Report Number: WUCSE-2006-60

2006-01-01

Use of gene expression profiling and machine learning to understand and predict primary graft dysfunction

Monika Ray, Sekhar Dharmarajan, Johannes Freudenberg, Weixiong Zhang, and Alexander G. Patterson

Use of gene expression profiling and machine learning to understand and predict primary graft dysfunction

Monika Ray[1], Sekhar Dharmarajan[3], Johannes Freudenberg[4], Weixiong Zhang[1,2*], G. Alexander Patterson[3]
[1]Washington University, Dept. of Computer Science and Engineering, St. Louis, MO 63130
[2]Washington University School of Medicine, Dept. of Genetics, St. Louis, MO 63110
[3]Washington University School of Medicine, Dept. of Cardiothoracic Surgery, St. Louis, MO 63110
[4]Cincinnati Children's Hospital Medical Centre, Biomedical Informatics, Cincinnati OH 45229
* Corresponding author

Abstract

Lung transplantation is the method of choice for the treatment of end-stage pulmonary diseases. A limited donor supply has dramatically increased the waiting time for transplant recipients. Approximately 4000 patients are currently on the transplant waiting list. Unfortunately, up to 10-20% of these patients will die from their underlying lung disease before an organ becomes available. Currently, only 10-20% of cadaveric donor organs offered for transplantation are judged to be acceptable under the current selection criteria. Of the donor lungs selected for transplantation, 15-30% of them fail due to primary graft dysfunction (PGD). PGD is a severe allograft ischemia-reperfusion (I/R) injury syndrome occurring in the hours following transplantation. It significantly affects morbidity as well as early and late mortality. This has resulted in an intense pressure to search for alternative selection criteria for selecting suitable donor lungs. In this study, we attempt to further our understanding of the gene products involved in PGD by observing the changes in gene expression across donor lungs that developed PGD versus those that did not. Our second goal is to use a machine learning technique - support vector machine, to distinguish donor lungs suitable for transplantation versus those that are not, based on the gene expression data. Results from microarray analysis produced a set of differentially expressed transcripts that were involved in signalling and apoptosis pathways. Various transcripts particular to stress-sensitive pathways were also identified. Results also indicate that the metallothionein gene, specifically metallothionein 3, may protect donor lungs from developing PGD. A classification accuracy of 70% was achieved, when a set of 100 differentially expressed transcripts was used to differentiate unsuitable donor lungs from suitable ones. This is the first such attempt to combine the identification of a molecular signature for PGD, using human samples, with machine learning methods for class (donor lung) prediction.

Introduction

Lung transplantation has gained widespread acceptance for the treatment of end-stage pulmonary diseases. However, two significant problems in clinical lung transplantation are a major shortage of donor organs and the incidence of primary graft dysfunction (PGD). PGD is a severe allograft ischemia-reperfusion (I/R) injury syndrome occurring in the hours following transplantation. It significantly af-

fects morbidity as well as early and late mortality. Improvements in operative techniques, donor management, and immunosuppressive protocols have decreased perioperative mortality to below 10% at most experienced lung transplant centers [1, 2]. The one- and five-year survival rates have improved to 76% and 49%, respectively [1]. These results, however, continue to lag behind those achieved for other solid organ transplants. The occurrence of PGD after lung transplantation significantly increases the duration of mechanical ventilation, hospital length of stay and short-term mortality after lung transplantation [3]. Survivors of PGD have a significantly protracted recovery with impaired physical function up to one year after transplantation and an increased risk of death extending beyond the first year after transplantation [3, 4].

The current criteria used to evaluate potential donor lungs appear to be inadequate at predicting how these lungs will function post-transplantation [5, 6, 7]. Donor organs are evaluated for lung transplantation on the basis of criteria that are primarily historically founded and largely arbitrary [8]. Relatively crude measures of lung function such as chest radiography, arterial oxygen tension on blood gases, and bronchoscopy are currently used to assess the quality of potential donor lungs. That these tools are inadequate in evaluating organs from prospective donors is evidenced by two recent developments. First, the liberalisation of the selection criteria and the use of 'marginal' donor lungs by many centers have not had a negative impact on outcome after transplantation [9, 10, 11]. A recent study showed no significant difference in a number of indices for infection and inflammation between donor lungs that were accepted and rejected for transplantation [7]. Second, the incidence of PGD or I/R injury, after transplantation remains unchanged at 10-20% despite the increased use of marginal donor lungs and improvements in all areas of lung transplantation [2, 4, 12]. The use of marginal donor lungs, extended graft cold ischemic times, recipient pathophysiology and current donor selection criteria have shown no correlation with the occurrence of PGD in most cases [13, 14, 15, 16, 17].

A limited donor supply has dramatically increased the waiting time for transplant recipients. Approximately 4,000 patients are currently on the transplant waiting list and this has resulted in intense pressure to search for alternative strategies. Unfortunately, up to 10-20% of these patients on the waiting list will die from their underlying lung disease before an organ becomes available. Currently, only 10-20% of cadaveric donor organs offered for transplantation are judged to be acceptable under the current selection criteria [18]. More biologically meaningful donor selection criteria may result in significant expansion of the number of lungs accepted from this potential donor pool [5].

The results of the above mentioned studies suggest that there may be complex, occult biological factors present in donor lungs which contribute to the development of PGD that are not detected by the current donor organ evaluation. Gene expression profiling is a powerful, high-performance tool of molecular biology that allows the analysis of the levels of expression of thousands of genes simultaneously. It has been previously used to study some transcripts involved in I/R using a rat model, but the study did not differentiate the suitable lungs from the unsuitable [19]. Therefore, this is the first report, according to our knowledge, where gene expression profiling has been used on human samples, along with the application of machine learning techniques to automatically distinguish unsuitable donor lungs from suitable donor lungs.

Our objective is two fold - the first is to obtain a set of genes involved in PGD and identify new gene products relevant to allograft transplantation; and the second is to use this set of genes for classification of donor lungs into PGD positive (i.e. lungs that develop PGD) or PGD negative (i.e. lungs that do not develop PGD) categories. The first objective would provide greater insights into the mechanism of PGD as well as extend the work of [19].

The set of genes identified as being involved in PGD can be designated as the 'molecular signature' of PGD. As many donor lungs that may be actually good are discarded by the current selection criteria employed by physicians, it would be useful to classify unseen donor lungs, using the molecular signature coupled with machine learning techniques, thereby increasing the possibility of having more available lungs for transplantation. This is the motivation behind our second objective.

Results and Discussion

The characteristics of the donor lungs are depicted in Table 1. The operative factors and the characteristics of the patients who have been identified with PGD versus those that have not, are shown in Table 2 and Table 3, respectively.

Pathways and gene products involved in PGD

First, the upregulated transcripts were analysed using the Ingenuity Pathway Analysis (IPA) software. There were 23 upregulated transcripts, of which 13 were focus genes. Focus genes are the genes that map onto the Ingenuity Pathways Knowledge Base (IPKB). The network generated from these genes are shown in Figure 1.

