Title: Systems Analysis of the Human Pulmonary Arterial Hypertension Lung Transcriptome

Authors: Robert S. Stearman, Quan M. Bui, Gil Speyer, Adam Handen, Amber R. Cornelius, Brian B. Graham, Seungchan Kim, Elizabeth A. Mickler, Rubin M. Tuder, Stephen Y. Chan, and Mark W. Geraci

ONLINE DATA SUPPLEMENT

Online Data Supplement

RNA Preparation and Microarray Data Collection:

Total RNA was prepared using the standard Trizol (Invitrogen) extraction with a Polytron homogenizer followed by chloroform phase separation, *i*-propanol precipitation, and column purification (RNEasy Mini-prep, Qiagen). Total RNA (>200 nucleotides) was quality controlled by UV ratios and BioAnalyzer RIN values (Agilent, RINs>5.0). RNA labeling and microarray data collection was done by the University of Colorado Denver genomics core using Affymetrix products and workflow, specifically the human HuGene1.0-ST microarrays. Microarray CEL files and associated data are available at NCBI GEO as GSE117261.

Partek Genomics Suite Analysis:

Affymetrix human HuGene1.0-ST microarrays were run and processed using their standard workflow at the University of Colorado Genomics Core. The resulting CEL files were imported into a Partek Genomics Suite 6.6 project along with the associated sample annotation and clinical/pathology data. The microarray data was collected in 3 different batches over a 4 year period as Pulmonary Hypertension Breakthrough Initiative study participants were enrolled. Each batch was composed of a mixture of Failed Donors (FD, controls) and Pulmonary Arterial Hypertension (PAH, all subtypes) patients. Principal Component Analysis (PCA) plots demonstrated a strong batch effect driven by the runs of the 3 groups of samples (Supplemental Figure E1) and the sex distribution (Male:Female) in the FD and PAH groups were highly skewed. After importing the CEL files into Partek using GC-RMA quantile normalization (1-3), batch and sex variable contributions were removed using an ANOVA model

(after “batch-remove” function). To evaluate of the “batch-remove” effects, an ANOVA model was run after correcting for just batch, generating 3,219 TranscriptIDs significantly associated with sex (p-value <0.05). After adding correcting for both batch and sex, there were zero TranscriptIDs associated with sex (p-value<0.05). A less restrictive p-value (rather than corrected q-value) was used to capture the highest number of potential sex-associated transcriptional changes. Finally, when our corrected dataset was queried using a human lung sex-based classifier (4), expression of sex-based genes did not significantly differ by sex, suggesting appropriate removal of lung sex-biased DE genes.

False Discovery Rate (FDR) calculated q-values were determined based on PAH versus FD gene expression differences (5, 6) and a q-value <0.001 was used. Large, high-quality microarray datasets can identify thousands of DE genes at a q-value <0.05, especially when comparing a tissue diseased state to a normal control. For the PHBI dataset, there were 5,308 TranscriptIDs that were DE at a q-value <0.05. In order to produce a more robust genelist, we chose to reduce the FDR to q-value <0.001 (Supplemental Figure E2). This cut-off should help eliminate false-positive DE genes at the expense of missing potentially interesting genes (false negatives). An alternative approach is to use a relaxed FDR q-value cut-off but impose a secondary criteria like fold-change threshold and/or overlap with additional datasets which we describe below (see (7) for a recent example using lung squamous cell carcinoma).

Independent Validation Methods:

We used 2 different approaches to validate the PAH classifier (Supplemental Figures E3 and E4). First, we used our PAH classifier to provide supervised analysis of an independent gene expression PAH versus control lung dataset ((8), Supplemental Figure E3). Second, we used a literature-derived PAH gene network (9) to provide supervised analysis of our PAH versus FD gene expression dataset (Supplemental Figure E4).

Supplemental Figure E3: The PAH classifier was used to supervise an independent PAH lung transcriptome microarray experiment (GSE15197 (8)) which used independently collected and process human lung tissue analyzed on a different microarray platform (Agilent). For this comparison, the Rajkumar *et al*. dataset included control (n = 13) and PAH only (n = 18) lung samples that were imported into Partek from GEO. The gene symbols from the PHBI classifier (n = 1,060 out of n = 1,140 TranscriptIDs) were used to overlap the full gene symbol list in GSE15197 (n = 893 matched). This subset genelist was then used for supervised clustering of GSE15197 (Supplemental Figure E3). The controls were grouped into 2 main branches separated from the main branch of PAH samples, while 2 PAH samples were misclassified as controls.

Supplemental Figure E4: A literature derived PAH gene network (341 human PAH-related genes made up of 293 protein-coding and 48 non-coding genes (9)) was used to supervise our PHBI microarray dataset. For this analysis, gene symbols were used to align the literature-derived network with our PHBI dataset (n = 112 gene symbols both overlapped and had q-value<0.05 in our dataset). As shown in supplemental Figure E4, the PHBI dataset was divided into 2 branches (PAH or FD) with only 1 sample misclassified.

