Enhanced exposure assessment and genome-wide DNA methylation in World Trade Center disaster responders

Pei-Fen Kuan[a], Zhongyuan Mi[d], Panos Georgopoulos[d], Dana Hashim[c], Benjamin J. Luft[b] and Paolo Boffetta[c]

DNA methylation has emerged as a promising target linking environmental exposures and cancer. The World Trade Center (WTC) responders sustained exposures to potential carcinogens, resulting in an increased risk of cancer. Previous studies of cancer risk in WTC-exposed responders were limited by the deficiency in quantitative and individual information on exposure to carcinogens. The current study introduces a new exposure-ranking index (ERI) for estimating cancer-related acute and chronic exposures, which aimed to improve the ability of future analyses to estimate cancer risk. An epigenome-wide association study based on DNA methylation and a weighted gene co-expression network analysis were carried out to identify cytosine-phosphate-guanosine (CpG) sites, modules of correlated CpG sites, and biological pathways associated with the new ERI. Methylation was profiled on blood samples using Illumina 450K Beadchip. No significant epigenome-wide association was found for ERI at a false discovery rate of 0.05. Several cancer-related pathways emerged in pathway analyses for the top ranking genes from epigenome-wide association study as well as enriched module from the weighted gene co-expression network analysis. The current study was the first DNA methylation study that aimed to identify methylation signature for cancer-related exposure in the WTC population. No CpG sites survived multiple testings adjustment. However, enriched gene sets involved in cancer, were identified in both acute and chronic ERIs, supporting the view that multiple genes play a role in this complex exposure. *European Journal of Cancer Prevention* **28:225–233**

European Journal of Cancer Prevention 2019, **28:**225–233

Keywords: cancer, environmental exposure, epigenome-wide association study, methylation, pathway analysis, World Trade Center

[a]Department of Applied Mathematics and Statistics, [b]Department of Medicine, Stony Brook University, Stony Brook, [c]Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York City, New York and [d]Environmental and Occupational Health Sciences Institute, Rutgers University, Piscataway, New Jersey, USA

Correspondence to Paolo Boffetta, MD, MPH, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1130, New York NY 10029, USA
Tel: + 1 212 824 7378; fax: + 1 212 849 2566; e-mail: paolo.boffetta@mssm.edu

Received 5 November 2017 **Accepted** 7 June 2018

Introduction

Concern has arisen about the potential for an increased risk of cancer among World Trade Center (WTC) responders, who sustained exposures to a complex mix of toxic chemicals that included multiple known and suspected human carcinogens (Lioy and Gochfeld, 2002). The combustion of jet fuel at high temperatures released soot, metals, benzene and other volatile organic compounds, and strong inorganic acids. The burning and subsequent collapse of the towers resulted in the release of particulate matter comprising asbestos, silica, cement dust, glass fibers, heavy metals like arsenic, beryllium, cadmium, chromium VI, nickel, polycyclic aromatic hydrocarbons, polychlorinated biphenyls, and polychlorinated dibenzofurans and dioxins (Clark *et al.*, 2001; Lioy and Gochfeld, 2002; Edelman *et al.*, 2003; Litten *et al.*, 2003; McGee *et al.*, 2003; Offenberg *et al.*, 2003). Evidence on increased cancer risk are emerging, in which the three cohort studies including WTC responders showed modest elevations in the risk of all cancers combined, with standardized incidence ratio ranging from 1.06 to 1.14 across studies, and substantial overlaps in the 95% confidence intervals (Zeig-Owens *et al.*, 2011; Li *et al.*, 2012; Solan *et al.*, 2013; Boffetta *et al.*, 2016).

The lack of quantitative, individual information on exposure to potential carcinogens is an important limitation of previous analyses of cancer risk in WTC-exposed populations. Two approaches can be considered, to overcome this limitation: an individualized, high-resolution assessment of circumstances of exposure experienced by WTC responders, and the use of biomarkers of exposure. The current study combines these two approaches by introducing a new exposure-ranking index for estimating cancer-related exposure, in an effort to improve the ability of future analyses to elucidate the cancer risk experienced by this population.

The epigenome acts as an interface between the genome and the environment. It is plastic, changing with environmental exposures (Feil and Fraga, 2012) thereby, regulating transcription. It has been suggested that

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website, www.eurjcancerprev.com.

DOI: 10.1097/CEJ.0000000000000460

perturbation in the epigenome in response to the environment is more stable than changes in the transcriptome (Barouki *et al.*, 2012) making epigenetic changes a potentially valuable tool for exposure assessment. One of the most studied epigenetic mechanisms is DNA methylation, a heritable epigenetic modification that does not change the underlying DNA sequence, and is involved in the regulation of gene expression (Plass and Soloway, 2002). The most common methylation site in mammals is a cytosine located next to a guanosine [cytosine-phosphate-guanosine (CpG)]. CpG islands are found mainly in the 5′ regulatory and promoter regions of genes; most are unmethylated in normal cells (Bird *et al.*, 1985).

Epigenetic changes in tumor tissues have been linked to specific environmental exposures. For instance, a gene-specific methylation in lung cancer was associated with tobacco smoking compared with alcohol consumption (Vaissiere *et al.*, 2009). Similarly, DNA methylation patterns were able to distinguish various environmental risk factors associated with hepatocellular carcinoma (Hernandez-Vargas *et al.*, 2010). Recently, specific epigenetic profiles have been identified in the methylome of peripheral blood DNA in patients exposed to a wide variety of environmental exposures including tobacco smoking (Wan *et al.*, 2012), benzene (Ji *et al.*, 2010), air pollution (Tarantini *et al.*, 2009), and arsenic (Bailey *et al.*, 2013). This holds the tremendous promise that these epigenetic changes in the blood may be exploited as biomarkers of exposure. In this study, we identify DNA methylation patterns associated with WTC exposure. The identified DNA methylation signature could serve not only as a biomarker of exposure, but may give us insight into the susceptibility to develop a variety of diseases including cancer.

