Recalibrating...“Omics” in ALI/ARDS: identification of novel targets for therapy

Claudia C. dos Santos  M.D. M.Sc. F.R.C.P.C

Assistant Professor of Medicine, Parker B. Francis Fellow, Clinician Scientist
Heart and Stroke Foundation

Scientist in the Keenan Research Centre of the Li Ka Shing Knowledge Institute
of St. Michael's Hospital

Saint Michael’s Hospital, Critical Care Department
Thoracic Surgery Research Laboratory
Interdepartmental Division of Critical Care

University of Toronto
No conflict to Declare
It's been over 10 years……where are we?

Unfulfilled Promise in Mapped Genome

By Sara Reeve on February 5, 2010 10:04 AM

Labeled the “blueprint of life” by leading scientists, the human genome sequence was going to usher in a new age of truly personalized and innovative medicine…. However, the promise has not yet been fulfilled…….
“Currently I have moved much closer to the idea of “genetic irrelevance” the idea that in the overwhelming majority of cases, our genes are of much less importance in determining our fate and that the environment in which we live and the lifestyle choices we make are of far greater importance.”

Rethinking the Promise of Genomics
By: Terry Grossman
Published: July 25, 2010
...Where are we?

Now.....

10 years ago.....the leap of faith!
Down-Regulation of the Euphoria
The Molecular Revolution
Unfulfilled promises of simplicity!

LIFE IS COMPLICATED

The more biologists look, the more complexity there seems to be. Erika Check Hayden asks if there’s a way to make life simpler.

Not that long ago, biology was considered by many to be a simple science, a pursuit of expedition, observation and experimentation. At the dawn of the twentieth century, while Albert Einstein and Max Planck were writing the critical laws of quantum mechanics, it turns out, is closer to 21,000, and biologists now know what many of those genes are. But at the same time, the genome sequence did what biological discoveries have done for decades. It opened the door to a vast labyrinth of new questions. Truly know an organism — or even a cell, an organelle or a molecular pathway — down to the finest level of detail?

Imagine a perfect knowledge of inputs, outputs and the myriad interacting variables, enabling a predictive model. How tantalizing this notion is depends somewhat on the scientist; some say it is enough to understand the basic principles that govern life, whereas others are still looking for that last variable.
Limitations in ‘omics’ technology

- Why have these studies not yielded novel clinically relevant therapies….yet?
- Difficult to make direct comparisons
- Different laboratory protocols, platforms, analysis
- Reliance on single mediators
- Different patient populations
- Small sample sizes
- Few, No, or inadequate cross-validation
- Prediction models prone to over-fitting
- Low prediction accuracies
- Less generalizable
PAUSE.
Let's Capitalize on the data we already have! And look again....
Rationale

- **Microarray meta-analysis (gene-gene)**
- Capitalizes on existing data
- Strategy substantially increases sample size
- Reduces the effect of noise
- Exploits data collected using different platforms and different model organisms
- Focuses on molecular signatures as opposed to individual molecules
- Increases generalizability
## Microarray Databases

- **Systematic search of data repositories**

<table>
<thead>
<tr>
<th>Database</th>
<th>Scope</th>
<th>Microarray experiment sets</th>
<th>Sample profiles</th>
<th>As of date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Expression Omnibus - NCBI</td>
<td>any curated MIAME compliant molecular abundance study</td>
<td>8094</td>
<td>205148</td>
<td>Mar-08</td>
</tr>
<tr>
<td>Stanford Microarray database</td>
<td>??</td>
<td>12742</td>
<td>?</td>
<td>Apr-08</td>
</tr>
<tr>
<td>Genevestigator database</td>
<td>Manually curated microarray data for expression meta-analysis</td>
<td>&gt;500</td>
<td>20500</td>
<td>July, 2008</td>
</tr>
<tr>
<td>ArrayExpress at EBI</td>
<td>Any curated MIAME or MINSEQE compliant transcriptomics data</td>
<td>4194</td>
<td>110731</td>
<td>May, 2008</td>
</tr>
<tr>
<td>UPenn RAD database</td>
<td>MIAMI compliant public and private studies, associated with ArrayExpress</td>
<td>~100</td>
<td>~2500</td>
<td>Sept. 1, 2007</td>
</tr>
<tr>
<td>UNC Microarray database</td>
<td>??</td>
<td>~31</td>
<td>2093</td>
<td>Apr-07</td>
</tr>
<tr>
<td>MUSC database</td>
<td>??</td>
<td>~45</td>
<td>555</td>
<td>Apr-07</td>
</tr>
<tr>
<td>caArray at NCI</td>
<td>Cancer data, prepared for analysis on caBIG</td>
<td>41</td>
<td>1741</td>
<td>Nov-07</td>
</tr>
<tr>
<td>UPSC-BASE</td>
<td>Data generated by microarray analysis within Umeå Plant Science Centre (UPSC)</td>
<td>~100</td>
<td>?</td>
<td>Nov-07</td>
</tr>
</tbody>
</table>
Extract data from studies

