Integrative Assessment of Chlorine-induced Acute Lung Injury in Mice

ONLINE DATA SUPPLEMENT

George D Leikauf, Hannah Pope-Varsalona, Vincent J Concel, Pengyuan Liu, Kiflai Bein, Annerose Berndt, Timothy M. Martin, Koustav Ganguly, An Soo Jang, Kelly A Brant, Richard A Dopico, Jr., Swapna Upadhyay, YP Peter Di, Qian Li, Zhen Hu, Louis J Vuga, Mario Medvedovic, Naftali Kaminski, Ming You, Danny C. Alexander, Jonathan E. McDunn, Daniel R. Prows, Daren L. Knoell, James P Fabisiak

Detailed Methods

Selection of mouse sex and strains: This study was performed in accordance with the Institutional Animal Care and Use Committee of the University of Pittsburgh (Pittsburgh, PA) and mice were housed under specific pathogen free conditions. Females (6-8 wk) mice were obtained from Jackson Laboratory (Bar Harbor, ME) and housed in barrier-protected microisolator cages and all procedures were approved by the University of Pittsburgh Institutional Animal Use and Care Committee. The main reason females were used is that male mice will fight when placed together in the inhalation chambers. Housing male mice alone (in isolation) produces a stressful environment (108) which could be a confounder in a genetic study of survival. Previous studies of acute lung injury following trauma/hemorrhagic shock in rats have demonstrated: a) females are less sensitive than males (109), b) females in estrous and proestrous are less sensitive than females in diestrous (110), and c) increased estrogen decreases sensitivity (111). Virgin female mice only go into menstrual cycling when housed with male mice or exposed to male urine (112). In this study, female mice were virgins, housed only with female mice, probably in a non-cycling state (much like the diestrous phase), and would not be afforded estrogen-protection. Mice in various stages of the menstrual cycle could have periods when estrogen protection is gained or lost, which would add to the variability of the survival time measured in this experiment. However, the within strain variability was low in this experiment with standard errors ≈10-15% of mean survival time for each mouse strain (See Figure 1).

Mouse strains were selected from the Tier 1-3 priority strain of the Mouse Phenome Project (MPP) and based on previous studies of lung biology. The priority strains for the Mouse Phenome Database (MPD) (113) were classified based on strain availability, research trends, community input, and resources, including the recent release of the NIEHS-Perlegen SNPs for 8.27 million genome-wide locations for a set of inbred strains (114). The Tier 1 contains the 16 NIEHS-Perlegen strains that were used in this study and included: 129X1/SvJ, A/J, AKR/J, BALB/cByJ, BTBR T+tf/J, C3H/HeJ, C57BL/6J, CAST/EiJ, DBA/2J, FVB/NJ, KK/HIJ, MOLF/EiJ, NOD/ShiLtJ, NZW/LacJ PWD/PhJ, and WSB/EiJ. The Tier 2 contains 10 strains of which the following 6 were used in this study: BUB/BnJ, C57BLKS/J, CBA/J, MRL/MpJ, SJL/J, and SM/J. These strains breed well, are widely used, and/or

have unique phenotypes. Tier 3 consists of 10 strains of which the following 8 were used in this study: C57BR/cdJ, C58/J, CZECHII/EiJ, LP/J, NON/ShiLtJ, PL/J, RIIIS/J, and SWR/J. We have selected 10 additional strains that have been used in other investigations of ovalbumin hyperreactivity (BALB/cJ) and lung growth/cancer (129S1/SvImJ, C57BL/10J, DBA/1J, JF1/Ms, I/LnJ, LG/J, PERA/EiJ, SEA/GnJ, and SPRET/EiJ). These 40 strains are well-distributed in the Mouse Family Tree (114) and have excluded strains that are difficult to breed (and become unavailable at times), hard to handle, or are redundant to other priority strains.

Chlorine exposure: High concentration chlorine can produce rapid and often lethal lung injury whereas low concentrations may cause a delayed pulmonary edema. Prolonged chlorine exposures may occur because chlorine moderate water solubility may not cause upper airway symptoms for several minutes. In addition, the density of chlorine is greater than that of air, causing it to remain near ground level and increasing exposure time. Chemical terrorist attacks can include multiple releases over longer periods. Industrial accident often can lead releases of inhalation hazards that vary in duration from 3-12 hours like that in Graniteville, SC (10, 115) or Bhopal (116) to several days like that in Cincinnati, Ohio (117). Long exposures are typically at low concentrations.

In a large population, such exposures could lead to a large number of fatalities and injuries, which could overwhelm the capacity of existing medical systems. Triage strategies must be carefully designed and revised when advances in diagnosis and therapy become available. One design issue is the determination of susceptible individuals requiring added attention, which can be assessed at two levels. The first assessment is when overt signs of susceptibility are evident. This group includes individuals manifesting overt signs of cardiopulmonary distress, or individuals that are very young or elderly. In field situations, such individuals generally are readily detected because they have physical characteristics consistent with their disease or status (e.g. age). Because such individuals can be a limited portion of the population, retaining such individuals for observation may not overtax existing capacities.