Network 1 primarily centres around tumor protein p53 (TP53). The focus genes are shown in red coloured shapes and more details on these nodes are given in Table 4. Figure 2 shows the location of the different gene products and the canonical pathways present in Network 1. The legend for the network is shown in Figure 7.

It is natural to expect many pathways related to apoptosis and cell signalling as over 50% of the donor lungs (PGD positive and PGD negative) were involved in some kind of trauma. Interestingly, a few transcripts identified are also cancer related genes. There is growing evidence of genetic parallels between lung development and several types of cancer [20, 21]. The authors of [22] have shown that Wnt signalling, cell cycle, and apoptosis pathways play important roles in lung development. We also have noticed an increased presence of genes in these pathways in our study (Figure 2).

Next, we analysed the 42 downregulated transcripts using IPA, and obtained 11 focus genes. The network created from these 11 genes is shown in Figure 3.

Network 2 shows a lot of activity around β-5 integrin (ITGB5) and GRB2-associated binding protein 2 (GAB2). The focus genes are shown in green coloured shapes and further description of these nodes are given in Table 5. Figure 4 shows the location of the different gene products and the canonical pathways present in Network 2. The legend for the network is shown in Figure 7.

Again, we observe similar pathways, as the ones present in Network 1, in Network 2. This is to be expected because a pathway can consist of up and downregulated genes.

Both the networks show the presence of nuclear factor-κB (NF-κB), stress-activated protein kinases /NH_2-terminal Jun kinase (SAPK/JNK) and p38 mitogen-activated protein kinase (MAPK) signalling pathways. NF-κB plays a vital role in mediating immune and inflammatory responses, and apoptosis. It regulates the expression of a large number of genes. Many of the gene products regulated by NF-κB in turn activate NF-κB, such as vascular endothelial growth factor (VEGF), and receptor for advanced glycation end product (RAGE). Activation of NF-κB involves the phosphorylation-induced, proteasome-mediated degradation of the inhibitory subunit - inhibitory protein κB. This protein is phosphorylated by an upstream serine kinase, which, in turn is phosphorylated and activated by additional upstream serine kinases. SAPK/JNK are members of the superfamily of MAP serine/threonine protein kinases. This family also includes p38 MAP kinases (p38 MAPK) and extracellular signal-related kinases (ERK) [23]. JNK/SAPK and p38 MAPK are known as stress-activated ki-

Table 1: Clinical donor characteristics

Characteristics	**PGD (n=7)**	**No PGD (n=24)**	p **value**
Age (years)	26.6 ± 8.9	24.0 ± 9.8	0.53
PaO2 (mm Hg)	406.7 ± 80.5	449.7 ± 80.0	0.17
Smoking history (pack-years)	1.5 ± 2.07	2.9 ± 6.32	0.59
Gender	71% M, 29% F	83% M, 17% F	0.78
Cause of death	57% Trauma, 43% NonTr	75% Trauma, 25% NonTr	0.66
Chest X-ray (CXR)	57% Abnl, 43% Nl	33% Abnl, 66% Nl	0.52

Table 2: Operative factors

Factors	**PGD (n=7)**	**No PGD (n=24)**	p **value**
Recipient Diagnosis	28.5% COPD, 28.5% CF, 43% Other	33% COPD, 33% CF, 33% Other	0.98
First lung ischemic time (min)	208 ± 44	240 ± 51	0.18
Second lung ischemic time (min)	330 ± 72	321 ± 51	0.69
Cardiopulmonary bypass (CPB)	72%	17%	0.02

Table 3: Outcomes of the patients with PGD and without PGD

Outcome	**PGD (n=7)**	**No PGD (n=24)**	p **value**
Days on ventilator	9.7 ± 11.7	2.0 ± 3.7	0.01
ICU stay (days)	11.3 ± 12.6	2.9 ± 3.6	0.006
Total length of stay (days)	20.3 ± 13	13.4 ± 8.1	0.09
Perioperative Mortality	28.5%	0%	0.02

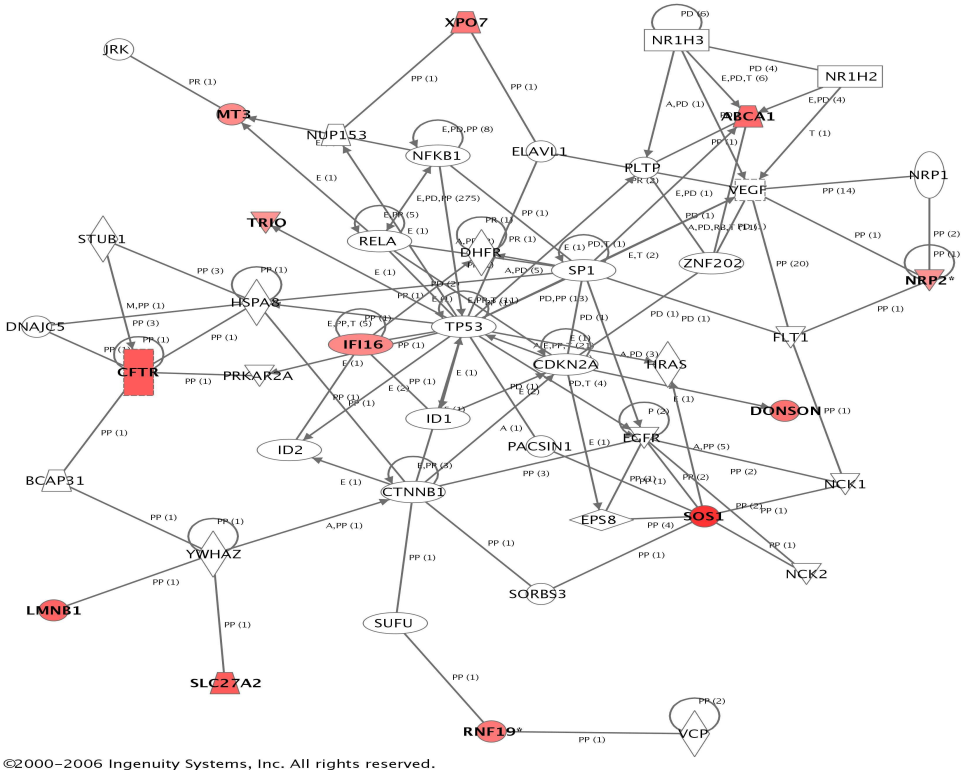


Figure 1: Network 1 - upregulated genes in PGD. This network primarily centres around tumor protein p53 (TP53). The focus genes are shown in red coloured shapes. Further details on the focus genes are provided in Table 4. The legend for this figure is Figure 7