- Only included studies where we had access to the **RAW** image files

  - This allowed us to re-analyze the data at the probe level:
    - Identify and remove arrays with poor quality
    - Check for batch effects
    - Filter out probes with poor quality spots
    - Aggregate technical replicates
    - Normalize data from different platforms using the same strategy

  .cel files – the raw image files

Hu et al. Submitted
Identify suitable studies

Studies involved ACUTE lung injury only in ventilated models Up to Feb 2008

<table>
<thead>
<tr>
<th>Rat</th>
<th>Mouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>RG-U34A</td>
<td>RG-U34A</td>
</tr>
<tr>
<td>RAE230V2</td>
<td>MOE430A</td>
</tr>
<tr>
<td>MG-U74A</td>
<td>Codelink 10k</td>
</tr>
<tr>
<td>Codelink 20k</td>
<td></td>
</tr>
</tbody>
</table>

+  
<table>
<thead>
<tr>
<th>Rat</th>
<th>Mouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>RG-U34A</td>
<td>RAE230V2</td>
</tr>
<tr>
<td>MOE430A</td>
<td></td>
</tr>
</tbody>
</table>

Hu et al. Submitted
Annotate and Resolve many-to-many relationships

- Summarize probe IDs – one probe ID for each EST
- Identify common probes
- Discard probes that are not shared between platforms and species
- If a probe matches to multiple gene IDs collapse data into a single expression value

Hu et al. Submitted
Building an Ortholog Database

Common “dictionary” for probes coming from different species and different platforms

Eukaryotic Gene Orthologues Database (Sequence Identity)
HGNC Comparison of Orthology Predictions Database (Functional Identity)
Manual Curation (BLAST)

Affymetrix Human U133 Plus 2.0 Microarray Platform

Mine databases to identify sequence and functional orthologs for ESTs on relevant microarray platforms
Match probes to human U133 Plus 2.0 platform to create master database

All probes were collapsed to Gene IDs defined by the Human U133 Plus 2.0 platform (contains 54,000 ESTs)

Hu et al. Submitted
Combine chips based on biological question

85 chips representing 9 experiments

3,363,450 data points

39,570 ortholog expressed sequence tags

Single-hit Model
HV vs. LV
Random Effects Model

Double-hit Model
MV+Inj vs. Inj
Random Effects Model

Classify individual chips based on mode of experimental lung injury. Calculate change in gene expression (effect size) for each comparison.

Random effects model to calculate mean effect size across all experiments, apply multiple comparisons correction.

Select significant genes by FDR (adj. p-value) and number of studies

Hu et al. Submitted
## Combining chips

<table>
<thead>
<tr>
<th>Method used to induce lung injury</th>
<th>Injury Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>No MV (NV) – spontaneous breathing or sham animals (surgical procedures but no lung injury)</td>
<td>Sham/Control</td>
</tr>
<tr>
<td>Inflammation alone (Inf): exposure to Lipopolysaccharide (LPS; intravenous [IV] or intraperitoneal [IP]) or acid aspiration (AA) alone</td>
<td>One Hit (Inflammation)</td>
</tr>
<tr>
<td>MV alone (MV): exposure to HV = high tidal volume $\geq 15$ ml/kg and/or $b&gt;1$ (overdistension) vs. LV = low tidal volume $\leq 12$ ml/kg and/or $b=1$ (optimal recruitment)</td>
<td>One Hit (Mechanical Ventilation)</td>
</tr>
<tr>
<td>MV combined with Inflammation (MV + Inf): where the first hit is exposure to with LPS or AA and the second is MV (HV/$b&gt;1$ or LV/$b=1$)</td>
<td>Two Hits (Inflammation + Mechanical Ventilation)</td>
</tr>
</tbody>
</table>

