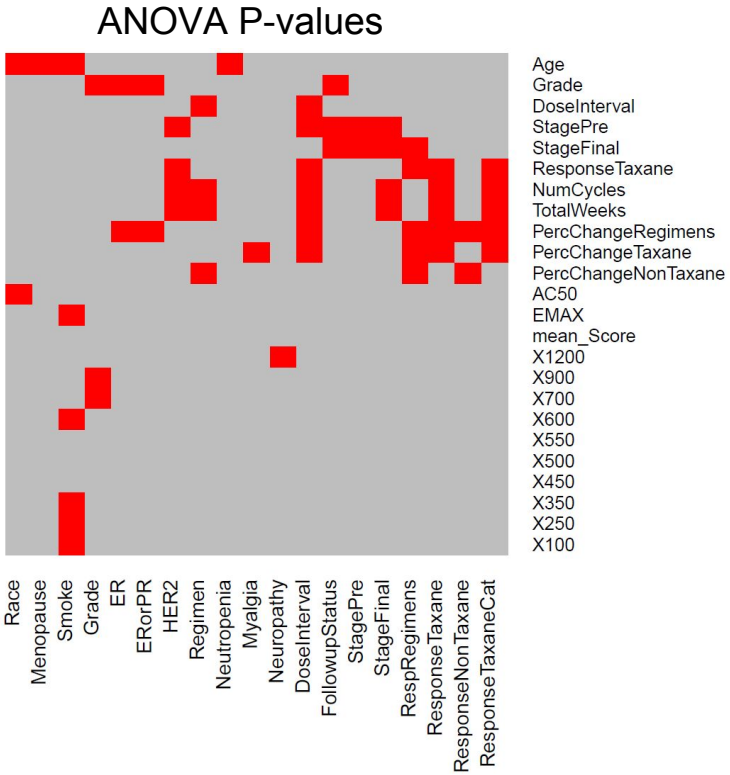
# Categorical Variable Associations



# Categorical/Continuous Variable Associations



# Significance of Clinical Trial Data to LCL Data

- With SLR we tested the same with IC50 as our regressor
  - Used False Discovery Rate (FDR)-controlling procedures (less strict that FWE) for the SLRs
  - To control FDR (proportion of false positives) we compare raw p-values to Benjamini-Hochberg critical value (BHCV) and set the FDR to 10%
- We also looked at MLR with IC50 as our regressor
- Using MANOVA, we investigated the significance of each clinical trial variable with dose-response curves as our regressor
  - Used Bonferroni correction for multiple comparisons

# Significance Testing Results: Individual Tests

## Dose-Response Curve

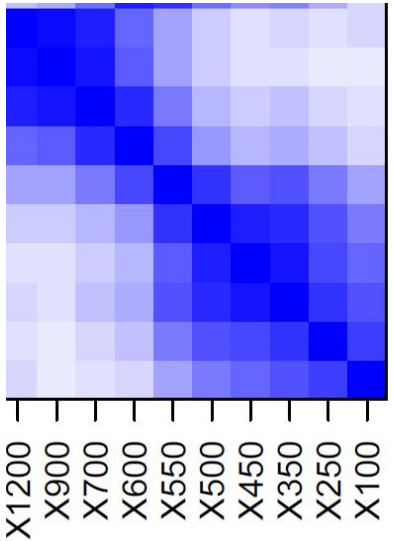
| Variable | P-Value |
|---|---|
| Race | 0.0049490 |
| Age | 0.3120007 |
| Menopause | 0.5107965 |
| Smoke | 0.0050083 |
| Grade | 0.5514878 |
| ER | 0.3770979 |
| ERorPR | 0.4621142 |
| HER2 | 0.9714836 |
| Regimen | 0.4028696 |
| NumCycles | 0.7709108 |
| Neutropenia | 0.7682658 |
| Myalgia | 0.6230537 |
| Neuropathy | 0.2966172 |
| DoseInterval | 0.8108977 |
| TotalWeeks | 0.4532002 |
| FollowupStatus | 0.7695808 |
| StagePre | 0.6687266 |
| StageFinal | 0.8365979 |
| ResponseTaxane | 0.9736503 |
| ResponseNonTaxane | 0.8850064 |