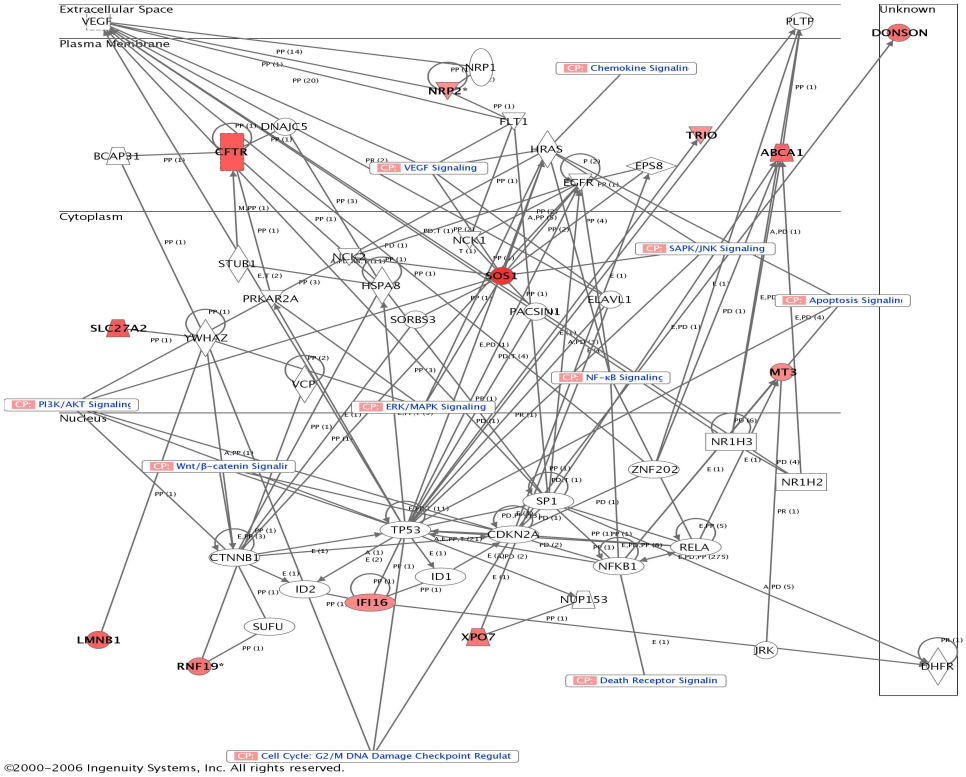


Figure 2: Network 1 with the canonical pathways overlaid. The focus genes are shown in red coloured shapes. The location of the different gene products are also depicted. Further details on the focus genes are provided in Table 4. The legend for this figure is Figure 7

Table 4: Description of the upregulated transcripts from our DE list present in Network 1

Symbol	Gene Name	Canonical Pathways	Role in Cell and diseases associated with
TP53	tumor protein p53	Apoptosis Signalling, Cell Cycle: G1/S Checkpoint Regulation, Cell Cycle: G2/M DNA Damage Checkpoint Regulation, Hypoxia Signalling in the Cardiovascular System, PI3K/AKT Signalling, SAPK/JNK Signalling, Wnt/β-catenin Signalling	apoptosis, growth, cell cycle progression, cell death, proliferation, G1 phase, transformation, senescence, G2 phase, quantity **Disease**: cancer, neoplasia, tumorigenesis, lymphoid cancer, lung cancer, head and neck cancer, breast cancer, brain cancer, squamous-cell carcinoma, metastasis, skin cancer, lymphomagenesis, Li-Fraumeni syndrome, liver cancer, hyperplasia, lung neoplasm, breast carcinoma, brain neoplasm, non-small cell lung cancer, dysplasia, etc.
IFI16	γ-interferon inducible protein 16	unknown	differentiation, apoptosis, proliferation, cell cycle progression, contact growth inhibition, morphology, accumulation, DNA damage response, G1/S phase transition, senescence
TRIO	triple functional domain (PTPRF interacting)	unknown	invasiveness, transformation, tumorigenicity, morphology, reorganisation
CFTR	cystic fibrosis transmembrane conductance regulator, ATP-binding cassette	unknown	binding, communication, whole-cell conductance, quantity, regulatory volume decrease, size, acidification, formation **Disease**: hepatic system disorder, inflammation, HIV infection, cystic fibrosis
LMNB1	lamin B1	unknown	apoptosis, disassembly, formation
SOS1	son of sevenless homolog 1 (Drosophila)	B Cell Receptor Signalling, EGF Signalling, ERK/MAPK Signalling, Estrogen Receptor Signalling, FGF Signalling, IGF-1 Signalling, IL-2 Signalling, IL-4 Signalling, IL-6 Signalling, Insulin Receptor Signalling, Integrin Signalling, Natural Killer Cell Signalling, Neuregulin Signalling, Neurotrophin/Trk Signalling, PDGF Signalling, PI3K/AKT Signalling, PPAR Signalling, PTEN Signalling, SAPK/JNK Signalling, T Cell Receptor Signalling, TGF-β Signalling, VEGF Signalling	transformation, growth, proliferation, morphology, invasion, ruffling, breakdown, maturation, reorganization, differentiation **Disease**: tumorigenesis, infection
DONSON	downstream neighbor of SON	unknown	unknown
SLC27A2	solute carrier family 27 (fatty acid transporter)	Fatty acid metabolism	unknown
RNF19	ring finger protein 19	unknown	biogenesis, cell death, quantity
NRP2	neuropilin 2	unknown	chemorepulsion, fasciculation, penetration, guidance, development, collapse, innervation, defasciculation. **Disease**: hemorrhage
ABCA1	ATP-binding cassette, sub-family A (ABC1)	unknown	binding, quantity, phagocytosis, engulfment, apoptosis, uptake, depletion **Disease**: tangier disease, atherosclerosis, microhemorrhage, cerebral amyloid angiopathy, hemorrhage, coronary artery disease, membranoproliferative glomerulonephritis, primary hypoalphalipoproteinemia
XPO7	exportin 7	unknown	unknown
MT3	metallothionein 3 (growth inhibitory factor (neurotrophic))	unknown	cell death, damage, inhibition, proliferation, growth **Disease**: limbic seizure, hypoxia, breast cancer

