# A gene expression landscape of drug-induced liver hepatotoxicity

*Héctor Rueda-Zárate*[1,2], *Iván Imaz-Rosshandler* [1], *Julieta Noguez-Monroy*[2] *and Claudia Rangel-Escareño* [1*]

1 Computational Genomics, National Institute of Genomic Medicine Mexico

2 Computational Sciences Graduate Program, Instituto Tecnológico y de Estudios Superiores de Monterrey C.C.M.

*corresponding author crangel@inmegen.gob.mx

## Introduction

The liver is one of the organs involved in biotransformation, chemical reactions that alter the structure, aqueous solubility and eventual disposition of non- nutritive compounds that enter into the organism. Xenobiotic biotransformation aims at controlling the toxication or detoxication of xenobiotic substances. However, during the biotransformation reactive intermediates may be produced, these could interact with critical cellular macromolecules and trigger the events that promote either tissue injury and cell death, permanent genomic changes, leading potentially to cancer.

Many currently and normally used drugs could affect the liver adversely in any combination of the reactions described. Liver injury can be classified as hepatocellular, cholestatic or mixed, based on criteria established by the Council for International Organizations of Medical Sciences (CIOMS) [14]. The drug-induced liver injury also known as DILI is classified as intrinsic and idiosyncratic hepatoxicity. The Intrinsic hepatotoxins cause hepatocellular damage and it is more related to other industrial agents more than it is to xenobiotics. However, xenobiotics are more closely related to idiosyncratic liver injury by its level of toxicity or to allergy reaction or other secondary effects. Toxic effects of drugs at all levels are extensively studied before these are administered to humans. The Toxicogenomics Project focuses on gene expression analyses in animals or in-vitro grown cells that have been exposed to the chemicals with the aim of understanding the molecular mechanisms of toxicity and eventualy be able to predict dangerous levels of toxicity.

## Materials and Methods

We used gene expression data from the Japanese toxicogenomics project (TGP), a 5-year project that was completed in 2007. TGPs database comprises nearly 18,000 Affymetrix microarrays testing 131 compounds, mainly medical drugs and their effect in the liver. All microarrays targeted the liver in both in vitro and in vivo experiments. All .CEL files were downloaded into a 32GB server for the analyses. A primary test on processing capabilities and algorithm complexities showed that up to a maximum of 400 microarrays could be pre-processed using R and biocoductor `affy` library and its dependencies at once on this server. Hence, the strategy for this analysis would have to be design in such a way that it loads only those sets of microarrays involved in the actual biological question.

## Strategy for integrative analysis

A map of how data are structured can be seen in Figure[1]. We need a strategy that will allow us to combine species(Hu, Rat), protocols (iVV, iVT), dosages (None, Low, Med, High) and time points. A mixture of differential expression analysis using `limma` and gene selection using ranking approach such as `timecourse` seems to be an appropriate beginning approach.

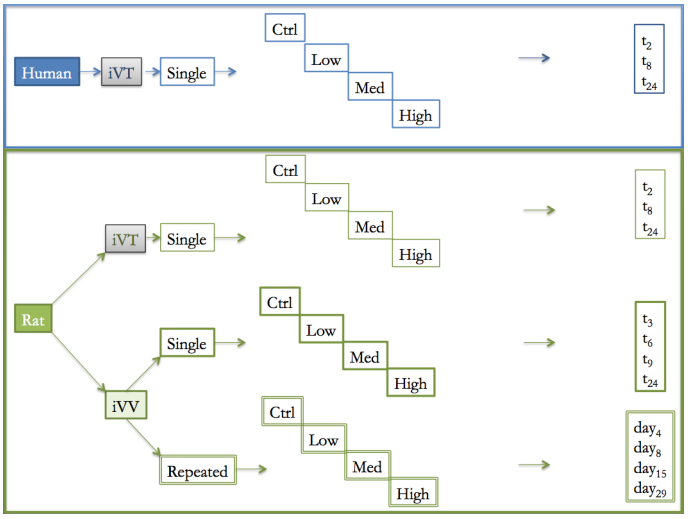


Figure 1. Group structure in TGP data

Just the human diagram (top portion in Figure [1]) would lead to up to 30 pairwise contrasts of interest per compound, roughly 117,900 comparisons if we were to use `limma` alone.

$$T_{contrasts} = N_{comp}\left[N_{timepts}\binom{N_{cond}}{2} + N_{cond}\binom{N_{timepts}}{2}\right]$$

That plus a similar number for RatiVT, also for RatiVV, plus repeated plus all cross-referenced contrasts. This, however, does not mean we cannot still do it. This approach should be taken on a more biological driven hypotheses rather than as massive computational analysis.

