Chu et al. Environmental Health (2016) 15:101
DOI 10.1186/s12940-016-0187-z

Environmental Health

RESEARCH Open Access

Gene expression network analyses in response to air pollution exposures in the trucking industry

Jen-hwa Chu[1*], Jaime E. Hart[2,3], Divya Chhabra[2], Eric Garshick[2,4], Benjamin A. Raby[2,5] and Francine Laden[2,3,6]

Abstract

Background: Exposure to air pollution, including traffic-related pollutants, has been associated with a variety of adverse health outcomes, including increased cardiopulmonary morbidity and mortality, and increased lung cancer risk.

Methods: To better understand the cellular responses induced by air pollution exposures, we performed genome-wide gene expression microarray analysis using whole blood RNA sampled at three time-points across the work weeks of 63 non-smoking employees at 10 trucking terminals in the northeastern US. We defined genes and gene networks that were differentially activated in response to $PM_{2.5}$ (particulate matter ≤ 2.5 microns in diameter) and elemental carbon (EC) and organic carbon (OC).

Results: Multiple transcripts were strongly associated ($p_{adj} < 0.001$) with pollutant levels (48, 260, and 49 transcripts for EC, OC, and $PM_{2.5}$, respectively), including 63 that were statistically significantly correlated with at least two out of the three exposures. These genes included many that have been implicated in ischemic heart disease, chronic obstructive pulmonary disease (COPD), lung cancer, and other pollution-related illnesses. Through the combination of Gene Set Enrichment Analysis and network analysis (using GeneMANIA), we identified a core set of 25 interrelated genes that were common to all three exposure measures and were differentially expressed in two previous studies assessing gene expression attributable to air pollution. Many of these are members of fundamental cancer-related pathways, including those related to DNA and metal binding, and regulation of apoptosis and also but include genes implicated in chronic heart and lung diseases.

Conclusions: These data provide a molecular link between the associations of air pollution exposures with health effects.

Keywords: Air pollution, Trucking industry, Gene expression, Network analysis

Background

Air pollution exposures, have been associated with a number of adverse health effects, including greater morbidity and mortality risks for cardiopulmonary diseases, and increased risk of lung cancer [1–6]. However, the underlying biological mechanisms have not been fully elucidated. Human studies of global changes in gene expression following controlled exposures [7], or using in vitro models [8, 9] have provided some insights in this regard, yet few studies have rigorously assessed the impact of air pollution on gene expression in real-life settings. For example, though observational studies have been conducted in individuals from geographic regions with differing levels of air pollution have suggested associations, [10] studies with more refined exposure measures have not been performed.

In this study, we characterized the cellular response induced by traffic-related air pollution exposures in a population of non-smoking US trucking industry employees. We performed genome-wide gene expression microarray analysis using whole blood RNA sampled at three time-points during the work week. We integrate

* Correspondence: jen-hwa.chu@yale.edu
[1]Section of Pulmonary, Critical Care and Sleep Medicine, Department of Internal Medicine, Yale University School of Medicine, New Haven, CT, USA
Full list of author information is available at the end of the article

BioMed Central

© The Author(s). 2016 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

these data with micro-environmental measures of occupational exposure to three pollutants –particulate matter ≤ 2.5 microns in aerodynamic diameter ($PM_{2.5}$), elemental carbon (EC), and organic carbon (OC) in $PM_{1.0}$ (particulate matter with a diameter of ≤1.0 μm). Our objective was to identify the genes and gene networks differentially activated in response to these exposures.

Methods

Study population

A total of 95 subjects were recruited from 10 trucking terminals in the northeastern US (CT, MA, MD, NJ, NY, and PA). The participants were workers whose job duties were characterized by different patterns of exposure: pick-up and delivery (P&D) drivers, with regular exposures to traffic; loading dock workers with regular exposures to propane forklifts and episodic exposures to diesel trucks and other vehicles in the terminal yard; office workers with no occupational traffic related exposures, and combination workers, who performed the job duties of a P&D driver or a dock worker, as needed.

The measurements took place between February 2009 and October 2010. Each subject was enrolled on the first day of the workweek following at least two days off. Whole blood samples were collected using PaxGene RNA tubes, three times from each subject: (1) before the first shift of the workweek (first day, AM draw); (2) at the end of the first shift (8–12 h later) on the same day (first day, PM draw); and (3) at the end of the last work shift of the same workweek (last day, PM draw). This design allowed us to assess the cross-shift effects after returning from work after at least 2 days off, and the cross-week effects (i.e., over 2–5 days). Our primary analyses were restricted to the 63 Caucasian non-smoking male workers with at least a single blood sample available. The majority of participants were excluded for being current active smokers ($n = 21$) given the known effects of smoking on gene expression [11] four participants were unable to provide a sufficient blood sample, two reported an active illness (cold or flu) at the time of blood draw, and one female and 7 non-white men were excluded. The final data set includes a total of 165 samples.

Measurement of traffic exposures

Micro-environmental samples of $PM_{2.5}$, and EC and OC in $PM_{1.0}$ were collected over the full workweek (24 h/day for 6–9 days) at each of the 10 terminals. Twelve-hour area samples were collected indoors in office spaces and terminal docks. Samples also were collected in the truck cabs of participating drivers during their work shifts on their first and last day of work. Detailed information on the exposure assessment for each of the three pollutants is described elsewhere [12]. Briefly, EC and OC were measured by collecting $PM_{1.0}$ on a 22-mm quartz tissue filter, preceded by a precision machined cyclone separator (SCC1.062 Triplex, BGI, Inc., Waltham, MA), which was then analyzed with thermal-optical carbon analyzer using the NIOSH 5040 method [13]. $PM_{2.5}$ was collected on a pre-weighed 37-mm Teflon filter (with a pore diameter of 0.2 μm) after passing through a precision-machined cyclone pre-selector to remove particles greater than 2.5 μm in aerodynamic diameter. After collection, the filter was reweighed to obtain the mass of $PM_{2.5}$ collected. The method was consistent with the EPA PQ200 Federal Reference Method [14, 15]. For each participant on each day, exposures to $PM_{2.5}$, EC, and OC were assigned as a weighted average of the time spent in each work location.

Gene expression data

Blood samples were stored at 4 °C on the day of collection until they were shipped overnight each day to our blood processing laboratory in Boston, MA in an insulated container with a cooler pack to keep samples chilled. Upon arrival, RNA was extracted using the Qiagen RNAEasy extraction kit, according to protocol and then stored at −80 °C until analysis. Gene expression profiling was conducted using the Illumina HumanHT-12 v4 Expression BeadChip, with RNA labeling and array hybridization performed according to protocol. Image capture was performed using the Illumina BeadArray Reader. Standard QC and preprocessing procedures were applied to remove failed samples ($n = 2$). Standard background correction and normalization procedures (Variance-Stabilizing Transform, [16]) were applied using the R package *lumi*. The final data set included information from 47,295 probes on 165 samples from 63 subjects.