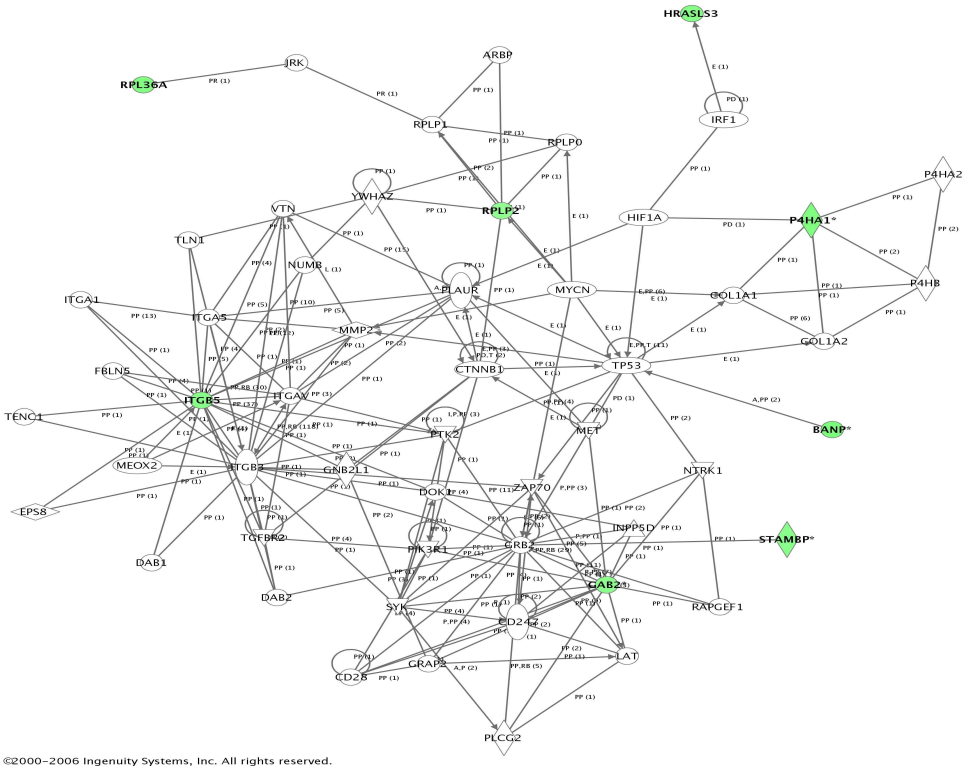


Figure 3: Network 2 - downregulated genes in PGD. This network shows a lot of activity around β-5 integrin (ITGB5) and GRB2-associated binding protein 2 (GAB2). The focus genes are shown in green coloured shapes. Further details on the focus genes are provided in Table 5. The legend for this figure is Figure 7

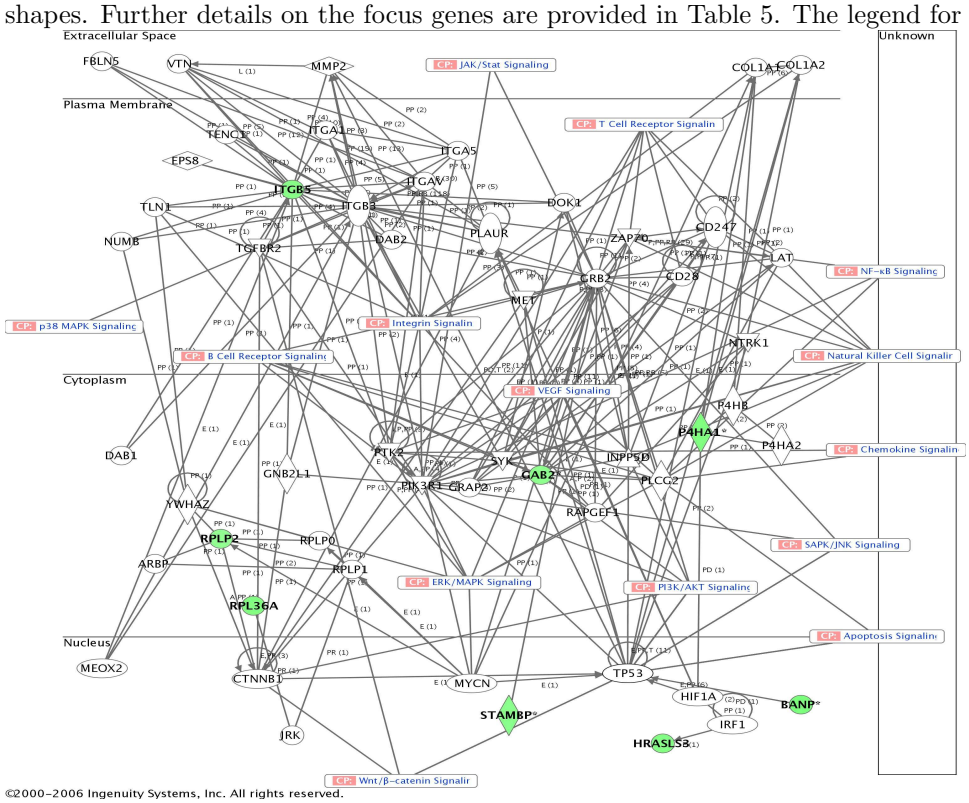


Figure 4: Network 2 with the canonical pathways overlaid. The focus genes are shown in green coloured shapes. The location of the different gene products are also depicted. Further details on the focus genes are provided in Table 5. The legend for this figure is Figure 7

Table 5: Description of the downregulated transcripts from our DE list present in Network 2

Symbol	Gene Name	Canonical Pathways	Role in Cell and diseases associated with
ITGB5	β-5 integrin	Integrin Signalling	binding, migration, adhesion, invasion, cell spreading, internalization, killing, cell death, outgrowth, anoikis **Disease**: tumorigenesis, melanoma, acute myeloid leukemia, glioblastoma multiforme, prostatic carcinoma
GAB2	GRB2-associated binding protein 2	B Cell Receptor Signalling, PI3K/AKT Signalling	proliferation, growth, differentiation, degranulation, phagocytosis, survival, adhesion, size, quantity, apoptosis **Disease**: systemic anaphylaxis, passive cutaneous anaphylaxis
STAMBP	STAM binding protein	unknown	apoptosis, cytostasis, proliferation, survival, cell death
BANP	BTG3 associated nuclear protein	unknown	cell cycle progression
P4HA1	procollagen-proline, 2-oxoglutarate 4-dioxygenase (proline 4-hydroxylase), alpha polypeptide I	Arginine and Proline Metabolism	unknown
RPLP2	ribosomal protein, large, P2	unknown	unknown
HRASLS3	HRAS-like suppressor 3	unknown	colony formation, apoptosis, proliferation **Disease**: tumorigenesis
RPL36A	ribosomal protein L36a	unknown	unknown

nases, and are responsive to numerous exogenous and endogenous stress-inducing stimuli, such as reactive oxygen species (ROS), oxidative stress, osmotic stress, proinflammatory cytokines, heat shock, and ultraviolet irradiation. Oxidative stress is defined as a persistent imbalance between the production of highly reactive molecular species (primarily oxygen and nitrogen) and antioxidant defences, finally resulting in tissue damage. There is evidence in literature that NF-κB, SAPK/JNK and p38 MAPK signalling pathways are stress-sensitive intracellular signalling systems, activation of which results in the increased expression of numerous gene products that cause cellular damage [24].

Gene products associated with stress-activated pathways emerged from both our study as well as the study in the rat model for ischemia-reperfusion injury [19]. As the experimental protocol, and animal model are different, one would not expect too much of an overlap. As suggested by the recent articles in *Nature Biotechnology* by the MicroArray Quality Control (MAQC) project [25], it is better to focus on pathways and broad functional relationships, rather than on individual genes. They state that 'even under the best circumstances, gene lists will still differ somewhat from person to person and place to place'. In our work, we have observed a good deal of overlap in the functional categories / pathways of the identified transcripts. As not all animal model studies translate well into human analysis, our investigation takes the study performed by [19] a step further by performing the analysis on human samples and showing consensus.