Hu et al. Submitted
Analysis Strategy

Two main comparisons are reported:

- **One-Hit Analysis**: Non-Injurious MV vs. Injurious MV
- **Two-Hit Analysis**: Inflammation vs. MV + Inflammation
The effect size $y_{ig}$ for gene $g$ in individual study $i$ was measured using the standardized mean difference

$$y_{ig} = \frac{(\bar{x}_{igt} - \bar{x}_{igc})}{S_{ig}^{pool}}$$ (1)

where $\bar{x}_{igt}$ and $\bar{x}_{igc}$ are the sample means of gene expression values for gene $g$ in treatment group $t$ (e.g. treatment) and group $c$ (e.g. control) of study $i$, respectively. $S_{ig}^{pool}$ is the pooled standard deviation. The estimated variance $s_{ig}^2$ of the unbiased effect size $y_{ig}$ is given by

$$s_{ig}^2 = (1/n_{it} + 1/n_{ic}) + y_{ig}^2 (2(n_{it} + n_{ic}))^{-1}$$ (2)

For a study with $n(n = n_t + n_c)$ samples, an approximately unbiased estimate of $y_{ig}$ is given by $y_{ig}^* = y_{ig} - 3y_{ig} / (4n - 9)$
For each gene $g$, the effect size $y_{ig}$ ($i=1,\ldots, I$) in $I$ studies was estimated using equation (1). We follow Choi et al. (2003) to place the estimated $\tilde{y}_{ig}$ into a hierarchical model and to test for differences between groups:

$$
\begin{align*}
\begin{cases}
y_{ig} = \theta_g + \varepsilon_{ig}, & \varepsilon_{ig} \sim N(0, s_{ig}^2) \\
\theta_g = \mu_g + \delta_g, & \delta_g \sim N(0, \tau_g^2),
\end{cases}
\end{align*}
$$

(3)

where $\tau_g^2$ is the between-study variability of gene $g$, $\mu_g$ means the average measure of differential expression across the $I$ studies for gene $g$. Here, $\tau_g^2$ and $\mu_g$ are gene-specific while $s_{ig}^2$ and $y_{ig}$ are gene and study-specific. $s_{ig}^2$ is the effect size variance of gene $g$, measuring the sampling error for the $i^{th}$ study. Using a random effects model (Choi et al. 2003), the meta-analysis estimate for $\mu_g$ can be calculated as:

$$
\hat{\mu}_g = \frac{\sum_{i=1}^{I} w_i y_{ig}}{\sum_{i=1}^{I} w_i}
$$

(4)
where the weights are given by \( w_i = \frac{1}{s_{ig}^2 + \tau^2} \) and \( \tau^2 \) is the between-study variability.

The variance of this estimator is obtained by

\[
\text{Var}(\hat{\mu}_g) = \frac{1}{\sum_{i=1}^{I} w_i} \tag{5}
\]

A test statistic to evaluate the treatment effect of gene \( g \) across all \( I \) studies can then be computed as

\[
z_g = \frac{\hat{\mu}_g}{\sqrt{\text{var}(\hat{\mu}_g)}} \tag{6}
\]
False Discovery Rates

- To determine the significance of each gene identified
  - Calculated a p-value corresponding to the z-statistics
  - Estimated a False Discover Rate (Benjamini and Hochberg 1995)
Selection of differentially expressed genes

(a) Relationship between adjusted p-value and mean effect size (MSE)
Overlap between significantly expressed genes

One-hit Model
HV vs. LV
299 genes

Two-hit Model
MV+Inf vs. Inf
723 genes

Both
40 genes

(FDR≤0.1) between one-hit and two-hit models

Hu et al. Submitted
Identification of Novel Genes

Identification frequency of significantly expressed genes from one-hit and two-hit models compared to the original individual studies.

Hu et al. Submitted
External Data Sets of Animal Models of Lung Injury

GSE11434, GSE9368, and GSE920829

Data sets downloaded from GEO containing experiments performed after Feb 2007

Unsupervised Clustering: No a priori information regarding group assignment is provided

Classification accuracy: % of samples correctly classified
Classification Accuracy in External Animal Data set:

100% Prediction accuracy

Hierarchical clustering of samples from mouse model of VILI using top 20 genes from single-hit model of meta-analysis as ranked by lowest FDR.