Statistical analysis

To maximize the power of our repeated measures of gene expression, we employed a mixed effect model that considered gene expression measures from all three blood draws, with the form:

$$\text{Expression} = \text{Exposure} + (1|\text{subject}) + \text{Confounders};$$

where the expression measurements were treated as repeated measures. Each exposure was considered separately, and to estimate the impact of long-term exposures, we used the average of the exposure measures from the first and last work shift for each participant. Personal factors were considered as potential confounders including age, and body mass index (BMI). Job title and terminal were not considered as confounders as they were assumed to be proxies for our measured

exposures. BMI was not associated with either the exposure or the outcome, therefore the only covariate included in the final model was age. Statistical significance was determined by estimating the False Discovery Rate (FDR) by permutation testing, to correct for possible *p*-value inflation introduced by the covariance of repeated measures. We also tested for the cross-shift effects (blood draw 1 vs. blood draw 2) and the cross-week effects (blood draw 1 vs. blood draw 3) within each subject, using the difference in the expression measurement between the blood draws as the response variable.

Gene expression changes to environmental perturbations are thought to arise through coordinated responses of specific gene networks that are often difficult to appreciate through single gene testing. We therefore applied Gene Set Enrichment Analysis (GSEA) using GSEA software from the Broad Institute to identify subsets of genes with shared function that were altered by exposure to vehicle exhaust from the Molecular Signature database (MSigDb [17]), which is a collection of annotated gene sets for GSEA analysis. The gene sets are categorized into different collections (C1-C7), and here we considered six functional categories of gene sets: C2 (curated gene sets), C3 (motif gene sets), C4 (computational gene sets), C5 (GO gene sets), C6 (oncogenic signatures), and C7 (immunologic signatures). The gene set enrichment analyses are based on a list of genes ranked by effect sizes from the linear mixed effect models for all three types of exposures were performed. Those gene sets that were significantly enriched for all three types of exposures (EC, OC, $PM_{2.5}$) were marked for further functional annotation. Significance was claimed at an FDR of 25 %, as recommended for GSEA. For these enriched gene sets, we also identified the genes that contributed most to the enrichment ("leading edges"). In addition, we also performed connectivity map analysis to identify additional genes connected to the genes correlated with exposure levels. Finally, we performed GSEA analyses on data from two other air pollution related Gene Expression Omnibus (GEO) datasets (GSE7462 [7] and GSE7543 [10]) to assess the generalizability of our results.

GSE7462 [7] is from a crossover, double-blind study of the effects of diesel exhaust inhalation compared to fresh air exposure on peripheral blood mononuclear cells (n = 23); and GSE7543 [10] is a study of differences in expression in peripheral blood samples collected from two regions of the Czech Republic with markedly different levels of pollution (n = 71).

Results

Selected characteristics of the study subjects, and the mean exposure levels of $PM_{2.5}$, EC and OC observed over the 5-day workweek, are presented in Table 1. The participants were 50.5 years old (SD = 8.4) on average, 52.4 % were former smokers, and pick-up and delivery drivers were the largest job group sampled (46.0 %). Out of the three pollutants, EC and $PM_{2.5}$ were modestly correlated (r^2 = 0.38), while EC-OC and $PM_{2.5}$-OC were not (r^2 = 0.02 and 0.07, respectively). We found no systematic difference in expression measurements or other demographic variables between the subjects with complete blood samples at all three collection times and the subjects with missing data (data not shown).

For the gene-level differential expression analysis, the tests for cross-shift and cross-week effects did not yield any significant results. Therefore we focus on the results from linear mixed-effect model. QQ plots contrasting the observed with permuted *p*-value distributions for the linear mixed effect model analyses are presented in Additional file 1: Figure S1, demonstrating excessive deviations of the observed results from expectation. We estimated the genomic inflation factor lambda for each analysis (EC = 1.09, OC = 1.39, and $PM_{2.5}$ = 1.15), to enable adjustment of our results for unobserved technical biases. Multiple transcripts were strongly correlated (padj < 0.001) with week-long average pollutant levels (EC n = 48, OC n = 260, $PM_{2.5}$ n = 49, see Fig. 1 for examples of most strongly correlated genes for each exposure), including 67 that were strongly correlated with at least two of three exposure measurements (See Additional file 2: Table S1), though no individual genes met our *a priori* threshold of statistical significance at an FDR < 0.1.

In contrast to the gene level analysis, GSEA revealed widespread differential responses with long-term exposures.

Table 1 Characteristics of the 63 white male trucking industry workers

	Total
Total no.	63
Age (years, mean ± SD)	50.5 ± 8.4
Past smoker (no. (%))	33 (52.4)
Primary job title (no.(%))	
Pick-up and Delivery Driver	29 (46.0)
Dockworker	12 (19.1)
Officeworker	15 (23.8)
Combination Workers[a]	7 (11.1)
Workweek average exposure (μg/m^3, mean ± SD)	
$PM_{2.5}$	11.5 ± 5.1
EC	0.6 ± 0.4
OC	8.8 ± 3.0

[a]Combination workers perform the jobs of pick-up and delivery drivers or dockworkers as needed