An exciting observation was that the metallothionein family of gene products was identified as being upregulated in both studies. We use the results from [19] and one can observe the expression level of metallothionein in microarray and RT-PCR in Figure 5. As can be seen, the levels of expression are much lower in microarray. However, RT-PCR confirms that it does have an increased expression. Hence, the rat study as well as ours do confirm the elevated expression of metallothionein.

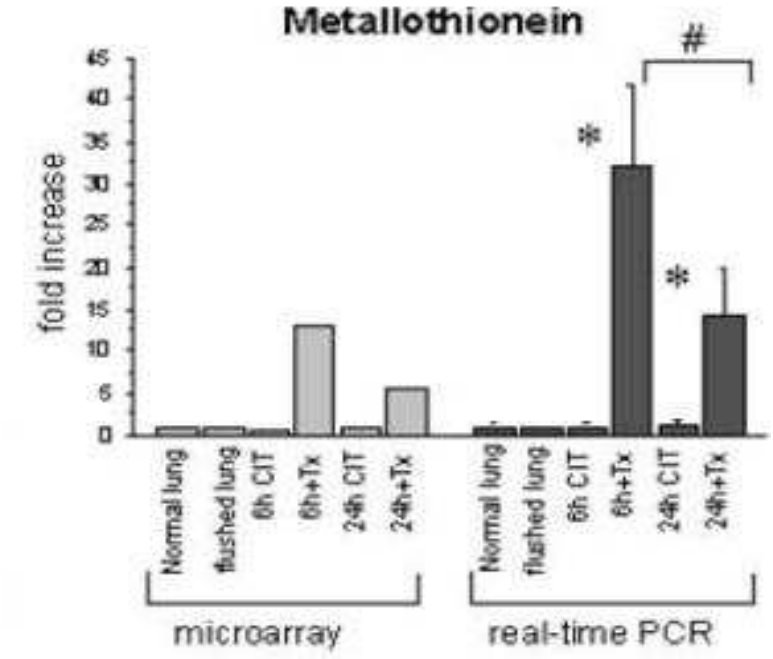


Figure 5: Metallothionein expression in microarray vs. RT-PCR in the rat study [19]. The level of expression of metallothionein is increased when verified by RT-PCR. This confirms that metallothionein does get upregulated.

In order to further study metallothionein, we extracted the metallotheionein 3 (MT3) pathway from Network 1. The MT3 pathway is shown in Figure 6.

Though the exact function of MT3 is not well known, there are a few studies that have explained the possible roles of metallothionein. A recent study has shown that metallothioneins have positive effects during the early phase of islet transplantation [26]. Another study has shown that the metallothionein gene is upregulated in wound margins particularly in regions of high mitotic activity [27]. These observations reflect its role in promoting cell proliferation and reepithelialiation. Furthermore, selected growth factors may modulate metallothionein gene expression and hence, the ability of cells to proliferate [27]. As can be seen from Figure 6, MT3 is connected to NF-κB1. In human fibroblasts, NF-κB protein consisting of p50 [NFKB1] and of p65 v-rel reticuloendotheliosis viral oncogene homolog A (RELA) increases expression of human MT3 mRNA. We already have discussed the importance of NF-κB in immune and inflammatory responses pathways. There is also an indirect relationship between MT3 and epidermal growth factor (EGF). EGF is involved in EGF signalling, ephrin receptor signalling, neuregulin signalling, and NF-κB signalling. EGF's role in the cell is proliferation, migration, mitogenesis, apoptosis, growth, chemotaxis, transformation, stimulation, S phase, and differentiation. All this information on MT3 indicates that it is a valuable gene associated with PGD and needs to be analysed in more detail. The overexpression of metallothionein may protect the lung graft from PGD. We feel that this is one of the most important insights into the mechanism of PGD.

Classification of donor lungs using SVM

The set of 100 ranked transcripts, obtained using RankGene, was used for the classification of donor lungs into PGD positive and PGD negative classes by SVM. The classifica-

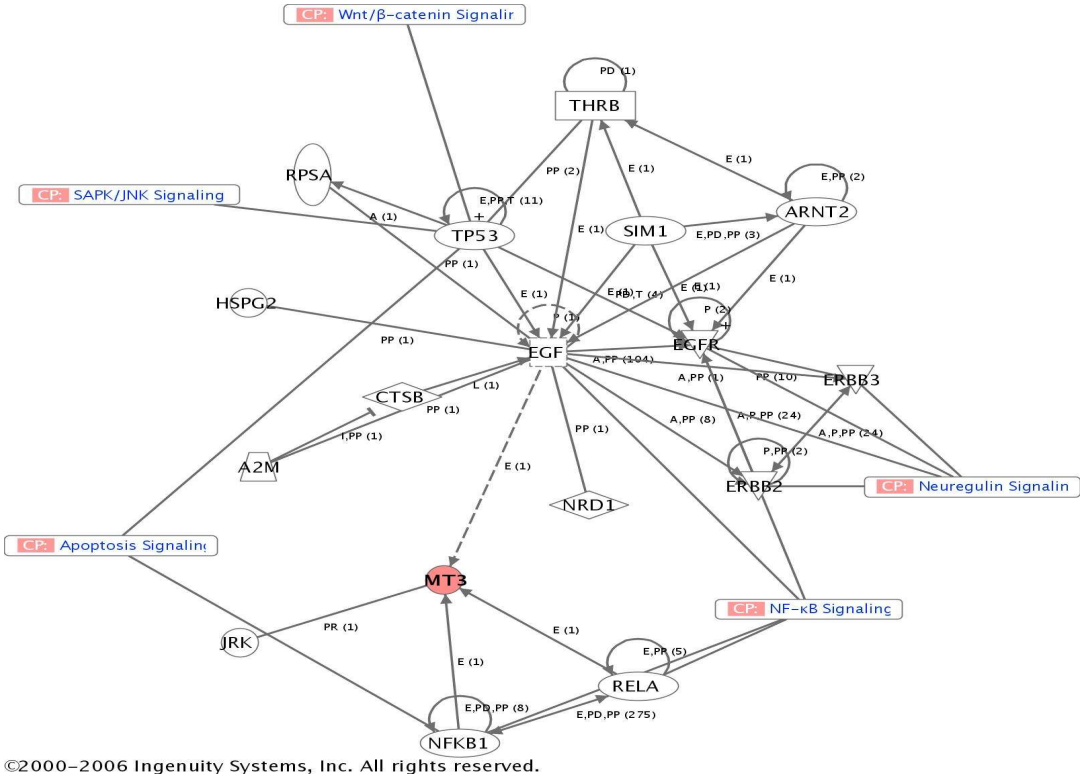
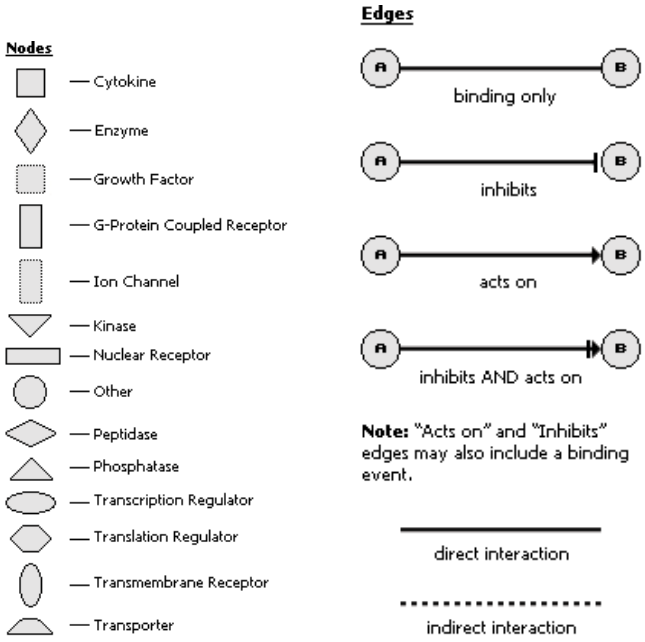