The second assessment level is when overt signs of susceptibility are not evident. These individuals may be genetically susceptible to specific chemicals that can induce acute lung injury, but

otherwise appear healthy shortly after exposure. In field situations, such individuals can have a silent phenotype and cannot be detected quickly because they have no obvious characteristics consistent with susceptibility. This can be a pressing problem in chlorine-induced acute lung injury because fatal pulmonary edema can be delayed (e.g., developing >24 hours later) in individuals that initially present with no or mild overt signs of symptoms. Thus, the model used in this study focuses on low-level, prolonged exposure that produces a delayed fatal pulmonary edema.

Chlorine (45 ppm x 24 h) exposures were conducted in single pass, laminar, dynamic 0.32 m^3 stainless steel inhalation chambers in HEPA filtered air. Air samples were monitored continuously with a direct reading instrument (Polytron 3000, Dräger Safety Inc., Pittsburgh, PA) during exposure, and airflows adjusted to maintain exposure within 5% of the target concentration. Chlorine (Matheson Tri-Gas, Montgomery, PA) was introduced into the chamber using flow meters from cylinders placed in a vented safety cabinet. During exposure, room air also was continuously monitored and no exposures above 0.5 ppm (current occupational threshold limit value) were observed.

Assessment of acute lung injury: To examine chlorine-induced changes in lung histology and, bronchoalveolar lavage protein, C57BLKS/J and C57BL/10J mice were exposed to filtered air (0 h, control) or chlorine (45 ppm for 6, 12 or, in the case of C57BL/10J 24h). Mice were killed immediately after chlorine exposure by intraperitoneal injection of pentobarbital sodium (100 mg/kg; Nembutal, Abbott Laboratories, Chicago, IL) and severing of the posterior abdominal aorta. To obtain tissue for mRNA analysis (n = 8 mice/strain/time), the diaphragm was punctured, and the chest cavity opened. Lungs were excised, frozen in liquid nitrogen, and stored (-70°C). To obtain tissue for histology (n = 3 mice/strain/time), the chest wall was left intact, a cannula was inserted into the trachea and the lung was instilled with phosphate-buffered saline containing 3.7% formaldehyde (pressure: 28 cm H_2O, Cat. No. SF100-4, Thermo Fisher Scientific, Pittsburgh, PA). The trachea was ligated, lungs removed, and the inflated lung was immersed in fixative (24 h, 4°C). Fixed lungs were washed with Dulbecco's Phosphate-Buffered Saline containing Ca^2+, Mg^2+, 6.1 mM D-glucose, and 0.33 mM sodium pyruvate (DPBS; Cat. No. 14287–080, Life Technologies, Carlsbad, CA), dehydrated through a series of graded

ethanol solutions (30-70%), and processed in paraffin blocks (Hypercenter XP, Shandon, Ramsey, MN). The lung tissue was sectioned (5 μm) and stained with hematoxylin and eosin.

To obtain bronchoalveolar lavage fluid (n = 5 mice/strain/time), a cannula was inserted into the trachea and the lung was instilled with Ca^{2+}, Mg^{2+}-free Hank's Balanced Salt Solution (Cat. No. 14175-095, Life Technologies) and the recovered lavage fluid placed on ice until protein analysis was performed. Total protein in cell-free supernatants was measured using a bicinchoninic acid assay (BCA; Cat. No. 23325, Thermo Scientific) using bovine serum albumin (BSA) as a standard.

Haplotype mapping analysis: Employing the resulting strain distribution pattern, whole-genome linkage disequilibrium analysis for survival time in chlorine was performed using single nucleotide polymorphisms (SNPs) from the National Institute of Environmental Health Sciences (NIEHS), Wellcome Trust Centre for Human Genetics (WTCHG) and the Broad Institute (which include SNPs from Roche and Genomic Institute of the Novartis Research Foundation). Currently, the NIEHS SNP database contains 8,272,574 SNP's, WTCHG SNP database contains 13,348 SNPs and the Broad SNP database contained 138,793 SNPs in commonly used mouse strains. The genomic positions of the SNPs for the two data sets were unified based on the latest NCBI mouse genome map (current build 37.1). These SNPs were edited to remove SNPs with <16 strains typed or without map information. The resulting informative (>1 million) SNPs will span the mouse genome at an average density of $\geq$5 kb/SNP. The SNP genotyping accuracy is over 99.8%.

