SCIENTIFIC REPORTS

OPEN

Classification and Regression Tree Approach for Prediction of Potential Hazards of Urban Airborne Bacteria during Asian Dust Events

Received: 18 January 2018
Accepted: 16 July 2018
Published online: 07 August 2018

Keunje Yoo[1,2], Hyunji Yoo[1], Jae Min Lee[3], Sudheer Kumar Shukla[4] & Joonhong Park[1]

Despite progress in monitoring and modeling Asian dust (AD) events, real-time public hazard prediction based on biological evidence during AD events remains a challenge. Herein, both a classification and regression tree (CART) and multiple linear regression (MLR) were applied to assess the applicability of prediction for potential urban airborne bacterial hazards during AD events using metagenomic analysis and real-time qPCR. In the present work, *Bacillus cereus* was screened as a potential pathogenic candidate and positively correlated with PM_{10} concentration ($p < 0.05$). Additionally, detection of the *bceT* gene with qPCR, which codes for an enterotoxin in *B. cereus*, was significantly increased during AD events ($p < 0.05$). The CART approach more successfully predicted potential airborne bacterial hazards with a relatively high coefficient of determination (R^2) and small bias, with the smallest root mean square error (RMSE) and mean absolute error (MAE) compared to the MLR approach. Regression tree analyses from the CART model showed that the PM_{10} concentration, from 78.4 μg/m^3 to 92.2 μg/m^3, is an important atmospheric parameter that significantly affects the potential airborne bacterial hazard during AD events. The results show that the CART approach may be useful to effectively derive a predictive understanding of potential airborne bacterial hazards during AD events and thus has a possible for improving decision-making tools for environmental policies associated with air pollution and public health.

Asian dust (AD) events, global dust transport events, have increased over the last 20 years due to global climate change and desertification[1–3]. East Asia is a major source region of global wind-blown dust aerosols. In spring and winter, dust uplifted from arid Asian areas is transported to northern China, Korea, Japan, and even as far as the western United States[1,2]. AD events are becoming less predictable due to an increase in the fraction of unanticipated dust particles derived from the newly formed deserts in western China and Mongolia[1,2]. Most previous studies have suggested that AD events result in increased occurrences of human diseases and environmental problems[1,2,4,5]. Therefore, AD events are recognized as a major social/environmental/clinical issue, with growing concern in East Asia[1].

Although biological agents in AD have received scant attention compared with physiochemical attributes, there is increasing evidence that exposure to bioaerosols during AD events may cause adverse health effects and severe diseases when pathogenic bacteria are involved[2,6,7]. To investigate their effects on public health during AD events, an appropriate methodology must define potential pathogens and employ an effective monitoring system[8,9]; however, there is sparse information on urban airborne bacterial communities[2,9]. Next-generation sequencing (NGS) can offer insights into the diversity and composition of airborne culturable and non-culturable

[1]Department of Civil and Environmental Engineering, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul, 03722, South Korea. [2]Department of Earth and Environmental Engineering, Columbia University, New York, NY, 10027, USA. [3]Department of Earth System Sciences, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul, 03722, South Korea. [4]Department of Built and Natural Environment, Caledonian College of Engineering, Seeb, Sultanate of Oman. Correspondence and requests for materials should be addressed to J.P. (email: parkj@yonsei.ac.kr)

Atmosphere environment parameters	AD events	Non-AD events	*p* value
PM_{10} (μg/m^3)	178 ± 97	66 ± 25	<0.001
Temperature (°C)	12.9 ± 5.9	16.8 ± 10.3	
Relative humidity (%)	42.2 ± 10.2	55.8 ± 12.9	
Wind speed (m/s)	3.1 ± 0.6	2.8 ± 1.1	
Duration of sunshine (hr)	6.0 ± 1.8	8.0 ± 2.2	
Evaporation (mm)	3.7 ± 2.4	3.1 ± 1.7	
Surface temperature (°C)	15.5 ± 7.1	17.5 ± 10.9	
Airborne bacterial parameters	**AD events**	**Non-AD events**	***p* value**
Bacterial abundance (copy numbers/m^3)	6.05E + 07 ± 1.00E + 06	3.22E + 05 ± 1.37E + 04	<0.001
Bacterial diversity (Shannon index)	4.21 ± 0.63	2.87 ± 0.41	
Relative abundance of potential pathogenic bacteria (%)	0.97 ± 0.32	0.55 ± 0.18	<0.05
Relative abundance of *B. cereus* group (%)	0.62 ± 0.18	0.19 ± 0.16	<0.05
bceT gene abundance (copy numbers/m^3)	4.27E + 04 ± 3.15E + 03	2.26E + 03 ± 2.44E + 02	<0.05

Table 1. Statistical summary of the data for the atmospheric environmental parameters and 730 airborne bacterial parameters between AD events (n = 10) and non-AD events (n = 45). The *p* values were calculated with t-test in SAS v. 9.2.

bacteria[7,10]. Research suggests that 16S rRNA gene-based NGS can successfully determine the abundance and diversity of potentially pathogenic bacteria for screening purposes in activated sludge, biosolids, drinking water, and soil[11–14].

