Statistical Issues in Predictive Toxicology

Edit Kurali[1], M.Sc.
Weimin Li[2], Roger Brown[3], Stacey Jones[4], Kay Tatsuoka[2], Michal Magid-Slav[2], Steve Clark[4], David McFarland[3], Daniela Ennulat[3], Patrick Wier[3]

GlaxoSmithKline
Edit.2.Kurali@gsk.com

1 Discovery Statistics
2 Discovery Bioinformatics
3 Safety Assessment
4 Discovery Technologies

Outline

- Toxicogenomics
- Review of Statistical Methods
- Case Study: Improving Detection of Liver Toxicity in Pre-clinical Development
- Summary

Toxicogenomics in Pre-clincial Development

- Marriage of genomics (DNA microarrays) with traditional toxicology
  - quantifies global gene expression change
  - highlights the cellular pathways involved
- Enables to gain insight into complex biologic responses to drugs
- Can enhance well-established toxicity biomarkers
  - liver enzymes in serum: ALT, AST, etc
- ➔Better decisions in candidate selection studies (go, no-go decisions)

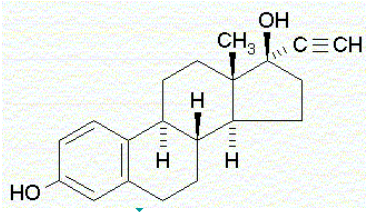
Data driven decision making in toxicology

Step1

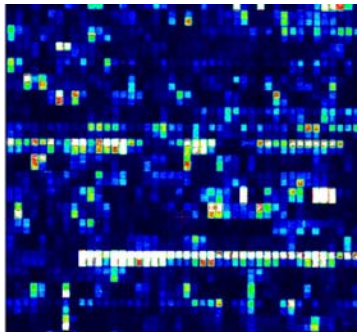
Clinical pathology

	ALP	ALT	AST	GGT	SBA	Tbili	Chol
	u/L	u/L	u/L	u/L	umol/L	mg/dL	mg/dL
1% methylcellulose	265.6	43.4	105.6	0.0	8.6	0.18	87.6
17-alpha-ethinylestradiol 30	536.4	38.4	93.6	0.2	10.2	0.14	*6.0*
17-alpha-ethinylestradiol 100	473.2	35.8	73.8	1.4	14.2	0.24	*3.4*
17-alpha-ethinylestradiol 300	523.5	*126.5*	153.0	*4.8*	14.0	*0.60*	*8.3*

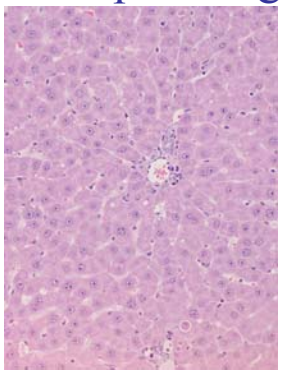
Structure/properties



Gene expression



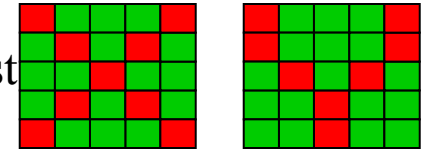
Histopathology



Step2

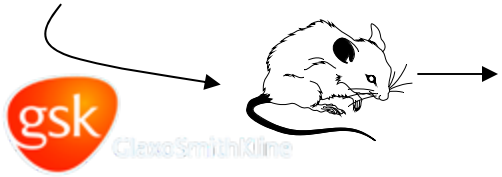
Analyze, Model, Validate/test

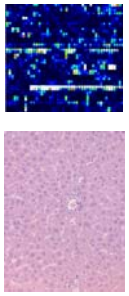
High Risk Low Risk

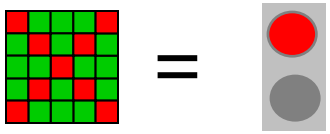


Step3

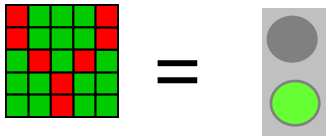
Candidate compound





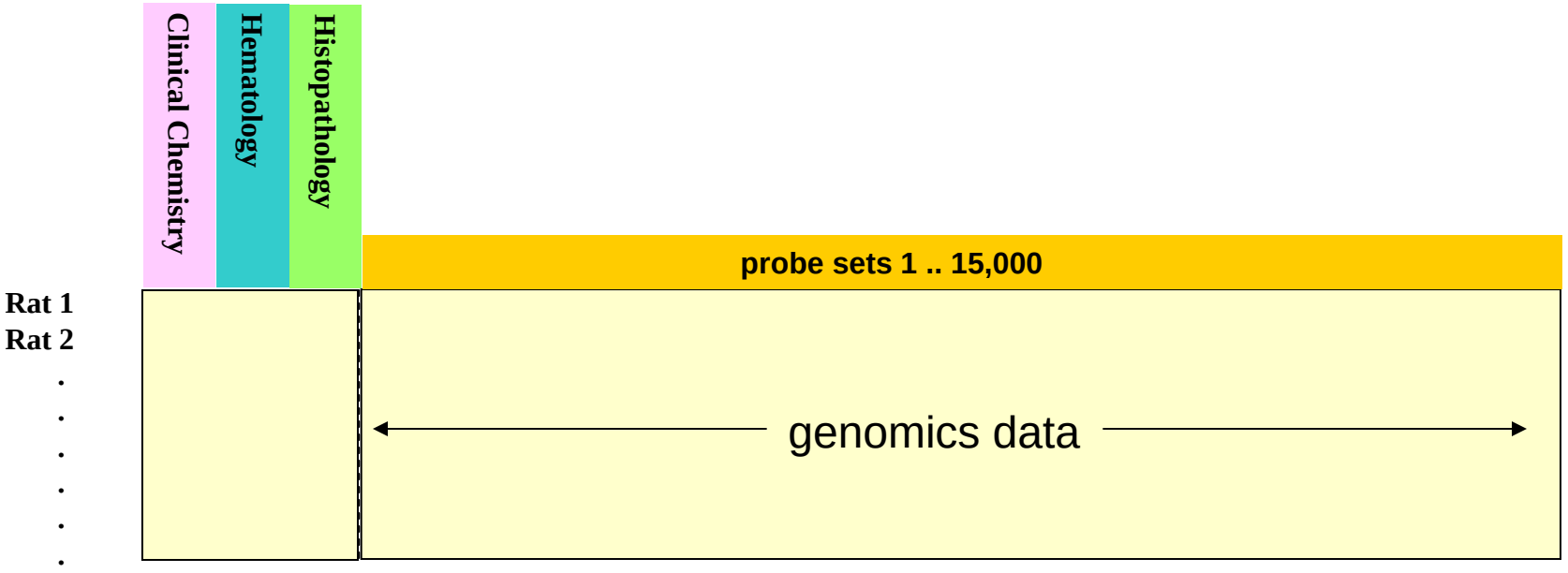


High Risk



Low Risk

High Dimensional Toxicogenomics Data

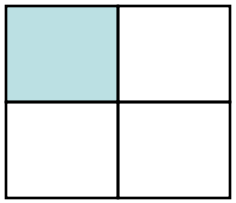


Need for dimension reduction

Statistical Methods
for Dimension Reduction

	Unsupervised	Supervised
Univariate	Univariate filtering, QC • Max./Min. Intensities • Nuisance variability 1	Hypothesis test based selection • t-test/ANOVA/Mixed Model • Correlation with the response • False discoveries 3
Multivariate	Overview of the data for global patterns • PCA • Clustering •Summary across variables 2	Multivariate predictive modeling • Shrinkage methods • Model averaging methods • Projection methods • False discoveries 4

Univariate Unsupervised Analysis

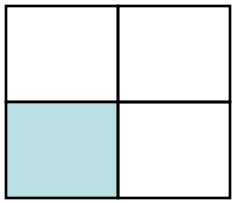


- Goal: To remove irrelevant or noisy variables without using the response information

- Methods
  - Intensity filter (signal too low)
    - Requires in-depth knowledge of the platform-Affymetrix, Taqman...
  - Nuisance variability filter (too much variation, “noisy genes”)
    - Variance components analysis
    - Robust statistical methods:
      - Summary by median, IQR
      - Variance estimation by Winsorizing

- Dimension reduction by removing non-informative variables

Multivariate Unsupervised Analysis

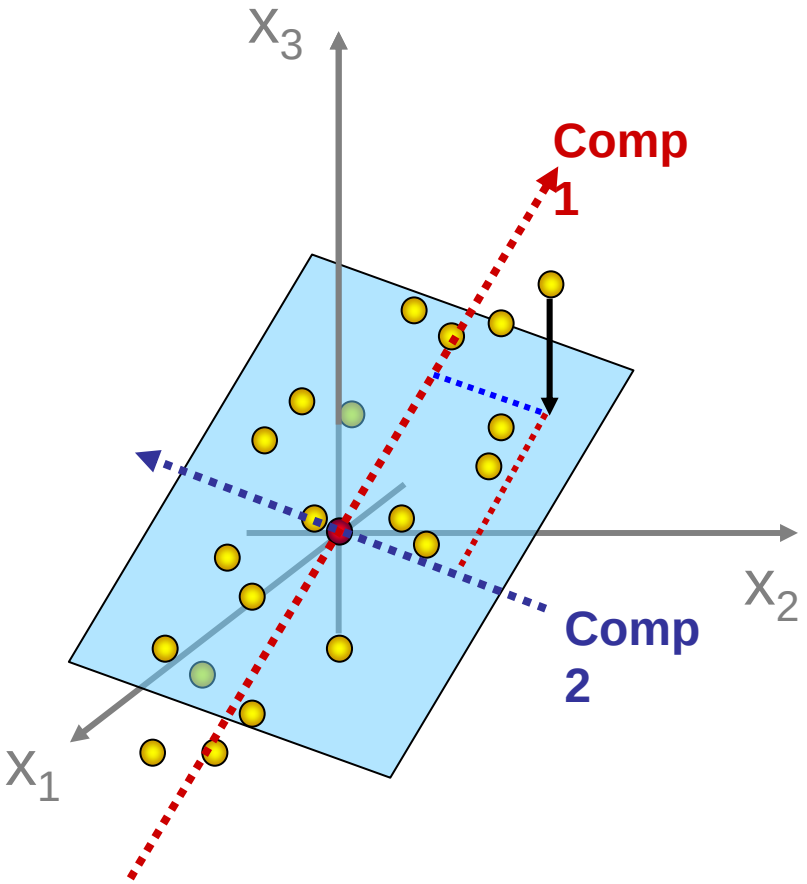


- Goal:
  - Overview the data for patterns and gross outliers
  - Find relationships among variables
- Method
  - Principal component analysis (PCA)
  - Cluster analysis
  - Summary measures
    - Toxicity Index: Rogatko et al, Clinical Cancer Research Vol. 10, 4645-4651, July 15, 2004
  - Dimension reduction
    - removing outlying subjects
    - grouping correlated variables into biologically meaningful categories
    - reducing number of variables into a few dimensions

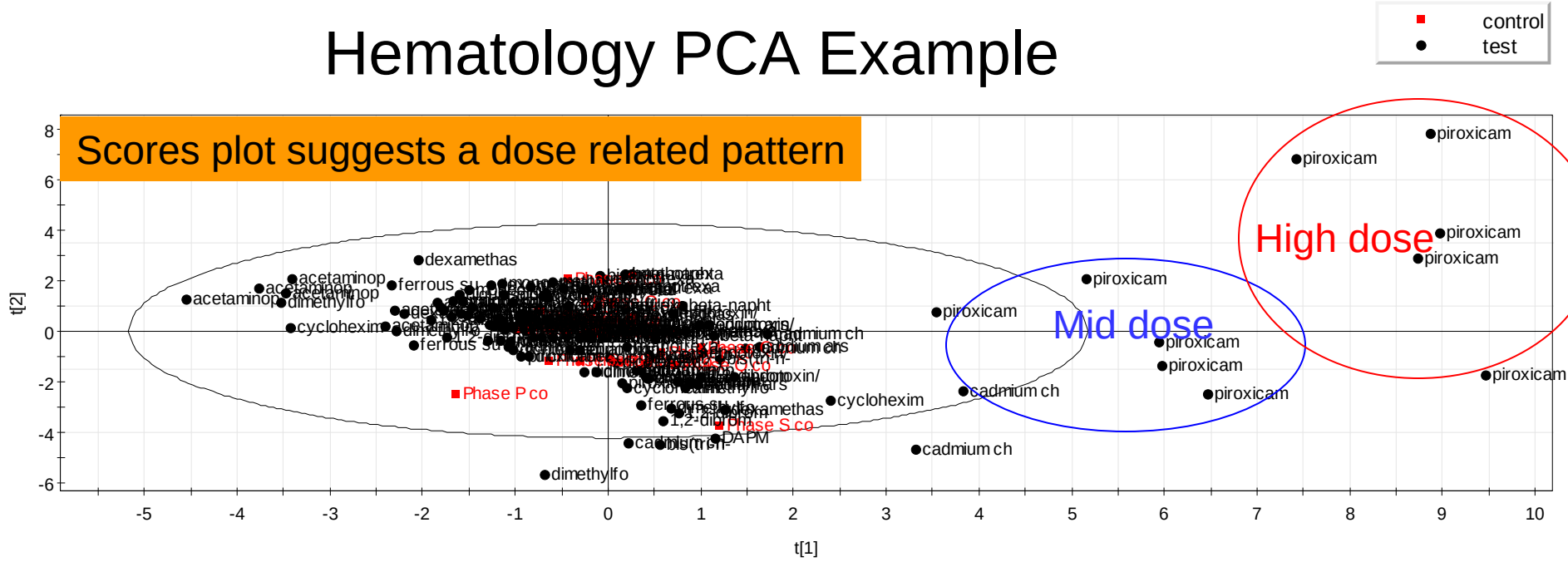
Principal Component Analysis:

Rotation and Projection

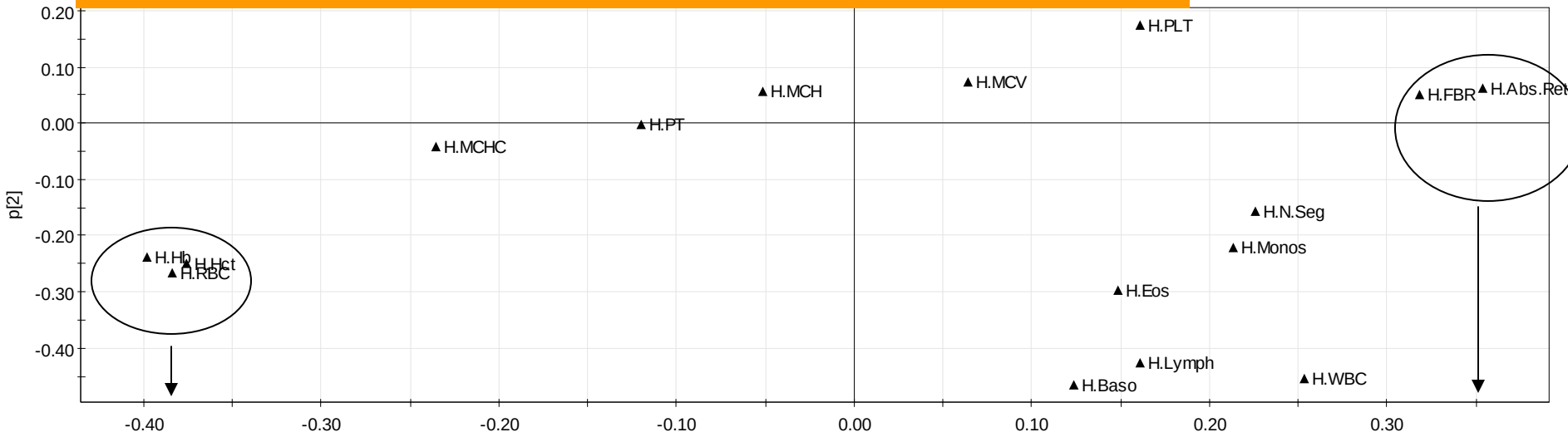
- Rotation results in a “new axes system”
  - Find direction of most variation (linear combination of original variables)
  - orthogonality
  - **Loadings** are the contribution of the original variables to the new axes
- Projection onto this new axes system:
  - **Scores** are the coordinates of the data projected to the new coordinate system
- Dimension reduction to only a few dimensions
  - Easy identification of data structure, patterns, outliers, etc…