Quantitative Real-Time Polymerase Chain Reaction:

Validation of differentially expressed (DE) mRNA levels predicted from the microarrays was done using gene specific primer/probes from Applied Biosystems (ABI). Briefly, 4μg of total lung RNA was reverse transcribed (High Capacity cDNA Archive Kit, ABI) and diluted to 10ng/μl based on the total RNA input. Two microliters of each cDNA was measured in duplicate (FastStart Universal Master Mix, Roche) across all the samples analyzed by microarrays (25 Failed Donors (FD) and 58 PAH lung samples). Ct values of 3 "housekeepers" (GAPDH, ACTB, B2M) were averaged and used to normalize the expression levels by $\Delta\Delta$Ct, which were converted into fold-change. The specific primer/probes used for each gene were indicated as optimum coverage.

We chose 5 phosphodiesterase (PDE) family members (PDE5A is a therapeutic target, (10)), two BMPs (BMP5 and BMP6, involved in BMPR signaling, (11)), and two prostaglandin D synthase enzymes (PTGDS and HPDGS, eicosanoid pathway enzymes, related to a treatment target (12)). Up-regulation of a number PDE family members was interesting for potentially identifying new drug targets, as PDE antagonist development is an active area of cancer research (13). BMPs are activating ligands for a number of membrane receptors, beyond BMPR2 which is often mutated in HPAH and IPAH. BMP5 and BMP6 up-regulation has important implications in discerning potential down-stream pathway activation. Consideration of prostaglandin D synthases antagonists for treatment of PAH is another potential avenue for therapeutic development. For these genes, the microarray dataset predicted up-regulation in PAH lung tissue compared to FD. Correspondingly, the qRT-PCR revealed up-regulation of these genes in the PAH lung, and their expression was significantly different from FD lung.

EDDY Analysis:

Considering complex molecular mechanisms underlying diseases such as cancer and PAH, the discovery of disease-specific therapeutic vulnerabilities will benefit significantly from the analysis of network-driven activities of a gene set rather than individual genes. A novel, network-based computational statistical approach, Evaluation of Differential DependencY (EDDY), combines pathway-guided and differential dependency analyses into a probabilistic framework (14, 15). It interrogates gene sets in related gene network catalogs and databases (*e.g.* REACTOME (16)) to test if inter-gene dependencies are significantly rewired between conditions. Probabilistic and gene-set assisted approaches together contribute to significantly higher sensitivity and specificity of EDDY, compared to other methods, such as GSEA and Gene Set Co-expression Analysis (GSCA) (14). EDDY's high sensitivity and specificity allow identification of network dependencies not evident with other tools (14), and was successfully applied to cancer studies (14, 15, 17). EDDY has uncovered biological network dependencies in glioblastoma (14) and adrenocortical carcinoma (18). Further application of EDDY could allow for analysis of topological characteristics of gene differential dependency networks and thus identify genes that play important, and hidden, roles in a specific condition. The algorithm website is publicly available (http://biocomputing.tgen.org/software/EDDY/). In recent analysis of datasets of >800 cancer cell lines exposed to 368 chemotherapies derived from the Cancer Cell Line Encyclopedia (CCLE) and the Cancer Therapeutics Response Portal (CTRP) (18-20), we identified pathways enriched with differential dependency among sensitive and non-sensitive cancer cell lines for each drug (21).

EDDY is a statistical approach that combines pathway-guided analysis and differential dependency analyses into a probabilistic framework (14, 15). It interrogates gene sets (pathways) in databases, such as REACTOME (16) or other resource that catalogs related gene networks, to test if gene dependencies are significantly rewired between conditions using a statistical test. In evaluating differential dependency, EDDY uses network likelihood distribution over multiple networks constructed via resampling and compares the distributions between the conditions, instead of just using the single, most probable network from each condition. The statistical significance of the divergence is then estimated using a permutation test. Probabilistic and gene-set assisted approaches together contribute to significantly higher sensitivity and specificity of EDDY, compared to other methods, such as Gene Set Co-expression Analysis (14). The method has been further improved by incorporating known gene interactions as prior knowledge. Further development of EDDY to allow for analysis of topological characteristics of gene differential dependency networks could identify genes that play important roles in biological signaling in a specific condition – hence, defining promising targets customized to a specific condition.

After statistically significant pathways and corresponding DDNs were identified by EDDY analysis, the role of each node was statistically evaluated based on its DDN topology. Specifically, the betweenness-centrality metric assesses a node's essentiality within a network (22) and is visualized in the condition-specific network through the node size. In each DDN, high essentiality mediators were identified as those with the largest betweenness-centrality difference between the two condition-specific networks and the size of the nodes in DDN

represents betweenness-centrality difference. The condition-specific rewiring metric identifies genes with a significant proportion of condition-specific edges assessed against the binomial distribution of these edges across the entire graph. In each DDN, these specificity mediators were identified, highlighting particularly highly altered roles between conditions. Both essentiality and condition-specific mediators are indicated by square nodes.