Materials and methods

Development of WTC 'exposure ranking indices'

WTC 'Exposure Ranking Indices' (WTC-ERIs) were developed for ranking potential exposures to hazardous substances related to post-9/11 activities of responders/workers. The calculation procedures and associated factors for estimating WTC-ERIs were adapted for the present work specifically to utilize information available for responders registered with the WTC General Responder Cohort (WTCGRC) (Dasaro *et al.*, 2015). The exposure-related information in the WTCGRC dataset, collected through the Exposure Activities Questionnaire (EAQ) (Woskie *et al.*, 2011; Solan *et al.*, 2013) was combined with the information developed through various studies that characterized exposure-relevant attributes and factors related to post-WTC activities (Lioy and Gochfeld, 2002; Lioy *et al.*, 2002; Huber *et al.*, 2004; Landrigan *et al.*, 2004; Yiin *et al.*, 2004; Stenchikov *et al.*, 2006). The WTC-ERI is an ordinal metric (Myatt, 2007; Brüggemann and Patil, 2011). WTC-ERI values represent numerical scores derived by a Multi-Criteria Decision Analysis (MCDA) procedure (Linkov and Moberg, 2011) involving factors related to date, duration, location, type of activity, microenvironment, type, and usage of personal protective equipment associated with each exposure event experienced by post-9/11 responders and workers. The index for a specific participant/worker is calculated by summing over the time-sequence of all WTC-related 'exposure events' for that participant/worker (where a typical exposure event is a work shift, though other types of events are possible). The MCDA score calculation procedure for the ERI was implemented primarily in Python 2.7 (*http://www.python.org*) for retrieving and analyzing information from the EAQ data available through the World Trade Center Health Program Data Center, complemented with a set of Matlab codes (*http://www.themathworks.com*), and is summarized in the following equation:

$$\mathrm{ERI}_w = F_{\mathrm{cloud},w} + \sum_{i=1}^{M} f^*_{\mathrm{time},i} \times f_{\mathrm{loc},i} \times f_{\mathrm{act},i} \times p_i \times \Delta t_i + \sum_{j=1}^{N} f^*_{\mathrm{time},j} \times \phi_{\mathrm{loc},j} \times \mathrm{f}_{\mathrm{act},j} \times p_j \times \mu_j \times \Delta\tau_j,$$

where, ERI_w is the ERI for a specific participant (responder or worker) identified by index w (who may be have been involved in different activities during different exposure events); $F_{\mathrm{cloud},w}$ is an exposure-related factor accounting for direct contact of the specific responder or worker with the dust cloud on 9/11/2001; f_{act} is an exposure-related factor accounting for each type of WTC-related activity (search and rescue, cleanup, etc.); $i = 1, \ldots M$ is a counting index for outdoor exposure events during day i, where $i = 1$ on 9/11/2001; $j = 1, \ldots N$ is a counting index for indoor (i.e. confined space) exposure events during day j, where $j = 1$ on 9/11/2001; f^*_{time} is an 'adjusted' exposure-related factor accounting for the time aspects of each exposure event [$f^*_{\mathrm{time}} = \mathrm{function}\ (f_{\mathrm{date},i})$ where $f_{\mathrm{date},i}$ is a factor associated with date relative to 9/11]; f_{loc} and ϕ_{loc} are exposure-related factors accounting for the locations where participant spend the majority of the exposure event (typically, but not exclusively, work shift) μ_j is a microenvironmental adjustment factor reflecting any information, if available, specific to the confined space settings of the exposure event p is a personal protective equipment (PPE) factor, reflecting the type and usage of PPE during the exposure event Δt_i, and $\Delta\tau_j$ is a factor reflecting the duration of exposure event during day i or j.

It should be noted that ERI for each participant is explicitly calculated as the equation below:

$$\mathrm{ERI} = \mathrm{ERI\_a} + \mathrm{ERI\_c},$$

where, ERI_a accounts for a participant's acute exposure (exposed during 9/11/2001 to 9/13/2001, $i, j \leq 3$); ERI_c accounts for a participant's chronic exposure (exposed during 9/14/2001 to 6/30/2002, $i, j > 3$).

Time-related/location-related factors

The values of the exposure factors related to the location, and the time period(s) spent at that location, for each worker/responder are summarized in Supplementary Table S1 (Supplemental digital content 1, *http://links.lww.com/EJCP/A208*).

Dust cloud factor

A special 'location' of particular concern is the area that was covered by the dust cloud on 9/11/2001. So, a worker-specific/responder-specific exposure-related factor ($F_{\text{cloud},w}$) accounting for a direct contact with the dust cloud on 9/11/2001, is used, and a numerical value of 0–500 is assigned for each of six different exposure scenarios (corresponding to situations ranging from no contact with the cloud to 'full immersion' without PPE).

Time-related factors

For the exposure-related factor (f^*_{time}) that accounts for the date, time, and the duration of each exposure event, the following equation is used:

$$f^*_{\text{time}} = \frac{t_{\text{day}}}{t_{\text{posD}}} \times t_{\text{shift}} \times f_{\text{date}},$$

where, t_{day} is the total number of working days during each time period; t_{posD} is the total number of possible working days during each time period derived from the first and the last day (if the first day or the last day is missing, it is assumed that the first day is 9/11/2001 and the last day is 6/30/2002); t_{shift} is the number of work shifts in a day (typically one work shift is 8 h); and f_{date} is a day-specific exposure-related factor accounting for severity of environmental and microenvironmental conditions on the date of each exposure event.

Location-related factors

Exposure-related factors (f_{loc} and ϕ_{loc}) accounting for the location of each exposure event consider differences in environmental and microenvironmental conditions in the general location areas identified on the map of Fig. 1 (Woskie *et al.*, 2011). These factors are assigned numerical values in the range of 1–5 for different WTC work areas identified in the WTCGRC dataset.

Activity-related factors

Exposure-related factors (f_{act}) that account for different types of WTC-related activities during an exposure event are assigned numerical values in the range of 1–10 for the activity categories identified in the WTCGRC. These categories include barge workers, carpenters, dock builders, electricians, glaziers, insulation workers, laborers, mechanics, roofers, truck drivers, military, canteen service, emergency medical technician, fire fighters, morgue workers, police officers, etc. It is important to note here that in cases where an activity type different from that corresponding to the job title is reported in the questionnaire (e.g. 'search and rescue' by a police officer), that activity is used to determine the appropriate factors for each of the activity categories listed in Supplementary Table 2 (Supplemental digital content 1, *http://links.lww.com/EJCP/A208*).