Hematology PCA Example



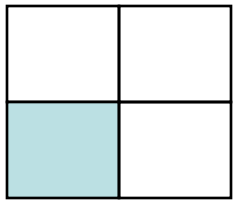
Loadings plot helps identify the underlying biological process:



Low red blood cell count

High reticulocyte count

Cluster Analysis



Discover groupings or patterns in the data

Distance Measures

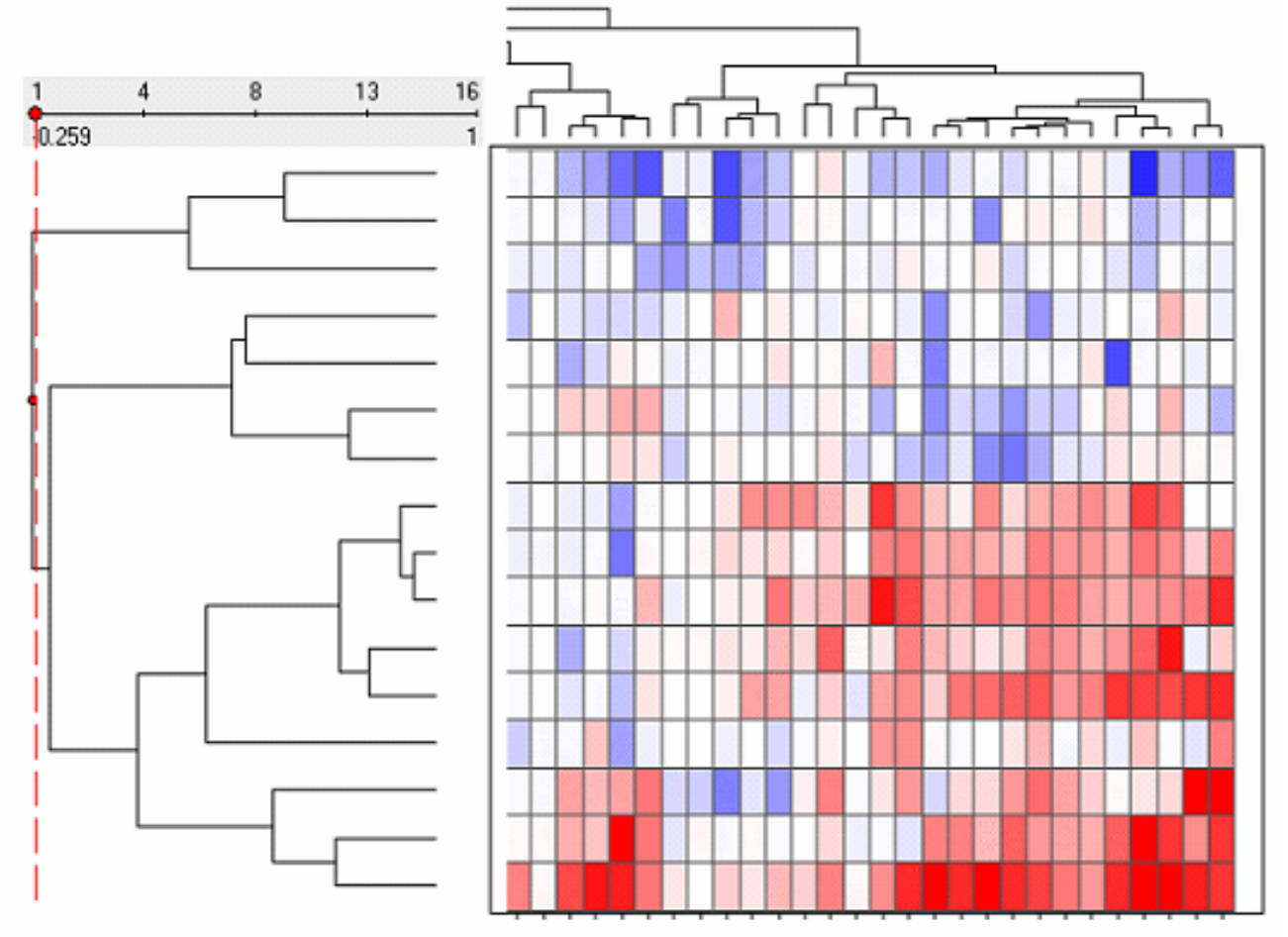
- Euclidean
- Correlation

Methods

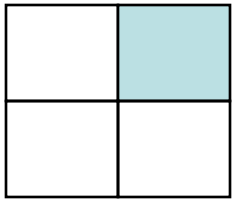
- Hierarchical
- Partitioning

Linkage

- Single
- Average

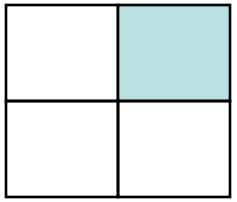


Univariate Supervised Analysis



- Goal:
  - Rank order variables in high dimensional studies
  - Reduce the number of variables for predictive modeling
- Analysis models
  - T-test
  - ANOVA
  - ANCOVA
  - Repeated measure model
  - Trend Analysis
- Issues to be considered
  - Transformations
  - False discoveries
- Dimension reduction through informed analysis

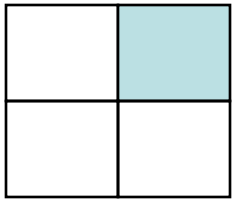
Trend Analysis



- Widely used in toxicology experiments:
  - response is often expected to be predicted by dose (ordered variable)

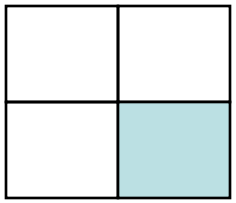
- Methods often used:
  - Jonckheere-Terpstra procedure
  - Cochran-Armitage trend test
  - Shirley's test
  - Williams test
  - Tukey's NOSTASOT

Trend Analysis



- Toxicogenomics
  - Dose-response:
    - Trend contrast analysis with respect to dose (vehicle, low, mid, high)
  - Time course:
    - Trend contrast analysis with respect to time to identify genes with biologically relevant time course pattern

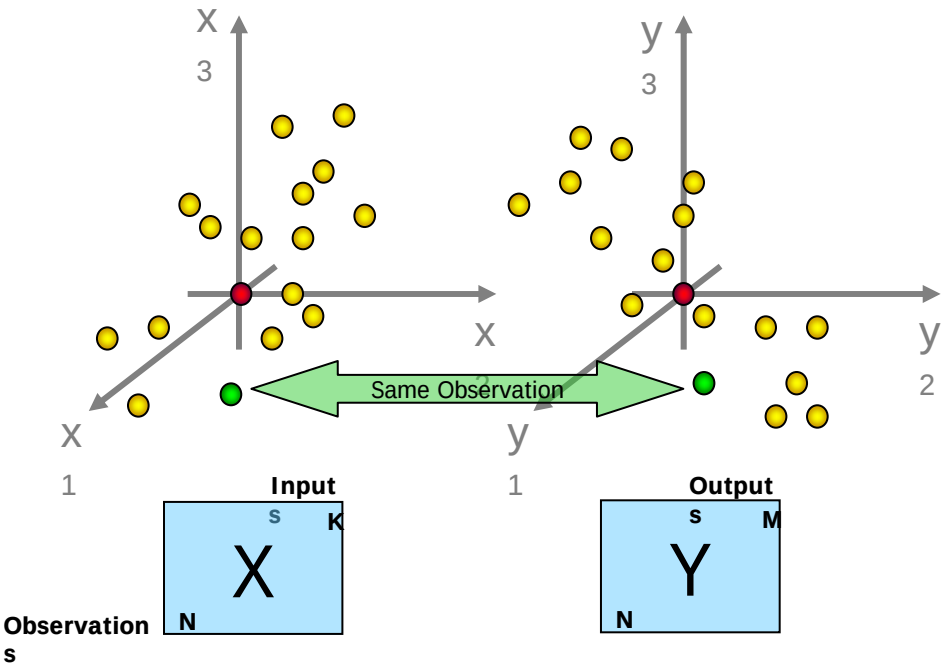
Multivariate Supervised Analysis

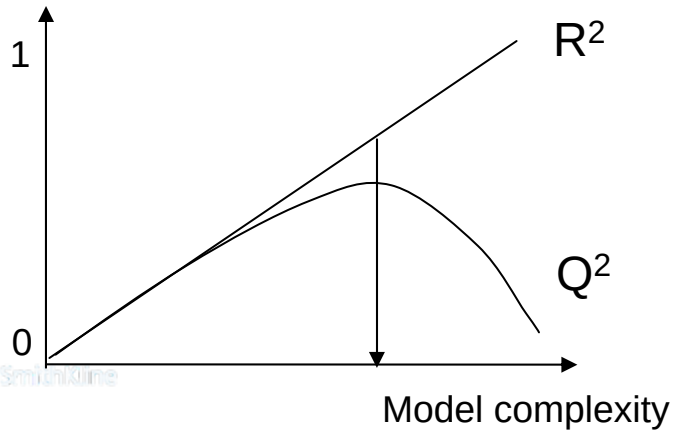


- Goal: Identify a small set of variables (including co-expressed variables) that are predictive of the endpoint of interest
- Types of Models
  - Projection methods
    - PLS/PLS-DA
  - Shrinkage regression models
    - Ridge regression
    - LASSO
    - Elastic Net
  - Model averaging approaches
    - CART
    - Random Forest
- Issue: False discoveries
- Dimension reduction through informed analysis and targeting at deriving a small set of predictive variables

Partial Least Squares Regression (PLS)

- Simultaneous PCA of X and Y
  - Constrained by maximizing the covariance between response and prediction components
- Diagnostics
  - goodness of fit (R^2)
  - predictive ability (Q^2)



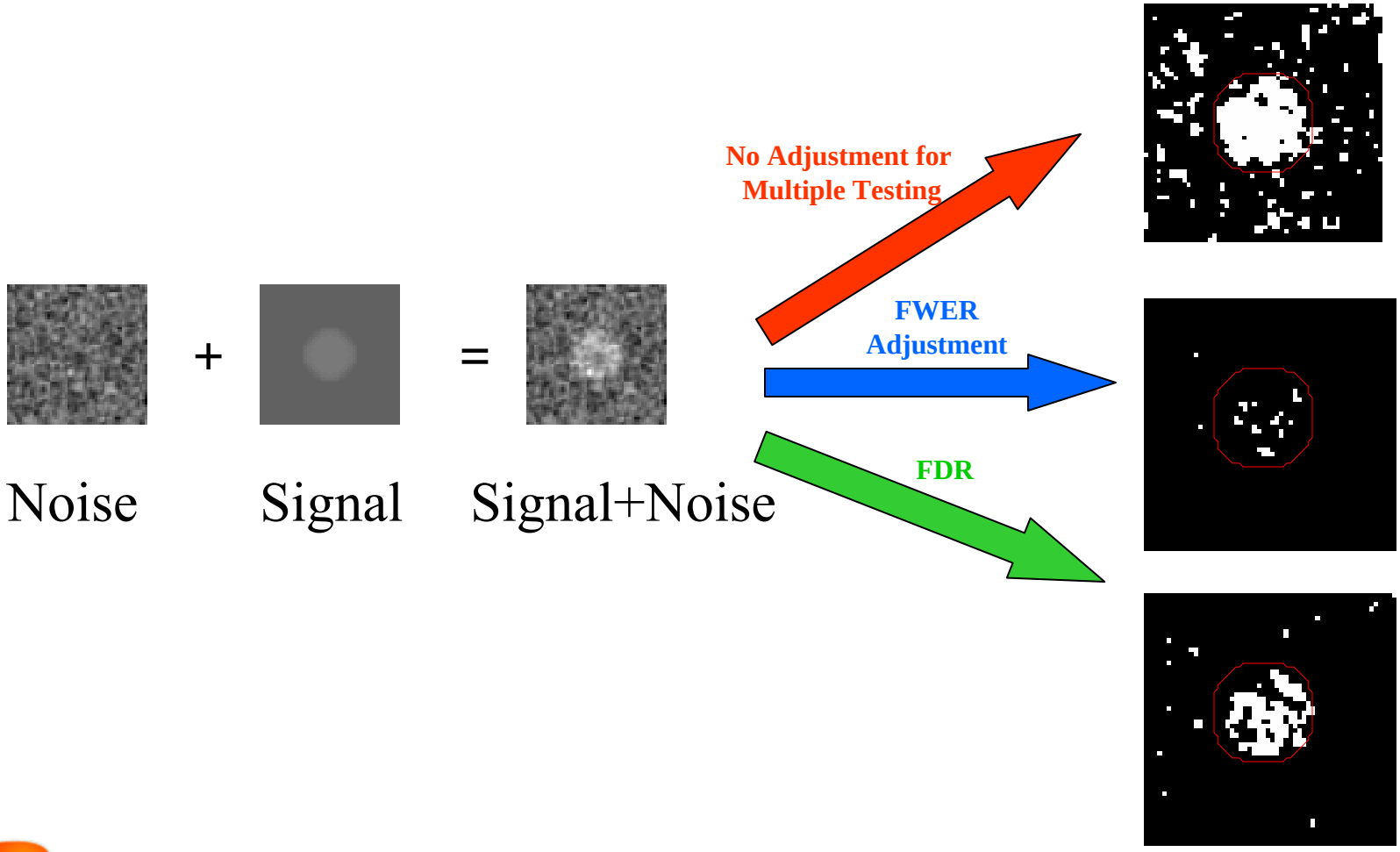


False Discoveries

- Univariate case:
  - large number of hypothesis tests ➔high chance of false positive results
- Multivariate case:
  - Overfitting, selection bias ➔ false clusters, biased estimates of prediction accuracy

Controlling the False Discovery Rate (FDR)

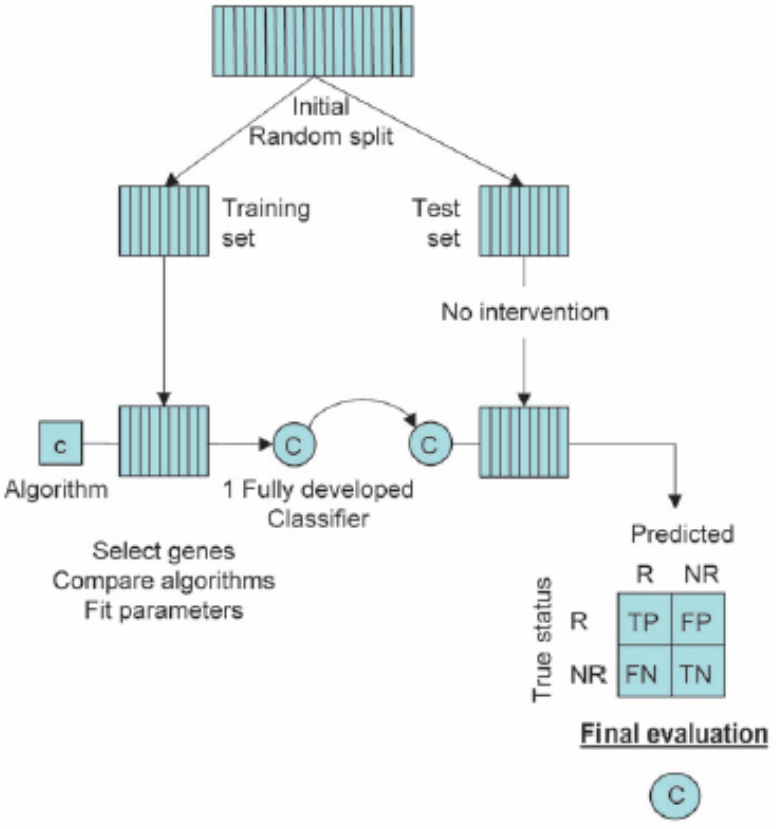
Benjamini and Hochberg, 1995



Multivariate Model Validation

A. Dupuy, R Simon (JNCI v99 (2), 2007)

- Improper validation is a common flaw in many published microarray studies
- Fundamental principle: the samples used for validation must not have been used in any way before being tested.