Identification of pathogens in bioaerosols requires long-term monitoring, and assessing bioaerosol risks to human health is time-consuming and costly. Instead, current real-time atmospheric environmental parameters are not only closely related to the occurrence of AD events but are also relatively faster and easier to analyze than detecting and assessing potential pathogens during AD events[15]. Therefore, modeling that depends on statistical analysis could be an alternative approach for exploring the relationship between airborne bacterial communities and atmospheric environmental conditions[16]. If certain relationships can be found between them it will then be possible to predict potential hazards one or two days in advance and more effectively protect public health[17,18]. Most importantly, reliable short-term prediction of potential airborne bacterial hazards may assist the authorities in managing atmospheric environmental policy for AD events. Despite the extensive research on physiochemical modeling studies during AD events[19,20], no specific research has so far been carried out to predict biological hazards during AD events.

Multiple linear regression (MLR) is one of the widely used statistical tools for finding an appropriate mathematical model and for determining the best-fitting coefficients of a model from the given data[16,18]. MLR generally provides good predictive capability in environmental studies, such as air quality prediction models[16,18], and can provide reasonable interpretation between dependent and predictor variables by statistical tests[21]. Machine learning and rule induction is a powerful statistical method for collecting, summarizing, and analyzing data from different perspectives into valuable and practical information to identify useful relationships[22,23]. As a representative machine learning method, the classification and regression tree (CART) has considerable advantages, including that it is nonparametric and is suitable for nonlinear structures and that it may be appropriate for solving complex, dynamic environmental problems from a small dataset[22,24]. Rule induction employed in CART can be used to find key rules on the basis of interactions between independent and dependent variables[22,23]. CART approaches have been used in environmental forecasting research to estimate urban air quality[18], determine groundwater pollution vulnerability[24], predict *in situ* dechlorination potential[25], predict water quality from wastewater treatment plants[26], assess microbial source tracking[27], and predict heavy metal sorption to soil[28]. Therefore, CART and MLR models could support decision-making and effective management of potential urban airborne bacterial hazards during AD events. However, no detailed comparison of the model performance has yet to be evaluated.

The aims of this study are to (1) compare the predictive abilities between MLR and CART approach for assessing potential airborne bacterial hazards during AD events, and (2) identify key atmospheric environmental parameters that significantly influence potential airborne bacterial hazards during AD events.

Results

Characterization of Atmospheric Parameters between AD and Non-AD Events. The average PM_{10} concentration of AD events was 178 μg/m^3, which was significantly (t-test, $p < 0.001$) higher, by 112 μg/m^3, than that of non-AD events (Table 1). Seasonal monitoring revealed that airborne bacterial abundance with PM_{10} concentrations was more than 10- to 50-fold higher during AD events, and non-AD events did not affect airborne bacterial abundance. Although studies[5,6] have indicated that atmospheric indicators such as temperature and relative humidity exhibit relatively high correlations during AD events, our monitoring results revealed no significant difference between AD and non-AD events. The parameters of the other air masses (e.g., wind speed, sunshine, evaporation, and surface temperature) displayed no differences between AD and non-AD events (Table 1).

Characteristics of Bacterial Communities between AD and Non-AD Events. The abundance of airborne bacteria was determined by qPCR, targeting the 16S rRNA gene in samples collected during the three study years. The 16 S rRNA gene copy numbers ranged from 4.85×10^3 to 2.58×10^8 gene copies/m^3. During AD events,