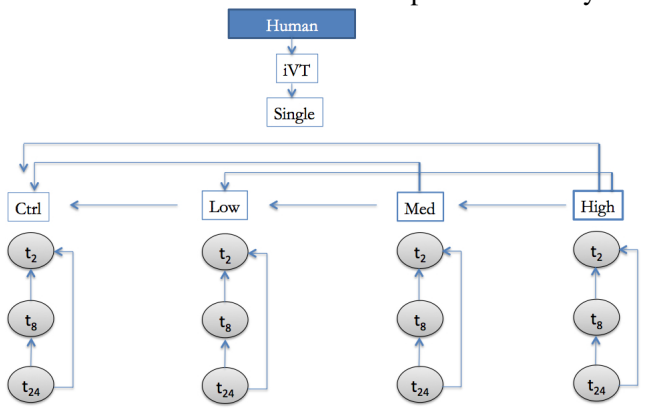


Figure 2. Diagram of contrasts within Human in-vitro samples per compound

Methods which rank genes (e.g. the MB statistic or the moderated Hotelling T2) perhaps provide easier access to genes whose absolute or relative expression varies over time. This approach is used for each of the four main paths in Figure [1] {HU.iVT, Rat.iVT.Single, Rat.iVV.Single,Rat.iVV.Repeated}. However, in order to see correlation of using animal model to infer potential toxicity in humans we would also need to take the differential expression approach. This will be applied to all groups with identical structure but different species. In our scenario that would only be {HU.iVT.Single, Rat.iVT.Single}.

The general structure of our strategy involves :
• Data storage and manipulation by using a relational database
• Raw-data analysis: quality metrics, normalization and background correction.
• Gene selection either by differential expression and/or by ranking method.
• Gene annotation and function. Conducting a gene set enrichment analysis on lists of ranked genes; a selection of orthologous genes using swissprotID and inParanoid database [11]. Simultaneously, research on drug toxicology to determine if compounds could be classified according to toxicity or type of liver damage.
• Machine learning approach: searching for possible patterns in data, clusters of compounds by unsupervised methods. Patterns in concentrations, patterns in the time-course results. • Tool development that will be available through R and Bioconductor.

## Data

The TGP data contains a collection of 17,657 Affymetrix$^{TM}$ microarrays from both in vitro and animal samples. Human samples were processed using Hgu133Plus2, animal samples were processed on the GeneChip Rat Genome 230 2.0 which is known to be a powerful tool for toxicology.

## MySQL database

Due to data complexity in terms of number of groups, labels for all barcoded data were stored as a relational MySQL database. This allows faster, easy and optimum access to a specific set of .CEL files for further analysis. Even though it was one table at first, the database will grow as more information is developed. It will be constantly normalized and designed to be scalable. Access to it is through R scripts, an example is shown here:

```
findMicroarrays(species=c("Rat"),
      expType=c("in vitro", "in vivo"),
      dose=c("Control", "Middle"),
      singleRepeat=c("NA", "Single", "Repeat"),
      compound=c("AA", "ACA", "WY"),
      sacTime=c("2 hr", "8 hr"),
      experiment=c("CAMDA13"),
      path="CELS/",
      orderBy="DOSE_LEVEL", conn=conexion)
```

This function collects files according to the specified parameters and it modifies the file names to match the conditions making all more easy to follow. The resulting query is shown below.

```
SELECT BARCODE FROM MICROARRAY WHERE...
    SPECIES IN ('Rat') AND TEST_TYPE IN ...
    ('iVT','iVV') AND DOSE_LEVEL IN ...
    ('Control','Low') AND SINGLE_REPEAT_TYPE...
    IN ('NA','Single','Repeated') AND ...
    COMPOUND_ABBREVIATION IN ...
    ('AA','ACA','WY') AND SACRIFICE_PERIOD...
    IN ('2 hr','8 hr') AND EXPERIMENT IN...
    ('CAMDA13') ORDER BY DOSE_LEVEL;
```