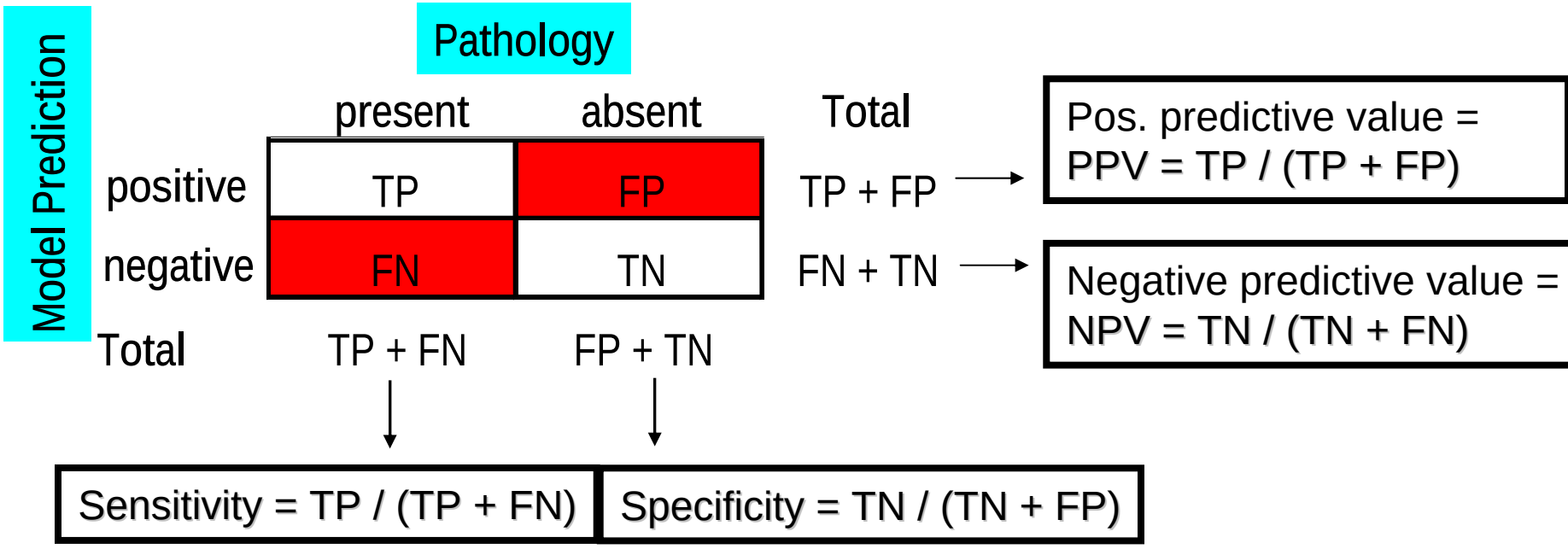


Model Validation Approaches

- Split sample (training set, test set)
  - Develop classifier in training set
  - Test set is only used to evaluate the classifier
- Cross validation
  - Iterative process
  - Ex: “leave-one-out”, k-fold CV
  - Gene selection steps need to be internal to the CV loop
- Dual-validation (CV + additional independent samples)

A. Dupuy, R Simon (JNCI v99 (2), 2007)

Model Evaluation

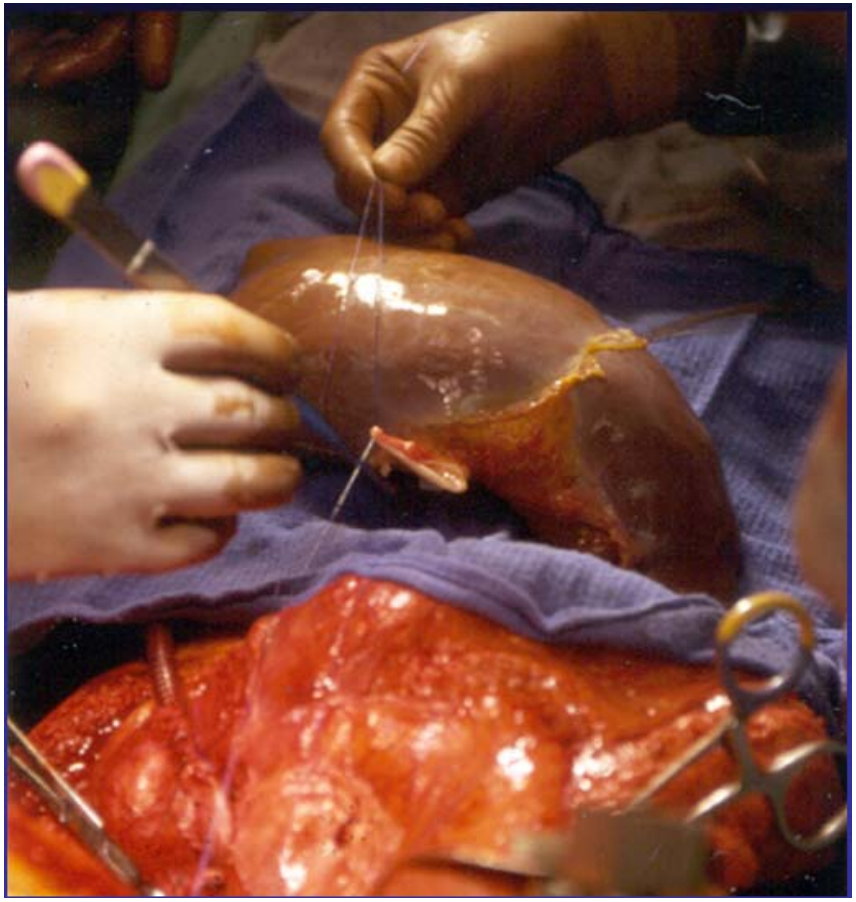


Case Study
Predictive Toxicology Project at GSK

- Goal: develop a gene panel to aid in screening for liver toxicity in candidate selection studies
- Stages of analysis
- Illustrate application of statistical methodology

Why Hepatotoxicity?

- In the United States, drug-induced liver injury (DILI) is the leading cause of acute liver failure (ALF)
  - disease of the developed word
- In the pharmaceutical industry, liver toxicity is the number one cause for
  - terminated development
  - non-approval
  - withdrawal
  - label warnings
- Earlier prediction is important!



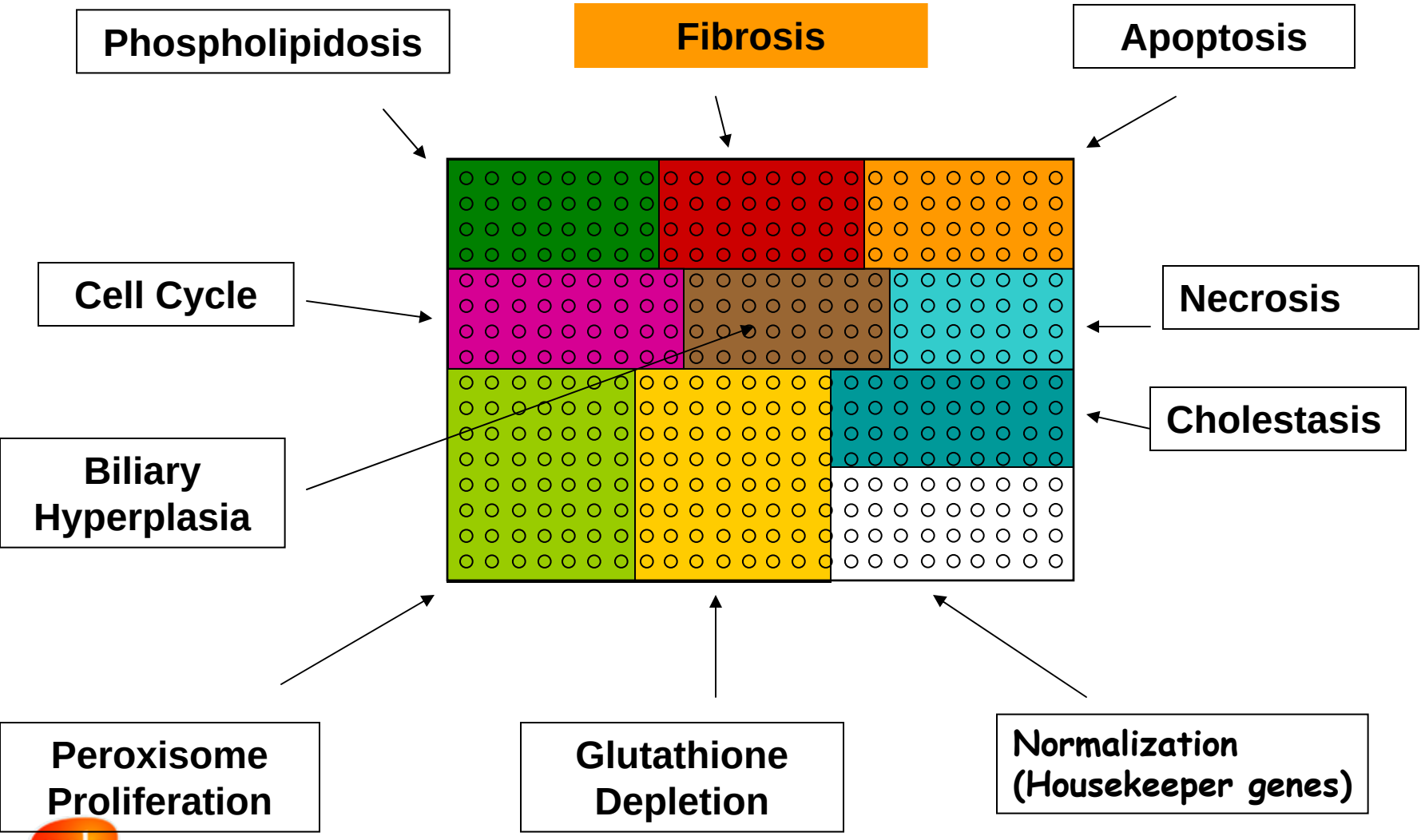
Toxicogenomics Data Collected at GSK

- > 200 compounds from literature manifesting wide variety of liver toxicities
  - multiple dose levels: vehicle, low, mid, high; admin. daily for 4 days
  - Time course study (8 weeks) with a single dose reference compound for fibrosis

- Traditional endpoints:
  - clinical chemistry (ALT, AST, etc)
  - Hematology (RBC, WBC, etc)
  - Histopathology of liver

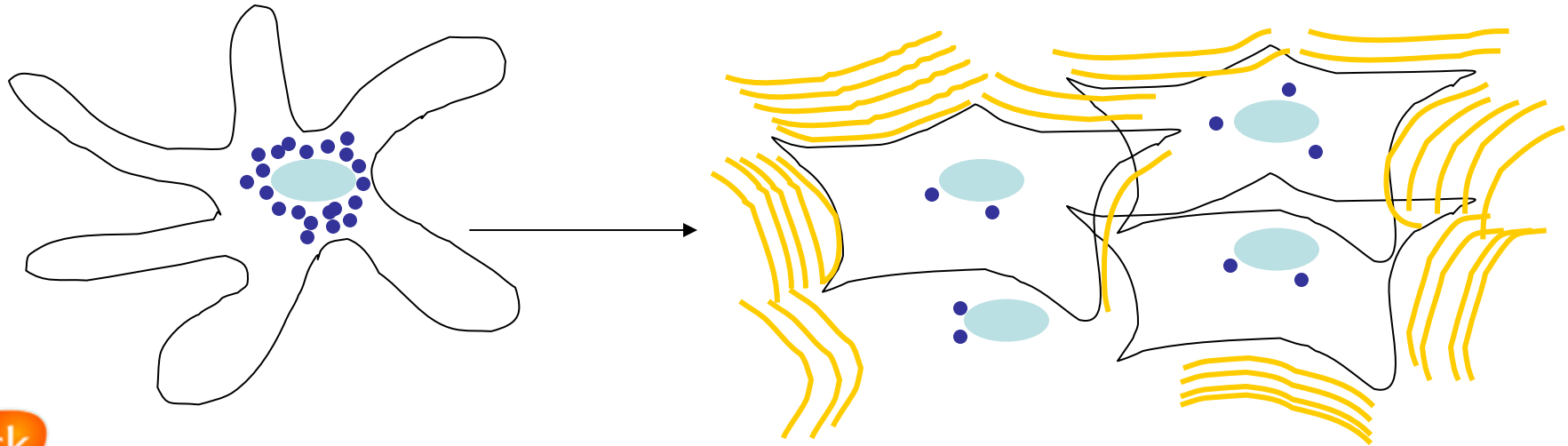
- Liver gene expression with Affymetrix rat microarray (~ 15,000 genes)

Identify Multiple Panels to Screen for Multiple Manifestations



Why Fibrosis?