To assess the association between individual SNPs and a transcriptional pathway using an informative SNP subset, a pairwise LD measure, r^2, was performed which is the squared correlation coefficient of alleles at two loci. To assess genomic background LD in the sample, r^2 was calculated for independent SNPs from different chromosomes. The association of transcriptional pathway with SNPs was tested by using two-sample Student's t statistic. The phenotypic data was transformed to approach normality by using the Box-Cox procedure prior to statistical analysis. A two-sided P value for each SNP was obtained to test the hypothesis of no association between the SNP and transcriptional pathway and presented as the negative 10-base logarithmic P value, i.e. -log (P). In addition, a sliding window of 3 SNPs was used to infer haplotypes at each locus. Haplotype-based association analysis

of transcriptional pathway was performed using Student's t statistic. In the haplotype analysis, a common variant was first identified and treated as one category and the other rare variants as another category using the R statistical package. The region in linkage disequilibrium was evaluated for functional relevance based on the role of the protein of the polymorphic genes in the identified region. For each gene, polymorphisms were assessed for differences in nonsynonymous coding within exons, or in consensus sequences of transacting elements contained in untranslated regulatory regions (UTR). Candidate genes were evaluated for possible biological plausibility based on current biomedical literature, with additional functional analyses performed in future studies.

An appropriate genome-wide threshold for declaring significant associations is crucial for the use of a dense SNP map. The conventional significance level LOD of 3.3 (corresponding to a point wise P value of 1×10^{-4}) for a genome-wide linkage study is not sufficiently rigorous and results in 4-7 false positives in a genome-wide linkage disequilibrium (LD) scan. A more stringent threshold was used. First a simulation procedure was implemented to establish a rigorous threshold for the genome-wide association analysis using our empirical data. The number of false positives for genome-wide association analysis at various nominal significance levels was estimated under different levels of trait heritability. This yielded one statistical association by chance in a genome-wide LD scan when the threshold is decreased to 1.6×10^{-5} or -log (P*)* = 4.8. Therefore we used a significance threshold of –log (P) > 4.8 and a suggestive threshold of 4.8 ≥ –log (P) > 4.0.

Transcriptomic (microarray) analysis: To perform a transcriptome-wide analysis of steady state mRNA levels, total lung mRNA was analyzed by microarray and selected changes confirmed by qRT-PCR (59, 60). Lung mRNA was isolated (Trizol method), reverse transcribed, and fluorescently labeled. Each sample was hybridized to a microarray containing 31,769 murine 70-mer oligonucleotides (Aligent). RNA quantity was initially assessed by spectrophotometer (Nanodrop ND-1000 Thermo Scientific, Wilmington, DE) and quality was assessed by electrophoresis (Agilent 2100 Bioanalyzer, Agilent Technologies, Santa Clara, CA). RNA (0.5 µg) was cyanine 3-labeled and cDNA synthesized (RNA Spike In – One Color, 5188-5282, Agilent Technologies). Labeled cRNA was transcribed from cDNA (Quick-Amp Labeling Kit - One Color, 5190-0442, Agilent Technologies). The

labeled cRNA was quantified (Nanodrop ND-1000 and Agilent 2100 Bioanalyzer), and hybridized (65°C, 17 h) (Gene Hybridization, 5188-5242, Agilent Technologies) onto the array (Whole Mouse Genome Kit 4x44K, G4122F, Agilent Technologies). Arrays were washed and scanned (DNA Microarray Scanner, G2505C, Feature Extraction Version 10.7.3.1, Agilent Technologies). Five microarrays were obtained for each strain (C57BLKS/J or C57BL/10J) and each time (0, 6, or 12 h) to yield a total of 30 microarrays.

Gene expression profiling was conducted using a hybridization microarray. The lung mRNA from exposed 0 (filtered air control), 6 and 12 h were analyzed using 1 microarray/mouse for the sensitive and resistant strain. Samples obtained from each mouse of these mice and an additional 3 mice/time/strain also used for subsequent qRT-PCR. Data normalization was performed for each microarray separately by subtracting channel specific local background intensities and centering the log-transformed intensities with a fitted local regression model (118). The statistical analysis was performed by fitting a mixed effects linear model for each gene separately with array effects assumed to be random while treatment and dye effects assumed to be fixed (119). This model was fitted for each gene and statistical significance of the differential expression between strains after adjusting for the array effect. Results were assessed by calculating p-values for the corresponding linear contrasts. Multiple hypotheses testing adjustment was performed by calculating false discovery rate to assess statistical significance (120).

qRT-PCR analysis: A total of 28 transcripts were assessed by qRT-PCR. Lung RNA (100 μg) from the divergent mouse strains (n = 5- 8 mice/strain) was reverse transcribed into first-strand cDNA with a High-Capacity cDNA Archive Kit (Applied Biosystems, Foster City, CA) in a 100-μl reaction volume. cDNA (10 μl) was used in a subsequent PCR reaction using 25 μl of TaqMan Universal PCR Master Mix (Applied Biosystems), 2.5 μl of each primer mixture, and 12.5 μl of RNAse-free water. Validated primer mixtures for each gene was obtained from Applied Biosystems and normalized to RPL32 or RPS18 (Applied Biosystems). Analysis was performed with an Applied Biosystems 7900HT System and the following conditions: 95°C for 10 min followed by 40 cycles of 95°C for 15 s and 60°C for 1 min. For relative quantification of expression of each gene between the strains, the comparative cycle number

threshold (C_T) method ($\Delta\Delta C_T$) was used: ΔC_T = C_T (transcript of the gene of interest) – C_T (e.g., RPL32), and this value was calculated for each sample. The comparative $\Delta\Delta C_T$ calculation involved finding the difference between each sample's ΔC_T and the mean ΔC_T for the most resistant strain. These values were then transformed to absolute values with a formula in which comparative expression level = $2^{-\Delta\Delta CT}$.