Personal protective equipment-related formulas and factors

Exposure-related factors accounting for PPE type and usage are specific to (i) respirators and (ii) dust masks. A total PPE factor is calculated during each exposure event by averaging of the factors for the PPE types used during the event:

$$p_i = \frac{\sum_{q=1}^{n} p_{i,q}}{n},$$

where, P_i is the total PPE factor, reflecting the type and usage of all types of PPE used during an exposure event i (range:1–0.1); $P_{i,q}$ is the PPE factor, reflecting the type and usage of a specific PPE type q used during the exposure event i; $q = 1, \ldots n$ is an index for different types of PPE.

The only information available for the usage of 'surgical/disposable mask' is the start date, so a constant PPE factor (0.55) for wearing a mask is assigned to the participant after and including their first day of wearing a mask.

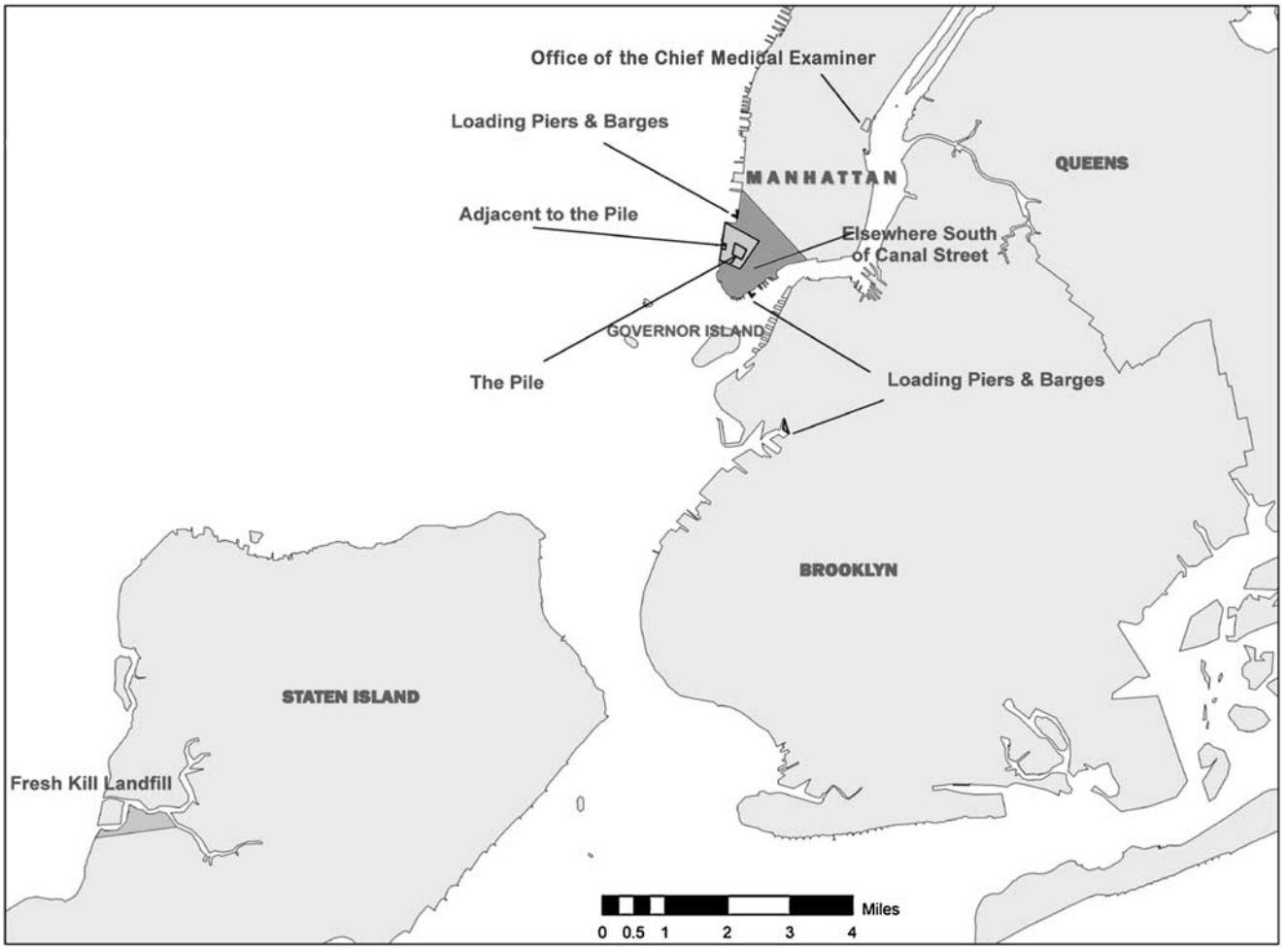
For respirators the following general relation is used:

$$p_{i,q} = 1 - f_{\text{resp\_type},q} \times f_{\text{freq\_G},q} \times \frac{f_{\text{maint\_R},q} + f_{\text{maint\_C},q} + f_{\text{maint\_clean},q}}{3} \times f_{\text{seal},q},$$

where, $f_{\text{resp\_type}}$ is the type of respirator used; $f_{\text{freq\_G}}$ is the frequency of respirator use; $f_{\text{maint\_R}}$ is the frequency of respirator replacement; $f_{\text{maint\_C}}$ is the frequency of respirator maintenance (cartridge replacement); $f_{\text{maint\_clean}}$ is the frequency of respirator cleaning; f_{seal} is the frequency of respirator seal check. A numerical value is assigned for each aspect of PPE (these values are listed in Supplementary Table S3, Supplemental digital content 1, *http://links.lww.com/EJCP/A208*), with the highest value indicating best protection, most frequent usage, and correct usage of PPE. The lowest value indicates worst or no protection, most infrequent usage, and the incorrect usage of the respirator. Information on the type of respirator (full face or half face) is only available for the first week (9/11/2001 to 9/18/2001); an average between full and half face respirator factors is used for days after 9/18/2001.

It should be noted that the quality of exposure-relevant information in the WTCGRC database varied considerably. Therefore, based on the evaluation of data gaps, uncertainties, and resolution for the records of each participant, these participant were classified in groups

Fig. 1



World Trade Center response and cleanup workforce locations (Woskie *et al.*, 2011).