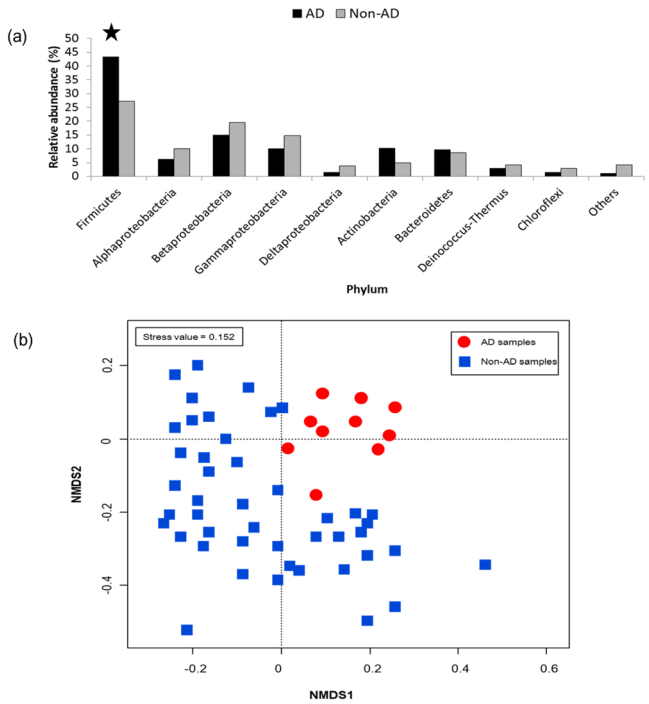

Figure 1. Relative abundance of airborne bacterial community structures between AD events and non-AD events (**a**) and non-metric multidimensional scaling (NMDS) ordination at the phylum level (**b**). Others indicate minor genus members with relative abundances <1.00%. *$p < 0.05$ (t-test in SAS v. 9.2).

the gene copy numbers (mean: 6.05×10^7 gene copies/m^3, Stdev: 1.00×10^6) increased remarkably compared to the non-AD (mean: 3.22×10^5 gene copies/m^3, Stdev: 1.37×10^4) levels ($p < 0.001$) (Table 1). Additionally, the bacterial 16 S rRNA gene copy numbers tended to correlate positively with PM_{10} concentration (Supplementary Fig. S1a). As indicated by the Shannon index (H′) values, airborne bacterial diversity significantly increased during AD events (Supplementary Fig. S1b). The increased airborne bacterial diversity during AD events and correlation with dust parameters suggest that dust events increase local airborne bacterial diversity.

AD and non-AD events were characterized by different bacterial taxa (Fig. 1). *Firmicutes* significantly increased with those for the non-AD events ($p < 0.05$) and composed the most dominant bacterial group during AD events (Fig. 1a). According to the NMDS plot, airborne bacterial structures of the AD samples were clustered together and separated from those of non-AD samples (Fig. 1b), indicating that AD events caused a significant shift in microbial community structures.

These results imply that although the nature of aerosol bacterial populations is variable, most airborne bacteria during AD events may be associated with particle size and air environmental conditions. A significant correlation between bacterial diversity and PM_{10} abundance during AD events suggested that desert dust might be the source of airborne bacteria[29]. According to the backward trajectory analysis (Supplementary Fig. S2), air masses during AD events contained microorganisms originating from the Gobi Desert that passed over China and the Yellow Sea to Seoul. However, air masses from non-AD events contained microorganisms transported from various directions near Korea. These results may support that the shift in airborne bacterial communities between AD and non-AD events is affected by the source of airborne bacteria and transport pathways (Supplementary Fig. S2).

Screening of Potential Pathogenic Bacteria Candidates. The sequences obtained using pyrosequencing were extracted by alignment with reference sequences, and all sequences were assigned at the species level (Supplementary Table S1). Potential pathogenic bacteria belonging to *Bacillus, Neisseria, Pseudomonas, Clostridium, Shigella, Acinetobacter, Ralstonia*, and *Staphylococcus* were detected in non-AD samples (Fig. 2), suggestive of the potential presence of bacterial hazards in urban bioaerosol environments, even though the 16 S rRNA gene sequence is limited in its ability to accurately determine pathogenicity[13,30]. The relative abundance of potential pathogenic bacteria candidates increased significantly during AD events and was positively correlated with PM_{10} concentration (Supplementary Fig. S1c). Compared with non-AD samples, significantly higher *Bacillus* (a potential pathogenic candidate) was detected in AD samples. In particular, *B. cereus* and *B. licheniformis* significantly increased ($p < 0.05$), suggestive of their potential as AD-specific bacterial pathogen candidates (Fig. 2). Although *B. licheniformis* was identified as an AD-specific candidate pathogen, the primer information on its pathogenic gene is insufficient for quantitative examination. Conversely, however, sufficient primer information of the pathogenic gene for *B. cereus* has been established previously. Therefore, we selected *B. cereus* as the AD-specific candidate pathogen.

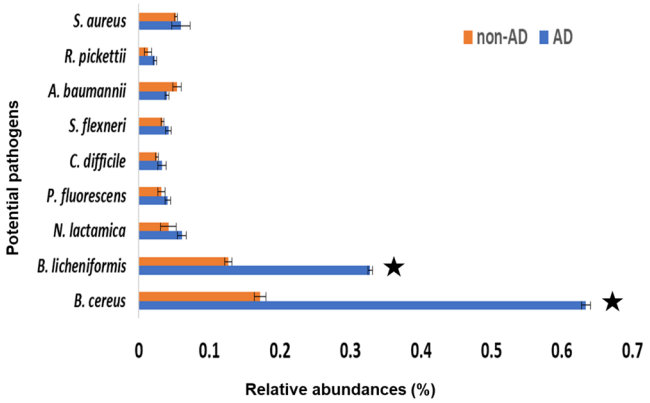

Figure 2. Relative abundance of potential pathogenic bacteria candidates among the total 16S rRNA gene sequence reads from the Pyrosequencing. * indicates $p < 0.05$ from t-test in SAS v.9.2.