The mRNA expression values were log2 transformed and quantized to values -1 (under-expressed), 0 (intermediate), and 1 (over-expressed). For each gene, median average deviation (MAD) was computed and used to determine under-expression (MAD < -1), over-expression (MAD > 1), and intermediate. Quantized expression array data from the samples were separated into groups of 25 failed donors (FD) and 58 PAH. EDDY then iterated over the full set of REACTOME pathways (16). Prior knowledge was mined from Pathway Commons (23). The recent porting of EDDY to graphical processing unit (GPU) accelerated this analysis.

Results identified 16 REACTOME pathways with statistically significant (p-value < 0.05) rewiring of gene dependencies representing a range of biological function – some known to be important in PAH and others not previously described. Longer teal colored bars indicate larger proportion of dependencies inferred for each gene network that has not been previously reported. For each significantly rewired pathway, differential dependency networks (DDNs) were generated, displaying the gene dependencies found in PAH (edges in red), in failed donor control (edges in blue), or found in both (edges in gray). In these DDNs, known functional interactions are denoted by solid lines, while statistical dependencies as calculated by EDDY analysis are displayed by

dashed lines. Qualitative RNA expression (up- or down-regulation) in PAH or failed donor control lungs was also determined and represented by color coding.

Supplemental Materials References

1. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003; 19: 185-193.
2. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 2003; 31: e15.
3. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003; 4: 249-264.
4. Dugo M, Cotroneo CE, Lavoie-Charland E, Incarbone M, Santambrogio L, Rosso L, van den Berge M, Nickle D, Pare PD, Bosse Y, Dragani TA, Colombo F. Human Lung Tissue Transcriptome: Influence of Sex and Age. *PLoS One* 2016; 11: e0167460.
5. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 2003; 100: 9440-9445.
6. Storey JD, Tibshirani R. Statistical methods for identifying differentially expressed genes in DNA microarrays. *Methods Mol Biol* 2003; 224: 149-157.
7. Li Y, Gu J, Xu F, Zhu Q, Ge D, Lu C. Transcriptomic and functional network features of lung squamous cell carcinoma through integrative analysis of GEO and TCGA data. *Sci Rep* 2018; 8: 15834.
8. Rajkumar R, Konishi K, Richards TJ, Ishizawar DC, Wiechert AC, Kaminski N, Ahmad F. Genomewide RNA expression profiling in lung identifies distinct signatures in idiopathic pulmonary arterial hypertension and secondary pulmonary hypertension. *Am J Physiol Heart Circ Physiol* 2010; 298: H1235-1248.
9. Zhao M, Austin ED, Hemnes AR, Loyd JE, Zhao Z. An evidence-based knowledgebase of pulmonary arterial hypertension to identify genes and pathways relevant to pathogenesis. *Mol Biosyst* 2014; 10: 732-740.
10. Maurice DH, Ke H, Ahmad F, Wang Y, Chung J, Manganiello VC. Advances in targeting cyclic nucleotide phosphodiesterases. *Nat Rev Drug Discov* 2014; 13: 290-314.
11. Goumans MJ, Zwijsen A, Ten Dijke P, Bailly S. Bone Morphogenetic Proteins in Vascular Homeostasis and Disease. *Cold Spring Harb Perspect Biol* 2018; 10.
12. Feldman J, Habib N, Radosevich J, Dutt M. Oral treprostinil in the treatment of pulmonary arterial hypertension. *Expert Opin Pharmacother* 2017; 18: 1661-1667.
13. Peng T, Gong J, Jin Y, Zhou Y, Tong R, Wei X, Bai L, Shi J. Inhibitors of phosphodiesterase as cancer therapeutics. *Eur J Med Chem* 2018; 150: 742-756.
14. Jung S, Kim S. EDDY: a novel statistical gene set test method to detect differential genetic dependencies. *Nucleic Acids Res* 2014; 42: e60.
15. Speyer G, Kiefer J, Dhruv H, Berens M, Kim S. Knowledge-Assisted Approach to Identify Pathways with Differential Dependencies. *Pac Symp Biocomput* 2016; 21: 33-44.
16. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, Matthews L, May B, Milacic M, Rothfels K, Shamovsky V, Webber M, Weiser J, Williams M, Wu G, Stein L, Hermjakob H, D'Eustachio P. The Reactome pathway Knowledgebase. *Nucleic Acids Res* 2016; 44: D481-487.
17. Zheng S, Cherniack AD, Dewal N, Moffitt RA, Danilova L, Murray BA, Lerario AM, Else T, Knijnenburg TA, Ciriello G, Kim S, Assie G, Morozova O, Akbani R, Shih J, Hoadley KA, Choueiri TK, Waldmann J, Mete O, Robertson AG, Wu HT, Raphael BJ, Shao L, Meyerson M, Demeure MJ, Beuschlein F, Gill AJ, Sidhu SB, Almeida MQ, Fragoso M, Cope LM, Kebebew E, Habra MA, Whitsett TG, Bussey KJ, Rainey WE, Asa SL, Bertherat J, Fassnacht M, Wheeler DA, Cancer Genome Atlas Research N,