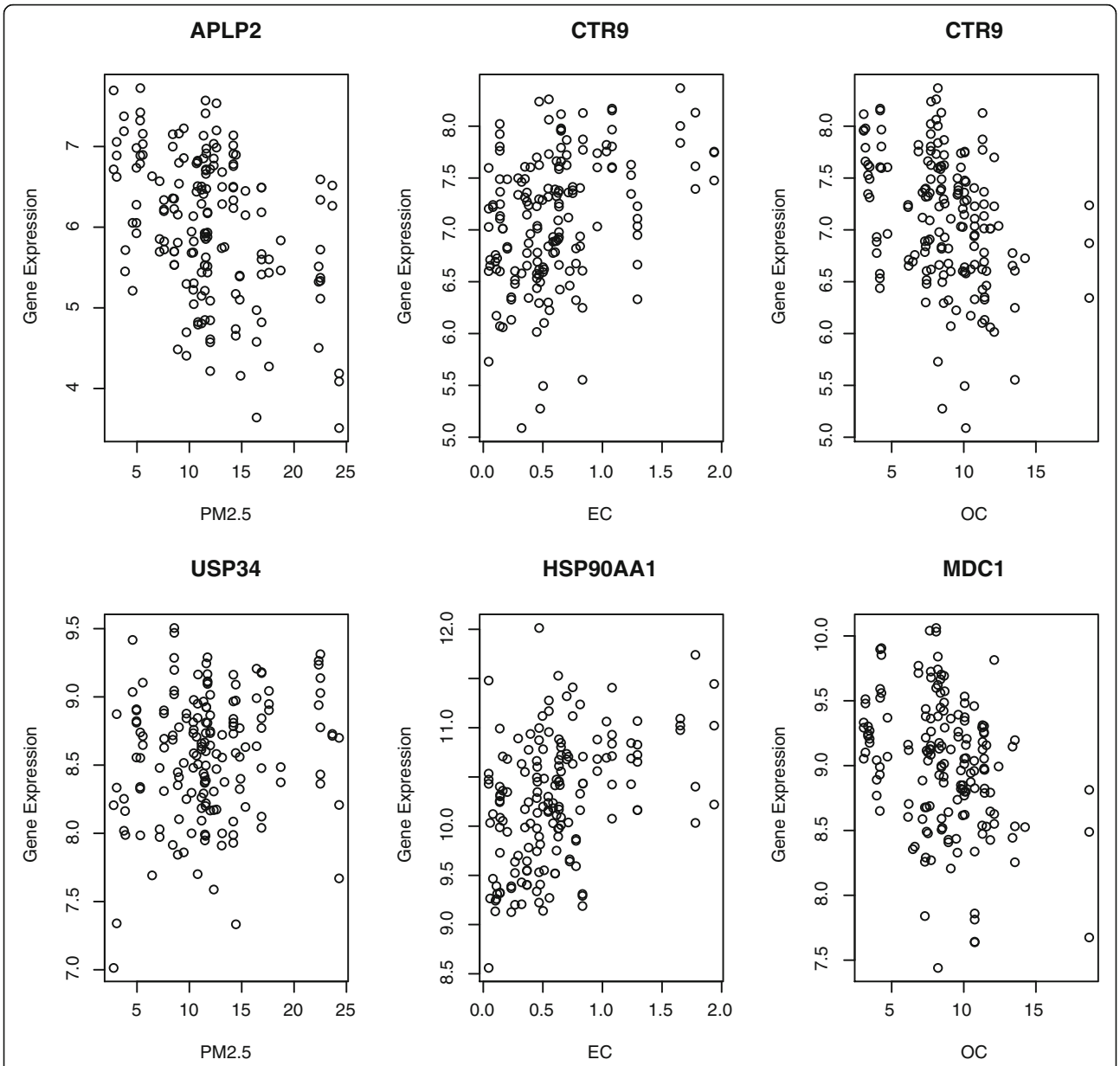

Fig. 1 Dot plots of most important expression profiles from core gene set. Left: Dot plots for genes most correlated with $PM_{2.5}$ (APLP2 and USP34); Middle: Dot plots for genes most correlated with EC (CTR9 and HSP90AA1); Right: Dot plots for genes most correlated with OC (CTR9 and MDC1)

A total of 6019 gene sets were significantly enriched (FDR < 0.25 and nominal p-value < 0.05) for any of the three exposure types (See Fig. 2), including 2384 gene sets from the C2 collection, 698 sets from the C3 collection, 586 from the C4 collection, 445 from the C5 collection, 128 from the C6 collection, and 1778 from the C7 collection. EC exposure was consistently correlated with the greatest number of enriched gene sets across all collections. Most notable was the great degree of gene set enrichment overlap across exposure types. Overall, 59.5 % (3580 of 6019) of gene sets were implicated in more than one exposure response, including 82.2 % of C7 collection sets, 64.5 % of C4 sets, 53.1 % of C6 sets, 49.4 % of C2 sets, 49.4 % of C3 sets, and 33.9 % of C5 sets (Fig. 2). Further inspection revealed that 20 % (1207) of gene sets were enriched across all three pollutants ($p < 10^{-16}$ for any sharing between pollutants, $p < 10^{-16}$ for sharing across all three pollutants).

From the regression and GSEA analyses, we defined a core set of 262 genes whose expression was modified by occupational exposure to vehicle exhaust and that were

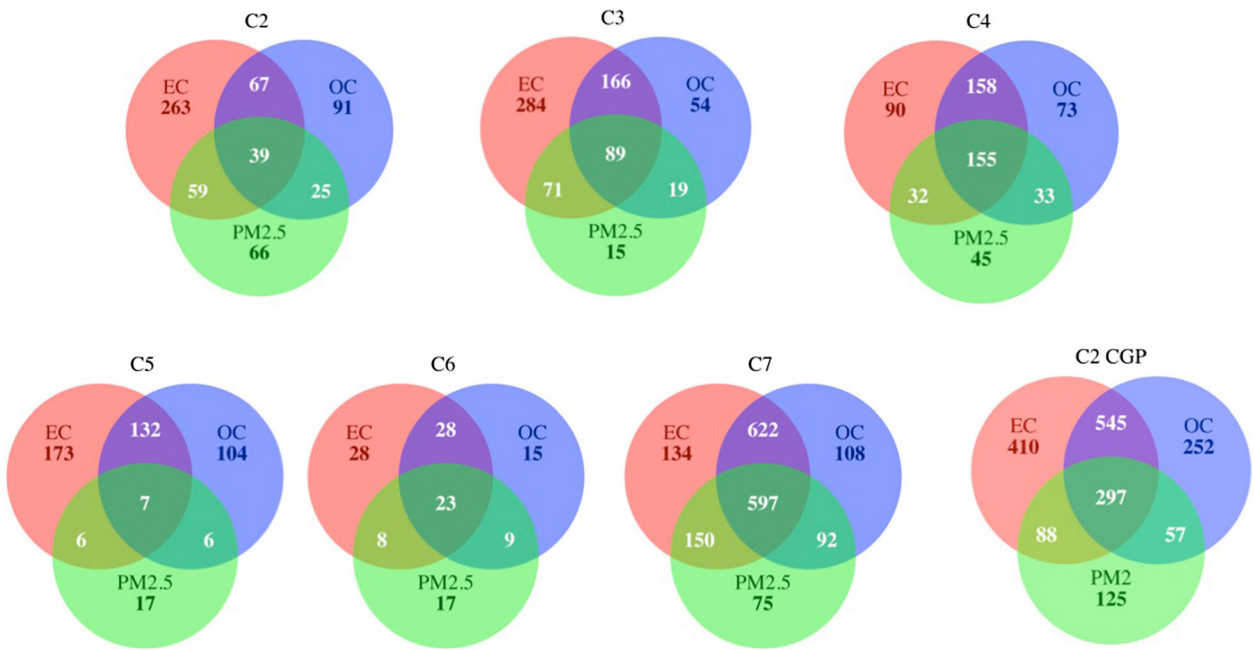

Fig. 2 Venn Diagrams for gene sets. The numbers of the gene sets that were significantly enriched with FDR < 0.25 and nominal *p*-values < 0.05. Note: C2 = curated gene sets (excluding CGP gene sets, see below), C3 = motif gene sets, C4 = computational gene sets, C5 = GO gene sets, C6 = oncogenic signatures, C7 = immunologic signatures, CGP = Chemical and genetic perturbations

most frequently represented in enriched gene sets that emerged from the GSEA (Additional file 2: Table S1). This core gene set included the 20 genes most differentially expressed for each pollution measure (OC, EC, and PM), those that were differentially expressed by two or more measures, genes that were enriched in at least 10 gene sets per MSigDb collection, and genes that were enriched in at least 20 sets common to all three exposure measures. To assess whether members of this core set were specific to our study, or have been implicated by others, we formally tested whether this core gene set was overrepresented in two previously published, independent air pollution gene expression datasets available through GEO.. We observed statistically significant enrichment of our selected gene set in both datasets ($p < 10^{-16}$ for each). Of the 262 genes, 114 were within the leading edge for the GSE7462 dataset, and 64 for GSE7543, with 25 genes common to both (Table 2). This common set of leading edge genes included multiple genes implicated in the interrelated processes of DNA binding (LEF1, MLH1, RBM5, STAT1, CITED2, APLP2, DDX3Y, ZNF589), metal binding (MAN1A1, PRIM1, RBM5, STAT1, ZFP161, and ZNF589), apoptosis regulation (CITED2, MLH1, RBM5, STAT1), and cancer pathways (HSP90AA1, LEF1, MLH1, STAT1).