In addition, transcripts and metabolites in biological pathways considered as hallmarks in lung injury were analyzed. These were determined once all the transcriptomic (microarray and qRT-PCR) and metabolomic analyses had been obtained to select from transcripts associated with cell death, NEF2L2 mediated oxidative stress response, lipid biogensis, proliferation, extracellular matrix remodeling, epithelial differentiation, epithelial-mesenchymal interactions, airway branching, endothelial proliferation, vascular mechanosensory function, vascular tension, vascular development, arterial-venous malformation, vascular guidance, and anti- and pro-angiogensis.

Metabolome sample preparation: Metabolome profiling samples were bar-coded using a laboratory information management system (LIMS) and kept frozen until assayed by LIMS-scheduled tasks. Data files were tracked and processed by their LIMS identifiers and archived to DVD at regular intervals. The sample preparation was automated with a liquid handler (MicroLab STAR®, Hamilton Robotics, Inc., Reno, NV). Sample extraction consisted of sequential organic and aqueous extractions. A recovery standard was introduced at the start of the extraction process. The resulting pooled extract was equally divided into a liquid chromatography (LC) fraction and a gas chromatography (GC) fraction. Samples were lyophilized (TurboVap®, Zymark, Claiper Life Science, Mountain View, CA), injection standards introduced, and suspended in injection solvent. In addition to controls and blanks, a well-characterized sample generated for quality control verification was randomly included multiple times allowing sample preparation and analytical variability to be assessed constantly.

Metabolomic profiling: This process involved: sample extraction, separation, detection, spectral analysis, data normalization, delineation of class-specific metabolites, pathway mapping, validation and functional characterization of candidate metabolites. All samples were randomized prior to mass spectrometric analyses to avoid any experimental drifts. A number of internal standards, including

injection standards, process standards, and alignment standards were used to assure QA/QC targets were met and to control for experimental variability. The reproducibility of the profiling process was addressed at two levels; one by measuring only instrument variation, and the other by measuring overall process variation. Instrument variation was measured from a series of internal standards (n=14 in this study) added to each sample just prior to injection. The median coefficient of variation (CV) value for the internal standard compounds was 3.9%.

To address overall process variability, metabolomic studies were augmented to include a set of nine experimental sample technical replicates (also called matrix, abbreviated as MTRX), which were spaced evenly among the injections for each day. Reproducibility analysis for the n=339 compounds detected in each of these nine replicate samples gave a measure of the combined variation for all process components including extraction, recovery, derivatization, injection, and instrument steps. The median CV value for the experimental sample technical replicates (tissue profiling part of this study) was 14.6%. The reproducibility of these experimental-sample technical replicates was assessed and the Spearman's rank correlation coefficient between pairs of technical replicates ranged from 0.93 to 0.97.

Liquid chromatography/mass spectroscopy (LC/MS): The LC/MS analysis is based on a Surveyor HPLC and a LTQ-FT mass spectrometer (Linear Ion Trap mass spectrometer with Fourier Transform, Thermo Fisher Corporation, Pittsburgh, PA). LTQ data was used for compound quantification and FT data was used to confirm the identity of specific compounds. The instrument continuously monitored positive and negative ions. The vacuum-dried sample was re-suspended in 100 µl injection solvent that contains $\geq$ 5 injection standards at fixed concentrations. Internal standards were used both to assure injection and chromatographic consistency. Chromatography involved an: water:acetonitrile (ACN) (0.1% trifluoroacetic acid) 5 - 100% gradient (8 min) followed by 100% ACN (8 min). Columns (Aquasil C-18, Thermo Fisher) were temperature controlled and were reconditioned after 50 injections.

Gas chromatography/mass spectrometry (GC/MS): For the metabolome profiling studies, the samples were re-dried under vacuum desiccation (24 h) and derivatized under dried nitrogen using *bis*

trimethylsilyl-triflouroacetamide (BSTFA). Samples were analyzed by GC/MS (Mat-95 XP, Thermo Fisher) using electron impact ionization and high resolution with a (5% phenyl)-methyl polysiloxane column with a 40 - 300°C gradient (16 min). Spectra were compared against libraries of authentic compounds.