Hammer GD, Giordano TJ, Verhaak RGW. Comprehensive Pan-Genomic Characterization of Adrenocortical Carcinoma. *Cancer Cell* 2016; 29: 723-736.
18. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jane-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi P, Jr., de Silva M, Jagtap K, Jones MD, Wang L, Hatton C, Palescandolo E, Gupta S, Mahan S, Sougnez C, Onofrio RC, Liefeld T, MacConaill L, Winckler W, Reich M, Li N, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR, Morrissey MP, Sellers WR, Schlegel R, Garraway LA. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012; 483: 603-607.
19. Rees MG, Seashore-Ludlow B, Cheah JH, Adams DJ, Price EV, Gill S, Javaid S, Coletti ME, Jones VL, Bodycombe NE, Soule CK, Alexander B, Li A, Montgomery P, Kotz JD, Hon CS, Munoz B, Liefeld T, Dancik V, Haber DA, Clish CB, Bittker JA, Palmer M, Wagner BK, Clemons PA, Shamji AF, Schreiber SL. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat Chem Biol* 2016; 12: 109-116.
20. Seashore-Ludlow B, Rees MG, Cheah JH, Cokol M, Price EV, Coletti ME, Jones V, Bodycombe NE, Soule CK, Gould J, Alexander B, Li A, Montgomery P, Wawer MJ, Kuru N, Kotz JD, Hon CS, Munoz B, Liefeld T, Dancik V, Bittker JA, Palmer M, Bradner JE, Shamji AF, Clemons PA, Schreiber SL. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discov* 2015; 5: 1210-1223.
21. Speyer G, Mahendra D, Tran HJ, Kiefer J, Schreiber SL, Clemons PA, Dhruv H, Berens M, Kim S. Differential Pathway Dependency Discovery Associated with Drug Response across Cancer Cell Lines. *Pac Symp Biocomput* 2017; 22: 497-508.
22. Freeman LC. A Set of Measures of Centrality Based on Betweenness. *Sociometry* 1977; 40: 35-41.
23. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, Schultz N, Bader GD, Sander C. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res* 2011; 39: D685-690.

Supplemental Figure Legends

Supplemental Figure E1. Principal component analysis (PCA) of the PHBI discovery cohort.

Affymetrix HuGene ST1.0 microarrays from batch 1-3 were imported into Partek Genome Studio and visualized by PCA plots. Supplemental Figure E1A shows the separation of batch 3 from batches 1 and 2 prior to adjustment. Supplemental Figure E1B shows the overlapping of all 3 batches after adjustment for Batch and Sex as co-variables in the ANOVA modeling used by Partek Genome Studio.

Supplemental Figure E2. Volcano plot highlighting PHBI classifier in relation to all TranscriptIDs measured.

All expression transcriptIDs are plotted as $\log_2$(Fold-change as PAH/FD) versus $\log_{10}$(q-value) in a volcano plot. The PHBI classifier (Affymetrix TranscriptIDs n = 1,140 q-value < 0.001) is highlighted in red circles while purple triangles include those that also have Fold-change >|1.5|.

Supplemental Figure E3. Supervised validation of the PHBI classifier on microarray datasets.

The PHBI classifier (Affymetrix TranscriptIDs n = 1,140 q-value < 0.001) was used for supervised hierarchical clustering to provide visualization of the validation datasets. A smaller PAH lung Agilent microarray dataset previously published (Rajkumar *et al*. 2010) was downloaded from GEO (GSE15197; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15197), imported into Partek, and the Agilent IDs were matched to their respective gene symbols. Of these, 13 were their lung controls (equivalent to FD) and 18 were PAH only. The remainder of their analysis were from PAH+IPF patient lungs, and were excluded from this analysis. The gene

symbols that overlapped with the PHBI classifier (n = 893 out of n = 1,140 total) were analyzed by Partek Genomic Suite ANOVA model and all results (regardless of p-value) were used for hierarchical clustering (Supplemental Figure E3).

Supplemental Figure E4. Supervised validation of the PHBI classifier using literature-derived PAH pathways.

Additional validation was completed using a literature-derived (key word searches along with PAH knowledge-based inquiry) genelist which was used to develop networks of potential pathways important to PAH disease. The gene symbols were used to select out the expression data from the full PHBI dataset (Zhao *et al*. 2014, n = 112 genes Supplemental Figure E4). The expression data in this subset was analyzed by Partek Genomic Suite ANOVA model and hierarchical clustering visualized the PHBI PAH and FD samples.