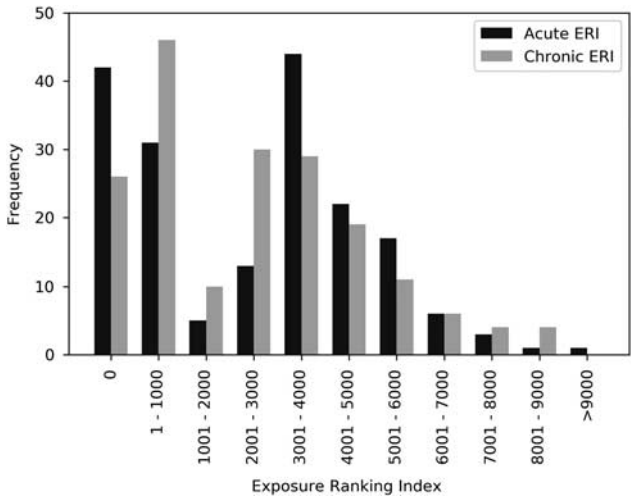
A–E, with group A including the participants with the most complete and unambiguous exposure data and group E including those with the highest uncertainties (see details in Supplementary Methods and in Supplementary Table S4, Supplemental digital content 1, *http://links.lww.com/EJCP/A208*, on the classification criteria and the numbers of participants assigned to each of the groups).

Participants

Participants were recruited through the Stony Brook WTC-Health Program, part of a consortium of Clinical Centers of Excellence in the New York metropolitan area established in 2002 to monitor and treat WTC-related conditions in responders to the WTC disaster (Herbert *et al.*, 2006). Enrollees with documented WTC experience were enlisted from extensive outreach efforts involving partnerships with volunteer organizations, labor unions, and public outlets. The current study was approved annually by the Committees on Research Involving Human Subjects at Stony Brook. University (IRB number: 604113). Written informed consent was obtained.

The EAQ data for 6110 participants in the Stony Brook database, who are part of the larger WTCGRC, were processed using the MCDA classification and scoring system developed for WTC-ERI. Sufficient and consistent information for developing 'unambiguous' estimates of ERI values was available for 2625 of these participants that were assigned to quality group A, whereas another 1390 participants with less complete information – though sufficient to develop rankings – were assigned to groups B and C (Supplementary Methods, Supplemental digital content 1, *http://links.lww.com/EJCP/A208*). The distribution of acute and chronic ERI values for all Stony Brook participants in group A is shown in Fig. 2 (not including zero acute ERI values for 632 participants). The ERI reflecting overall exposure of responders from post-WTC activities was calculated as the sum of the acute and chronic ERI values. High and

Fig. 2



Distribution of ERI_a (acute ERI) and ERI_c (chronic ERI) values for 503 Stony Brook participants with complete and consistent exposure-relevant information in the Exposure Activities Questionnaire (EAQ) component of the WTC-HPDC dataset. ERI, exposure-ranking index; WTC, World Trade Center.

low ERIs were defined as ERI of up to 356 and ERI of at least 5599, where the values 356 and 5599 were the top and bottom 10 percentile, respectively (Solan *et al.*, 2013).

A subset of 185 responders with ERI values in groups A–C, assessed between February 2012 and March 2014 were included in DNA methylation profiling. All participants provided blood samples for the epigenetics assays. Inclusion criteria were signed consent, sufficient English language skills to participate in a diagnostic interview, and male sex. We included only males because females show notably different methylation patterns from males, and very few responders were females. Participants were 51.3 years of age on average at the time of blood drawing, predominantly White (83.2%) and nonsmokers (95.7%) (Table 1).

Illumina infinium human methylation 450K beadchip

Blood samples were obtained from each participant by venipuncture and sent to the Roswell Park Cancer Institute for DNA extraction. Genomic DNA was isolated from 0.3 ml of whole blood using the Qiagen BioRobot Universal System and the QIAamp DNA blood BioRobot MDx Kit (Qiagen, Valencia, California, USA) following the manufacturer's recommended protocol. DNA methylation profiling was performed by the Roswell Park Cancer Institute using the Human Methylation 450K BeadChip (Illumina Inc., San Diego, California, USA). DNA extraction and methylation profiling were done blinded to group assignment. Overall, 500 ng of high quality genomic DNA measured by picogreen quantitation (Life technologies, Grand Island, New York, USA) was bisulphite converted, amplified, fragmented, and hybridized to the Illumina Infinium Human Methylation 450K Beadchip using standard Illumina protocol (Illumina Inc.). Data were processed using Illumina's GenomeStudio methylation module (v1.9.0) (Illumina Inc.).

Table 1 Clinical characteristics

ERI	Low ($n=69$)	High ($n=116$)	P value
Age [mean (SD)]	54.3 (8.1)	49.5 (6.3)	< 0.001
Race [*n* (%)]			
White	61 (0.88)	93 (0.80)	0.213
Other	8 (0.12)	23 (0.20)	
Smoker [*n* (%)]			
Yes	3 (0.04)	5 (0.04)	0.999
No	66 (0.96)	111 (0.96)	

The P values were computed from t-test (for age) and χ^2-test (for race and smoking status) comparing high to low ERI.
ERI, exposure-ranking index.

Data preprocessing and normalization

The 450K BeadChip methylation data from the GenomeStudio were imported into R (*http://cran.r-project.org*). Preprocessing of methylation data at the 485 557 CpG sites were performed as follows. CpG sites with detection P value of more than 0.001 are set to missing and CpG sites with more than 20% missing were filtered. β-Mixture quantile normalization (Teschendorff *et al.*, 2013) was applied to the β values for correction of bias due to the type I and type II probes. Nonspecific, cross-hybridized CpG sites (Chen *et al.*, 2013; Price *et al.*, 2013), CpG sites overlapping with a SNP, and CpG sites mapping to repeat regions were filtered. The final data consisted of 375 223 CpG sites.

Estimation of blood cell type proportions

Cell type proportions have been implicated in DNA methylation analysis of whole blood samples (Houseman *et al.*, 2015). The proportions of CD8T, CD4T, natural killer, B cell, monocytes, and granulocytes were estimated using the R packages minfi and FlowSorted. Blood.450 (Aryee *et al.*, 2014) on the basis of the procedures described in Houseman *et al.* (2012). We normalized the sum of the proportions per sample to one, and included five out of six estimated cell types as adjustment factor in our epigenome-wide association study (EWAS). In addition, two-sample t-tests were used to assess the association between each cell type and ERI.