Figure 6: Network 3 - Metallothionein pathway. In human fibroblasts, NF-κB protein consisting of p50 [NFKB1] and of p65 v-rel reticuloendotheliosis viral oncogene homolog A (RELA) increases expression of human MT3 mRNA. The overexpression of metallothionein may protect the lung graft from PGD. The legend for this figure is Figure 7



(a) (b)

A **A**ctivation / Deactivation
RB Regulation of Binding
PR Protein-MRNA binding
PP Protein-Protein binding
PD Protein-DNA binding
B **B**inding (only appears prior to IPA 3.0)
E **E**xpression
I **I**nhibition
L Proteo**L**ysis
M Biochemical **M**odification
O **O**ther (only appears prior to IPA 3.0)
P **P**hosphorylation / Desphosphorylation
T **T**ranscription
LO **Lo**calization

(c)

Figure 7: Network Legend (a)Key for nodes in the network, (b)Key for edges in the network, (c)Key for edge labels in the network

tion accuracy of SVM in differentiating the two classes was 70%. This indicates that this set of transcripts does play a vital role in distinguishing unsuitable and suitable donor lungs.

The SVM did better at identifying the suitable lungs (i.e. low false negative). Considering that the motivation behind using machine learning for the selection of suitable donor lungs was to detect those that otherwise would have been discarded, this observation is promising. The unsuitable donor lungs were more often misclassified and this can be attributed to the fact that there were very few unsuitable donor lungs in the dataset (16 unsuitable lungs versus 34 suitable lungs) and subsequently, an even smaller number in the training set. Furthermore, our dataset had been pre-selected by physicians based on clinical criteria. Hence, the dataset did not have *truly* unsuitable donor lungs, i.e., lungs considered unsuitable by clinical criteria. Obviously, certain lungs that passed the selection criteria, developed PGD. In essence, these were lungs that *seemed* to be good by the current clinical criteria. Hence, the gene expression patterns of the unsuitable donor lungs are very similar to the patterns of suitable lungs. In fact, when the gene expression values of the DE transcripts were compared between PGD positive and PGD negative lungs, the difference was marginal. These observations are not surprising as both sets of lungs were considered suitable by clinical criteria, and therefore the difference between them would be very minimal. After all, the lung transplant centre at Washington University, Saint Louis, is the largest in the world and also considered as one of the best in the United States of America.

The SVM had difficulty in recognising some unsuitable donor lungs as it was not being trained on the gene expression pattern of a large number of unsuitable donor lungs, or, for that matter, on a large number of *truly* unsuitable donor lungs. Given the fact that we had only 50 samples, in which we did not have truly unsuitable lungs, the classification performance is good. Increasing the sample size in both categories would lead to a more accurate and possibly larger set of DE transcript involved in PGD, as well as improved classification results.

As the differences at the macroscopic level between PGD positive and PGD negative donor lungs are minimised after employing the clinical selection criteria, gene expression profiling would help in emphasising whatever small differences there may be. SVMs are capable of using these marginal differences to identify suitable and unsuitable donor lungs. This is where machine learning plays a valuable role - *assisting* physicians and not necessarily overruling them. Hence, machine learning methods, such as SVMs, can be used in conjunction with clinical criteria to identify unsuitable donor lungs, thereby further decreasing the chances of using donor lungs that would develop PGD.

Conclusion

The incorporation of biological information into donor lung evaluation, based on studies such as this one, may deem many of the excluded organs as suitable for transplantation, directly impacting the mortality of patients on the lung transplant waiting list. Studies show that 15-30% of patients develop clinically significant primary graft dysfunction (PGD) after lung transplantation. PGD is the single most significant factor in determining perioperative morbidity and mortality and has a devastating impact on outcome following lung transplantation. It is the primary factor determining duration of mechanical ventilatory support and length of ICU and hospital stay following lung transplantation. Perioperative mortality rates for those with clinically significant PGD are as high as 40-60%. One year survival rates fall from 69% to 40% and 2-year rates from 66% to 27% in those who suffer significant PGD. Furthermore, those that survive complications of PGD endure lengthy hospitalisation periods and a protracted and often compromised recovery, evidenced by inferior exercise tolerance and pulmonary function testing and the inabil-

ity to achieve independent lifestyles. Moreover, PGD is now being identified as a risk factor for acute and chronic rejection.

In this study, gene expression profiling of donor lung samples was used to determine gene products that are associated with the development of PGD after transplantation. It also resulted in analysing possibly relevant pathways involved in PGD. When biological markers were used to differentiate between PGD positive and PGD negative lungs, a good classification accuracy was achieved. The incorporation of biological markers into donor organ evaluation will have a significant impact on outcomes after lung transplantation, by potentially expanding the donor pool of organs selected for transplantation and by identifying lungs at risk for the development of PGD post-transplant, which would allow pre-treatment of these high risk organs or matching of these organs to relatively lower risk recipients. Further identification and elucidation of genetic markers in donor lungs associated with PGD could have a significant impact on lowering the incidence and preventing the morbidity and mortality of PGD after lung transplantation. Our results indicate that we have successfully achieved both our objectives.

Materials and Methods

Donor Lung Sampling

From August 2003 to January 2005, biopsies of 50 donor lungs used for bilateral sequential lung transplantation at Washington University School of Medicine were obtained from the anterior right middle lobe or lingula immediately prior to cold-flushing and these samples were immediately snap-frozen in liquid nitrogen and then stored in a -70 Celsius freezer until used for analysis. Specimens were sampled using standard techniques for open lung wedge biopsy. An area of lung tissue approximately 1 x 1 cm was isolated and excised using 2 staple lines from a 30 mm EndoGIA stapler (US Surgical, Norwalk, CT). This protocol was approved by the Human Studies Committee and Institutional Review Board at Washington University School of Medicine and protection of human subjects, namely recipients, was afforded by detailed informed consent before entrance into this research protocol.