Hu et al. Submitted
Prospective Validation:

One-Hit

B qRT-PCR

Gadd45a
Irx3
Hk2

Fold change in gene expression (relative to Gapdh and baseline non-ventilated)

CD79a
HoxB5

LV HV LV+B HV+B

Two-Hits

Csad
Prx
Tspan3
Btg3
Kcnj8

Fold change in gene expression (relative to Gapdh and baseline non-ventilated)

LV HV LV+B HV+B

Pneumonia (48 hrs), HV ventilation (x 4hrs)

Hu et al. Submitted
Can we validate data in human cells?

**Beas2b Cells**

**One-Hit**

**Two-Hits**

**SAEC Cells**

Hu et al. Submitted
Can we use this data to detect lung injury in human samples?

- Major problem!
  - ACUTE samples (<8 hours)
  - Lung tissues

\[ \text{Microarray experiments} \]

- Clinically relevant samples of ACUTE lung injury?
- Access to lung tissues?

Hu et al. Submitted
Lung Transplantation Data Sets

Donor Lungs

Usually have ALI
Short duration of ALI
Ventilated for short periods of time
Access to tissue
Validation in human samples

Donor

Recipient

Acute Primary Graft Failure
1. Can the meta-analysis predicted “injury” expression profiles be detected in donor lungs?

2. Do they predict those patients that will go on to develop PGF?
Microarray data from donor lungs

- Washington data set (GSE8021)
  

- 50 Affymetrix chips (U133 Plus 2)

- RNA extracted immediately following organ procurement

- Prior to cold-flushing

![Diagram showing 50 chips, 16 PGFs, and 34 No-PGFs]
Misclassification-Penalized Posteriors (MiPP)

- Prediction performed in Washington data set:
- Split the data set:
  - 33 samples $\rightarrow$ Training set
  - 17 samples $\rightarrow$ Test set
- Used MiPP to sequentially select genes
- One gene at the time
- Linear Discriminant Analysis
- Determine the smallest genes combination to inform regarding sample classification

Hu et al. Submitted
MiPP calculates the posterior probability of correct sample classification.

MiPP score is the sum of posterior probabilities of correct classification

MINUS

The number of incorrect classification.

Combination of genes with best MiPP score is the most informative.
# Identification of Lung Injury Profiles in Human Donor Lungs

<table>
<thead>
<tr>
<th>ONE-HIT</th>
<th>TWO-HITS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prediction Accuracy</strong></td>
<td><strong>Prediction Accuracy</strong></td>
</tr>
<tr>
<td>88.9% in the training group</td>
<td>82.4% in the training group</td>
</tr>
<tr>
<td>77.6% in the test group</td>
<td>71.3% in the test group</td>
</tr>
</tbody>
</table>
Probabilities were calculated based on five-fold cross-validation on 50 samples.

- Accuracy 86%
- Compared to 57.7% using traditional P:F ratios

Samples are plotted by index number, and the vertical axis represents the probability of classification into either the PGF (red) or GOOD (green) group.
Receive operating characteristic (ROC) of PGF and GOOD samples using the 15-gene prediction model.

ONE-HIT List

- Sensitivity of 75%
- Specificity of 91.2% for the 50-sample dataset

ROC curve is drawn based on the five-fold cross-validation of the 50 human donor lung samples using the 15-gene prediction model.

Hu et al. Submitted
Conclusions

- Clinically relevant translational biology approaches
- Capitalizing of available (electronic) data
- Hypothesis driven vs. Fishing for a Hypothesis
- Validation – prospective validation
- The challenge and hope for the future

Can we ‘filter out mouse genes’?