Statistical method for epigenome-wide association study

To identify CpG sites associated with each phenotype, separate multple linear regression for each CpG site was first fitted on logit transformed β values $[\log(\beta/(1-\beta))]$ as response, and ERI, adjusting for age, race, smoking status, and cell types. Statistical significance for CpG association with ERI was assessed by the Wald test. A false

discovery rate (FDR) (Benjamini and Hochberg, 1995) was used to account for multiple testings.

Pathway and gene ontology analyses

Pathway and gene ontology analyses were carried out using gometh function in Bioconductor package missMethyl (Phipson *et al.*, 2016). As the number of CpG sites mapping to each gene varied in the Methylation 450K BeadChip, pathway and gene ontology analyses would be biased and inaccurate (Geeleher *et al.*, 2013). Gometh accounted for the varying number of CpG sites per gene by providing a previous probability for each gene based on gene length, followed by a modified hypergeometric test for over-representation of a gene set (Young *et al.*, 2010). We tested for over-representation among the top 500 CpG sites from EWAS, against the background list of 375 223 CpG sites. In all, 290 KEGG pathways (minimum and maximum number of genes for each gene set were 15 and 500, respectively) were tested. Gene sets significant at FDR of less than 0.05 were reported.

Weighted gene co-expression network analysis

The weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath, 2008) was used to identify modules of correlated CpG sites on the logit transformed β values. The Pearson correlation matrix was raised to the 10th power to achieve scale-free topology. The minimum module size was set as 30, and the cut-offs for splitting and merging modules were 2 and 0.25, respectively. The methylation profiles for each module were represented by the eigenCpG. The association between module eigenCpG and ERI was performed using a two-sample *t*-test, and the *P* values were corrected by FDR. Over-representation analysis described above was also carried out on the CpG sites within the identified modules from WGCNA.

Results

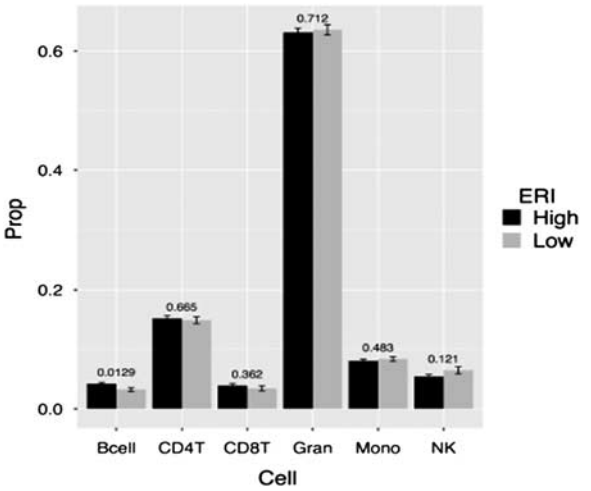
Cell type proportions

Figure 3 displayed the association between the estimated cell type proportions and ERI. At nominal *P* value 0.05, B cell was higher in high ERI compared with low ERI group.

Epigenome-wide association analysis

The volcano plots in the top panel of Supplementary Fig. S1 (Supplemental digital content 1, *http://links.lww.com/EJCP/A208*) showed an approximately equal amount of hypermethylation and hypomethylation patterns comparing high to low ERI. EWAS with ERI did not identify statistically significant CpG sites at an FDR of 0.05. Two CpG sites were significant at a nominal *P* value of 0.0001. Both CpG sites were hypermethylated and mapped to 3′-UTR of *GABRA4* and gene body of *TUBB*, respectively. Top 10 CpG sites were provided in Supplementary Table S5 (Supplemental digital content 1, *http://links.lww.com/EJCP/A208*).

Fig. 3



Barplots of estimated cell type proportions. Each barplot depicts the mean value (*y*-axis) and one SE. *P* values on each cell type were computed from two-sample *t*-tests. ERI, exposure-ranking index; Gran, granulocyte; Mono, monocyte; NK, naturak killer cell.

Pathway and gene ontology analyses

Table 2 displayed the enriched KEGG pathways at FDR 0.05 for the top 500 CpG sites from EWAS with ERI. Twenty-one KEGG pathways were found to be enriched among the top 500 CpG sites associated with ERI, including several cancer-related pathways, that is, *PPAR*, *MAPK*, Ras, and *PI3K–Akt* signaling pathways.

Weighted gene co-expression network analysis

Overall, 109 modules were identified from the WGCNA analysis. The eigenCpG of two modules (M1 and M2) were associated with ERI at nominal *P* value of less than 0.05, but did not survive FDR control. M1 and M2 contained 388 and 59 CpG sites, respectively. Twenty-two KEGG pathways were identified to be enriched among the 388 CpGs sites from M1 (Table 3), including B-cell-receptor signaling pathway, hematopoietic cell lineage, primary immunodeficiency and cytokine–cytokine receptor interaction, chemokine signaling pathway, and several other cancer-related pathways. No statistically significant pathway was identified for M2 module.

Discussion

While it might still be too early to see the full effect of WTC exposure on cancer incidence (Boffetta *et al.*, 2016), it is important to assess the presence in this population of biomarkers that may be related to cancer risk. The current study introduced a new exposure index for estimating cancer-related exposure in the WTC cohort. The ERI allowed ranking quantitatively WTCGRC members in terms of their acute and chronic WTC exposure, and provided a refinement of previous approaches on the basis of qualitative information (Wisnivesky *et al.*, 2011; Solan *et al.*, 2013).