## IC50

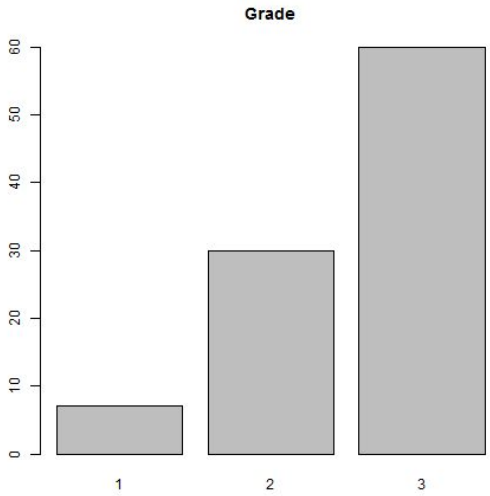
| Variable | P-Value | Rank | BHCV |
|---|---|---|---|
| Race | 0.0499910 | 1 | 0.0075 |
| Smoke | 0.0896301 | 2 | 0.0150 |
| Grade | 0.1184657 | 3 | 0.0225 |
| Regimen | 0.2367851 | 4 | 0.0300 |
| FollowupStatus | 0.3194446 | 5 | 0.0375 |
| ER | 0.3759591 | 6 | 0.0450 |
| ERorPR | 0.5139868 | 7 | 0.0525 |
| Menopause | 0.5287281 | 8 | 0.0600 |
| StageFinal | 0.5451857 | 9 | 0.0675 |
| DoseInterval | 0.5621745 | 10 | 0.0750 |
| Neutropenia | 0.6075599 | 11 | 0.0825 |
| HER2 | 0.6924021 | 12 | 0.0900 |
| ResponseTaxane | 0.7618464 | 13 | 0.0975 |
| StagePre | 0.7766474 | 14 | 0.1050 |
| Age | 0.7805672 | 15 | 0.1125 |
| Neuropathy | 0.7869654 | 16 | 0.1200 |
| NumCycles | 0.7889688 | 17 | 0.1275 |
| ResponseNonTaxane | 0.8585856 | 18 | 0.1350 |
| Myalgia | 0.8604907 | 19 | 0.1425 |
| TotalWeeks | 0.8818844 | 20 | 0.1500 |