- 8th leading cause of death in US
- Develops after repeated and persistent insult due to a toxic agent (ex. alcohol)
- Repair process ➔ stellate cells are activated ➔ fibrous scars formed ➔ disrupted architecture ➔ loss of liver function (irreversible cirrhosis)
- Histopathology not easy to detect



Quiescent stellate cell

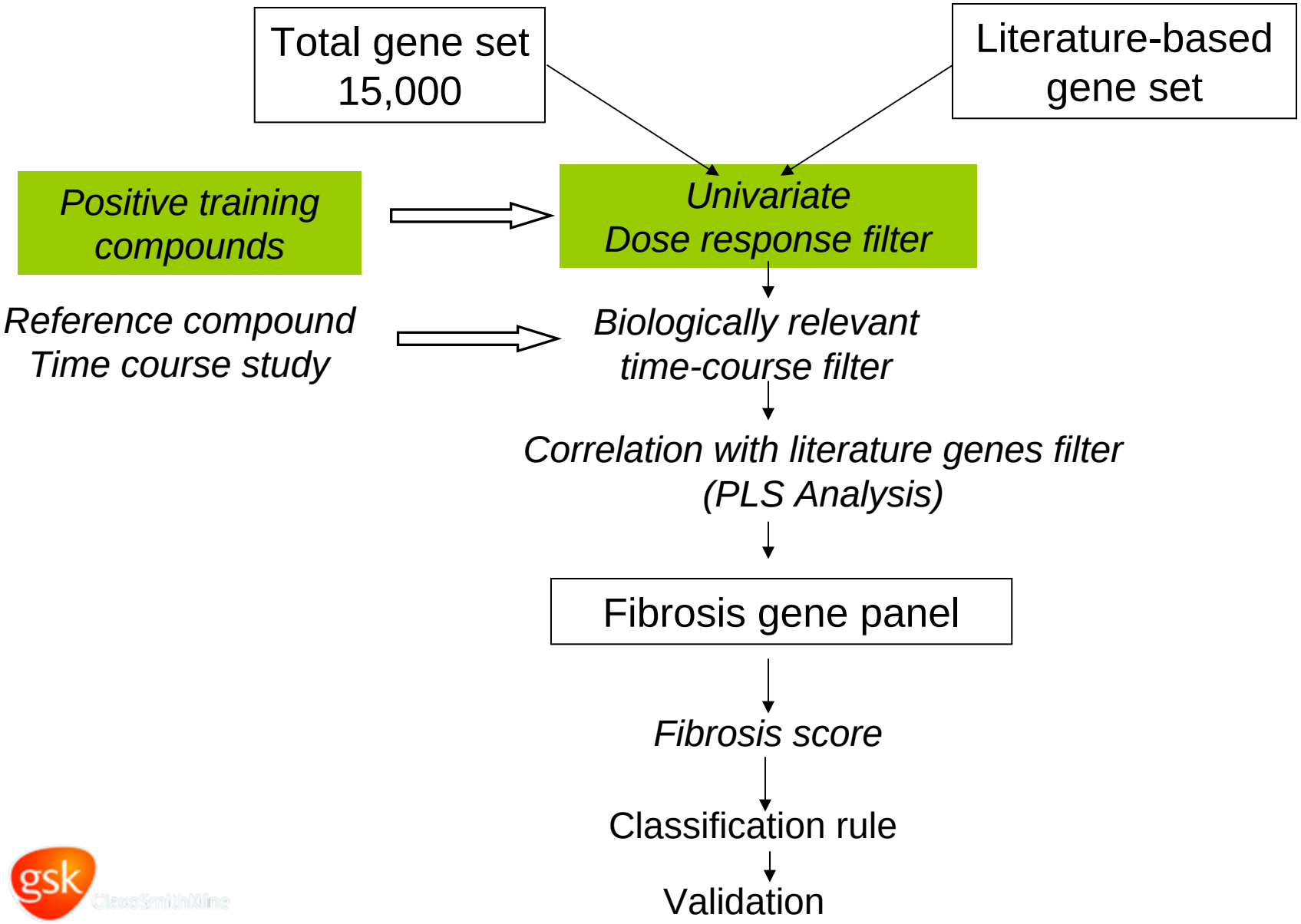
Activated myofibroblast

The challenges

- How to use rat 4 day studies to address a chronic (weeks to months) manifestation?
  - Chronic time course study with reference compound for fibrosis
- How to distinguish genes specific for fibrosis/HSC activation from genes involved in non-specific processes?
  - Careful selection of training compounds
  - Use of fibrosis specific genes identified from literature
- How to reduce the dimensionality of the data?
  - Using statistical methods for dimension reduction

required cross-disciplinary team effort

Stages of Analysis

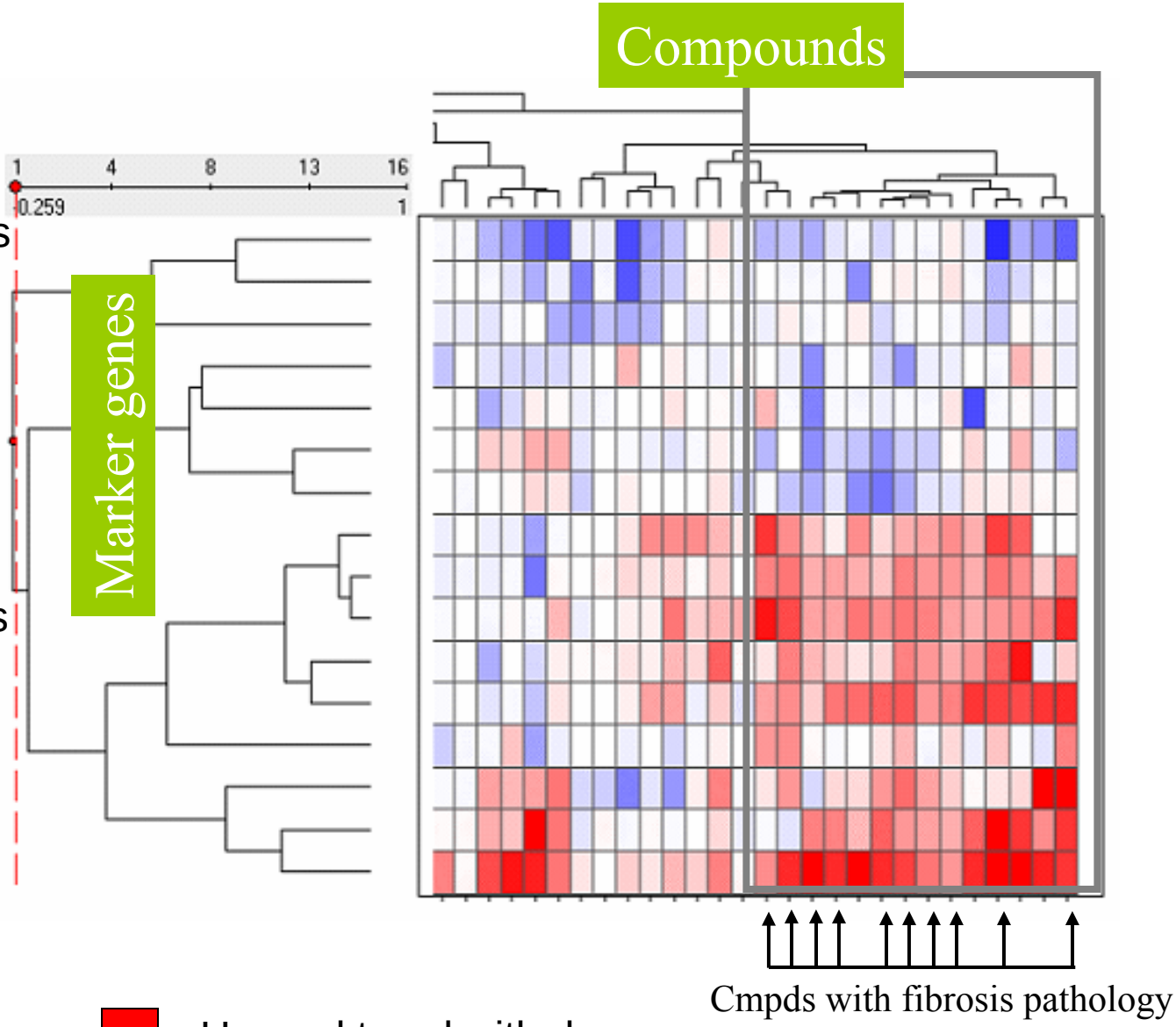


Fibrosis Positive Training Compounds

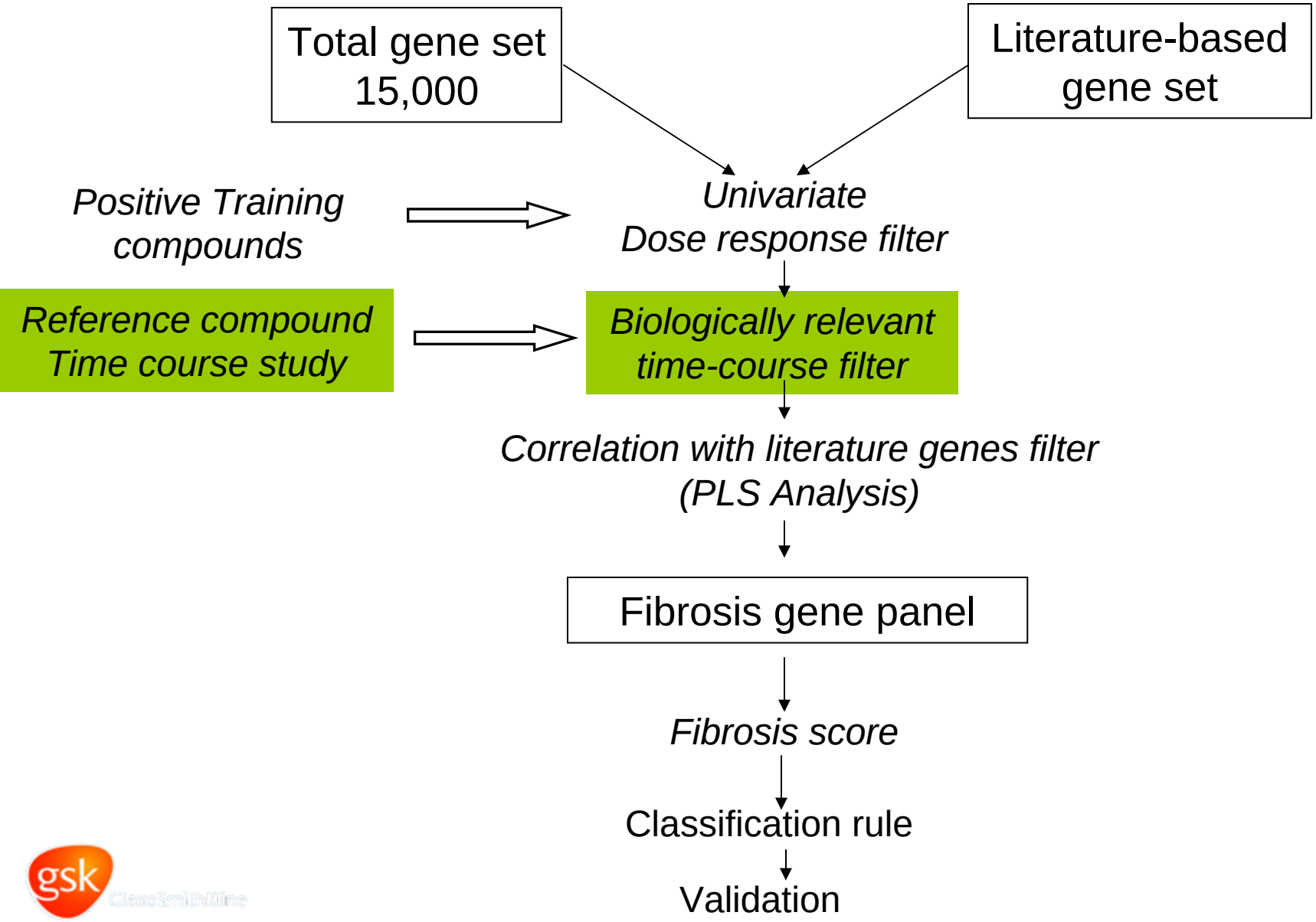
- Select fibrosis positive compounds
  - histopathology
  - literature evidence
  - dose related trend in key marker genes from literature (via clustering)
- Divide fibrosis positive set into training and test compounds
- Screen genes for dose related trend in training compounds

Hierarchical Clustering of Compounds and Literature Genes

- 2 way clustering of dose response scores (p-value transform)
- Additional fibrosis positive compounds were identified based on expression profiles of fibrosis marker genes