Table 2 List of significant KEGG pathways (FDR < 0.05) for ERI among the top 500 CpG sites

Pathways	*n*	DE	FDR	Gene
Neuroactive ligand-receptor interaction	270	10	4.76E−05	CHRND, ADRA1D, GABRA2, GABRA4, GALR1, MCHR1, GRIN2A, GRM7, HTR5A, PTGFR
Human papillomavirus infection	309	10	0.00172	COL1A1, IFNAR2, NOTCH4, HES6, CCND1, MPP5, THBS1, TSC1, WNT3A, CREB3L1
Thiamine metabolism	15	3	0.00305	AK4, AK5, TPK1
Calcium signaling pathway	180	7	0.00438	ADRA1D, GRIN2A, HTR5A, PTGFR, CACNA1H, CACNA1G, MCU
PPAR signaling pathway	72	4	0.00753	FABP5, ACSL3, SLC27A6, ANGPTL4
Amyotrophic lateral sclerosis	51	4	0.00753	GRIN2A, NEFM, NEFH, SOD1
Phagosome	143	5	0.0104	ATP6V0E2, TUBB, TAP2, THBS1, COLEC11
Synaptic vesicle cycle	63	4	0.0104	CPLX2, CPLX1, ATP6V0E2, STX1A
PI3K–Akt signaling pathway	325	8	0.0108	COL1A1, ANGPT1, IFNAR2, CCND1, THBS1, TSC1, CREB3L1, FGF19
Proteoglycans in cancer	203	6	0.032	GPC1, IHH, MIR21, CCND1, THBS1, WNT3A
Ras signaling pathway	224	6	0.0346	RGL1, ANGPT1, GRIN2A, MAPK10, RASAL2, FGF19
Nicotine addiction	40	3	0.0377	GABRA2, GABRA4, GRIN2A
Pathways in cancer	394	8	0.038	CTBP2, CTNNA2, MAPK10, RARA, RARB, CCND1, WNT3A, FGF19
Choline metabolism in cancer	98	4	0.0389	DGKG, MAPK10, SLC22A5, TSC1
MAPK signaling pathway	254	6	0.0486	MAPT, ECSIT, MAPK10, CACNA1H, CACNA1G, FGF19
Hedgehog signaling pathway	46	3	0.0486	ADRBK2, IHH, CCND1
Hippo signaling pathway	153	5	0.0486	YAP1, CTNNA2, CCND1, MPP5, WNT3A
Circadian entrainment	96	4	0.0486	GRIN2A, PRKG1, CACNA1H, CACNA1G
Ubiquitin mediated proteolysis	136	4	0.0492	UBE4B, ERCC8, PARK2, UBE2B
Platelet activation	123	4	0.0492	COL1A1, PRKG1, P2RY12, FERMT3
Taste transduction	79	3	0.0492	HCN4, GABRA2, GABRA4

DE, number of genes in our list overlapping with the pathway; ERI, exposure-ranking index; FDR, false discovery rate; gene, list of overlapping genes; *n*, size of the pathway.

Table 3 List of significant KEGG pathways (FDR < 0.05) for M1 module

Pathways	*n*	DE	FDR	Gene
B-cell-receptor signaling pathway	70	9	0	BLNK, INPP5D, NFATC1, NFKBIE, VAV2, CD19, CD22, CD79B, CD81
Hematopoietic cell lineage	90	6	1.29E-05	FCER2, ITGB3, KIT, IL1R2, CD19, CD22
Primary immunodeficiency	36	5	1.29E-05	TNFRSF13B, BLNK, JAK3, RAG1, CD19
Cytokine–cytokine receptor interaction	246	7	0.000149	CCR6, TNFRSF13B, IL10RA, KIT, CXCR5, TGFB1, IL1R2
Chemokine signaling pathway	179	7	0.000243	CCR6, ADRBK1, GNG3, GRK6, JAK3, CXCR5, VAV2
Epstein–Barr virus infection	196	7	0.000425	FCER2, IL10RA, JAK3, NFKBIE, USP7, CD19, HDAC4
Regulation of actin cytoskeleton	211	7	0.00158	DIAPH1, CYFIP1, CYFIP2, ITGB3, SSH3, VAV2, ARHGEF7
mTOR signaling pathway	150	6	0.00179	FLCN, LRP5, TBC1D7, RPTOR, RPS6KA2, SLC7A5
Osteoclast differentiation	121	5	0.00179	LILRA2, BLNK, ITGB3, NFATC1, TGFB1
Transcriptional misregulation in cancer	179	6	0.00521	PTCRA, PAX5, CDK14, TCF3, IL1R2, LDB1
Cholinergic synapse	112	5	0.00708	PIK3R5, GNG3, KCNQ1, CAMK2A, KCNQ4
PI3K–Akt signaling pathway	325	7	0.0128	PIK3R5, GNG3, ITGB3, JAK3, KIT, RPTOR, CD19
Th17 cell differentiation	105	4	0.0155	JAK3, NFATC1, NFKBIE, TGFB1
cAMP signaling pathway	198	5	0.0291	ADORA2A, NFATC1, VAV2, CAMK2A, ACOX3
Morphine addiction	91	4	0.0291	ADRBK1, GABRD, GNG3, GRK6
FoxO signaling pathway	129	4	0.0341	PRKAG2, RAG1, TGFB1, USP7
Tuberculosis	163	4	0.0346	IL10RA, TGFB1, CAMK2A, KSR1
Apelin signaling pathway	136	4	0.0426	PIK3R5, GNG3, PRKAG2, HDAC4
Proteoglycans in cancer	203	5	0.0426	ANK1, ITGB3, TGFB1, VAV2, CAMK2A
Inositol phosphate metabolism	73	3	0.0448	INPP5J, INPP5D, ITPKB
SNARE interactions in vesicular transport	33	2	0.0457	VAMP1, STX7
MicroRNAs in cancer	281	5	0.0457	FOXP1, ITGB3, RPTOR, ST14, HDAC4

DE, number of genes in our list overlapping with the pathway; ERI, exposure-ranking index; FDR, false discovery rate; gene, list of overlapping genes; *n*, size of the pathway.

A limitation of the present application of the ERI framework is that the available exposure data were not sufficient to attribute separate exposure to specific agents that contributed to the mixture of WTC contaminants. Although certain measurements for specific agents at various locations are available, the resolution of exposure-relevant time/location data for the study participants was too 'coarse' to allow more specific assessments and in any case strong correlations between the agents could have hampered any such effort.