These 25 genes, which were derived from our expression analyses of three exposure measures and were also enriched in the two independent GEO datasets, represent the most highly reproducible (i.e. most robust) gene set to emerge from our analysis. To determine the interrelationship among this core set of genes, we applied GeneMANIA network analysis [18], the results of which are displayed in Fig. 3. The derived network consisted of 24 of the 25 core genes (the sole core gene not incorporated into the network, CTAGE5, is a known pseudogene) and 20 additional genes that were pulled in by GeneMANIA. Though these 44 genes have been implicated in numerous biological and cellular processes, several processes were statistically significantly overrepresented among these lists, including those of DNA binding (13 of 44 genes, $p = 0.01$), cell surface receptor linked signal transduction (11 genes, $p = 0.01$), and cancer (8, $p = 1.1 \times 10^{-4}$). The derived network demonstrated substantial evidence of interconnectivity: a total of 131 gene-gene interactions were observed, with each of the 25 core genes interacting on average with 6.95 other genes (range 1–15). Genes (nodes) with the highest number of interactions included GNAQ (15 connections), CTR9 (13), HSP90AA1 (12), MLH1, FNBP4, ACP1, MDC1 (each with 10 connections), and LEF1, LPIN1, and RBM5 (9 connections each). Together, these ten hub genes linked to all but two of the other genes in the network, and five (ACP1, HSP90AA1, LEF1, MLH1, and RBM5) are common to the major cancer-related pathways identified above (DNA, metal binding, and apoptosis regulation).

Table 2 List of 25 genes in the core gene set that were differentially expressed in two independent GEO data sets

Gene symbol	Gene name	Number of interactions	Number of pathways	Mean (SD) log2 expression			*p*-value			
				Trucker	Dock	Office	ANOVA	$PM_{2.5}$	EC	OC
GNAQ	Guanine nucleotide binding protein, q polypeptide	15	18	5.96 (1.17)	7.04 (0.91)	6.15 (1.74)	0.0004	0.1535	0.0116	0.0017
CTR9	Ctr9, Paf1/RNA polymerase II complex component	13	7	7.01 (0.64)	7.44 (0.49)	7.10 (0.61)	0.0037	0.5095	6.11E-06	5.59E-06
HSP90AA1	Heat shock protein 90 kDa alpha, class A member 1	12	0	10.24 (0.66)	10.43 (0.35)	10.13 (0.74)	0.1687	0.8814	1.36E-08	0.0353
MLH1	mutL homolog 1	10	19	8.66 (0.29)	8.83 (0.21)	8.65 (0.39)	0.0184	0.0037	0.0025	0.0547
FNBP4	Formin binding protein 4	10	23	10.36 (0.49)	10.86 (0.29)	10.44 (0.62)	1.10E-05	0.1515	0.0063	0.0004
ACP1	Acid phosphatase 1	10	13	8.71 (0.52)	9.22 (0.28)	8.69 (0.56)	3.53E-06	0.0107	1.95E-06	0.0003
MDC1	Mediator of DNA-damage checkpoint 1	10	17	8.93 (0.49)	9.32 (0.35)	8.93 (0.54)	0.0004	0.1663	0.0530	5.75E-05
LEF1	Lymphoid enhancer-binding factor 1	9	42	11.17 (0.45)	11.41 (0.32)	11.00 (0.46)	0.0006	0.0291	0.0096	0.0001
LPIN1	Lipin 1	9	28	9.69 (0.52)	10.17 (0.36)	9.78 (0.65)	9.96E-05	0.1880	0.0006	0.0005
RBM5	RNA binding motif protein 5	9	15	11.73 (0.30)	12.02 (0.22)	11.77 (0.39)	4.36E-05	0.0333	0.0101	0.0008
MAN1A1	Mannosidase, alpha, class 1A, Member 1	8	4	5.26 (0.97)	5.85 (0.75)	5.46 (1.06)	0.0138	0.2408	2.35E-05	2.06E-06
ZFP161	Zinc finger protein 161	8	17	8.08 (0.51)	8.28 (0.26)	8.05 (0.60)	0.1081	0.1419	0.0046	0.0030
HSPA8	Heat shock 70 kDa protein 8	7	21	12.49 (0.24)	12.66 (0.17)	12.47 (0.32)	0.0031	0.0096	1.81E-05	0.0081
PRIM1	Primase 1	6	39	7.70 (0.40)	7.88 (0.24)	7.56 (0.44)	0.0045	0.0632	3.47E-05	0.1771
STAT1	Signal transducer and activator of transcription 1	6	0	11.86 (0.60)	11.49 (0.34)	11.79 (0.51)	0.0066	0.6718	0.0821	2.14E-06
CITED2	Cbp/p300-interacting transactivator 2	4	19	9.62 (0.48)	9.84 (0.38)	9.70 (0.42)	0.0728	0.0201	0.0069	0.0112
APLP2	Amyloid beta precursor-like protein 2	4	0	6.01 (0.98)	5.84 (0.65)	6.21 (0.77)	0.2322	5.65E-08	0.3388	0.8264
DDX3Y	DEAD box polypeptide 3, Y-linked	4	16	2.48 (1.72)	2.84 (1.48)	2.70 (1.77)	0.5375	0.0574	0.0386	0.0191
HLA-DMA	Major histocompatibility complex, class II, DM alpha	4	0	12.50 (0.46)	12.18 (0.44)	12.45 (0.54)	0.0069	0.7922	7.36E-06	9.93E-06
DOCK9	Dedicator of cytokinesis 9	2	18	2.95 (1.48)	4.37 (1.19)	3.25 (1.82)	7.74E-05	0.1116	0.0001	0.0002
EML3	Echinoderm microtubule associated protein like 3	2	17	11.70 (0.39)	11.56 (0.33)	11.57 (0.34)	0.0787	0.0175	0.0133	0.0006
GPBAR1	G protein-coupled bile acid receptor 1	2	0	10.24 (0.60)	9.91 (0.48)	10.31 (0.81)	0.0227	0.9312	0.0002	7.27E-07
USP34	Ubiquitin specific peptidase 34	2	22	8.49 (0.39)	8.89 (0.31)	8.55 (0.58)	5.32E-05	0.0366	0.0095	0.0077
ZNF589	Zinc finger protein 589	1	0	9.00 (0.49)	9.02 (0.36)	8.66 (0.40)	0.0004	3.60E-08	0.2132	0.5830
CTAGE5	CTAGE family, member 5 pseudogene	0	0	3.26 (1.05)	4.10 (0.99)	3.68 (1.27)	0.0008	0.5163	0.5885	0.0008