RNA Isolation

Single isolates of donor lung samples were homogenised in the presence of RNAzolB and finally dissolved in RNase-free H2O. 25 g of total RNA was treated with DNase using the Qiagen RNase-free DNase kit and samples were further purified using RNeasy spin columns (Qiagen, Valencia, CA). Total RNA treated with DNase was dissolved in RNase-free H2O to a final concentration of 0.2 g/ l. RNA quality was assessed by 1% agarose gel electrophoresis in the presence of ethidium bromide. Samples that did not reveal intact and approximately equal 18S and 28S ribosomal bands were excluded from further study.

cDNA Synthesis and Gene Expression Profiling

This study used commercially available high-density microarrays (Affymetrix, Santa Clara, CA) that produce gene expression levels on 22,278 probe sets (Affymetrix Human Genome U133Av2.0 Array). Each donor lung biopsy was analysed on a different GeneChip. Preparation of cDNA, hybridisation, and scanning of the arrays were performed according to the manufacturer's instructions. The arrays were scanned using the Affymetrix GeneArray scanner. Image analysis was performed with the Affymetrix GeneChip software. We also performed a quality control test on the dataset and results indicated that the dataset quality was good.

Data

The data from all 50 gene chips was normalised using the GCRMA method developed by [28]. The 50 donor lung samples were divided into two groups - those that developed PGD after transplantation (PGD positive) and those that did not (PGD negative). PGD was defined as T0 Grade III dysfunction according to International Society for Heart and Lung Transplantation criteria - that is, a ratio of partial pressure of arterial oxygen (PaO_2) to fraction of inspired oxygen (FiO_2) less than 200 on the first arterial blood gas in the intensive care unit after transplantation (generally 4-6 hours after actual reperfusion) [29]. Sixteen samples were classified as PGD positive according to this definition and the remaining thirty-four were PGD negative.

Transcripts Selection

As data quality is important, we calculated the quality of our dataset using the R package 'affyQCReport' [30] and results were favourable. We then proceeded to the next step in our study - the identification of differentially expressed (DE) transcripts. The objective was to find a set of DE transcripts/probes that could be used as a molecular signature for the condition. DE transcript extraction falls into two broad categories - *wrapper* methods and *filter* methods. In wrapper transcript selection methods, the DE transcript identification phase is integrated with the classification phase. In filter methods, the DE transcript extraction phase is independent of the classification phase. In this study, we used two packages for the identification of DE transcripts - RankGene [31], and significance analysis of microarrays (SAM) [32].

RankGene is a programme for analysing gene expression data, feature selection and ranking genes based on the predictive power of each gene/transcript to classify samples into functional or disease categories. It supports eight different measures for quantifying a gene's ability to distinguish between classes. For our analysis, we used the t-statistics measure of predictability. SAM is an open-source software which identifies DE genes based on the change in gene expression relative to the standard deviation of repeated measurements [32]. It uses the false discovery rate (FDR) and q-value method presented in [33] to select genes. The q-value is analogous to the p-value and is corrected, through a permutation process, for the natural variability of the expression data. The q-value of a transcript is the FDR for the transcript list that includes that transcript and all transcripts that are more significant. SAM also provides a tail strength (TS) value which measures the deviation of each p-value from its expected value. Therefore, large positive TS values indicate evidence against the null hypothesis, i.e., there are more small p-values than one would expect by chance [34].

We first ran RankGene on the complete set of probes. Since we were interested in the most highly DE transcripts, we chose to take the top 100 transcripts from the ranked list for further analysis. On this list of 100 DE transcripts, we applied SAM. SAM output 81 up and down regulated transcripts based on a FDR of 0% and a TS of 92.7%. After averaging the values of and removing multiple probes matching to the same gene name, 23 upregulated and 42 downregulated transcripts were obtained. These sets of up and down regulated transcripts were used for further analysis in Ingenuity Pathway Analysis software.

Pathway analysis

Ingenuity Pathway Analysis (IPA) (www.ingenuity.com) was used to perform pathway analysis on the two sets of DE transcripts - upregulated and downregulated, to identify networks of genes that are known to interact functionally. IPA uses the Ingenuity Pathways Knowledge Base(IPKB) which contains large amounts of indivually modelled relationships between objects (e.g., genes, proteins and mRNAs) to dynamically generate significant biological/gene expression networks and pathways. The identified DE transcripts from our analysis that are mapped onto the IPKB are called 'focus genes'. These are used as starting points for building the networks. First, IPA queries the IPKB for interactions between the focus genes and all other genes stored in IPKB and then generates a set of networks/pathways with a maximum of 35 genes. A p value for each network is calculated according to the user's list of DE genes. This is accomplished by comparing the number of focus genes that are present in a given pathway, relative to the total number of occurrences of those genes in all pathways stored in IPKB. The score of the network is shown as the negative logarithm of the p value, indicating the likelihood of the focus genes in a network being found together by random chance. In our study, we further analysed networks that had a network score of 10 or higher. This network analysis is an exploratory *in silico* approach and does not necessarily indicate that the pathway or network actually exists.

Support Vector Machines

Originally developed by Vapnik [35], the support vector machine (SVM) is a statistical learning tool which has been extensively used for binary classification with great success. Ranging from classification of cancer [36] to determination of haemodialysis dosage [37], SVMs have proven to be an effective tool in a wide-range of applications.

SVM was used for the classification of patient samples into PGD positive or PGD negative categories. The dataset consisted of 50 patient samples and 100 transcripts (ranked transcripts from RankGene). Following is the manner in which SVM was used. The dataset is divided into training and test(unseen by the classifier) sets. The test set is

also the validation set because although the user knows the classes of the samples in the test set, the classifier does not see the samples in the test set while it is training. The SVM is trained on the training set. The classifier performance is measured by the prediction accuracy on the test set. It is quite well known that the set of significant genes (SG) from a particular set of training data is very often very different from one chosen from a different set of training data. Obtaining a SG set from the complete dataset (i.e. from all 50 patient samples), leads to a selection bias. In order to avoid selection bias, an external cross-validation (CV) was performed i.e. the classifier performance was measured using only the set of genes (i.e. a subset of the 100 transcripts) obtained from the training set and not from the complete dataset of 50 patients. Ten fold CV was carried out rather than leave-one-out (LOO) CV as the variability in the list of SG is much lower with 10 fold CV and this is what is preferred.

Acknowledgement

The authors would like to thank Seth D. Crosby, M.D., at the Genome Sequencing Centre at Washington University School of Medicine for his help with Ingenuity Pathway Analysis. This research was funded by NIH-NRSA 5F32 HL074687 (S. Dharmarajan), NIH R01 HL41281 (G.A. Patterson) and by NSF grants EIA-0113618 and IIS-0535257 (W. Zhang).

References

[1] EP Trulock, LB Edwards, DO Taylor, MM Boucek, BM Keck, and MI Hertz. Registry of the international society for heart and lung transplantation: Twenty-second official adult lung and heart-lung transplant report - 2005. *J Heart Lung Transplant*, 24:956–967, 2005.