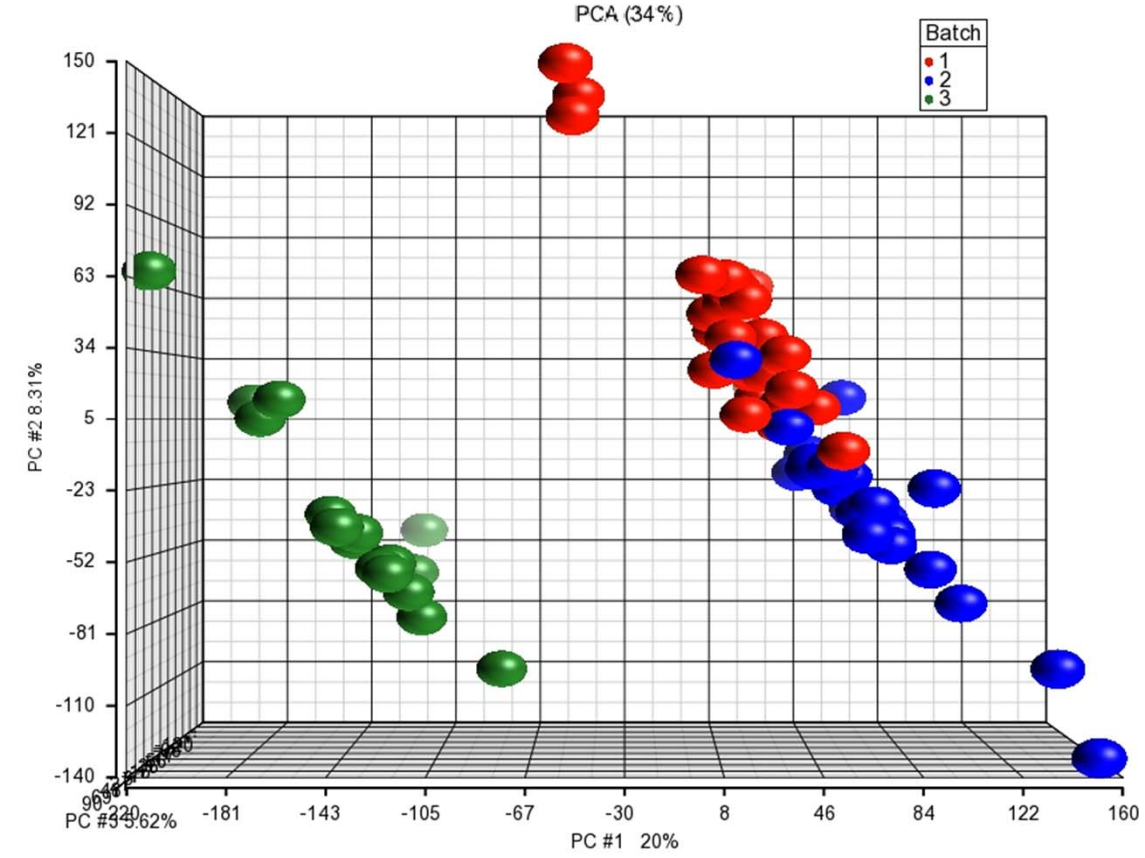
Supplemental Figure E5. qRT-PCR validation of selected predicted up-regulated genes in PAH.

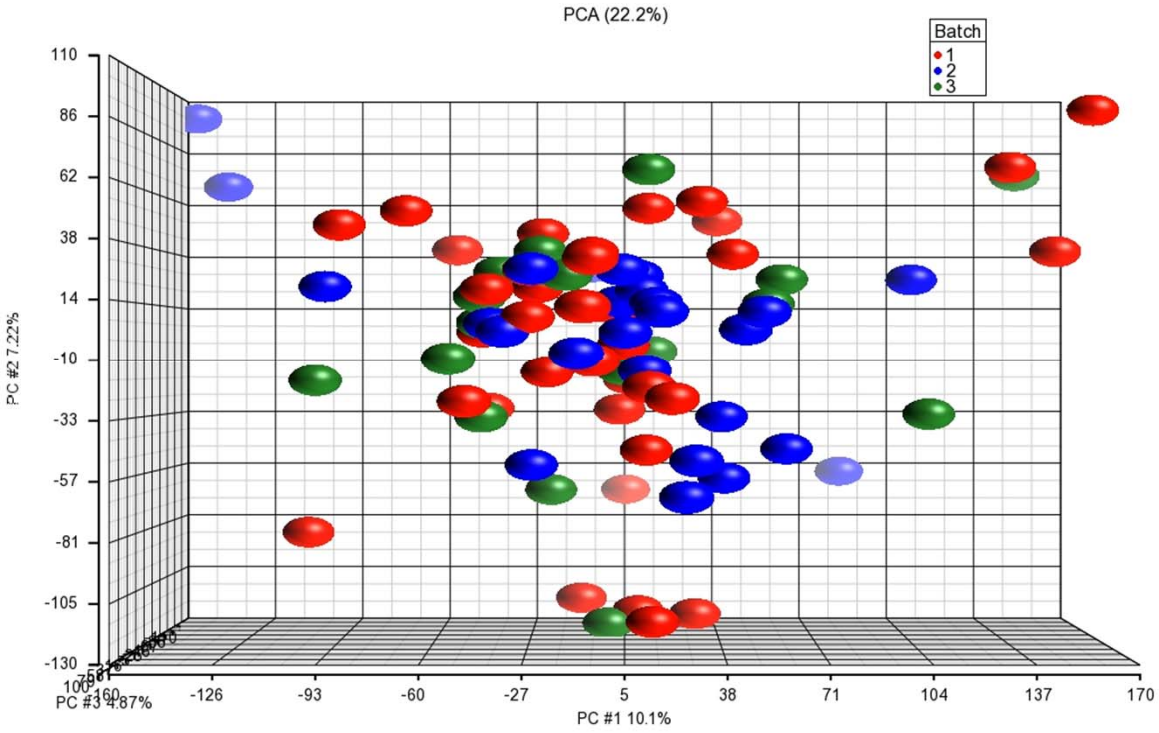
Nine genes were tested by qRT-PCR including 5 PDEs, 2 genes involved in PGD_2 synthesis, and 2 BMPs. All were in the PHBI classifier (q-value < 0.001) and predicted to be up-regulated in PAH lung tissue. All 9 genes were found up-regulated by qRT-PCR using Student's t-test for significance.