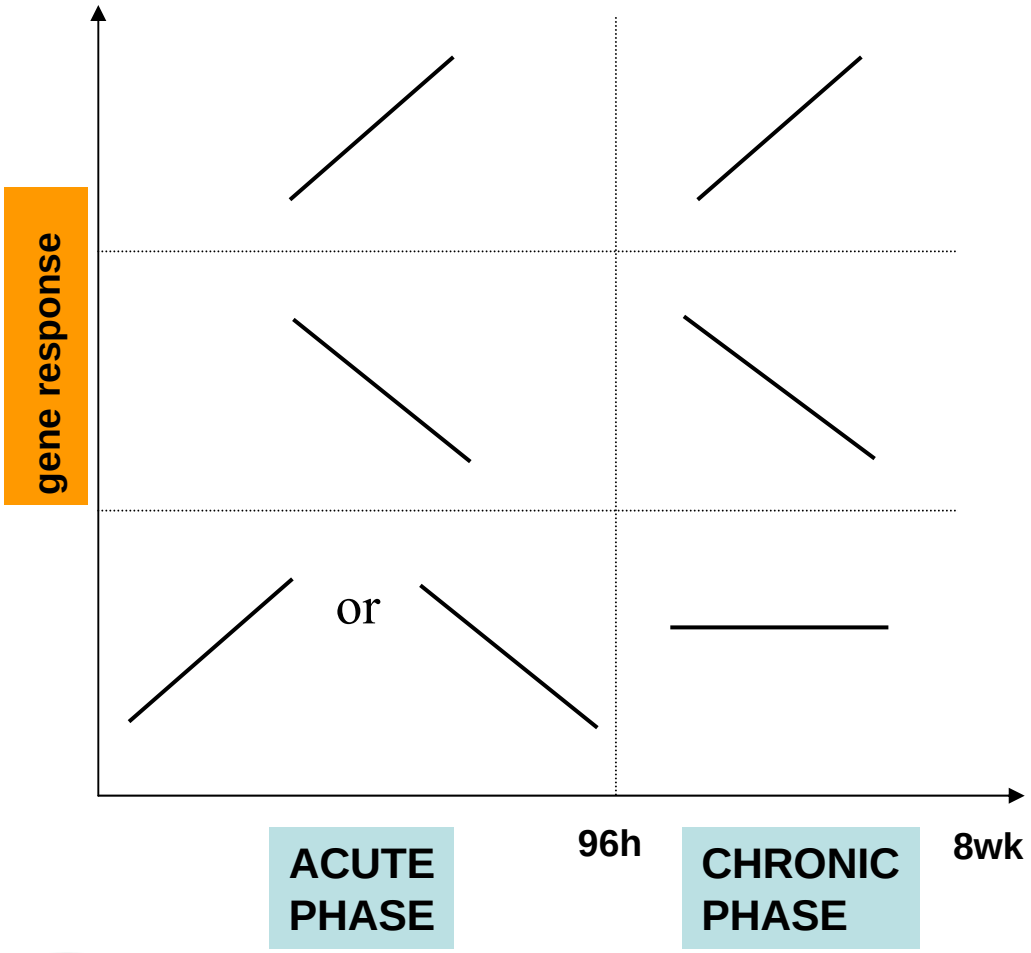


Stages of Analysis



Screen genes for biologically relevant time course pattern

using trend contrast FDR p-values

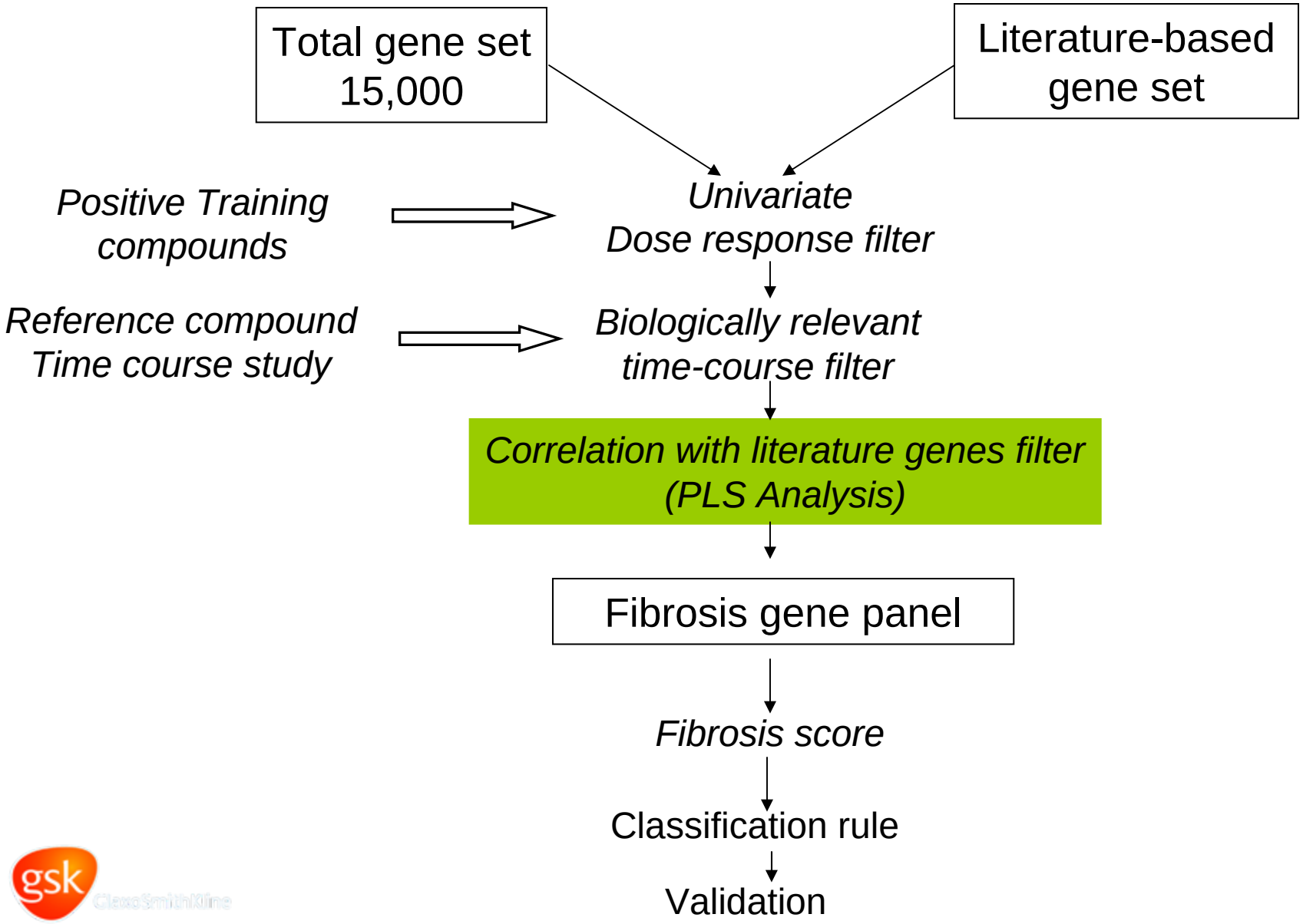


Biologically relevant

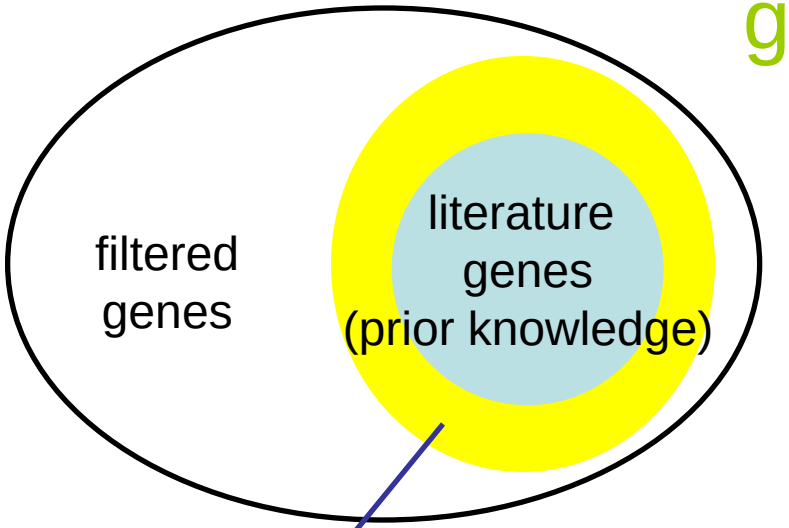
Biologically relevant

Not biologically relevant for fibrosis (acute response)

Stages of Analysis



PLS analysis:
finding genes correlated with literature genes



Any genes “correlate” with the “literature genes”??

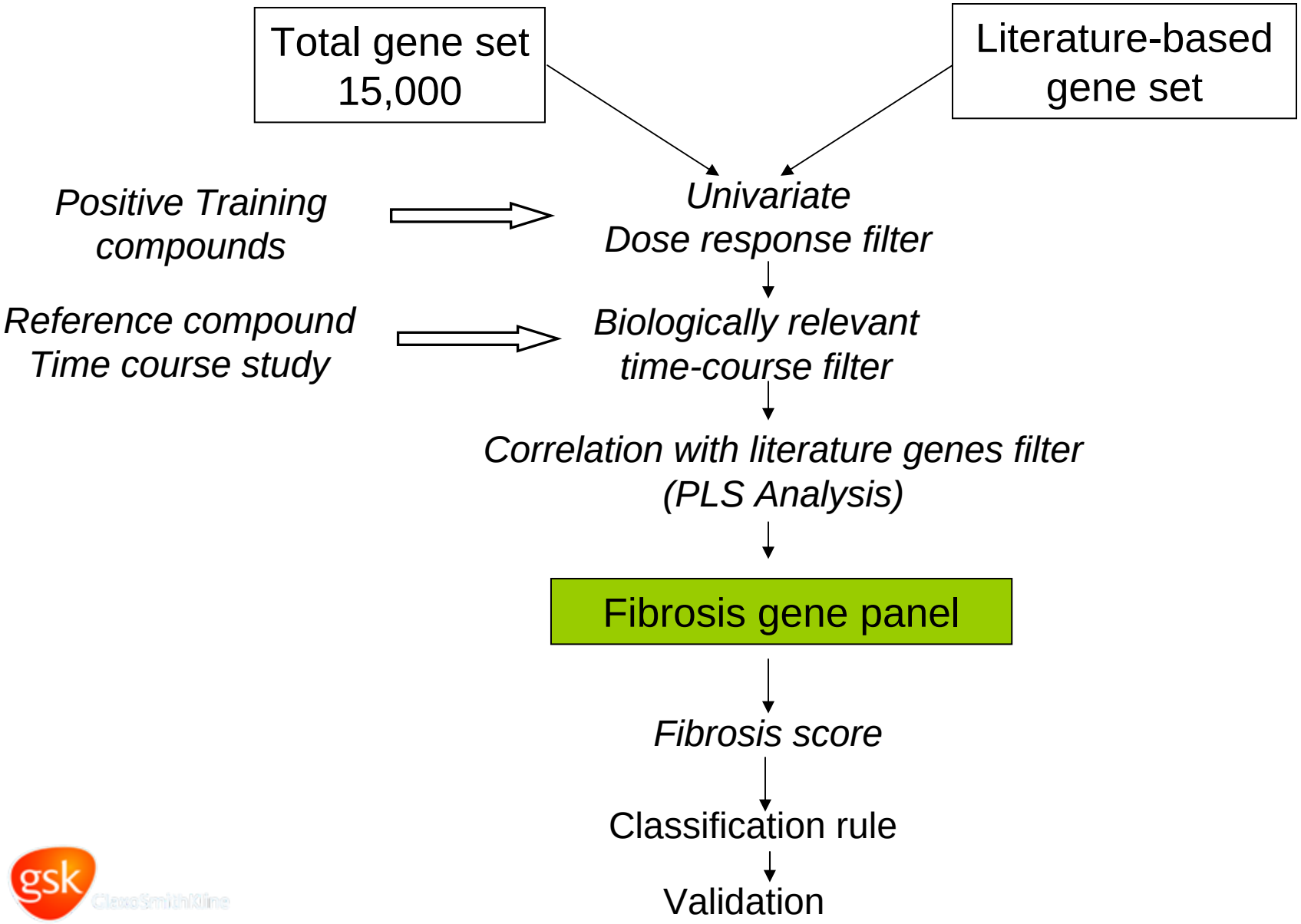
Partial Least Squares (PLS) analysis:

- Compromise of PCA and multiple regression
- Simultaneous dimension reduction of X and Y variables
- Constrained by maximizing the covariance between response and prediction components

$$X = TP' + E$$
$$Y = UC' + F$$

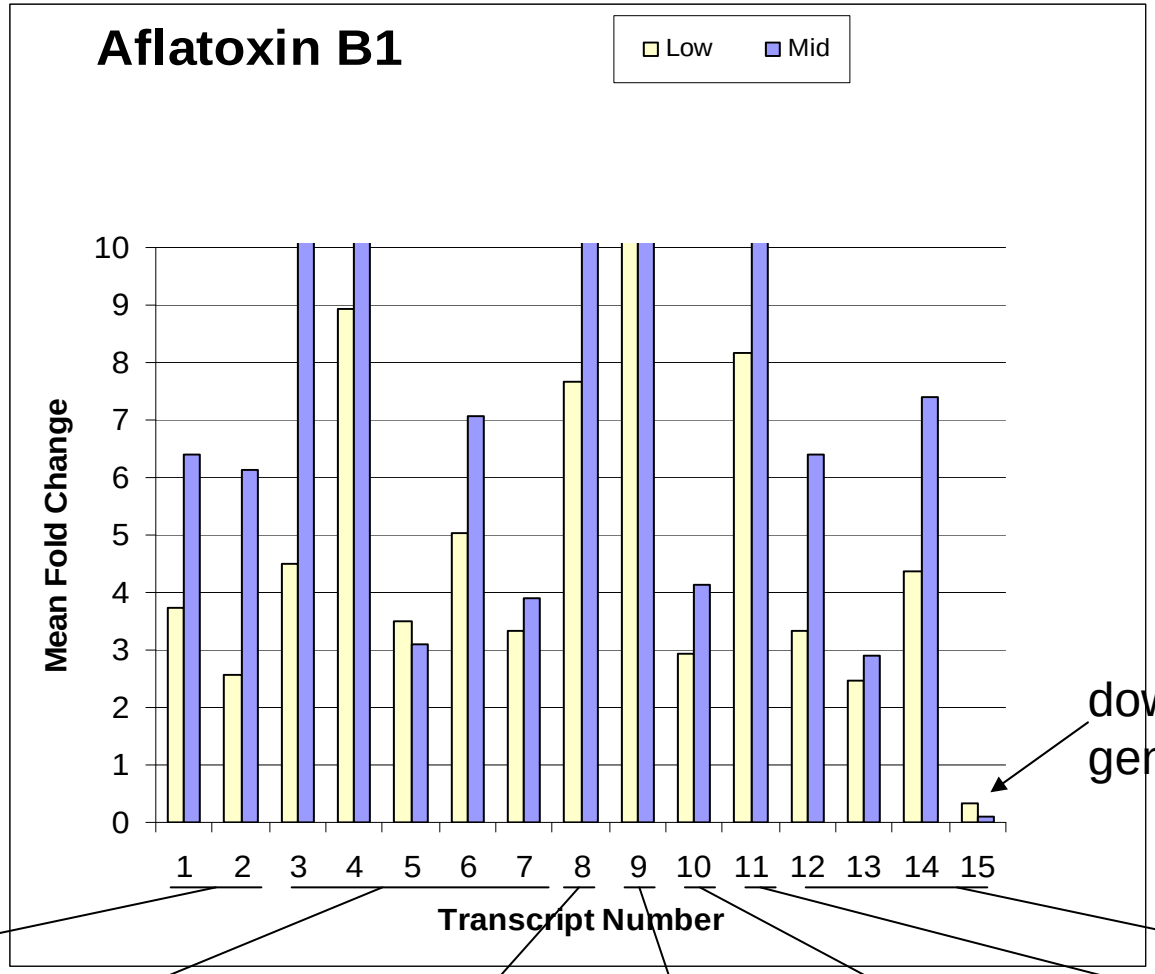
scores loadings residuals

Stages of Analysis



Fibrosis Gene Panel:

Diverse Fibrosis Specific Biological Processes Represented



down-regulated gene

collagen component

stellate cell activation

↓ collagen proteolysis

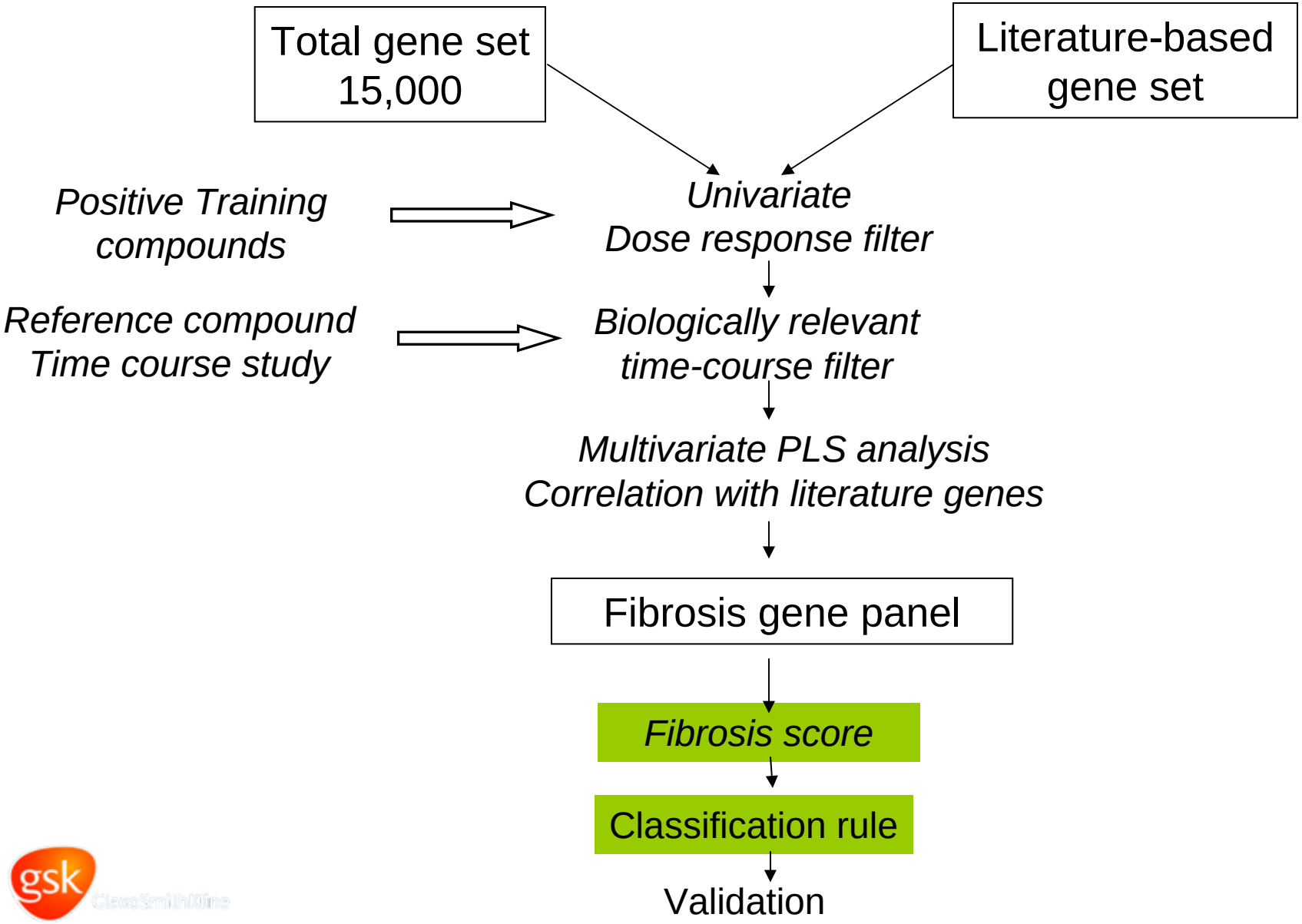
degrades elastin

conn. tissue remodeling

↑ expression in cirrhosis

?