[2] BF Meyers, de la Morena, SC Sweet, EP Trulock, TJ Guthrie, Mendeloff EN, Huddleston C, Cooper JD, and Patterson GA. Primary graft dysfunction and other selected complications of lung transplantation: A single-center experience of 983 patients. *J Thorac Cardiovasc Surg*, 129:1421–1429, 2005.

[3] JD Christie, JS Sager, SE Kimmel, VN Ahya, C Gaughan, NP Blumenthal, and RM Kotloff. Impact of primary graft failure on outcomes following lung transplantation. *Chest*, 127:161–165, 2005.

[4] JD Christie, RM Kotloff, VN Ahya, G Tino, A Pochettino, C Gaughan, E DeMissie, and SE Kimmel. The effect of primary graft dysfunction on survival after lung transplantation. *Am J Respir Crit Care Med*, 171:1312–1316, 2005.

[5] AJ Fisher, JH Dark, and PA Corris. Improving donor lung evaluation: a new approach to increase organ supply for lung transplantation. *Thorax*, 53:818–820, 1998.

[6] D Weill. Donor criteria in lung transplantation: an issue revisited. *Chest*, 121:2029–2031, 2002.

[7] AJ Fisher, SC Donnelly, G Pritchard, JH Dark, and PA Corris. Objective assessment of criteria for selection of donor lungs suitable for transplantation. *Thorax*, 59:434–437, 2004.

[8] AE Frost. Donor criteria and evaluation. *Clin Chest Med*, 18:231–237, 1997.

[9] S Sundaresan, J Semenkovich, L Ochoa, G Richardson, EP Trulock, JD Cooper, and GA Patterson. Successful outcome of lung transplantation is not compromised by the use of marginal donor lungs. *J Thorac Cardiovasc Surg*, 109:1075–1080, 1995.

[10] SM Bhorade, W Vigneswaran, MA McCabe, and ER Garrity. . liberalization of donor criteria may expand the donor pool without adverse consequence in lung

transplantation. *J Heart Lung Transplant*, 19:1199–1204, 2000.

[11] E Gabbay, TJ Williams, AP Griffiths, LM Macfarlane, TC Kotsimbos, DS Esmore, and GI Snell. Maximizing the utilization of donor organs offered for lung transplantation. *Am J Respir Crit Care Med*, 160:265–271, 1999.

[12] JD Christie, JE Bavaria, HI Palevsky, L Litzky, NP Blumenthal, LR Kaiser, and RM Kotloff. Primary graft failure following lung transplantation. *Chest*, 114:51–60, 1998.

[13] SM Fiser, IL Kron, SM Long, AK Kaza, JA Kern, DC Cassada, DR Jones, MC Robbins, and CG Tribble. Influence of graft ischemic time on outcomes following lung transplantation. *J Heart Lung Transplant*, 20:1291–1296, 2001.

[14] RC King, OAR Binns, F Rodriguez, RC Kanithanon, TM Daniel, WD Spotnitz, CG Tribble, and IL Kron. Reperfusion injury significantly impacts clinical outcome after pulmonary transplantation. *Ann Thorac Surg*, 69:1681–1685, 2000.

[15] RJ Novick, LE Bennett, DM Meyer, and JD Hosenpud. Influence of graft ischemic time and donor age on survival after lung transplantation. *J Heart Lung Tranplant*, 18:425–431, 1999.

[16] BF Meyers, J Lynch, EP Trulock, TJ Guthrie, JD Cooper, and GA Patterson. Lung transplantation: a decade of experience. *Ann Surg*, 230:362, 1999.

[17] AJ Boujoukos, GD Martich, JD Vega, RJ Keenan, and BP Griffith. Reperfusion injury in single-lung transplant recipients with pulmonary hypertension and emphysema. *J Heart Lung Transplant*, 16:440–448, 1997.

[18] EP Trulock. Lung transplantation: recipient selection. *Chest Surg Clin N Am*, 3:1–18, 1993.

[19] M Yamane, M Liu, H Kaneda, S Uhlig, TK Waddell, and S Keshavjee. Reperfusion-induced gene expression profiles in rat lung transplantation. *American Journal of Transplantation*, 5:21602169, 2005.

[20] AE Bonner, WJ Lemon, and M You. Gene expression signatures identify novel regulatory pathways during murine lung development: implications for lung tumorigenesis. *J Med Genet.*, 40(6):408–17, 2003.

[21] Genetic parallels found between lung development and lung cancer. *http://www.sciencedaily.com/releases/2006/07/0607* July 10, 2006.

[22] AE Bonner, WJ Lemon, TR Devereux, RA Lubet, and M You. Molecular profiling of mouse lung tumors: association with tumor progression, lung development, and human lung adenocarcinomas. *Oncogene*, 23(5):1166–1176, 2004.

[23] JM Kyriakis and J Avruch. Sounding the alarm: protein kinase cascades activated by stress and inflammation. *J Biol Chem.*, 271(5):2431324316, 1996.

[24] JL Evans, ID Goldfine, BA Maddux, and GM Grodsky. Are oxidative stress-activated signaling pathways mediators of insulin resistance and β-cell dysfunction? *Diabetes*, 52(1):1–8, 2003.

[25] MAQC Consortium. The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24:1151 – 1161, 2006.

[26] X Li, H Chen, and PN Epstein. Metallothionein protects islets from hypoxia and extends islet graft survival by scavenging most kinds of reactive oxygen species. *J Biol Chem*, 279:765771, 2004.

[27] ABG Lansdown. Metallothioneins: potential therapeutic aids for wound healing in the skin. *Wound Repair and Regeneration*, 10(3):130–132, 2002.

[28] Z Wu, RA Irizarry, R Gentleman, FM Murillo, and F Spencer. A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.*, 99:909–917, 2004.

[29] JD Christie, M Carby, R Bag, P Corris, M Hertz, and D Weill. Report of the ishlt working group on primary lung graft dysfunction part ii: Definition. a consensus statement of the international society for heart and lung transplantation. *J Heart Lung Transplant*, 24:1454–1459, 2005.

[30] C Parman and C Halling. affyqcreport: A package to generate qc reports for affymetrix array data. *www.bioconductor.org*, 2006.

[31] Y Su, TM Murali, V Pavlovic, M Schaffer, and S Kasif. Rankgene: Identification of diagnostic genes based on expression data. *Bioinformatics*, 19(12):1578–1579, 2003.

[32] VG Tusher, R Tibshirani, and G Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA*, 98:5116–5121, 2001.

[33] JD Storey. A direct approach to false discovery rates. *J. R. Statist. Soc. B*, 64(3):479498, 2002.

[34] J Taylor and R Tibshirani. A tail strength measure for assessing the overall univariate significance in a dataset. *Biostatistics*, 7(2):167181, 2006.

[35] V Vapnik. Support vector machines. *Statistical learning theory, John Wiley and Sons Inc.*, 1998.

[36] T Furey, N Cristianini, N Duffy, D Bednarski, M Schummer, and D Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–14, 2000.

[37] M Ray and N Atray. Support vector machine for determining dose of dialysis. *Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2004.