Discussion

This study represents the first attempt to characterize the molecular impact of air pollution using micro-environmental measures of exposure. Four primary observations were made. First, GSEA demonstrated widespread evidence of pathway-specific changes in gene expression, with 6019 gene sets demonstrating enrichment for traffic-related air-pollution correlated genes. The enriched gene sets implicated a range of cellular responses and pathways. Several – including oxidative stress responses [19], interferon-mediated in amatory responses to viral infection [20], and hypoxemia-induced responses [21] – are known response mechanisms for dealing with the likely biological consequences of air pollutant exposures. Second, we found that despite the relatively low within-subject correlation of the measures of exposure, there was a striking degree of shared gene set enrichment across pollutants (59.5 % for any sharing, 20 % for sharing across all three exposures). The degree of overlap across the exposures, while far more than

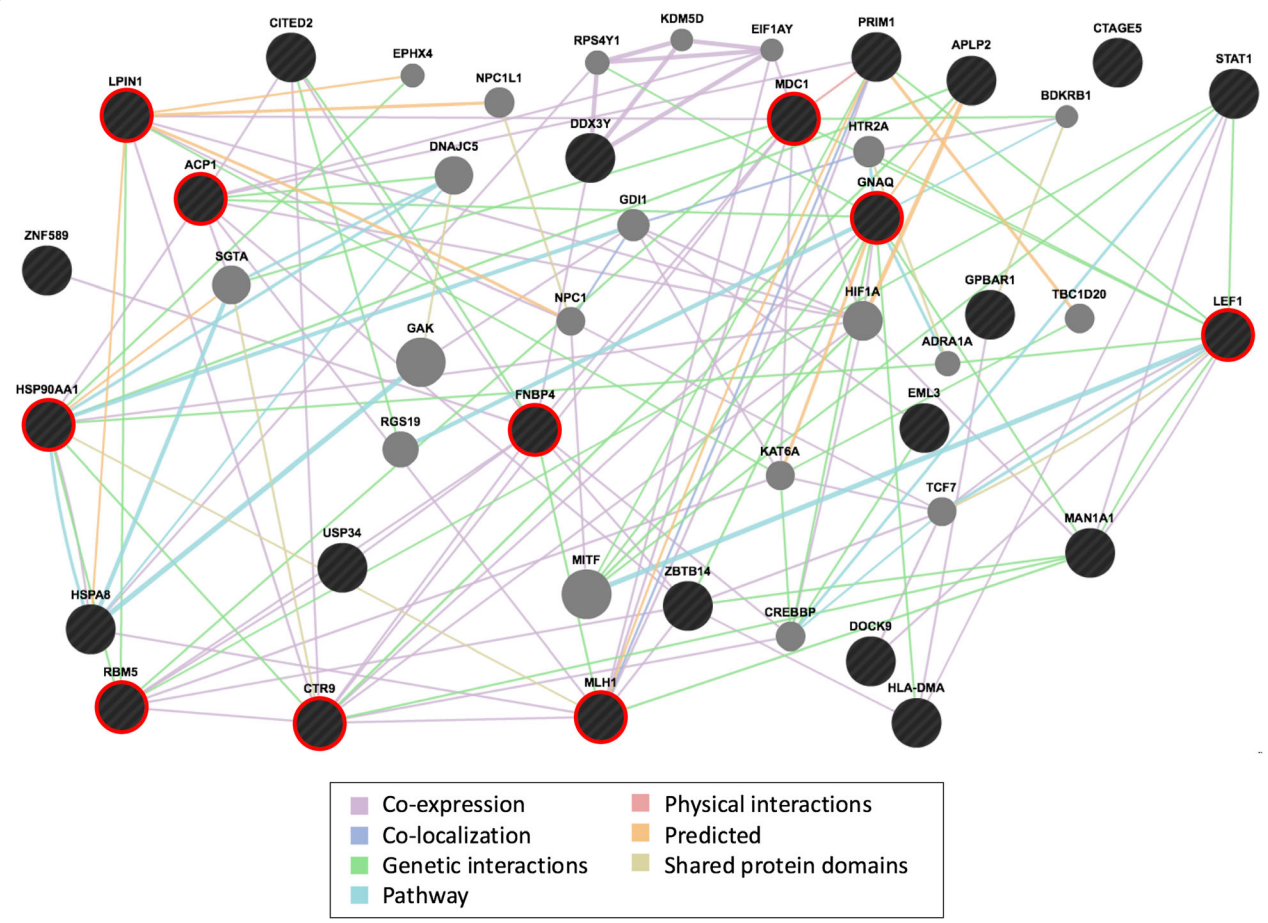

Fig. 3 Interconnectivity of particulate induced genes

would be expected by chance, was not surprising because the three exposure metrics (EC, OC and $PM_{2.5}$) are representing different properties of the same source. These overlapping gene sets implicated a core set of genes and pathways, suggesting a set of common molecular responses to such exposure. In addition, inspection of the gene memberships within the leading edges of these overlapping gene sets suggested evidence of overlap of critical genes, both across pathways and across pollutants. However, the 80 % of non-overlapping gene sets also suggest independent mechanisms may be more related to individual exposures. Third, our connectivity analysis defined a sub-network of interconnected genes at the heart of this shared response. Though several individual components of the network have been previously implicated in anti-oxidative and other protective cellular responses to pollution [22], our findings focused greater attention on these genes and their neighbors as central role players underlying these responses. Finally, we demonstrated the generalizability of our results beyond the trucking industry to the general population by demonstrating that the core set of overlapping genes that emerge from our analyses were also overrepresented in two independent air pollution expression datasets.