# Manova: Challenges

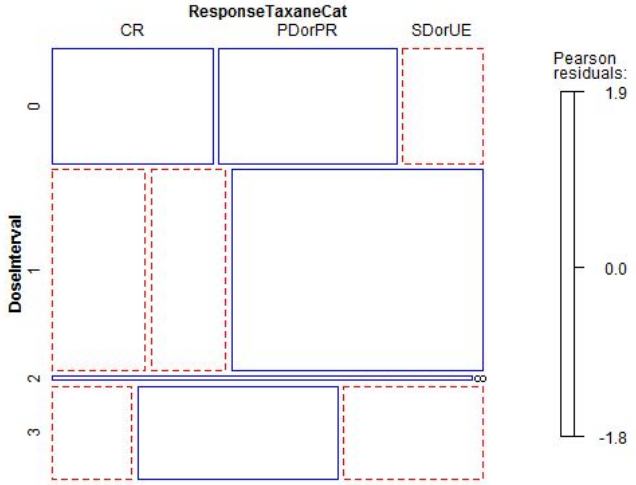
Highly Correlated Responses




Highly Imbalanced Classes



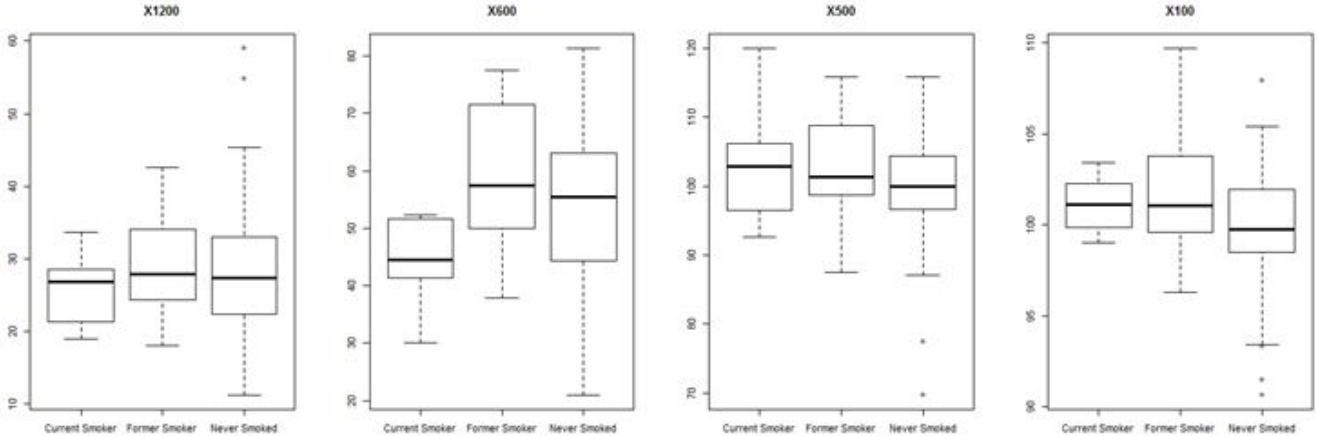

Empty Cells



# Manova: Contrasts

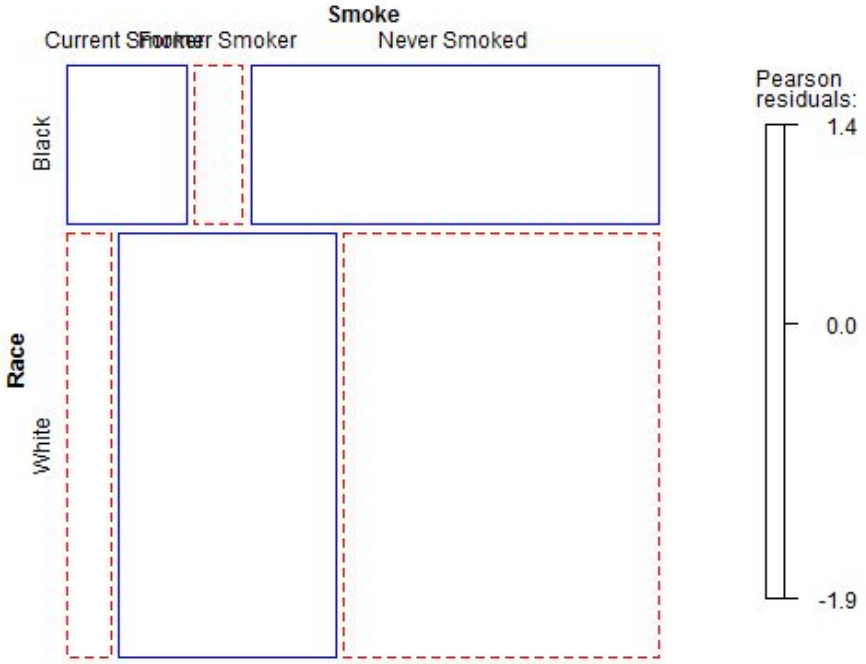
| Contrast | X1200 | X600 | X500 | X100 | p-value |
|---|---|---|---|---|---|
| **Former Smoker - Never Smoked** | 0.17 | 4.15 | 3.53 | 1.9 | 0.22 |
| **Former Smoker - Current Smoker** | 3.5 | 13.78 | 0.23 | 0.41 | 0.042 |
| **Never Smoked - Current Smoker** | 3.33 | 9.62 | -3.3 | -1.48 | 0.039 |