Isotope dilution GC/MS quantification of target metabolites: Residual water was removed from samples by forming an azeotrope [100 µl dimethylformamide (DMF)] followed by vacuum drying. Samples were injected using an ion column injector into a gas chromatograph (6890N, Agilent Technologies, Inc., Santa Clara, CA) equipped with a 15-m DB-5 capillary column (0.2 mm ID; film thickness, 0.33 µm; J & W Scientific, Folsom, CA) interfaced with a mass detector (5975 MSD, Agilent). The *t*-butyl dimethylsilyl derivatives were quantified by selected ion monitoring (SIM), using isotope dilution electron-impact ionization GC/MS. The m/z for native and labeled molecular peaks included: 158 and 161 (sarcosine), 406 and 407 (cysteine), 432 and 437 (glutamic acid), 297 and 301 (thymine), and 218 and 219 (glycine), respectively. Assessment of citric acid was performed on the GC-MS in the full scan mode. The area under the peak with m/z 591 was measured and normalized to ^{13}C glycine peak (m/z: 219, internal standard) in each run followed by normalization to the tissue weight. The resulting value was scaled down by dividing by a constant value of 1 million and used to plot the relative box plots. Relative area counts for each compound were obtained by manual integration of its chromatogram peaks using Xcalibur software (Thermo Fisher).

Metabolomic library: Mass spectral data were compared to a library created using approximately 800 commercially available compounds that were acquired and registered into Metabolon LIMS (Research Triangle Park, NC). All compounds were analyzed at multiple concentrations under the same conditions as the experimental samples, and the characteristics of each compound were registered into a LIMS-based library. The same library was used for both the LC and GC platforms for determination of their detectable characteristics. These were then analyzed using custom software packages. Initial data visualization used SAS and Spotfire. To capture threshold levels, sample minimum was used to impute values because the mass spectrometer threshold for detection may differ between samples.

Additional bioinformatic analyses: To examine temporal transcriptomic and metabolomic profiling, transcripts or metabolites were clustered according to expression with time and strain, and evaluated by suite of functional hierarchies programs including LRpath (121). LRpath is a logistic regression-based method for identifying predefined sets of biologically related genes enriched with (or depleted of) differentially expressed transcripts in microarray experiments. We functionally related the odds of gene set membership with the significance of differential expression, and calculated adjusted P-values as a measure of statistical significance. LRpath displays robust behavior and improves statistical power compared with tested alternatives. It is applicable in experiments involving two or more sample types, and accepts significance statistics of the investigator's choice as input.

We will also analyze the microarray datasets using CLEAN (122). Our objective was to identify the specific features that are critical in acute lung injury. To accomplish this we employed a combination of classical supervised machine learning approaches and Bayesian modeling approaches based on the context-specific infinite mixtures framework to identify specific temporal transcriptional signatures. We performed both the gene-expression-centric and pathway-centric modeling of patterns of gene expression. In the gene-expression-centric analysis, we identified genes conferring differential transcriptional response by modeling gene expression data. Next, groups of genes identified were placed into functional clusters constructed based on the traditional sources of such information and by custom clusters characterizing transcriptional targets of key regulatory genes. In the pathway-centric approach, the evidence of transcriptional effects on specific pathways was directly assessed using the Gene Set Enrichment Analysis (GSEA). In either approach the space of potentially discriminatory features ultimately includes gene expression patterns and the gene labels derived from relevant functional categories/pathways. In either the expression-centric or pathway-centric analysis, genes whose expression levels are affected by the compounds tested were identified in the statistical analysis.

A key feature of this model-based clustering approach is its robustness with respect to heterogeneity of clusters of co-expressed genes within different context. In our case a context consisted of a time-course gene expression profiles, measured with a strain of mice. In this interpretation, each gene's expression profile consists of 250 estimates of differential gene expression

from the control, organized in 50 contexts. Traditional clustering approaches would be incapable of performing such global analysis because it is almost certain that clusters of co-expressed genes will not be uniform across all 50 contexts. Unless such context specificity is explicitly accounted for, potentially important patterns of co-expression induced only within a subset of experimental conditions would likely be drowned in the random fluctuations introduced by the noninformative contexts. For example, genes forming a cluster of co-expressed genes within contexts defined by only of a subset of dosages of a single agent within the resistant strain could be an important determinant of the resistance to the effects of this agent. Yet, such local patterns would almost certainly be lost in the traditional cluster analysis whenever they do not persist over a sufficiently large number of experimental conditions. Another key feature of our analysis is its ability to produce meaningful assessment of statistical significance of identified clusters. Since the underlying statistical model accounts for all sources of uncertainty in the data and Bayesian, including the averaging over models with all possible number of global and local clusters, the posterior probabilities of co-expression produced by the computational procedure are valid measures of statistical significance of induced patterns.

In addition to identifying patterns of expression using pre-specified contexts, we used the Bayesian infinite module networks to hierarchically cluster individual contexts into larger meta-contexts based on homogeneity of clusters of co-expressed genes within individual contexts. Results of this analysis allowed for a direct identification of gene clusters with discriminative patterns between different compounds. By forming larger meta-contexts, the precision of the cluster analysis improved and the level of similarities between expression patterns within individual contexts established. This overall approach to identifying discriminative features has significant advantages over the traditional methods. For example, it is not clear whether a differential expression of the same gene at different time point in the overall progression should be treated as a common or a distinctive feature of two transcriptional profiles. Such differences could be results of a minor random fluctuation in gene expression measurements and cause the differential expressions at a time not to be statistically significant. Similarly, a gene could be associated with one agent but not another because its differential expression does not reach the statistical significance cut-off due to random fluctuations in the data. In this respect,

statistically significant clusters of genes co-expressed within whole context or multiple contexts was much more robust features for discriminating effects of different agents.