The major finding of our analysis is the characterization of a core network of interconnected genes common to all three air pollution measures (Fig. 3) that also form a common gene set that is enriched in two previously published air pollution-expression datasets. This core network consists of genes that are members of fundamental cancer-related pathways, including those related to DNA and metal binding, apoptosis regulation, and cell surface receptor linked signal transduction. Of potential greatest significance was the observation that 10 of the core genes formed connections with all but 2 of the 44 genes making up the network. Among these 10 hubs are 5 – ACP1, HSP90AA1, LEF1, MLH1, and RBM5 - common to the major identified cancer-related pathways. Though these genes have promiscuous function, they are all implicated in tumor pathobiology, providing a potential etiological link between the known associations of chronic air pollution exposure and lung cancer mortality [2, 4], including studies assessing diesel exhaust exposure in the trucking

industry and other populations [6, 23, 24]. For example, acid phosphatase 1 (ACP1), whose expression was strongly correlated with $PM_{2.5}$ exposure in our cohort and was a leading edge member of 13 gene sets enriched across all three exposure types, is a low-molecular-weight protein tyrosine phosphatase (LMW-PTP) with both pro- and anti-oncogenic functions (reviewed in [25]). Two common isoforms have been described, both were correlated with $PM_{2.5}$ exposure in our dataset ($p = 4.7 \times 10^{-4}$ and 9.8×10^{-3}) despite the fact that they have differing patterns of cellular localization - one isoform localizes to the cytoskeleton, the other to the cytoplasm. ACP1 interacts with many proteins implicated in tumor progression, including janus kinase [26], 31 β-catenin [27], and the ephrin A2 receptor (EPHA2, also implicated in our core network, [28]), among others, and ACP1 was one of eight genes whose combined expression in peripheral blood has been suggested as a predictive signature of stage I lung adenocarcinoma [29]. The oncogenic relevance of the other four hub genes – including the heat shock protein HSP90AA1 [30–32], the lymphoid-enhancer binding factor LEF1 [21, 33–35], the DNA mismatch repair gene MLH1 [36–38], and the tumor suppressor RNA-binding motif protein RBM5 [39–41] - have been reviewed elsewhere [30–32, 36–41], further supporting this network as a molecular link between air pollution exposure and lung cancer risk.

Though the most prominent and consistent findings emerging from our analysis delineate network submodules implicated in cancer pathogenesis, the expression signatures that emerged also included a large collection of genes implicated in other diseases, including myocardial and cerebral ischemic injury, sudden cardiac death, and chronic obstructive pulmonary disease, particularly STAT1 [42–44], FZD2 [45, 46], GCLM [47], CD63 [48], and SP4 [49]. Given that all of these genes were members of the core gene expression set (Additional file 2: Table S1), and many were among the most highly connected hub genes, they represent important biological targets in the pathogenesis of these most common pollution-related diseases. Although the study was not designed to assess relationships with clinical disease, we applied a disease connectivity analysis using the "set analyzer" tool in the Comparative Toxicogenomic Database (http://ctdbase.org) and the list of 248 genes identified from our study (Additional file 3: Table S2). Of 20 diseases with Bonferoni-corrected p-values < 0.01, 12 of the were in the categories of cancer, lymphatic disease, and immune system diseases, consistent with our analysis using a gene-expression based network analysis.

Evaluation of strengths of our findings and their contribution to our current understanding of the adverse consequences of air pollution must consider several important strength and limitations relative to prior work. Unlike prior studies, measures of exposure were collected for all study subjects in real-life work settings over a workweek, providing more accurate estimates for analysis. Importantly, the exposure levels experienced by these workers (such as the truck drivers) overlap with ambient exposures experienced by the general public who would be likely to experience similar on-road exposures during such activities as commuting, making our results applicable to a wider population. Our repeated measures design provides for more accurate estimates of gene expression compared to single time-point studies, and is more robust to outliers. In addition, our sampling was performed at 10 sites within the Northeastern United States, providing good representation across the range of pollutant exposures. Lastly, we note the strong evidence of enrichment of our core gene sets in two previously published studies, providing strong evidence of the reproducibility and generalizability of our findings to other populations.

Several limitations, however, must be recognized. First, there are some limitations in the study design. Due to the demographics of employment of the trucking industry in the Northeast, our analysis was restricted to white men. The levels of physical activity during work, which differ between job titles, was not available. Heterogeneity in cell compositions, such as white blood cell counts, could not be addressed due to lack of data. We did examine the white blood cell marker genes based on the *cellmix* R package [50], and found no associations with exposures, either at gene level or gene set (pathway) level. Therefore, we do not believe that cell composition would bias our results. Second, the number of individual involved (63) was small, and the study might be underpowered, even with the repeated design. Third, although we examined the impacts of pollution in an occupational setting, the levels of pollution were low, likely due to decreasing pollution emissions in the trucking industry in recent years [51], and these low levels may explain the relatively modest number of observed changes in gene expression in our study. Therefore, we may not have detected additional genes of importance. Our reliance on GSEA mitigates this concern somewhat, but not completely. Lastly, our analysis relies on a one-week sampling of both exposures and expression measures, with the assumption that these observations are representative of more long-term processes. Though longer, more repeated sampling designs might provide more representative findings, two lines of evidence suggest that the efforts to collect such data would add only incrementally. First we have previously demonstrated in this industry that short windows (one week) of exposure sampling are representative of exposures measured at other time periods [52]. On average, our study subjects have been employed in the same position for 19 years. It is thus likely that the exposure estimates generated from

this study are similar to what would be observed over longer periods of time (months to years). Additionally, we found no evidence of cross-week differential expression, providing some reassurance that the gene expression measures reflect the individual global patterns of gene expression measured over longer time periods. Finally, our finding that our core set of pollution-correlated genes was also prominent in two independent (albeit limited) datasets, suggests that our results may be generalizable and relevant to other exposed populations.

Conclusions

In summary, we have characterized the molecular impact of traffic-related air pollution, and have identified a sub-network of interconnected genes implicated in cancer pathogenesis and related processes that are consistently perturbed in response to air pollution exposure. These data provide greater insights into the adverse health consequences of traffic-related air pollution.

Additional files

Additional file 1: QQ plots after permutation for gene-level analysis. (PDF 4315 kb)

Additional file 2: Table S1. List of the core set of 262 genes from differential expression analysis and GSEA. (DOCX 15 kb)

Additional file 3: Table S2. List of enriched diseases in the Comparative Toxicogenomic Database using the core set of 248 genes from differential expression analysis and GSEA. (DOCX 14 kb)

Abbreviations

BMI: Body mass index; COPD: Chronic obstructive pulmonary disease; DNA: Deoxyribonucleic acid; EC: Elemental carbon; FDR: False discovery rate; GEO: Gene Expression Omnibus; GSEA: Gene set enrichment analysis; OC: Organic carbon; P&D: Pick-up and delivery; PM2.5: Particulate matter ≤ 2.5 microns in diameter; QC: Quality control; RNA: Ribonucleic acid; SD: Standard deviation

Acknowledgments

We thank all subjects for their participation in this study.

Funding

This work is supported by grant R01 ES016284 from the National Institutes of Health. JC is supported by NIH/NHLBI grant K99HL114651.

Availability of data and materials

Microarray expression data described in this paper has been deposited in the NCBI Gene Expression Omnibus (GEO) under accession code GSE83864.

Authors' contributions

JEH, EG and FL conceived and designed the experiments. JEH, EG, BAR and FL collected the samples and performed the experiments. JC, DC and BAR analyzed the data. JC, JEH and BAR wrote the manuscript. All authors have read and approved the manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Participants provided informed consent for biologic sample collection, completion of detailed questionnaires, and analysis of RNA, and were compensated at the end of the protocol for their participation. The protocol was approved by the Institutional Review Board of the Brigham and Women's Hospital and the Human Subjects Committee of the Harvard T.H. Chan School of Public Health.