And the resulting sample names are shown below.

```
 [1] "Rat.iVT.Control.NA.2 hr.WY-1"...
 [3] "Rat.iVT.Control.NA.8 hr.AA-3"...
[13] "Rat.iVT.Low.NA.2 hr.AA-13"...
[15] "Rat.iVT.Low.NA.8 hr.AA-15"...
[21] "Rat.iVT.Low.NA.8 hr.WY-21"...
```

## Low-level analysis

Gene expression microarray raw data for subsets of samples collected through the database were pre-processed in the R statistical environment. A quality control tests

were run on randomly selected sets, showed a constant behavior of $MM-probes > PM-probes$ in a range from 22-30% causing serious concerns about using MAS5.0 algorithm for background correction. In fact, this analysis showed also that RMA [9] led to bimodal distribution indicating that background adjustment was unnecessary. Data were normalized using quantile normalization [4] , summarization was done using medianpolish. All methods from the Bioconductor `affy` library.

### Timecourse analysis

Genes were ranked based on large absolute or relative amounts of change over time as a function of the drug concentration in relation to their replicate variances. For every selected subset, genes were classified according to a multivariate empirical Bayes statistic for replicated microarray time course data MB statistics implemented in the `timecourse` package [21].

### Human-Rat orthologues

The human, mouse and rat genomes encode a very similar number of genes. Human-Rat share roughly 89 to 90% of genes [8] with a majority that have persisted without deletion or duplication since the last common ancestor. The most important aspect is perhaps that almost all human genes known to be associated with disease have orthologues in the rat genome. However, their rates of synonymous substitution are significantly different from the remaining genes. Hence, even though the high correlation we are also conducting an orthology analysis through related proteins using the InParanoid database [11]. This databases information is based on information about swissprotID. We are also exploring the ENSEMBLE database for this purpose. More tables are added to our database so gene-to-gene information could be generated.

### Gene set enrichment analysis

After using timecourse approach, lists of genes of interest are generated. We may end up with way too many genes to examine in proper detail. Hence, a good way of comparing conditions is thorugh a gene set enrichment analysis that could tell us about cellular mechanisms behind different lists. The idea is to identify pathways affected by highly ranked genes in Human iVT and compare to those found in Rat iVT and Rat iVV. Tools used for this approach involve DAVID [5], GSEA [6] and BiNGO [3]

### Machine Learning Approach

Even though we have access to a quite impressive sample size, this number is fastly diluted by the number of groups in the study. If we see Figure 1, we have four main groups {HuiVT, RatiVT, Rat iVV, RatiVV-Rep}, between 119 and 131 xenobiotics, and between 3 and 4 time points. So we basically have only either two or three replicates in each group for statistical assessment. The question we would like to address here is: What can we learn from data?
Hence, an unsupervised hierarchical clustering would allow to see sets of genes that follow a similar profile between the main groups. Properly validated these sets of genes could potentially be used as markers for in-vitro human models avoiding the need of performing animal model approaches. Class discovery and clustering validation can also be tested using Consensus Clustering method [18]

## Results

Database implementation and data retrieval through R made all of the timecourse analyses time efficient. Only 48 compounds were selected since these are found in all four groups. Results from timecourse ranked all genes and only the top 50 from each group and each compund (2400 approx.) were selected for further analysis. Below there is a list of the 25 most common genes where column labeled as Count represents the number of times that gene was present across drugs, time and concentration in three groups: {HU iVT, Rat iVT, Rat iVV}.