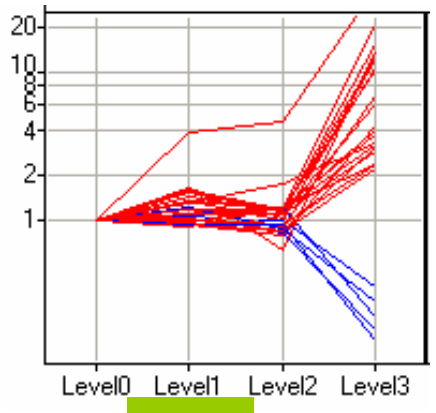
Stages of Analysis



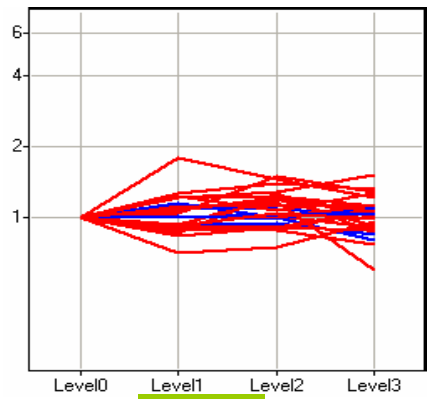
Fibrosis Score and Classification Rule

- Score is a summary measure of a compound's effect on the gene panel
  - Higher score means higher risk for fibrosis
  - Threshold determined by sensitivity/specificity analysis in discriminating positive training and control compounds
- Classification rule:
  - score > threshold ➔ high fibrosis risk
  - score < threshold ➔ low risk for fibrosis

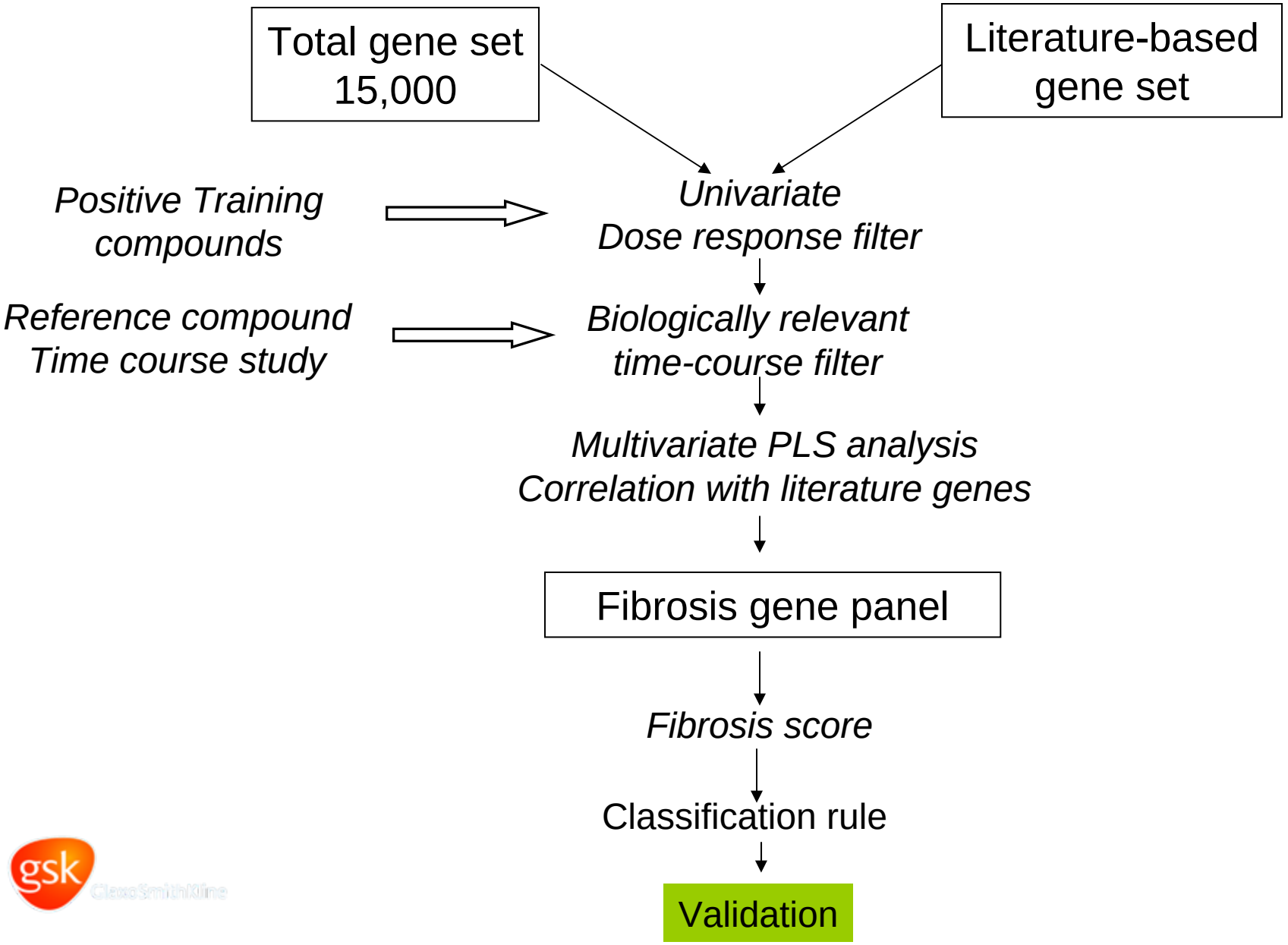
High risk compound



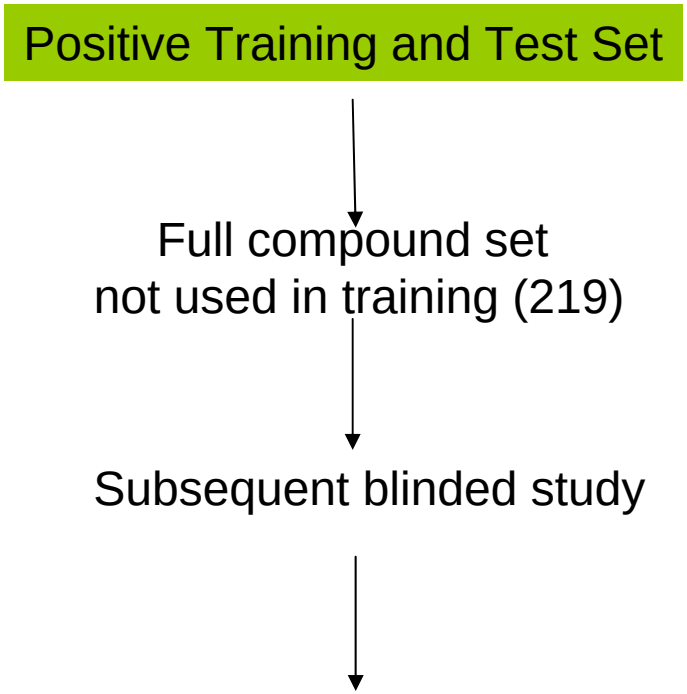
Low risk compound



Stages of Analysis

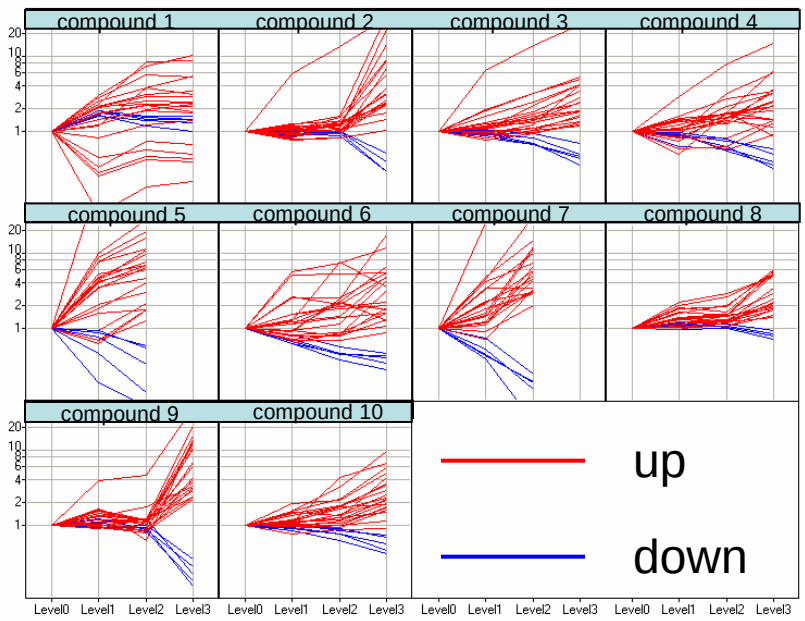


Model Validation



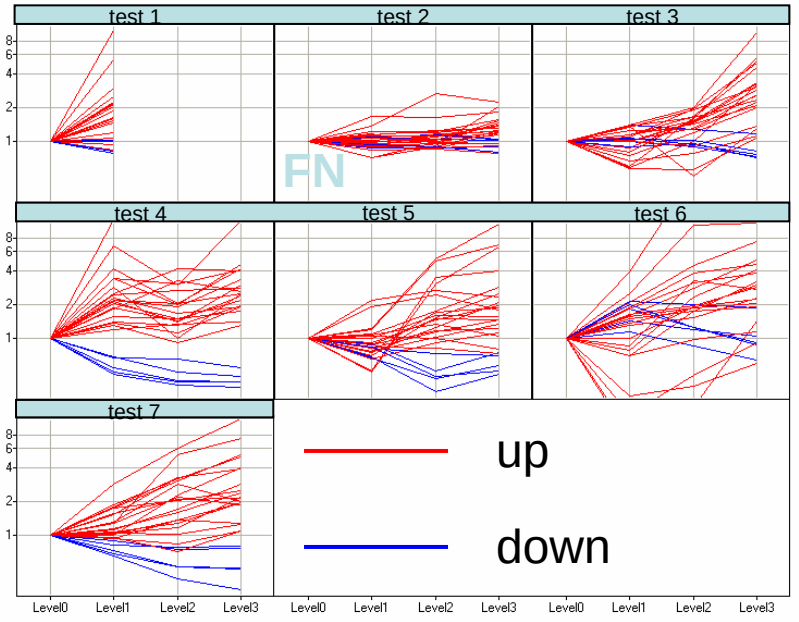
Classifier Performance Visualization

Fibrosis Positive Training Set



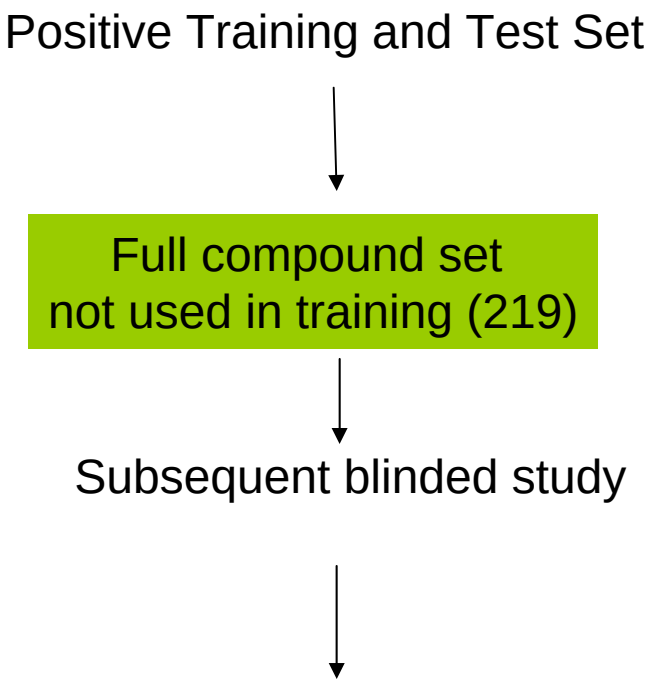
100% accuracy - not proper assessment of performance