Supplemental Figure E6. Tumor Necrosis Factor is a potential master upstream regulator in PAH.

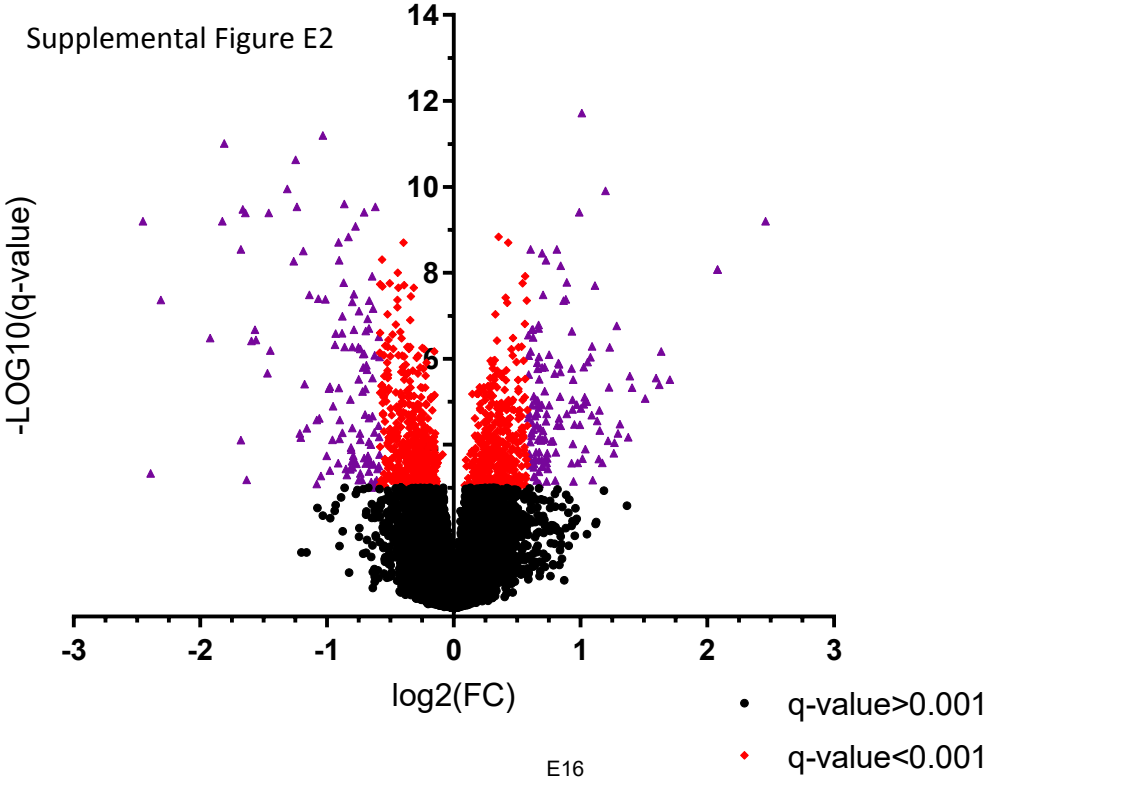
The PHBI classier, including fold-change expression and q-values, was imported into Ingenuity Pathway Analysis. Tumor Necrosis Factor (shown as alias TNFA) was identified as the top ranked upstream regulator (IPA pathway enrichment, p-value 1.24E-14; see also Table 3 and Supplemental Table 4) with 147 connected genes. Connected genes are displayed alphabetically from ABCG2 (roughly 9 o'clock) to WISP2 in a clockwise direction.