Author details

[1]Section of Pulmonary, Critical Care and Sleep Medicine, Department of Internal Medicine, Yale University School of Medicine, New Haven, CT, USA. [2]Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [3]Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [4]Pulmonary, Allergy, Sleep, and Critical Care Medicine Section, VA Boston Healthcare System, Boston, MA, USA. [5]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [6]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA, USA.

Received: 31 March 2016 Accepted: 24 October 2016
Published online: 03 November 2016

References

1. Brook RD, Rajagopalan S, Pope 3rd CA, Brook JR, Bhatnagar A, Diez-Roux AV, Holguin F, Hong Y, Luepker RV, Mittleman MA, et al. Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association. Circulation. 2010;121(21):2331–78.
2. Hamra GB, Guha N, Cohen A, Laden F, Raaschou-Nielsen O, Samet JM, Vineis P, Forastiere F, Saldiva P, Yorifuji T, et al. Outdoor particulate matter exposure and lung cancer: a systematic review and meta-analysis. Environ Health Perspect. 2014;122(9):906–11.
3. Hoek G, Krishnan RM, Beelen R, Peters A, Ostro B, Brunekreef B, Kaufman JD. Long-term air pollution exposure and cardio- respiratory mortality: a review. Environ Health. 2013;12(1):43.
4. Loomis D, Grosse Y, Lauby-Secretan B, El Ghissassi F, Bouvard V, Benbrahim-Tallaa L, Guha N, Baan R, Mattock H, Straif K, et al. The carcinogenicity of outdoor air pollution. Lancet Oncol. 2013;14(13):1262–3.
5. Pope 3rd CA, Dockery DW. Health effects of fine particulate air pollution: lines that connect. J Air Waste Manage Assoc. 2006;56(6):709–42.
6. Vermeulen R, Silverman DT, Garshick E, Vlaanderen J, Portengen L, Steenland K. Exposure-response estimates for diesel engine exhaust and lung cancer mortality based on data from three occupational cohorts. Environ Health Perspect. 2014;122(2):172–7.
7. Peretz A, Peck EC, Bammler TK, Beyer RP, Sullivan JH, Trenga CA, Srinouanprachnah S, Farin FM, Kaufman JD. Diesel exhaust inhalation and assessment of peripheral blood mononuclear cell gene transcription effects: an exploratory study of healthy human volunteers. Inhal Toxicol. 2007;19(14):1107–19.
8. Libalova H, Uhlirova K, Klema J, Machala M, Sram RJ, Ciganek M, Topinka J. Global gene expression changes in human embryonic lung fibroblasts induced by organic extracts from respirable air particles. Part Fibre Toxicol. 2012;9:1.
9. Watterson TL, Sorensen J, Martin R, Coulombe Jr RA. Effects of PM2.5 collected from Cache Valley Utah on genes associated with the inflammatory response in human lung cells. J Toxicol Environ Health A. 2007;70(20):1731–44.
10. van Leeuwen DM, Pedersen M, Hendriksen PJ, Boorsma A, van Herwijnen MH, Gottschalk RW, Kirsch-Volders M, Knudsen LE, Sram RJ, Bajak E, et al. Genomic analysis suggests higher susceptibility of children to air pollution. Carcinogenesis. 2008;29(5):977–83.
11. Paul S, Amundson SA. Differential Effect of Active Smoking on Gene Expression in Male and Female Smokers. J Carcinog Mutagen. 2014;5:1000198.
12. Smith TJ, Davis ME, Reaser P, Natkin J, Hart JE, Laden F, Heff A, Garshick E. Overview of particulate exposures in the US trucking industry. J Environ Monit. 2006;8(7):711–20.
13. Casinelli ME, OCP. Elemental Carbon (Diesel Particulate) 5040. In: NIOSH Manual of Analytical Methods. 4th ed. Ohio: National Institute for Occupational Safety and Health; 1998.
14. Tainio M, Tuomisto JT, Hanninen O, Aarnio P, Koistinen KJ, Jantunen MJ, Pekkanen J. Health effects caused by primary fine particulate matter (PM2.5) emitted from buses in the Helsinki metropolitan area, Finland. Risk Anal. 2005;25(1):151–60.