Fibrosis Positive Test Set

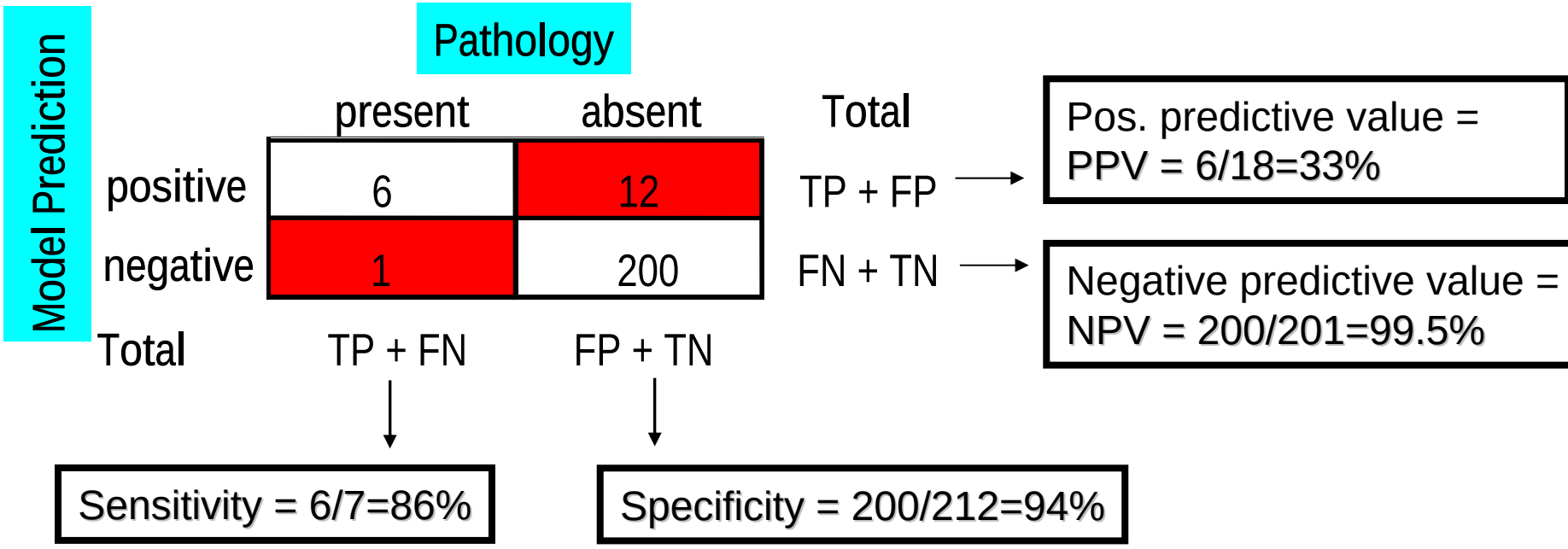


1 false negative: 86% accuracy

Model Validation

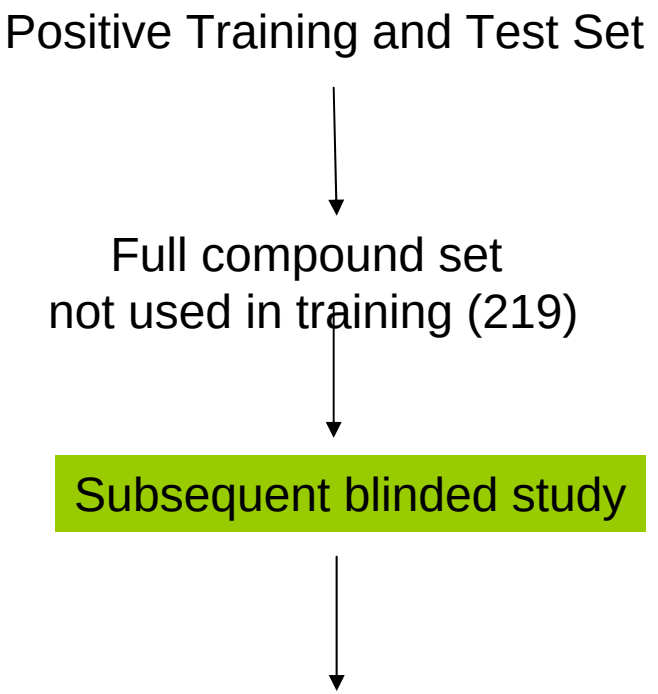


Model Evaluation on All Compounds not used in Training



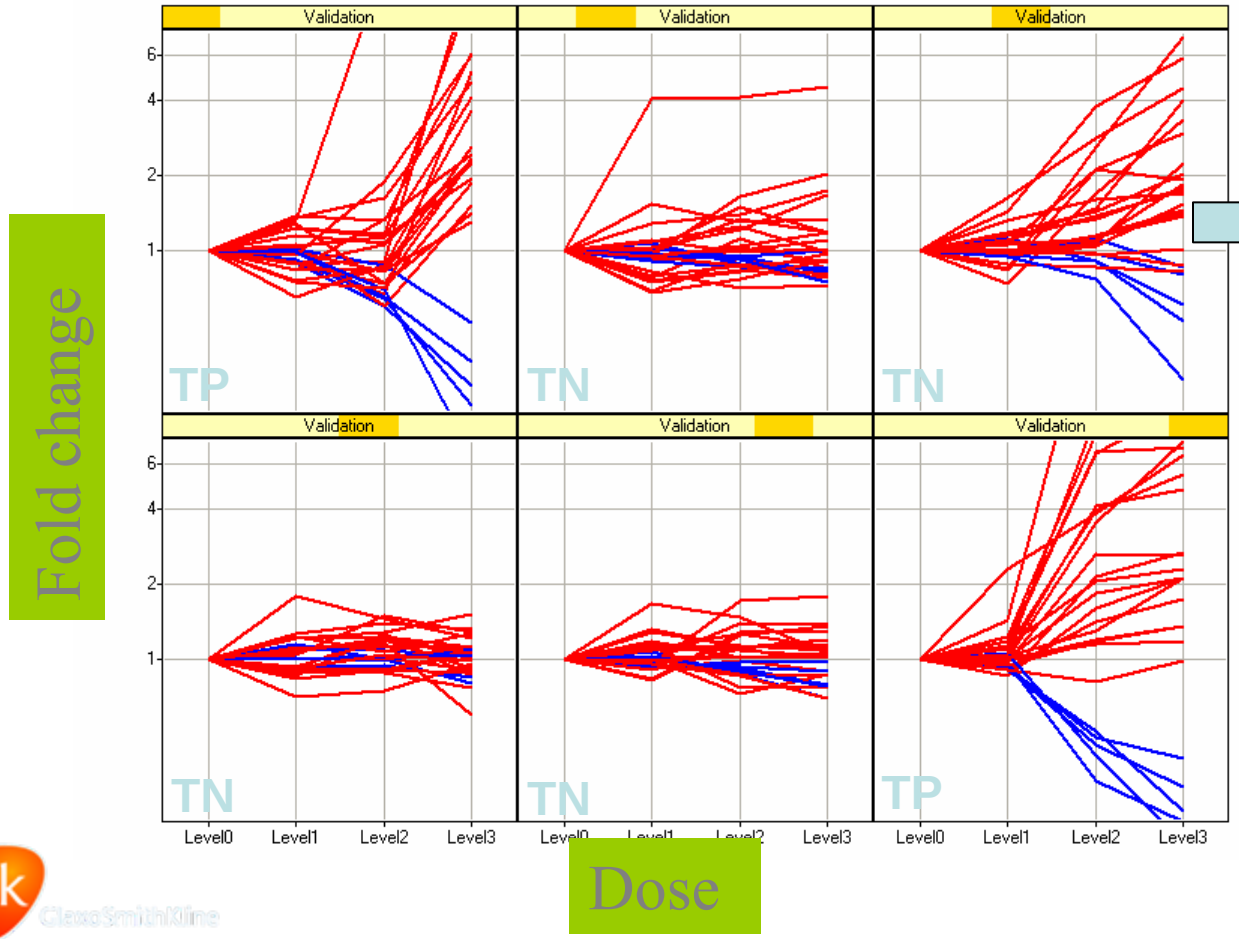
Note: PPV and NPV take prevalence into account

Model Validation



Predictive performance of the gene panel

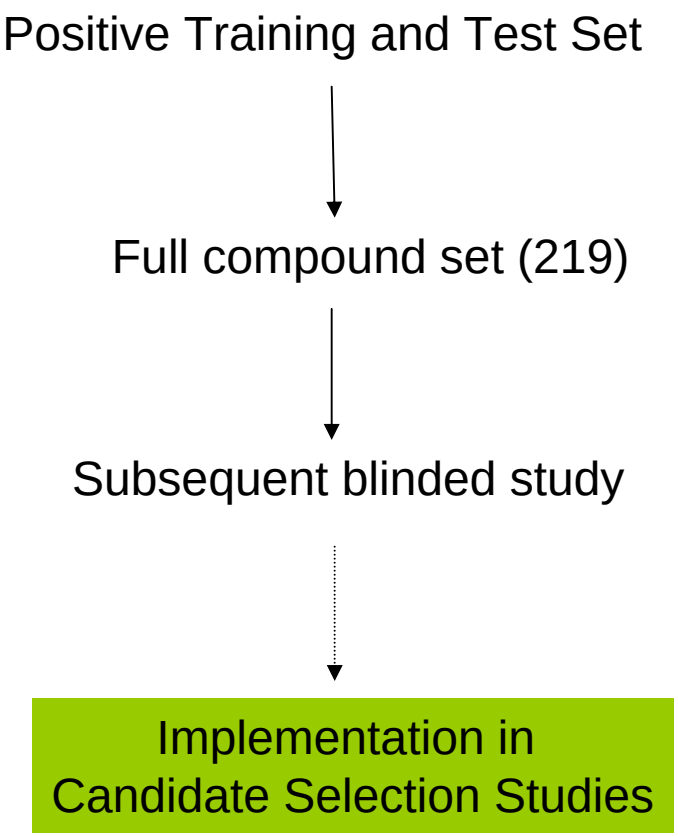
Examples from a blinded study with compound identity hidden



- No induction of two key fibrogenesis genes
- A true negative!

TP: true positive
TN: true negative

Model Validation



Summary

- Development of toxicity biomarkers requires
  - well designed study
  - lots of data
  - cross-disciplinary team effort
    - biology/toxicology
    - bioinformatics
    - statistics
- Proper validation is important

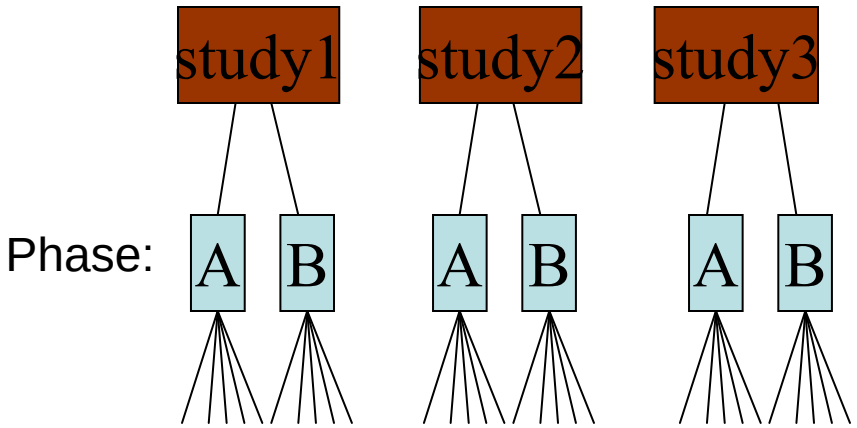
Acknowledgements

Kim Roland, Krista Stayer, Mark Tirmenstein, Jeffrey Ambroso, Holly Jordan, Chandi Elangbam, Gianni Dal Negro, Federica Crivellente, Lucinda Weir, Helen Billings, Sarah Nesfield, Maria Beaumont, Paul Trennery, Michael Santostefano, KB Tan, Ryan Boyle, Yifen Chen, Jessica Schreiter, Mike Trower, Mary Brawner, Georgina Paolazzo, Melissa Bertraiux, Kevin Kershner, Jessica Shroeck, Elizabeth Docherty, Derk Bergsma, Sujoy Ghosh, Qi Wang, Klaudia Steplewski, Erin Sharpe, Julie Keller, Ashley Hughes, Emma Akuffo, Jeff Hill, Paul Cutler, Isro Gloger, Louisa Bill, Mike Lutz, Patrick Warren, Mike Lonetto, Jacob Angert, Junping Jing, Hannah Muthyala, Mike Italia, JoAnn Betts, Leli Sarov-Blat, Marian Birkeland, Dilip Rajagopalan, Prakash Dev, Dave Mack, Christine Debouck, David Searls

Lei Zhu, Kwan Lee, Katja Remlinger, Paul McAllister, Amit Bhattacharyya


THANK YOU

Robust Variance Component Analysis



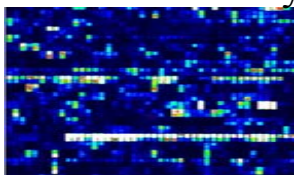
- Vehicle only data to identify noisy genes (large variation due to noise factors)
  - Initial variance estimates by Winsorizing: i.e. moving outlying points toward the rest of the data (remove effects of outliers)
  - Final estimates by REML
  - SPLUS method="winsor"

Hepatotoxicity Knowledge Base (HTKB)

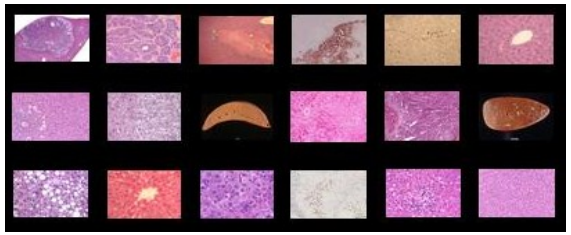
Input: Reference Data Base

Reference compounds

Liver Microarray



Clinical & Histopathology

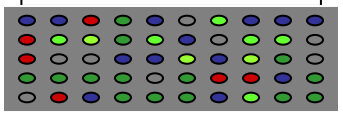


Analysis: Teams & Technology

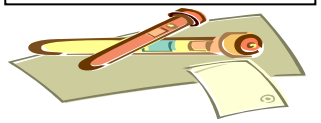
Output: Practical Preclinical & Clinical Applications

Liver Tox Target Panel

Liver Tox Gene Panel



Liver Tox Biomarkers



Liver Tox Candidate SNPs

Lead Optimiz ⇒ | *Cand Select* ⇒ | ⇒ *Drug Development* ⇒

Multiple Testing

- Large number of hypothesis tests (15,000) => large number of false positives just by chance (~ 750 false positive genes)
  - Traditional approach: Bonferroni adjustment
    - $p_{Bonf} = N*p$
    - Assumes independence
    - Controls the family-wise type I error rate
    - Too conservative => too many false negatives
  - Resampling methods (Westfall an Young, 1993)
    - p_R = estimates the likelihood of obtaining the uncorrected p-value by chance
    - Don't assume independence
    - Control the family-wise type I error rate
    - Still conservative

False Discovery Rate
Benjamini and Hochberg (1995)

- Controls the false discovery rate (proportion of false positives within the set of genes declared as positive)
  - Strikes a balance between too many false positives and negatives
  - Available in Proc Multtest in SAS
  - Popular choice for genomic experiments

1 Error Rates

Real Setup	Expressed? No	Expressed? Yes		Test Result No	Test Result Yes	
No	U	V	m_0	12402	8	12410
Yes	T	S	m_1	15	63	78
	A	R	m	12417	71	12488

1. Familywise Error Rate (e.g. Bonferroni): $\text{FWER} = \Pr\{V > 0\}$
2. False Discovery Rate: $\text{FDR} = E\left\{\frac{V}{R}\middle| R > 0\right\} \cdot \Pr\{R > 0\}$
   Positive False Discovery Rate: $\text{pFDR} = E\left\{\frac{V}{R}\middle| R > 0\right\}$
   In microarray experiments, it is reasonable to assume FDR=pFDR, since $\Pr\{R > 0\} = 1$.

2 Approaches to FDR

2.1 Testing Approach

- Benjamini–Hochberg (BH) Procedure

  (Benjamini and Hochberg 1995) proposed FDR concept and BH procedure, which controls FDR at $m_0\alpha/m$. For a given $0 < \alpha < 1$, let

$$i_0 = \max\left\{i : P_{(i)} \leq \frac{i}{m}\alpha\right\}$$

  Then BH rejects hypotheses corresponding to $P_{(1)}, \ldots, P_{(i_0)}$, if i_0 exists, otherwise retain all null hypotheses.

Example with $\alpha = 0.1$ and $m = 5$

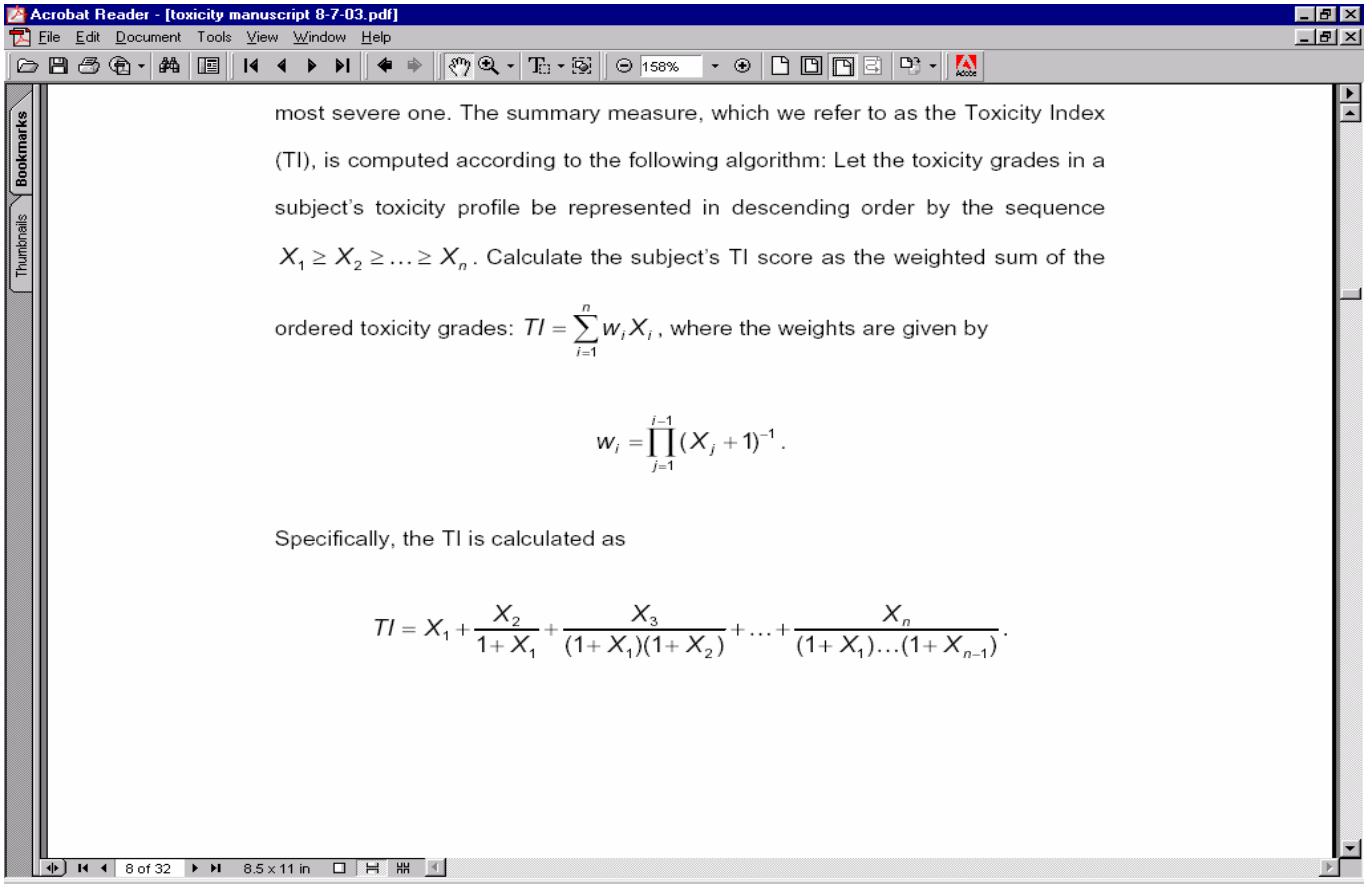
i	1	2	3	4	5
$i\alpha/m$	0.02	0.04	0.06	0.08	0.10
$P_{(i)}$	0.0092	0.0108	0.0243	0.0912	0.1941