Allows ‘recalibration’ of mouse data based on clinically relevant human samples
Acknowledgements

- Mentorship
  - Arthur S. Slutsky
  - Mingyao Liu

- Contributors
  - Pingzhao Hu

Gene Expression Omnibus

And all my colleagues that have generously contributed by depositing their data in the repository
Summary

- RNA from lungs - animal models of ALI/ARDS
- Only ventilated models (clinically relevant)
- Microarray meta-analysis – pooling of data to increase sample size
- Used a Random effects model to compare effect sizes
- Compared samples based on pathophysiological principles (Ventilation +/- Inflammation)
- Identified 2 lists containing potential predictors
- Validated lists in external animal models
- Validated list in human cells (primary and transformed)
- Validated lists in clinically relevant model of ALI/ARDS

- 15 gene predictors of likelihood of developing PGF
<table>
<thead>
<tr>
<th>One-hit Model</th>
<th>Two-hit Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Term</strong></td>
<td><strong>Term</strong></td>
</tr>
<tr>
<td>Starch and Sucrose Metabolism</td>
<td>LPS/IL-1 Mediated Inhibition of RXR Function</td>
</tr>
<tr>
<td>No. of Genes: 6</td>
<td>No. of Genes: 17</td>
</tr>
<tr>
<td>P-value: 4.57E-5</td>
<td>P-value: 1.23E-6</td>
</tr>
<tr>
<td>NF-κB Signaling</td>
<td>NRF2-mediated Oxidative Stress Response</td>
</tr>
<tr>
<td>No. of Genes: 8</td>
<td>No. of Genes: 15</td>
</tr>
<tr>
<td>P-value: 5.50E-5</td>
<td>P-value: 1.35E-6</td>
</tr>
<tr>
<td>VEGF Signaling</td>
<td>Integrin Signaling</td>
</tr>
<tr>
<td>No. of Genes: 6</td>
<td>No. of Genes: 18</td>
</tr>
<tr>
<td>P-value: 9.33E-5</td>
<td>P-value: 6.16E-6</td>
</tr>
<tr>
<td>IL-10 Signaling</td>
<td>Aryl Hydrocarbon Receptor Signaling</td>
</tr>
<tr>
<td>No. of Genes: 5</td>
<td>No. of Genes: 13</td>
</tr>
<tr>
<td>P-value: 1.70E-4</td>
<td>P-value: 1.15E-5</td>
</tr>
<tr>
<td>Hepatic Cholestasis</td>
<td>β-alanine Metabolism</td>
</tr>
<tr>
<td>No. of Genes: 7</td>
<td>No. of Genes: 10</td>
</tr>
<tr>
<td>P-value: 2.40E-4</td>
<td>P-value: 4.47E-5</td>
</tr>
<tr>
<td>Galactose Metabolism</td>
<td>Alanine and Aspartate Metabolism</td>
</tr>
<tr>
<td>No. of Genes: 4</td>
<td>No. of Genes: 9</td>
</tr>
<tr>
<td>P-value: 3.39E-4</td>
<td>P-value: 4.47E-5</td>
</tr>
<tr>
<td>Integrin Signaling</td>
<td>Xenobiotic Metabolism Signaling</td>
</tr>
<tr>
<td>No. of Genes: 8</td>
<td>No. of Genes: 17</td>
</tr>
<tr>
<td>P-value: 4.57E-4</td>
<td>P-value: 6.03E-5</td>
</tr>
<tr>
<td>Toll-like Receptor Signaling</td>
<td>Glutathione Metabolism</td>
</tr>
<tr>
<td>No. of Genes: 4</td>
<td>No. of Genes: 10</td>
</tr>
<tr>
<td>P-value: 4.68E-4</td>
<td>P-value: 6.61E-5</td>
</tr>
<tr>
<td>IL-6 Signaling</td>
<td>Butanoate Metabolism</td>
</tr>
<tr>
<td>No. of Genes: 5</td>
<td>No. of Genes: 10</td>
</tr>
<tr>
<td>P-value: 8.13E-4</td>
<td>P-value: 1.55E-4</td>
</tr>
<tr>
<td>Hepatic Fibrosis / Hepatic Stellate Cell Activation</td>
<td>Valine, Leucine and Isoleucine Degradation</td>
</tr>
<tr>
<td>No. of Genes: 6</td>
<td>No. of Genes: 9</td>
</tr>
<tr>
<td>P-value: 8.91E-4</td>
<td>P-value: 8.91E-4</td>
</tr>
</tbody>
</table>
Defining the clinical problem

- Limitations of clinical criteria to identify patients with ALI/ARDS
- The superimposed effect of VILI and other confounders on clinical phenotype
- Paucity of diagnostic tests
- Intra-observer variability
- Discrepancies between clinical diagnosis and autopsy findings
Why is List 1 better than list 2?

Could it be that Inflammation is ‘non-specific’ but VILI is not?