*bonferroni correction for multiple testing



# 2-Way Manova

```
Type II MANOVA Tests: Pillai test statistic
           Df test stat approx F num Df den Df  Pr(>F)
Smoke       2   0.19138   2.1164      8    160 0.03711 *
Race        1   0.13079   2.9717      4     79 0.02429 *
Smoke:Race  2   0.15168   1.6412      8    160 0.11702
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



# Model Fitting

Model:

Response: Recovery status (none, partial, full)

Baseline covariates: age, menopause, estrogen receptor, treatment regimen, number of cycles, total weeks, pre-experiment cancer stage, race, smoking status, Her2 (human epidermal growth factor receptor 2)

Additional covariates of interest: IC-50, dose response curve values at each of 10 doses (100, 250, …, 1200)

Does adding LCL data improve the ordinal logistic regression model of Recovery Status?

# Model Fitting Results

Evaluate models via AIC = 2p - 2ln(L)

| Model | AIC |
| --- | --- |
| Baseline | 173.8224 |
| Add IC-50 to the model | 175.1712 |
| Add dose-response curve values to the model | 186.7834 |

# Regression Trees

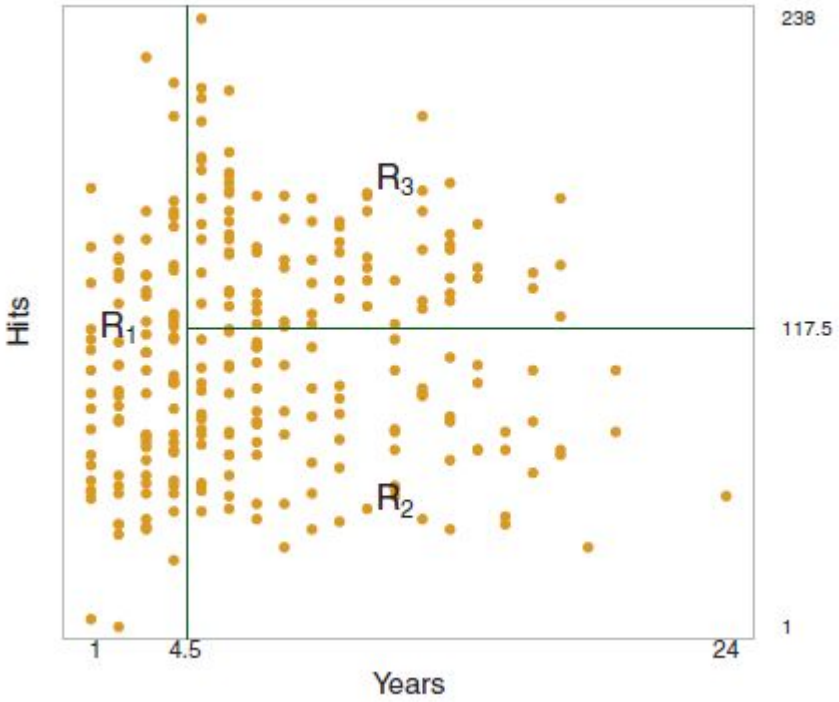
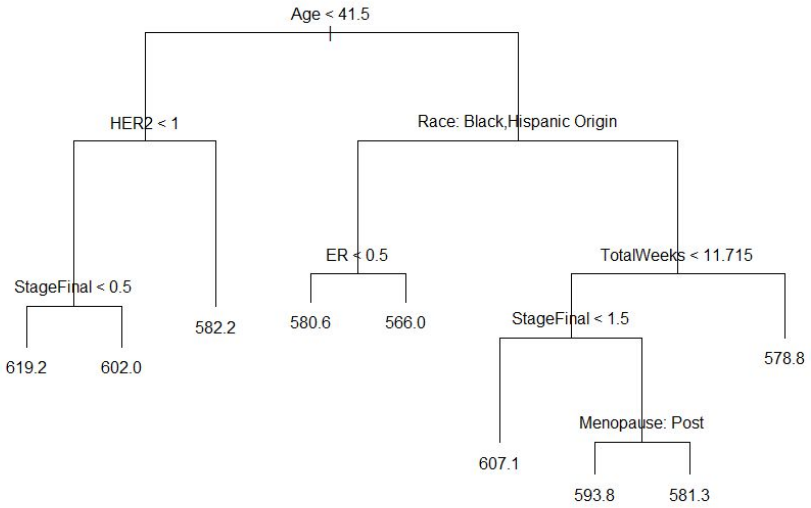
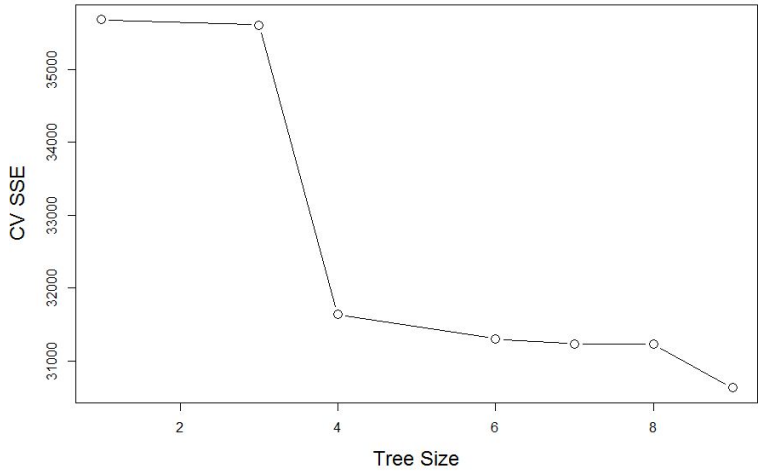



Figure from Introduction to Statistical Learning by Hastie et al.

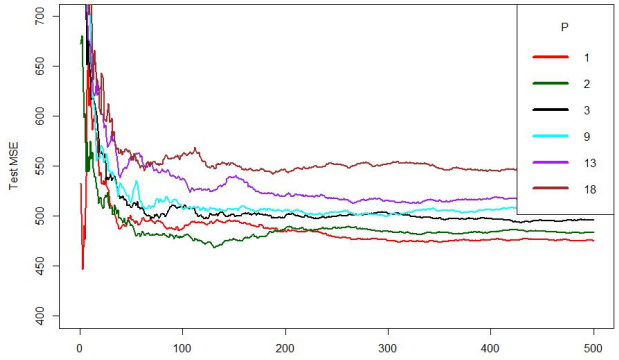
# Regression Trees

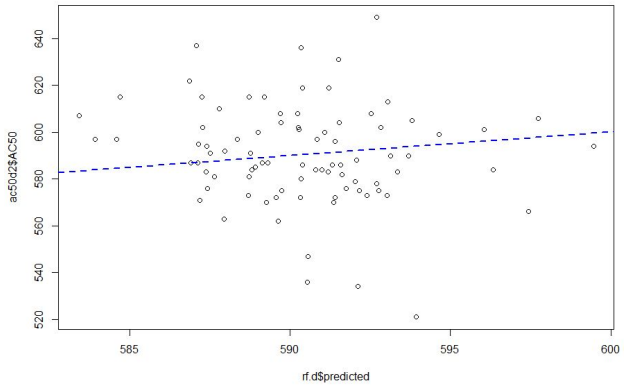


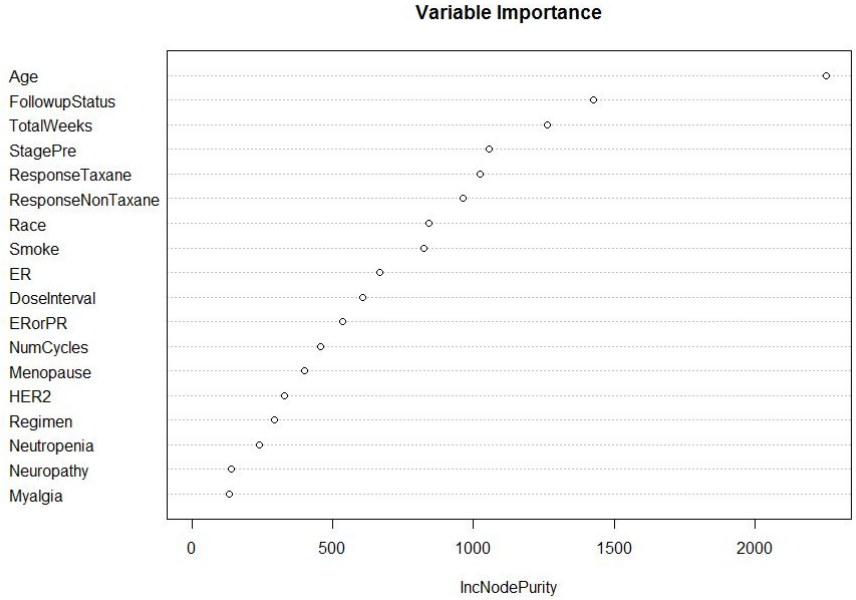




Average difference between actual and predicted IC50 for held out data: 55.3

# Random Forests




Average difference between actual and predicted IC50 for held out data: 21.9

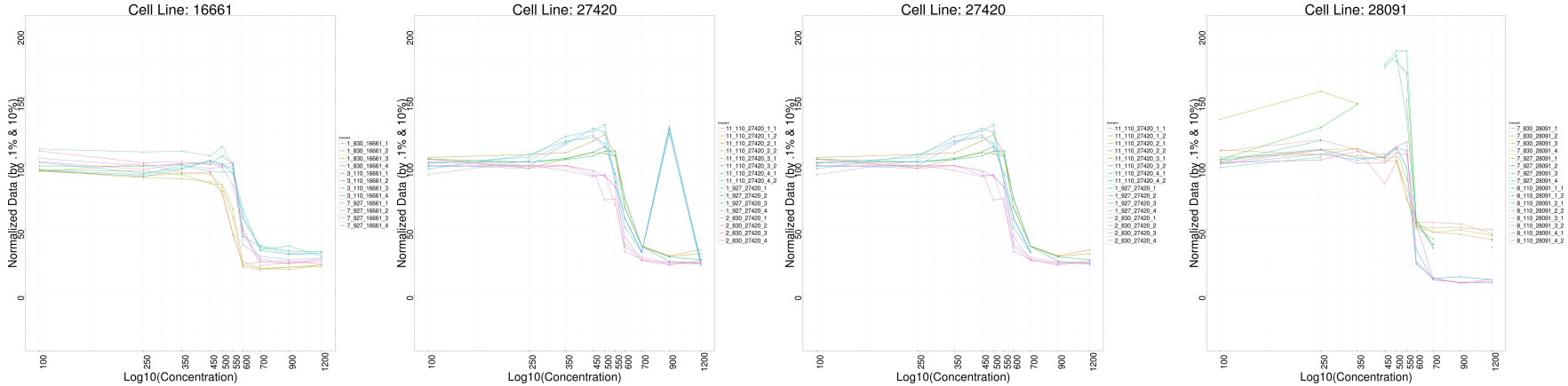
# Future Directions/Limitations

- Issues with QC pipeline
- Use model selection to find best linear model for predicting clinical outcome and IC50 response
  - Cross Validation error for comparison with tree based methods
- Regression trees and random forest model for predicting clinical outcome
  - Ordinal responses are tricky for tree based methods

# Conclusions

- Significance tests indicated that the variables smoking and race are associated with dose-response curves
- Regression trees showed some clinical outcome variables were important predictors of IC50 values after controlling for other patient data
    - Novel result with data that has never been collected before

# Issue With QC Pipeline




- May need to use alternate strategies
  - Regression smoothing
  - Median of replicates

# Outline

- Background: LCLs and Breast Cancer Treatments
- Research Questions
- Data
- Methods
  - Associations
- Data Exploration