This was the first DNA methylation study that aimed to identify methylation signature for cancer-related exposure in the WTC population. Although no CpG site was significant after adjustment for multiple testings, pathway analysis revealed enriched gene sets involved in cancer. Specifically, several cancer-related KEGG pathways were identified among the top ranking CpG sites associated with ERI and CpG sites within enriched module from WGCNA, including PPAR, Ras, MAPK, PI3K–Akt, mTOR, and chemokine signaling, as well as pathways, proteoglycans, choline metabolism, and microRNAs in cancer; although different overlapping genes were implicated. Ras proteins are involved in cellular signal transduction and the *Ras* genes (HRas, KRas and NRas) are the most common oncogenes in cancer (Fernandez-Medarde and Santos, 2011). In contrast,

MAPK pathways are involved in stress signaling, and abberant MAPK pathways lead to uncontrolled growth and tumerigenesis (Dhillon *et al*., 2007). Akt has been shown to be the hub of signaling pathway implicated in tumorigenesis (Fresno Vara *et al*., 2004). These pathways are involved in the major hallmarks of cancer, including cell cycle, survival, motility, and genomic instability (Hanahan and Weinberg, 2011). The identified cancer-related enriched gene sets support the importance of pathway analysis compared with single-gene approach, that is, the search for the combined effect of multiple genes acting in concert in this complex exposure.

Limitations

The current study had several strengths, including the use of an enhanced exposure assessment approach and the first and the largest EWAS sample to date in environmental exposure study. Nonetheless, our findings must be considered in the context of several limitations. First, the methylation data were profiled on whole blood collected more than 10 years post-9/11 attack. Our analysis was carried out under the assumption that DNA methylation changes from high exposure are persistent over time, as methylation is known to be a stable marker (Riggs, 1989). Second, unexposed control group was not available in this study. We sought to address this issue by selecting extreme exposure groups within our populations, that is, top 10% versus bottom 10%. It is important to note that a comparison with an external population might be affected by selection bias and confounding, which are less likely to occur in internal comparisons within a relatively homogeneous population. Third, the sample size of our study is relatively small for EWAS. Fourth, our methylation analysis was performed in DNA samples derived from whole blood cells and was thus a mix of cell types. We controlled for the cell types heterogeneity using the state-of-the-art statistical method, but future work needs to isolate and examine each cell type individually.

Conclusion

The current study aimed to provide a better understanding of the relationship between epigenetic alteration and WTC-related exposure, with potential relevance to cancer risk. Enriched gene sets were involved in several biological pathways associated with cancer, including Ras, MAPK, and PI3K–Akt signaling pathways. Taken together, this provides biological evidence supporting a possible association between exposure and risk of cancer among the WTC responders.

Acknowledgements

The authors gratefully acknowledge the support of the first responders and rescue/recovery workers for participating in this study. They also thank the staff of the Stony Brook World Trade Center Health Program and the World Trade Center Health Program Data Monitoring Center for ongoing support. They are also grateful to Sean Glenn at the Roswell Park Cancer Institute for his contributions to the methylation assessment. The findings and conclusions in this article are those of the authors and do not represent the official position of NIOSH, the CDC, or the US Public Health Service.

The analysis was supported by grants nos 200-2011 41815 (PI: Paolo Boffetta), U01 OH010987 (PI: Paolo Boffetta), U01 OH010416-01 (PI: Benjamin J. Luft) from the National Institute for Occupational Safety and Health (NIOSH)/ Centers for Disease Control and Prevention (CDC) and P30 ES005022 from NIH-NIEHS (ZY. Mi and P. Georgopoulos).

Conflicts of interest

There are no conflicts of interest.

References

Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of infinium DNA methylation microarrays. *Bioinform* **30**:1363–1369.

Bailey KA, Wu MC, Ward WO, Smeester L, Rager JE, García-Vargas G, *et al.* (2013). Arsenic and the epigenome: interindividual differences in arsenic metabolism related to distinct patterns of DNA methylation. *J Biochem Mol Toxicol* **27**:106–115.

Barouki R, Gluckman PD, Grandjean P, Hanson M, Heindel JJ (2012). Developmental origins of non-communicable disease: implications for research and public health. *Environ Health* **11**:42.

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* **57**:289–300.

Bird A, Taggart M, Frommer M, Miller OJ, Macleod D (1985). A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **40**:91–99.

Boffetta P, Zeig-Owens R, Wallenstein S, Li J, Brackbill R, Cone J, *et al.* (2016). Cancer in World Trade Center responders: findings from multiple cohorts and options for future study. *Am J Ind Med* **59**:96–105.

Brüggemann R, Patil GP (2011). *Ranking and prioritization for multi-indicator systems: Introduction to partial order applications*. New York, NY: Springer Science & Business Media.

Chen Y-A, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, *et al.* (2013). Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**:203–209.

Clark RN, Green RO, Swayze GA, Meeker G, Sutley S, Hoefen TM, *et al.* (2001). *Environmental studies of the World Trade Center area after the September 11, 2001 attack*. Hoboken, NJ: USGS.

Dasaro CR, Holden WL, Berman KD, Crane MA, Kaplan JR, Lucchini RG, *et al.* (2015). Cohort Profile: World Trade Center Health Program General Responder Cohort. *Int J Epidemiol* **46**:e9.

Dhillon AS, Hagan S, Rath O, Kolch W (2007). MAP kinase signalling pathways in cancer. *Oncogene* **26**:3279–3290.

Edelman P, Osterloh J, Pirkle J, Caudill SP, Grainger J, Jones R, *et al.* (2003). Biomonitoring of chemical exposure among New York City firefighters responding to the World Trade Center fire and collapse. *Environ Health Perspect* **111**:1906.

Feil R, Fraga MF (2012). Epigenetics and the environment: emerging patterns and implications. *Nat Rev Genet* **13**:97–109.

Fernandez-Medarde A, Santos E (2011). Ras in cancer and developmental diseases. *Genes Cancer* **2**:344–358.

Fresno Vara JA, Casado E, de Castro J, Cejas P, Belda-Iniesta C, Gonzalez-Baron M (2004). PI3K/Akt signalling pathway and cancer. *Cancer Treat Rev* **30**:193–204.

Geeleher P, Hartnett L, Egan LJ, Golden A, Raja Ali RA, Seoighe C (2013). Gene-set analysis is severely biased when applied to genome-wide methylation data. *Bioinformatics* **29**:1851–1857.

Hanahan D, Weinberg RA (2011). Hallmarks of cancer: the next generation. *Cell* **144**:646–674.