Table 1. Top 50 genes in each compound and number of times they are present

| Hu iVT | Count | Rat iVT | Count | Rat iVV | Count |
|---|---|---|---|---|---|
| CYP3A5 | 49 | CXCL3 | 45 | TXNRD1 | 23 |
| CYP3A4 | 34 | CYP1A1 | 32 | ACOT2 | 17 |
| NEAT1 | 31 | SLC7A11 | 19 | HSDL2 | 14 |
| RRM2 | 29 | SOX4 | 19 | CCND1 | 13 |
| RRAD | 27 | PDK4 | 17 | SREBF1 | 13 |
| CYP2C9 | 23 | ANGPTL4 | 16 | DUSP6 | 12 |
| CYP1A1 | 20 | HSDL2 | 16 | SRXN1 | 12 |
| MALAT1 | 20 | ACAT3 | 14 | PTPRF | 11 |
| ATF3 | 19 | HMGCS2 | 13 | STAC3 | 11 |
| TSKU | 18 | NREP | 13 | ANKH | 10 |
| ANGPTL4 | 17 | CD36 | 12 | ATP1B1 | 10 |
| ARL14 | 17 | CPT1A | 12 | HAMP | 10 |
| CYP1A2 | 17 | SERPINB9 | 12 | HSPB8 | 10 |
| PCK1 | 17 | ACOT2 | 11 | PPCS | 10 |
| C19orf80 | 16 | DHRS3 | 11 | SLC13A4 | 10 |
| CYP3A7 | 14 | TAGLN | 11 | TBC1D15 | 10 |
| EGR1 | 14 | FASN | 10 | TM2D3 | 10 |
| GDF15 | 14 | HSP90AB1 | 10 | CAR14 | 9 |
| PPP1R15A | 14 | LSS | 10 | GCLC | 9 |
| TRIB3 | 14 | AKR1D1 | 9 | PKLR | 9 |
| FAM13A | 13 | CYP26B1 | 9 | CXCL12 | 8 |
| IFRD1 | 13 | FABP7 | 9 | MGLL | 8 |
| MAFF | 13 | GDE1 | 9 | PDK4 | 8 |
| RPL38 | 13 | PEX11A | 9 | ACACA | 7 |

The level of correlation or intersection between these genes and assuming same symbol indicates orthologues

is shown in the Venn diagram in Figure 3.

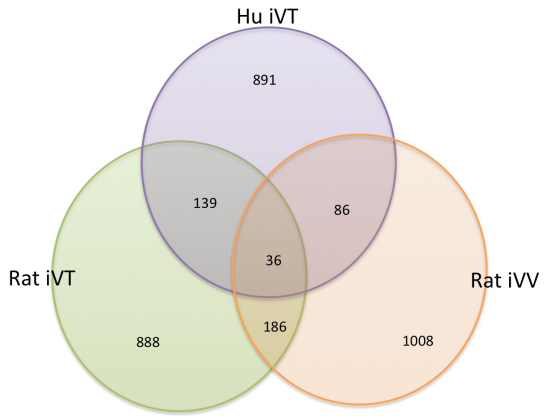


Figure 3. Genes found in all three groups and some of the 48 compounds were 36: {*PEX11A, PDK4, ANGPTL4, SERPINB9, CYP1A1, LSS, FASN, TRIB3, CREM, SWI5, PPP1R3B, PIR, NREP, HMGCR, ABCD3, RDX, TIPARP, SQLE, NQO1, HSPH1, YPEL5, EGR1, PTPRF, MDM2, JUN, BHLHE40, LDLR, TSKU, IFRD1, GCLM, SGK1, RRM2, EFNA1, IRF7, BCL6, INHBE* }

Since there is a vast amount of information for an abstract, we concentrated on one drug and pursued a more detailed analysis.

### Case study intrinsic DILI: Acetaminophen

Acetaminophen toxicity is the leading drug-related cause. At low doses, the drug is conjugated to water-soluble metabolites in the liver and is excreted in the urine. At higher doses, glutathione depletion leads to saturation of the conjugation mechanism, leaving the parent compound to be metabolised to toxic intermediates. Moreover, toxicity risk increases if there exists chronic alcohol consumption, obesity, or drugs that induce the P-450 cytochrome system lowering the toxic threshold of acetaminophen [16], [17]. Timecourse results for this compound are shown in Figure 4.

We observe that patterns are different on each group even though is the same compound. We should consider however, that ranking is determine by genes with changes across time as a function of concentration also including that replicates do not vary much. A collective view of this including the top 100 genes can be seen in Figure 5 where we observe that overall gene profiles are different. This simply suggests that performing a timecourse analysis do not exclude the usual between groups differential expression approach. We would like to point out that this approach can also be done and has been done through the database.

One interesting remark is about the effect of high concentration effect at time 24 hrs. It is not clear whether gene is down-regulated as a response to high concentration or if we are facing a cell viability issue and the cell simply dies. We performed a gene set enrichment analysis and found out apoptosis pathways are significant for some drugs.