Supplemental Figure E1

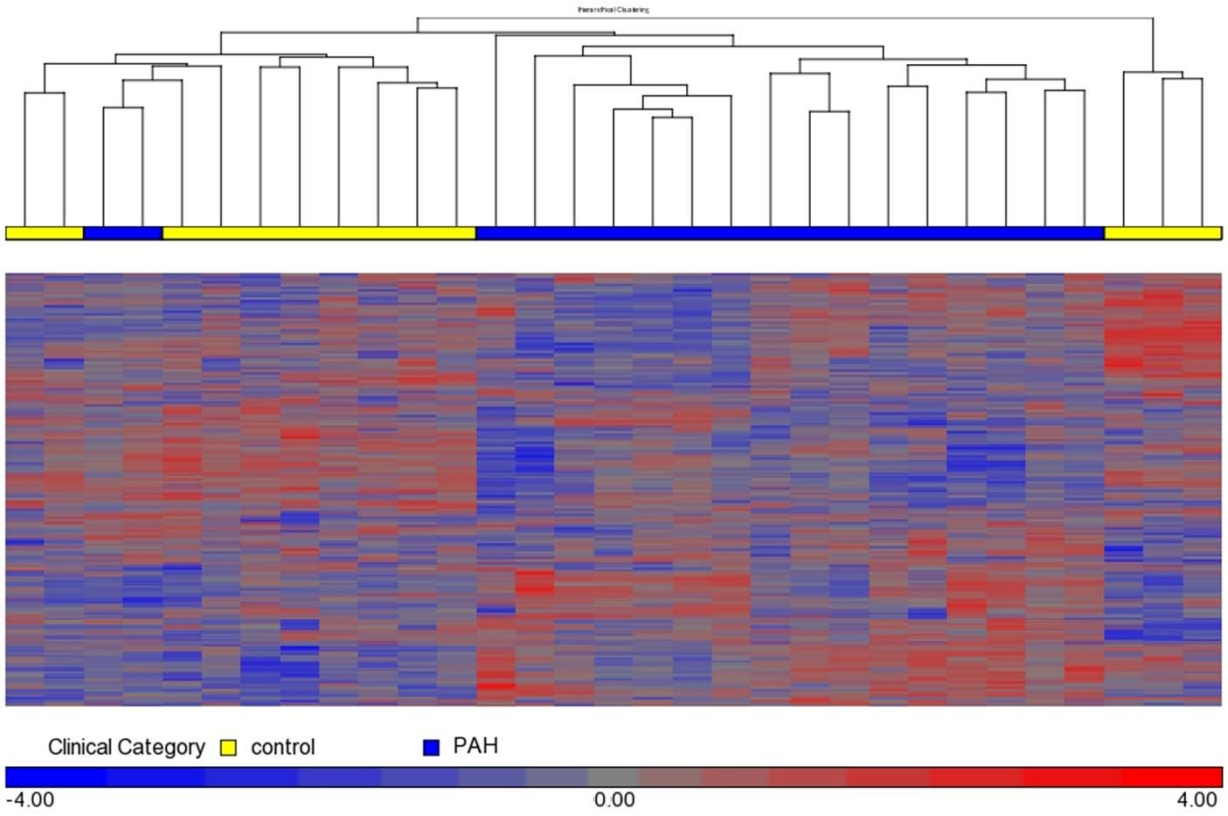




Supplemental Figure E2

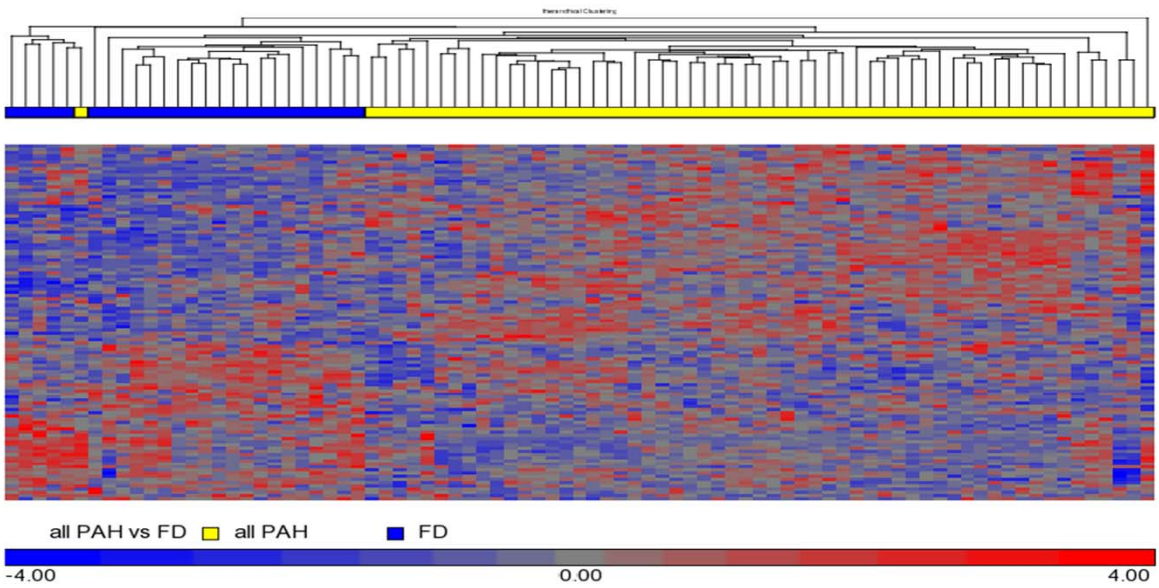


Supplemental Figure E3



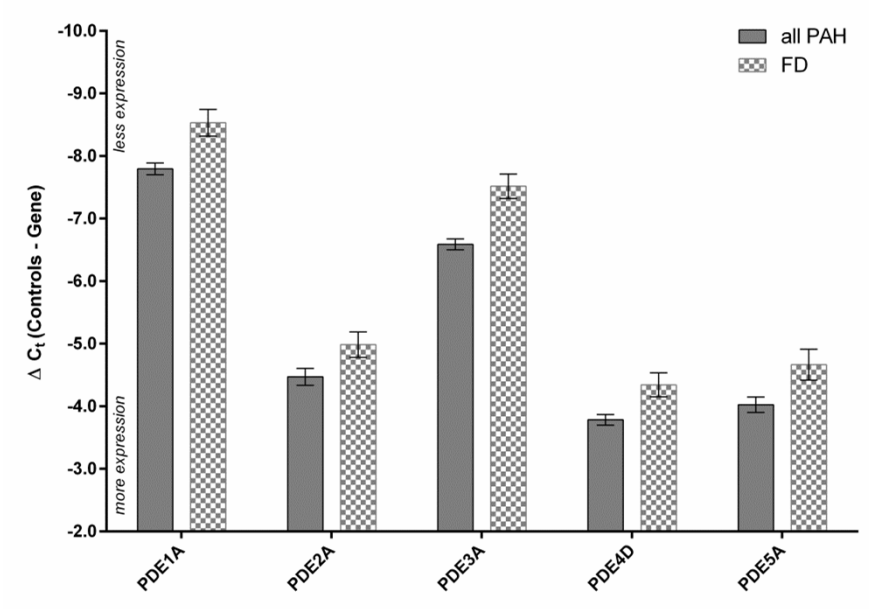
Validation: Performance of PHBI Classifier on the Rajkumar Dataset (n=893)

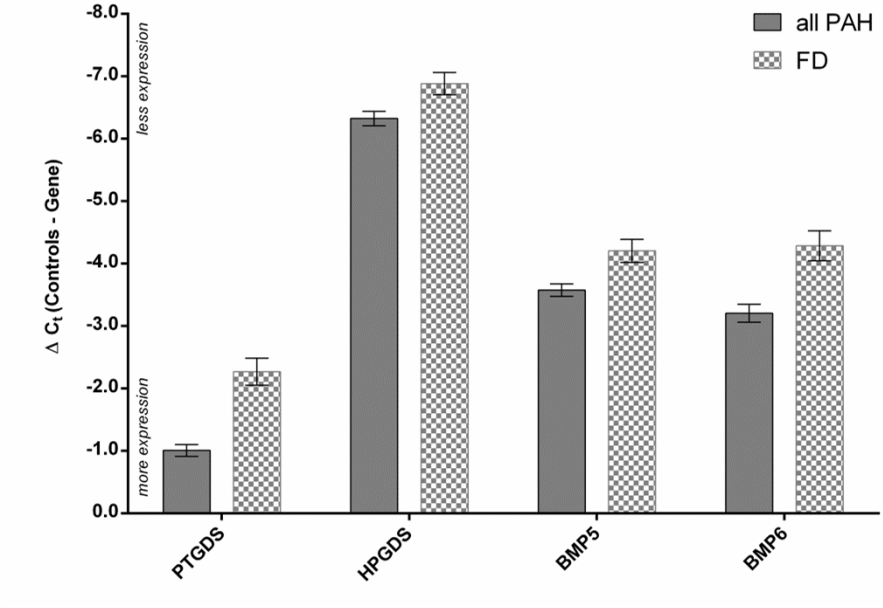
Supplemental Figure E4



Validation : Performance of Zhao PAH Gene Network (n=112; q-value<0.05)

Supplemental Figure E5





Gene	Fold Change PAH/FD	ttest (p-value)
PDE1A	1.67	3.305E-03
PDE2A	1.43	3.951E-02
PDE3A	1.90	1.170E-04
PDE4D	1.47	1.181E-02
PDE5A	1.56	1.135E-02
PTGDS	2.40	7.288E-06
HPGDS	1.47	9.388E-03
BMP5	1.55	1.843E-03
BMP6	2.12	1.200E-04

Supplemental Figure E6