Compromise between global t and gene-specific t

Cui and Churchill, Genome Biology 2003, 4:210

- Two extremes of t-statistics:
  - Global t = $\frac{R_g}{SE}$
  - Gene-specific t = $\frac{R_g}{SE_g}$
- Compromise between the two extremes:
  - SAM t = $\frac{R_g}{c + SE_g}$, where c is chosen to minimize the CV
  - Efron's 90% rule t = $\frac{R_g}{c + SE_g}$ where c is the 90th percentile of the global standard error
  - Regularized t = $\frac{R_g}{\sqrt{\frac{v_0 SE^2 + (n-1)SE_g^2}{v_0 + n - 2}}}$

Dimension Reduction of Histopathology Data: Toxicity Index



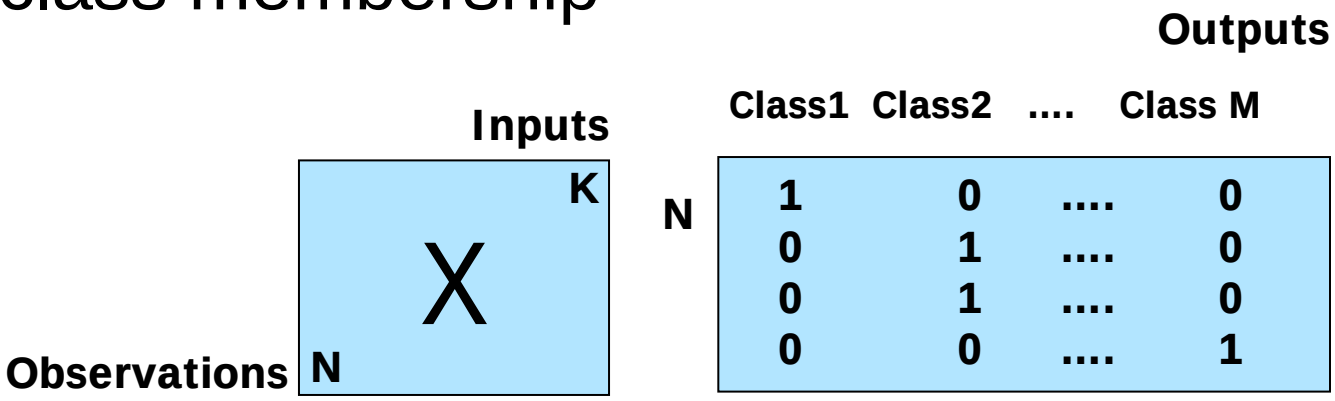
Rogatko at al, Clinical Cancer Research Vol. 10, 4645-4651, July 15, 2004

Toxicity Index - Properties

- The final score is between 0 and 5
  - Integer part of the score equals the highest histopath score of the animal
  - Fractional part indicates additional, lower grade toxicities
  - 1 higher grade score has a larger weight than several lower grade scores
- Convenient summary across biologically meaningful groupings of histopathology scores
- Can be modeled by standard analysis methods (ANOVA, PLS, etc)

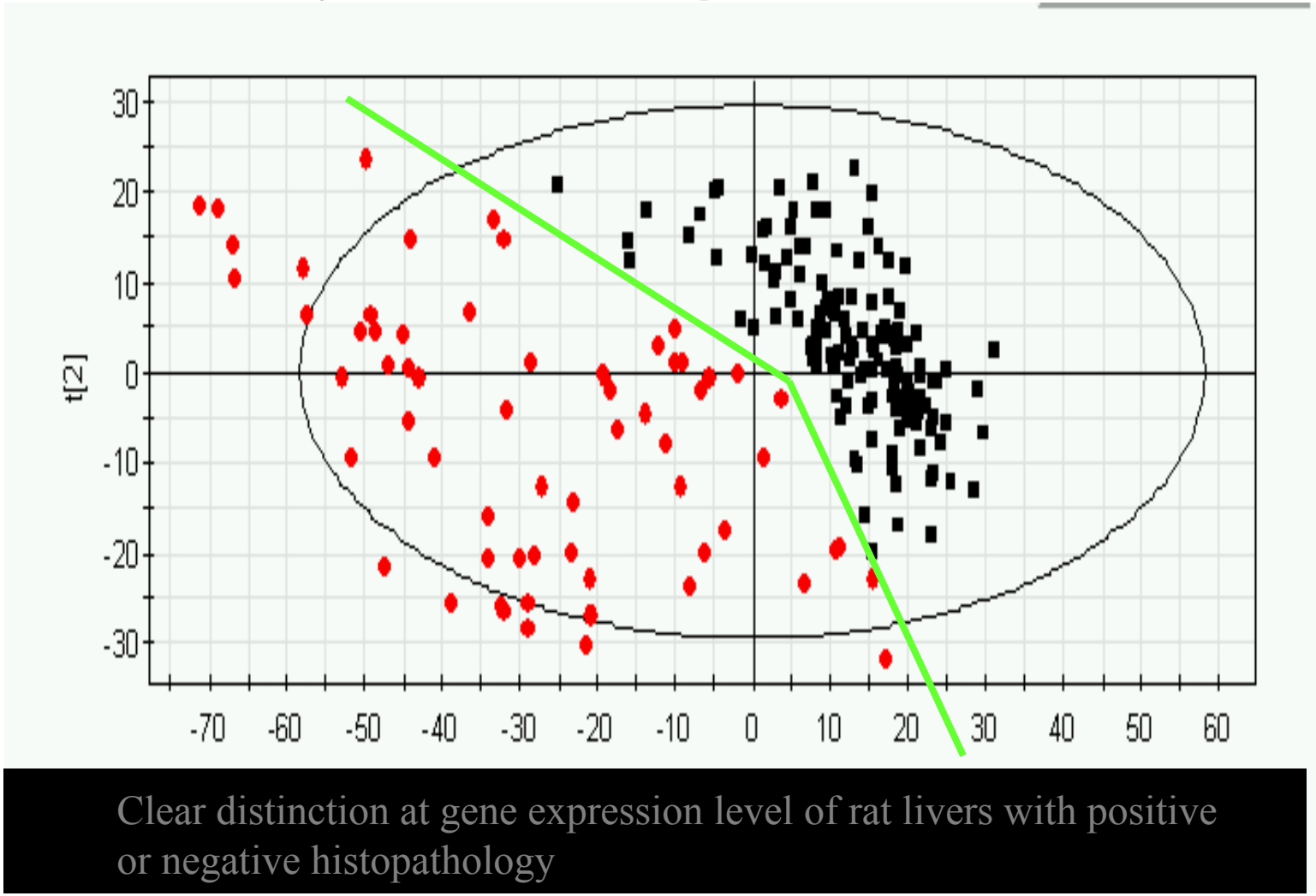
Partial Least Square Discriminant Analysis (PLS-DA)

- Known groupings or classes
- Y matrix of dummy variables indicating class membership



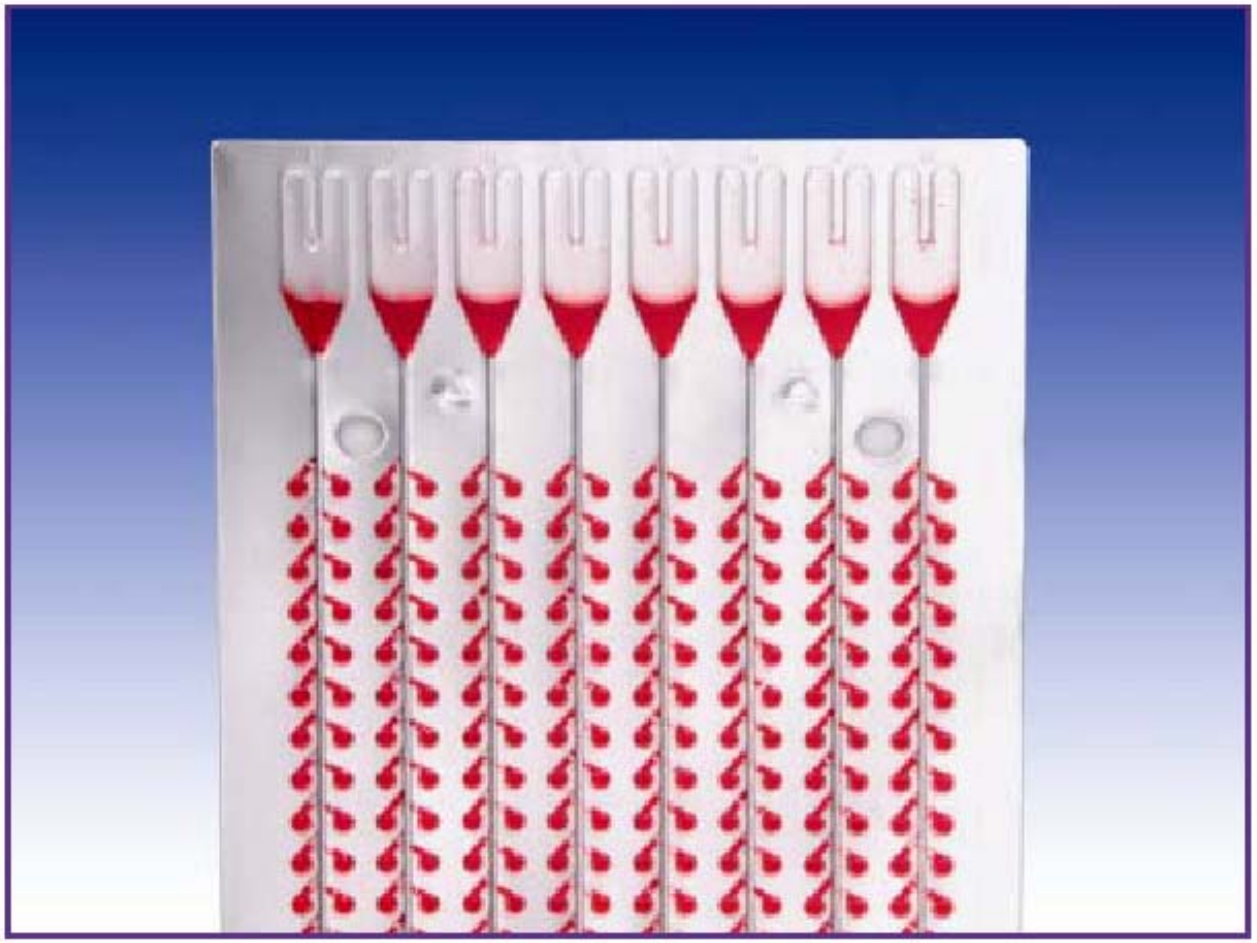
- PLS model built on new Y matrix

PLS-DA Example:
Discriminating positive and negative rat hepatotoxicants



Clear distinction at gene expression level of rat livers with positive or negative histopathology

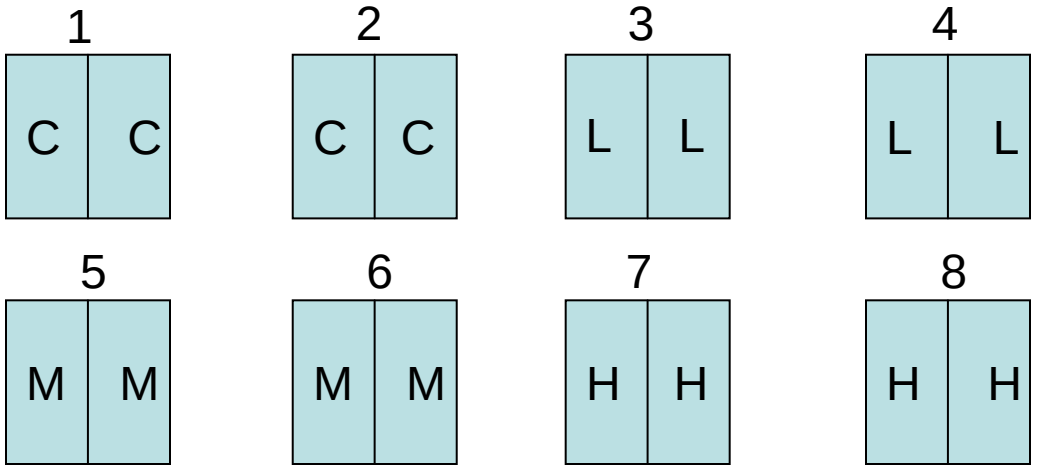
HepatoTaq panel: Supporting the New Technology



ABI Microfluidic Card v.2. low density array containing 150 liver toxicity specific genes
New version can accommodate 2 samples per card
50% Cost reduction

Original treatment to card allocation:

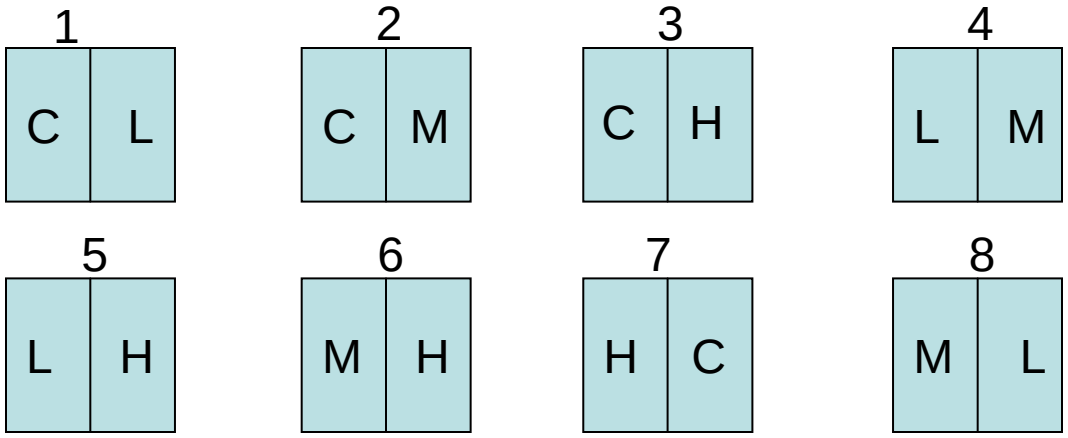
Design 1



C: Vehicle Control
L: Low dose
M: Mid dose
H: High dose

Improved treatment to plate allocation implemented:

Design 2



The advantage of Design 2 is that animals from the same treatment group are placed across 4 cards instead of 2. Simulations showed that this will reduce the bias in the treatment mean estimate due to card effects by at least 30% over the bias in Design 1