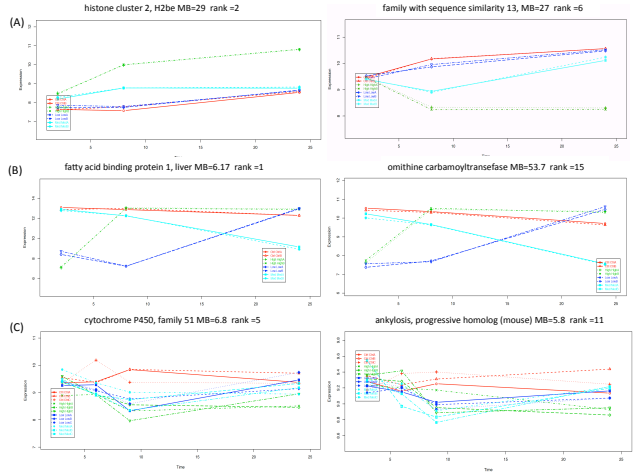


Figure 4. Top ranked genes in Acetaminophen (APAP): Top row (A) Human iVT, middle row (B) Rat iVT, bottom Row(C) Rat iVV Red = control, blue = Low, Cyan = Med and Green = High

Hierarchical clustering does not show interesting patterns in terms of gene profiles. However, among the three groups the Human iVT plots shows more interpretable results. As we can see that for most genes patterns of up-reglation occur at time 24 hrs.

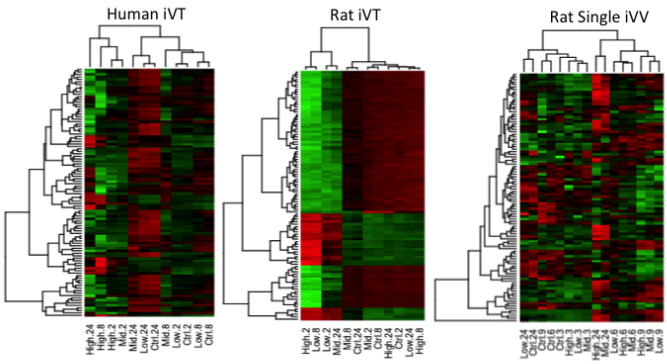


Figure 5. Hierarchical cluster plots for three groups.

Consensus clustering [18] on the selected 48 compounds are shown in Figure 6. Here we observe two interesting

patterns. It seems that after 2 hours most compounds tend to cluster into smaller number of groups (left panel top and bottom), but after 24 hours patterns are more heterogenous. This suggests that after 24 hours what we see is a putative different drug dependent metabolic process and biotransformation. Our hypothesis is that there may exist a molecular sub-classification of drugs based on gene expression profiles. We will further explore this by combining the Drug versus Disease data R- package and two databases DrugBank and ChemmineR

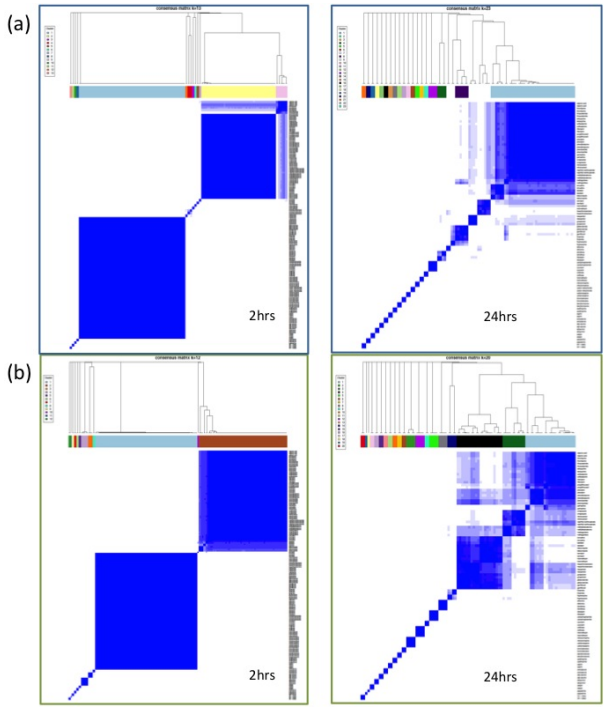


Figure 6. Consensus clustering for High concentration (a) Human iVT at 2 and 24hours.; (b) Rat iVT at 2 and 24 hours

## Discussion

We have performed a broad analysis of this data set that has led us to pursue various hypotheses. Some of them are presented here and many others are currently under revision. It seems there is plenty of room for more discoveries and at this point we can only see the potential but not the end of the road. For instance, there is still work in progress for Rat in-vivo with repeated samples, a more specific gene set enrichment analysis, an extensive exploration of mechanisms for drug classification and feature selection using machine learning approaches among others.