15. Yanosky JD, MacIntosh DL. A comparison of four gravimetric fine particle sampling methods. J Air Waste Manage Assoc. 2001;51(6):878–84.
16. Lin SM, Du P, Huber W, Kibbe WA. Model-based variance-stabilizing transformation for Illumina microarray data. Nucleic Acids Res. 2008;36(2):e11.
17. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.
18. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 2010;38(Web Server issue):W214–20.
19. Hartz AM, Bauer B, Block ML, Hong JS, Miller DS. Diesel exhaust particles induce oxidative stress, proinflammatory signaling, and P-glycoprotein up-regulation at the blood–brain barrier. FASEB J. 2008;22(8):2723–33.
20. Siegel PD, Saxena RK, Saxena QB, Ma JK, Ma JY, Yin XJ, Castranova V, Al-Humadi N, Lewis DM. Effect of diesel exhaust particulate (DEP) on immune responses: contributions of particulate versus organic soluble components. J Toxicol Environ Health A. 2004;67(3):221–31.
21. Xu X, Kherada N, Hong X, Quan C, Zheng L, Wang A, Wold LE, Lippmann M, Chen LC, Rajagopalan S, et al. Diesel exhaust exposure induces angiogenesis. Toxicol Lett. 2009;191(1):57–68.
22. Kimura J, Nguyen ST, Liu H, Taira N, Miki Y, Yoshida K. A functional genome-wide RNAi screen identifies TAF1 as a regulator for apoptosis in response to genotoxic stress. Nucleic Acids Res. 2008;36(16):5250–9.
23. Garshick E, Laden F, Hart JE, Davis ME, Eisen EA, Smith TJ. Lung cancer and elemental carbon exposure in trucking industry workers. Environ Health Perspect. 2012;120(9):1301–6.
24. Garshick E, Laden F, Hart JE, Rosner B, Smith TJ, Dockery DW, Speizer FE. Lung cancer in railroad workers exposed to diesel exhaust. Environ Health Perspect. 2004;112(15):1539–43.
25. Alho I, Costa L, Bicho M, Coelho C. The role of low-molecular-weight protein tyrosine phosphatase (LMW-PTP ACP1) in oncogenesis. Tumour Biol. 2013;34(4):1979–89.
26. Lee JK, Edderkaoui M, Truong P, Ohno I, Jang KT, Berti A, Pandol SJ, Gukovskaya AS. NADPH oxidase promotes pancreatic cancer cell survival via inhibiting JAK2 dephosphorylation by tyrosine phosphatases. Gastroenterology. 2007;133(5):1637–48.
27. Taddei ML, Chiarugi P, Cirri P, Buricchi F, Fiaschi T, Giannoni E, Talini D, Cozzi G, Formigli L, Raugei G, et al. Beta-catenin interacts with low-molecular-weight protein tyrosine phosphatase leading to cadherin-mediated cell-cell adhesion increase. Cancer Res. 2002;62(22):6489–99.
28. Kikawa KD, Vidale DR, Van Etten RL, Kinch MS. Regulation of the EphA2 kinase by the low molecular weight tyrosine phosphatase induces transformation. J Biol Chem. 2002;277(42):39274–9.
29. Rotunno M, Hu N, Su H, Wang C, Goldstein AM, Bergen AW, Consonni D, Pesatori AC, Bertazzi PA, Wacholder S, et al. A gene expression signature from peripheral whole blood for stage I lung adenocarcinoma. Cancer Prev Res. 2011;4(10):1599–608.
30. Shen H, Zhu H, Song M, Tian Y, Huang Y, Zheng H, Cao R, Lin J, Bi Z, Zhong W. A selenosemicarbazone complex with copper efficiently down-regulates the 90-kDa heat shock protein HSP90AA1 and its client proteins in cancer cells. BMC Cancer. 2014;14(1):629.
31. Wang J, Cui S, Zhang X, Wu Y, Tang H. High expression of heat shock protein 90 is associated with tumor aggressiveness and poor prognosis in patients with advanced gastric cancer. PLoS One. 2013;8(4):e62876.
32. Chu SH, Liu YW, Zhang L, Liu B, Li L, Shi JZ, Li L. Regulation of survival and chemoresistance by HSP90AA1 in ovarian cancer SKOV3 cells. Mol Biol Rep. 2013;40(1):1–6.
33. Li TW, Ting JH, Yokoyama NN, Bernstein A, van de Wetering M, Waterman ML. Wnt activation and alternative promoter repression of LEF1 in colon cancer. Mol Cell Biol. 2006;26(14):5284–99.
34. Niemann C, Owens DM, Schettina P, Watt FM. Dual role of inactivating Lef1 mutations in epidermis: tumor promotion and specification of tumor type. Cancer Res. 2007;67(7):2916–21.
35. Gutierrez A, Sanda T, Ma W, Zhang J, Grebliunaite R, Dahlberg S, Neuberg D, Protopopov A, Winter SS, Larson RS, et al. Inactivation of LEF1 in T-cell acute lymphoblastic leukemia. Blood. 2010;115(14):2845–51.
36. Bronner CE, Baker SM, Morrison PT, Warren G, Smith LG, Lescoe MK, Kane M, Earabino C, Lipford J, Lindblom A, et al. Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. Nature. 1994;368(6468):258–61.
37. Papadopoulos N, Nicolaides NC, Wei YF, Ruben SM, Carter KC, Rosen CA, Haseltine WA, Fleischmann RD, Fraser CM, Adams MD, et al. Mutation of a mutL homolog in hereditary colon cancer. Science. 1994;263(5153):1625–9.
38. Suter CM, Martin DI, Ward RL. Germline epimutation of MLH1 in individuals with multiple cancers. Nat Genet. 2004;36(5):497–501.
39. Oh JJ, West AR, Fishbein MC, Slamon DJ. A candidate tumor suppressor gene, H37, from the human lung cancer tumor suppressor locus 3p21.3. Cancer Res. 2002;62(11):3207–13.
40. Zhang L, Zhang Q, Yang Y, Wu C. The RNA recognition motif domains of RBM5 are required for RNA binding and cancer cell proliferation inhibition. Biochem Biophys Res Commun. 2014;444(3):445–50.
41. Fushimi K, Ray P, Kar A, Wang L, Sutherland LC, Wu JY. Up-regulation of the proapoptotic caspase 2 splicing isoform by a candidate tumor suppressor, RBM5. Proc Natl Acad Sci U S A. 2008;105(41):15708–13.
42. Carroll CJ, Sayan BS, Bailey SG, McCormick J, Stephanou A, Latchman DS, Townsend PA. Regulation of myocardial interleukin-6 expression by p53 and STAT1. J Interferon Cytokine Res. 2013;33(9):542–8.
43. McCormick J, Suleman N, Scarabelli TM, Knight RA, Latchman DS, Stephanou A. STAT1 deficiency in the heart protects against myocardial infarction by enhancing autophagy. J Cell Mol Med. 2012;16(2):386–93.
44. Southworth T, Metryka A, Lea S, Farrow S, Plumb J, Singh D. IFN-gamma synergistically enhances LPS signalling in alveolar macrophages from COPD patients and controls by corticosteroid-resistant STAT1 activation. Br J Pharmacol. 2012;166(7):2070–83.
45. Heijink IH, de Bruin HG, van den Berge M, Bennink LJ, Brandenburg SM, Gosens R, van Oosterhout AJ, Postma DS. Role of aberrant WNT signalling in the airway epithelial response to cigarette smoke in chronic obstructive pulmonary disease. Thorax. 2013;68(8):709–16.
46. Laeremans H, Hackeng TM, van Zandvoort MA, Thijssen VL, Janssen BJ, Ottenheijm HC, Smits JF, Blankesteijn WM. Blocking of frizzled signaling with a homologous peptide fragment of wnt3a/wnt5a reduces infarct expansion and prevents the development of heart failure after myocardial infarction. Circulation. 2011;124(15):1626–35.
47. Kobayashi T, Watanabe Y, Saito Y, Fujioka D, Nakamura T, Obata JE, Kitta Y, Yano T, Kawabata K, Watanabe K, et al. Mice lacking the glutamate-cysteine ligase modifier subunit are susceptible to myocardial ischaemia-reperfusion injury. Cardiovasc Res. 2010;85(4):785–95.
48. Cha JK, Jo WS, Shin HC, Bae HR, Ho JM, Kim JW. Increased platelet CD63 and P-selectin expression persist in atherosclerotic ischemic stroke. Platelets. 2004;15(1):3–7.
49. St Amand TR, Lu JT, Zamora M, Gu Y, Stricker J, Hoshijima M, Epstein JA, Ross Jr JJ, Ruiz-Lozano P, Chien KR. Distinct roles of HF-1b/Sp4 in ventricular and neural crest cells lineages affect cardiac conduction system development. Dev Biol. 2006;291(2):208–17.
50. Gaujoux R, Seoighe C. Cell Mix: a comprehensive toolbox for gene expression deconvolution. Bioinformatics. 2013;29(17):2211–2.
51. Davis ME, Hart JE, Laden F, Garshick E, Smith TJ. A retrospective assessment of occupational exposure to elemental carbon in the U.S. trucking industry. Environ Health Perspect. 2011;119(7):997–1002.
52. Davis ME, Laden F, Hart JE, Garshick E, Blicharz A, Smith TJ. Predicting changes in PM exposure over time at U.S. trucking terminals using structural equation modeling techniques. J Occup Environ Hyg. 2009;6(7):396–403.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central