We used more traditional statistical and machine learning methods to identify discriminatory features in gene expression patterns. The expression profiles were summarized by using traditional singular decomposition methods such as Principal Components Analysis (PCA) and Partial Least Squares (PLS). The groups of genes contributing significantly to major Eigen values/eigenvectors obtained in PCA were assessed with respect to their correlation with different agents and strains. PCA was performed both on global variance-covariance matrix and variance covariance matrices derived for different subsets of expression measurements. In the case of PLS, the significant Eigen values/eigenvectors are identified in a supervised fashion so that they are guaranteed to correlate with different groupings of interest. Unfortunately, as any other supervised machine learning approaches, PLS are susceptible to over fitting. We balanced the unsupervised nature of PCA and supervised nature of PLS to ensure the reproducibility of conclusions drawn from such analysis.

We are focusing on the differences that could be used in understanding physiological and ultimately pathological events that occur during acute lung injury. Through our bioinformatic approaches, we deciphered the transcriptional signatures specific to strain differences. By selecting strains with varied sensitivity, we contrasted the resulting transcriptional events to better understand class specific responses.

Expression candidate transcript in mouse embryonic lung or protein in human lung: To initially assess the identified candidate genes, GenePaint.org database (123) of in situ transcript hybridization in mouse embryos and Human Protein Atlas database (124) of protein expression were interrogated for evidence of lung expression. Transcripts for BCKDHB, CALN1, IGFBP5, KLF4, NEO1, ROR1, SLCO4C1, TNFRSF19, and TNS1 were evident in mouse lung at embryonic day 14 (**Fig. 2, Supplemental Fig.S4**). Immunohistological staining for BCKDHB, KLF4, NCOR2, PACS1, PML, ROR1, SACS, SLC38A4, SLCO4C1, TNS1, and TTK was present in human airway epithelium (**Fig. 2, Supplemental Fig.S5**). Immunohistological staining for TNFRSF19 was evident in human alveolar macrophage (**Supplemental Fig.S5**).

Supplemental References

E1 Innos J, Philips MA, Raud S, Lilleväli K, Kõks S, Vasar E. Deletion of the Lsamp gene lowers sensitivity to stressful environmental manipulations in mice. Behav Brain Res. 228:74-81, 2012.

E2 Ananthakrishnan P, Cohen DB, Xu DZ, Lu Q, Feketeova E, Deitch EA. Sex hormones modulate distant organ injury in both a trauma/hemorrhagic shock model and a burn model. Surgery. 137:56-65, 2005.

E3 Caruso JM, Xu DZ, Lu Q, Dayal SD, Deitch EA. The female gender protects against pulmonary injury after trauma hemorrhagic shock. Surg Infect (Larchmt). 2:231-40, 2001.

E4 Doucet D, Badami C, Palange D, Bonitz RP, Lu Q, Xu DZ, Kannan KB, Colorado I, Feinman R, Deitch EA. Estrogen receptor hormone agonists limit trauma hemorrhage shock-induced gut and lung injury in rats. PLoS One. 5:e9421, 2010.

E5 Stelck RL, Baker GL, Sutherland KM, Van Winkle LS. Estrous cycle alters naphthalene metabolism in female mouse airways. Drug Metab Dispos. 33:1597-602, 2005.

E6 Grubb SC, Maddatu TP, Bult CJ, Bogue MA. Mouse phenome database. Nucleic Acids Res. 237(Database issue):D720-30, 2009.

E7 Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA, Beilharz EJ, Gupta RV, Montgomery J, Morenzoni MM, Nilsen GB, Pethiyagoda CL, Stuve LL, Johnson FM, Daly MJ, Wade CM, Cox DR. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature. 448:1050-3, 2007.

E8 Duncan MA, Drociuk D, Belflower-Thomas A, Van Sickle D, Gibson JJ, Youngblood C, Daley WR. Follow-up assessment of health consequences after a chlorine release from a train derailment--Graniteville, SC, 2005. J Med Toxicol. 7:85-91, 2011.

E9 Mishra PK, Samarth RM, Pathak N, Jain SK, Banerjee S, Maudar KK. Bhopal Gas Tragedy: review of clinical and experimental findings after 25 years. Int J Occup Med Environ Health. 22:193-202, 2009.

E10 Keyes SE. Re-creation and worse case scenario of accidental release of styrene gas from a rail car. (Viewed at http://etd.ohiolink.edu/view.cgi/Keyes%20Sarah%20Elizabeth.pdf?ucin1259077613 February 22, 2012). Master Thesis, University of Cincinnati, 2009. pp.2-3.

E11 Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica 12:111-39, 2002.

E12 Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS. Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol. 8:625-37, 2001.