## Conclusions

The Japanese Toxicogenomics Project (TGPJ) is a combined efford between the National Institute of Health Sciences and 17 pharmaceutical companies. The purpose of the study and its results will impact drug development and toxicology research wordwide. A database fed by new gene information was created. In this work we propose an interactive model for analyses that uses a database that can be queried with specific biological questions. Then a collection of R functions will perform low-level analysis; classification providing a set of genes of interest either by timecourse, concentration or contrast specific approach; and data mining. We are currently working on an R package, as well as a manual for the scripts.

# References


[1] Uehara T, Ono A, Maruyama T, Kato I, Yamada H, Ohno Y, Urushidani T., The Japanese toxicogenomics project: application of toxicogenomics. Mol Nutr Food Res. 54(2):218-27, 2010.

[2] Chen, M., et al., FDA-approved drug labeling for the study of drug-induced liver injury (DILI). Drug Discov Today, 2011. 16(15-16): p. 697-703.

[3] BiNGO: A Biological Network Gene Ontology tool. http://www.psb.ugent.be/cbd/papers/BiNGO/Home.html

[4] Bolstad, B.M., Irizarry R. A., Astrand M., and Speed, T.P. **(2003)**. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. *Bioinformatics* 19(2):185-193.

[5] DAVID: Bioinformatics Database. *Nature Protocols* 2009; 4(1):44 and *Nucleic Acids Res.* 2009;37(1):1 http://david.abcc.ncifcrf.gov/

[6] Tamayo, et al. (2005, PNAS 102, 15545-15550) and Mootha, Lindgren, et al. *Nat Genet* 2004 34, 267-273. http://www.broadinstitute.org/gsea/index.jsp

[7] Gentleman R., Carey V., HUber W., Irizarry R., Dudoit S. Bioinformatics and Computational Biology Solutions Using R and Bioconductor **2005** Statistics for Biology and Health - Springer.

[8] Richard A. Gibbs and George M. Weinstock et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution *Nature* 428, 493-521 (1 April 2004)

[9] Irizarry Rafael A. , Bolstad Benjamin M., Collin Francois , Cope Leslie M., Hobbs Bridget and Speed Terence P. **(2003)**. Summaries of Affymetrix GeneChip probe level data *Nucleic Acids Research* 31(4):e15.

[10] Irizarry, RA, Hobbs, B, Collin, F, Beazer-Barclay, YD, Antonellis, KJ, Scherf, U, Speed, TP **(2003)** Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics* Vol. 4, Number 2: 249-264.

[11] http://inparanoid.sbc.su.se

[12] Wit Ernest and McClure John. **(2004)**. Statistics for Microarrays Design, Analysis and Inference. *John Wiley & Sons Ltd.*

[13] R Ramachandran, S Kakar Histological patterns in drug-induced liver disease. *Journal of Clin Pathology(2008)*

[14] Watkins PB and Seeff LB. Drug-induced liver injury: summary of a single topic clinical research conference. *Hepatology* 43:61831 (2006)

[15] Williams R. Classification, etiology, and considerations of outcome in acute liver failure. Semin Liver Dis 1996;16:3438

[16] Lee WM. Acute liver failure. Clin Perspect Gastroenterol 2001;2:10110.

[17] Larson AM, Polson J, Fontana RJ, et al. Acetaminophen-induced acute liver failure: results of a United States multicenter, prospective study. Hepatology 2005;42:136472.

[18] Stefano Monti , Pablo Tamayo , Jill Mesirov , Todd Golub. Consensus clustering – A resampling-based method for class discovery and visualization of gene expression microarray data. *Journal Machine Learning*. Volume 52 Issue 1-2, July-August 2003, 91 - 118

[19] Zhijin Wu, Rafael A. Irizarry, Robert Gentleman, Francisco Martinez Murillo, and Forrest Spencer **(2004)** A Model Based Background Adjustment for Oligonucleotide Expression Arrays. *Johns Hopkins Univ, Dept. of Biostatistics Working Papers. Working Paper 1.*

[20] Vapnik, V. N. *The Nature of Statistical Learning Theory* (2nd Ed.), Springer Verlag, 2000

[21] Yu Chuan Tai and Terence P. Speed. A multivariate empirical Bayes statistic for replicated microarray time course data. *Ann. Statist.* Volume 34, Number 5 (2006), 2387-2412.