Herbert R, Moline J, Skloot G, Metzger K, Baron S, Luft B, *et al.* (2006). The World Trade Center disaster and the health of workers: five-year assessment of a unique medical screening program. *Environ Health Perspect* **114**:1853–1858.

Hernandez-Vargas H, Lambert M-P, Le Calvez-Kelm F, Gouysse G, McKay-Chopin S, Tavtigian SV, *et al.* (2010). Hepatocellular carcinoma displays distinct DNA methylation signatures with potential as clinical predictors. *PLoS ONE* **5**:e9749.

Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, *et al.* (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**:1.

Houseman EA, Kim S, Kelsey KT, Wiencke JK (2015). DNA methylation in whole blood: uses and challenges. *Curr Environ Health Rep* **2**:145–154.

Huber A, Georgopoulos P, Gilliam R, Stenchikov G, Wang S-W, Kelly B, *et al.* (2004). Modeling air pollution from the collapse of the world trade center and assessing the potential impacts on human exposures. *Environ Manager* **2004**:35–40.

Ji Z, Zhang L, Peng V, Ren X, McHale C, Smith M (2010). A comparison of the cytogenetic alterations and global DNA hypomethylation induced by the benzene metabolite, hydroquinone, with those induced by melphalan and etoposide. *Leukemia* **24**:986–991.

Landrigan PJ, Lioy PJ, Thurston G, Berkowitz G, Chen LC, Chillrud SN, *et al.* (2004). Health and environmental consequences of the world trade center disaster. *Environ Health Perspect* **112**:731–739.

Langfelder P, Horvath S (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**:559.

Li J, Cone JE, Kahn AR, Brackbill RM, Farfel MR, Greene CM, *et al.* (2012). Association between World Trade Center exposure and excess cancer risk. *JAMA* **308**:2479–2488.

Linkov I, Moberg E (2011). *Multi-criteria decision analysis: environmental applications and case studies*. Boca Raton, FL: CRC Press.

Lioy PJ, Gochfeld M (2002). Lessons learned on environmental, occupational, and residential exposures from the attack on the World Trade Center. *Am J Ind Med* **42**:560–565.

Lioy PJ, Weisel CP, Millette JR, Eisenreich S, Vallero D, Offenberg J, *et al.* (2002). Characterization of the dust/smoke aerosol that settled east of the World Trade Center (WTC) in lower Manhattan after the collapse of the WTC 11 September 2001. *Environ Health Perspect* **110**:703.

Litten S, McChesney DJ, Hamilton MC, Fowler B (2003). Destruction of the World Trade Center and PCBs, PBDEs, PCDD/Fs, PBDD/Fs, and chlorinated biphenylenes in water, sediment, and sewage sludge. *Environ Sci Technol* **37**:5502–5510.

McGee JK, Chen LC, Cohen MD, Chee GR, Prophete CM, Haykal-Coates N, *et al.* (2003). Chemical analysis of World Trade Center fine particulate matter for use in toxicologic assessment. *Environ Health Perspect* **111**:972.

Myatt GJ (2007). *Making sense of data: a practical guide to exploratory data analysis and data mining*. Hoboken, NJ: John Wiley & Sons.

Offenberg J, Eisenreich S, Chen L, Cohen M, Chee G, Prophete C, *et al.* (2003). Persistent organic pollutants in the dusts that settled across lower Manhattan after September 11, 2001. *Environ Sci Technol* **37**:502–508.

Phipson B, Maksimovic J, Oshlack A (2016). missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics* **32**:286–288.

Plass C, Soloway PD (2002). DNA methylation, imprinting and cancer. *Eur J Hum Genet* **10**:6.

Price EM, Cotton AM, Lam LL, Farré P, Emberly E, Brown CJ, *et al.* (2013). Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin* **6**:1.

Riggs AD (1989). DNA methylation and cell memory. *Cell Biochem Biophys* **15**:1–13.

Solan S, Wallenstein S, Shapiro M, Teitelbaum SL, Stevenson L, Kochman A, *et al.* (2013). Cancer incidence in world trade center rescue and recovery workers, 2001–2008. *Environ Health Perspect* **121**:699–704.

Stenchikov G, Lahoti N, Diner DJ, Kahn R, Lioy PJ, Georgopoulos PG (2006). Multiscale plume transport from the collapse of the World Trade Center on September 11, 2001. *Environ Fluid Mech* **6**:425–450.

Tarantini L, Bonzini M, Apostoli P, Pegoraro V, Bollati V, Marinelli B, *et al.* (2009). Effects of particulate matter on genomic DNA methylation content and iNOS promoter methylation. *Environ Health Perspect* **117**:217.

Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, *et al.* (2013). A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29**:189–196.

Vaissiere T, Hung RJ, Zaridze D, Moukeria A, Cuenin C, Fasolo V, *et al.* (2009). Quantitative analysis of DNA methylation profiles in lung cancer identifies aberrant DNA methylation of specific genes and its association with gender and cancer risk factors. *Cancer Res* **69**:243–252.

Wan ES, Qiu W, Baccarelli A, Carey VJ, Bacherman H, Rennard SI, *et al.* (2012). Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. *Hum Mol Genet* **21**:3073–3082.

Wisnivesky JP, Teitelbaum SL, Todd AC, Boffetta P, Crane M, Crowley L, *et al.* (2011). Persistence of multiple illnesses in World Trade Center rescue and recovery workers: a cohort study. *Lancet* **378**:888–897.

Woskie SR, Kim H, Freund A, Stevenson L, Park BY, Baron S, *et al.* (2011). World Trade Center disaster: assessment of responder occupations, work locations, and job tasks. *Am J Ind Med* **54**:681–695.

Yiin L-M, Millette JR, Vette A, Ilacqua V, Quan C, Gorczynski J, *et al.* (2004). Comparisons of the dust/smoke particulate that settled inside the surrounding buildings and outside on the streets of southern New York City after the collapse of the World Trade Center, September 11, 2001. *J Air Waste Manage Assoc* **54**:515–528.

Young MD, Wakefield MJ, Smyth GK, Oshlack A (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* **11**:R14.

Zeig-Owens R, Webber MP, Hall CB, Schwartz T, Jaber N, Weakley J, *et al.* (2011). Early assessment of cancer outcomes in New York City firefighters after the 9/11 attacks: an observational cohort study. *Lancet* **378**:898–905.