E13 Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Royal Stat Soc, Series B. 57, 289–300, 1995.

E14 Sartor MA, Leikauf GD, Medvedovic M. LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. Bioinformatics. 25:211-7, 2009.

E15 Freudenberg JM, Joshi VK, Hu Z, Medvedovic M. CLEAN: CLustering Enrichment ANalysis. BMC Bioinformatics. 10:234, 2009.

E16 Visel A, Thaller C, Eichele G. GenePaint.org: an atlas of gene expression patterns in the mouse embryo. Nucleic Acids Res. 32:D552-6, 2004.

E17 Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, Wernerus H, Björling L, Ponten F. Towards a knowledge-based human protein atlas. Nat Biotechnol. 28:1248-50, 2010.

Supplemental Tables

Supplemental Table S1. Single nucleotide polymorphisms (SNP) associated with survival in chlorine (45 ppm) in mice as determined by haplotype mapping of 40 inbred strains.

Supplemental Table S2. Nonsynonymous single nucleotide polymorphisms in candidate genes associated with chlorine-induced acute lung injury.

Supplemental Table S3. Lung transcript levels of candidate genes with single nucleotide polymorphisms associated with survival during chlorine-induced acute lung injury in sensitive C57BLKS/J and resistant C57BL/10J mouse strains.

Supplemental Table S4. Transcript for transcription factors of putative 5'UTR (promoter) DNA binding sites with single nucleotide polymorphisms associated with candidate genes for chlorine-induced acute lung injury in mice.

Supplemental Table S5. Transcriptome in the lung of sensitive (C57BLKS/J) and resistant (C57BL/10J) mouse strains without or during chlorine-induced acute lung injury.

Supplemental Table S6. Small molecule metabolome in the lung of sensitive (C57BLKS/J) and resistant (C57BL/10J) mouse strains without or during chlorine-induced acute lung injury.

Supplemental Table S7 Lung metabolomic profiling pathway analysis comparing sensitive C57BLKS/J and resistant C57BL/10J mouse strains following chlorine exposure.

Supplemental Figure Legends

Supplemental Figure S1. Transcript levels of candidate genes that differed between the C57BL/10J and C57BLKS/J and mouse strains following chlorine exposure. Mice were exposed to filtered air (Control, 0h), or to chlorine (45 ppm) for 6 or 12 h, lung mRNA isolated, and transcript expression levels determined by quantitative real time polymerase chain reaction. ***Abbreviations***: **SLC35A5**: solute carrier family 35, member A5; **SLCO4C1**: solute carrier organic anion transporter family, member 4C1; **ST8SIA4**: ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase; **UBC**: ubiquitin C. *Significantly different from strain-matched control as determined by analysis of variance (ANOVA) with an all pairwise multiple comparison procedure (Holm-Sidak method). †Significantly different between the sensitive C57BLKS/J and resistant C57BL/10J mouse strain at indicated times as determined by analysis of variance (ANOVA) with an all pairwise multiple comparison procedure (Holm-Sidak method). Values are mean ± SE (n = 8 mice/strain/time).

Supplemental Figure S2. Pathways enrichment in the sensitive C57BLKS/J mouse lung following chlorine exposure. (**A**) Pathways with increased transcripts included Src homology-3 domain, transcription factor activity, and cell death pathway. (**B**) Pathways with decreased transcripts in protein transport, translation, and vasculature development pathway. Displayed are individual transcripts that are members of a pathway significantly enriched in C57BLKS/J mouse lung as determined by LRpath analysis of microarray data. Value are means ± SE (n = 5 arrays/strain/time).

Supplement Figure S3. Pathways enrichment in the resistant C57BL/10J mouse lung following chlorine exposure. (**A**) Pathways enriched transcripts that increased following exposure more in the resistant C57BL/10J mouse lung included cell adhesion, cytoskeleton organization, and protein catabolic process pathway. (**B**) Pathways enriched transcripts that decreased following exposure more in the resistant C57BL/10J mouse lung included RNA binding, transcription, and mitochondrion pathway. Noteworthy contrasts between strains included that the transcription factor activity pathway, which contained transcripts that increased more in sensitive C57BLKS/J lung whereas the transcription pathway had transcripts that decreased more in resistant C57BL/10J lung. Similarly, the protein transport pathway had transcripts that decreased more in sensitive C57BLKS/J whereas the protein catabolic process pathway had transcripts that increased more in resistant C57BL/10J lung. Displayed are individual transcripts that are members of a pathway significantly enriched in C57BLKS/J mouse lung as determined by LRpath analysis. Value are means ± SE (n = 5 arrays/strain/time).

Supplemental Figure S4. **Localization of transcripts for candidate genes in mouse embryo lung tissue by in situ hybridization. (*Left Panels*):** Mouse embryo (embryonic day 14.5) in situ hybridization using antisense riboprobes obtained from GenePaint.org database (Visel 2004). Tissue sections (thickness: 20µm and inter-section distance: 100µm) were stained with cresyl violet (Nissl-method) and digitally scanned (original magnification: 5x). Note the mouse strains are indicated and differ from the C57BLKS/J or C57BL/10J so that appearance is considered an inclusion but not an exclusion criterion. **(*Right Panels*)**: In situ hybridization detects transcript in mouse embryo lung (original magnification: 20x). Increased transcript riboprobe intensity in the lung was detected for BCKDHB, IGFBP5, KLF4, NEO1, ROR1, SLC35A5, SLCO4C1, TNFRSF19, and TNS1. Moderate to low intensity was noted for SLC35A5, SLC38A2, ST88SIA4,and TNS1 ***Abbreviations***: **AACS**: acetoacetyl-CoA synthetase; **ADAMTS19**: a disintegrin-like and metallopeptidase (reprolysin type) with thrombospondin type 1 motif, 19 **BCKDHB**: branched chain ketoacid dehydrogenase E1, beta polypeptide; **CYP11A1**: cytochrome P450, family 11, subfamily a, polypeptide 1; **DOCK10**: dedicator of cytokinesis 10; **FRY**: furry homolog (Drosophila); **IGFBP5**: insulin-like growth factor binding protein 5, **KLF4**: Kruppel-like factor 4 (gut); **NEO1**: neogenin; **PACS1**: phosphofurin acidic cluster sorting protein 1; **IKBKAP**: inhibitor of kappa light polypeptide enhancer in B-cells, kinase complex-associated protein; **NCOR2**: nuclear receptor co-repressor 2; **NEO1**: neogenin; **PACS1**: phosphofurin acidic cluster sorting protein 1; **PML**: promyelocytic leukemia; **PPIP5K2**: diphosphoinositol pentakisphosphate kinase 2; **ROR1**: receptor tyrosine kinase-like orphan receptor 1; **SACS**: sacsin; **SEMA7A**: sema domain,

immunoglobulin domain (Ig), and GPI membrane anchor, (semaphorin) 7A; **SLC35A5**: solute carrier family 35, member A5; **SLC38A2**: solute carrier family 38, member 2; **SLC38A4**: solute carrier family 38, member 4; **SLCO4C1**: solute carrier organic anion transporter family, member 4C1; **ST8SIA4**: ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase; **TNFRSF19**: tumor necrosis factor receptor superfamily, member 19; **TNS1**: tensin 1; **TTK**: Ttk protein kinase.

Supplemental Figure S5. **Localization of transcripts for candidate genes in human lung tissue by immunohistochemical analysis. (*Left Panels*)**: Human airway epithelium staining for indicated protein by immunochemistry for indicated proteins obtained from the Human Protein Atlas (Uhlen 2010). In the airway epithelium, strong staining was detected for BCKDHB, IKBKAP, KLF4, NCOR2, SEMA7A, SLC38A4, TNS1, and UBC and moderate to weak staining was detected for AACS, CYP11A1, NEO1, PML, SACS, SLC38A2, SLCO4C1, and TTK. **(*Right Panels*)**: Human alveolar region staining for indicated proteins. In the alveolar epithelium, strong staining was detected for AACS, KLF4, NCOR2, NEO1, PACS1, SEMA7A, SLC38A4, SLCO4C1, TNS1, and TTK and moderate to weak staining was detected in BCKHDB, PML, PPIP5K2, SACS, SLC38A2, and UBC. In alveolar macrophage, moderate staining was detected for BCKDHB, DOCK10, NCOR2, NEO1, PACS1, PML, PPIP5K2, SACS, SLC38A2, SLC38A4, SLCO4C1, TNFRSF19, TNS1, TTK, and UBC. ***Abbreviations***: **AACS**: acetoacetyl-CoA synthetase; **BCKDHB**: branched chain ketoacid dehydrogenase E1, beta polypeptide; **CYP11A1**: cytrochome P450 , family 11, subfamily a, polypeptide 1; **DOCK10**: dedicator of cytokinesis 10: **IGFBP5**: insulin-like growth factor binding protein 5, **IKBKAP**: inhibitor of kappa light polypeptide enhancer in B-cells, kinase complex-associated protein; **KLF4**: Kruppel-like factor 4 (gut); **NCOR2** : nuclear receptor co-repressor 2; **NEO1**: neogenin; **PACS1**: phosphofurin acidic cluster sorting protein 1; **PML**: promyelocytic leukemia; **PPIP5K2**: diphosphoinositol pentakisphosphate kinase 2; **SACS**: sacsin; **SEMA7A**: sema domain, immunoglobulin domain (Ig), and GPI membrane anchor, (semaphorin) 7A; **SLC35A5**: solute carrier family 35, member A5; **SLC38A2**: solute carrier family 38, member 2; **SLC38A4**: solute carrier family 38, member 4; **SLCO4C1**: solute carrier organic anion transporter family, member 4C1; **TNFRSF19**: tumor necrosis factor receptor superfamily, member 19; **TNS1**: tensin 1, **TTK**: Ttk protein kinase; **UBC**: ubiquitin C.