Target	Subset		Performance Indexes		
			RMSE	MAE	R^2
Bacterial abundance (16S rRNA gene copies)	MLR	Training	8.67	7.43	0.76
		Test	15.7	12.2	0.68
	CART	Training	6.48	4.04	0.81
		Test	10.2	8.02	0.70
Bacteria diversity (Shannon index)	MLR	Training	15.4	10.7	0.65
		Test	23.3	15.9	0.58
	CART	Training	8.14	5.25	0.78
		Test	13.2	10.8	0.66
Relative abundance of potential pathogenic bacteria	MLR	Training	14.2	12.3	0.72
		Test	22.8	17.2	0.61
	CART	Training	9.01	5.87	0.78
		Test	14.4	10.4	0.71
Relative abundance of *B. cereus*	MLR	Training	18.4	12.6	0.70
		Test	26.1	19.2	0.58
	CART	Training	7.80	4.79	0.82
		Test	11.3	7.25	0.77
bceT gene abundance	MLR	Training	16.4	10.3	0.66
		Test	23.5	16.1	0.54
	CART	Training	8.48	6.04	0.78
		Test	12.3	9.07	0.75

Table 2. Performance indicators for the developed predictive MLR and CART models.

The abundance of *bceT* gene copy numbers ranged from 3.27×10^4 to 1.15×10^5 gene copies/m^3 during AD events (Table 1). *BceT* gene copy numbers exhibited a similar trend as the relative abundance of potential pathogenic bacteria (Supplementary Fig. S1c) and were significantly higher during AD events ($p < 0.05$).

Assessment of Prediction Performance for AD Events. After demonstrating that airborne bacterial parameters, in particular bacterial hazards, increased significantly ($p < 0.05$) during AD events, we used AD-specific airborne bacterial parameters to evaluate whether the MLR and CART models could achieve good performance in reflecting AD event characteristics. According to the performance indexes, the CART approaches outperformed the MLR approaches (Table 2). Most airborne bacterial parameters yielded good correlations between predicted and real-time measured values in the CART model (Table 2). The estimates of the relative abundance of potential pathogenic bacteria, *B. cereus* populations, and *bceT* gene abundance for AD events displayed relatively good fits ($R^2 = 0.71–0.77$) with the least bias and smallest RMSE (11.3–14.4) and MAE (7.25, 10.4) in the test set results (Table 2). CART and rule induction effectively reproduced variations in airborne bacterial parameters using on-site measurement data, in particular the relative abundance of *B. cereus* populations and *bceT* gene abundance during AD events (Table 2).

Identification of Important Variables Associated with Airborne Bacterial Parameters. The CART and rule induction method has outstanding advantages in terms of identifying independent variables that

may significantly influence its dependent variables and in providing rule induction between the independent and dependent variables[23].

To induct a rule between the atmospheric environmental input variables and target variables (airborne bacterial parameters), we performed a CART-based tree analysis. The final regression trees generated by rule induction with the airborne bacterial parameters for each child node of this tree in the training dataset were shown (Fig. 3, Supplementary Fig. S3). With respect to the independent variables, the first split of the tree was defined as the PM_{10} subject (Fig. 3a). Fourteen datasets were clustered with PM_{10} concentrations $\geq$78.4 μg/m^3, and the remaining twenty-four datasets were clustered with PM_{10} concentrations <78.4 μg/m^3. Higher PM_{10} subjects were segregated based on the temperature subject (Fig. 3a). Figure 3b was constructed for the relative abundance of *B. cereus* as predictors. The first split of the tree was defined with respect to the PM_{10} subject, and the nodes were segregated with relative humidity and temperature as the subject (Fig. 3b). All figures can be interpreted in the same way (Fig. 3, Supplementary Fig. S3). A relative importance ranking of individual parameters for airborne bacterial hazards was possible (Supplementary Table S2). PM_{10}, relative humidity, and temperature took precedence over the other parameters and were deemed essential parameters for predicting the airborne bacterial hazard potential.

Discussion

Recently, the East Asian region's climatic conditions such as scarce rains and droughts have boosted the persistence of atmospheric bioaerosols[1]. Therefore, it is important to integrate this process into air quality modeling systems intended for air quality planning and assessment in order to assess impacts on human health[31] and ecosystems[32]. Although it is recognized that dust particles contain pathogens, in most cases the potential hazards or risks associated with them is still largely unclear[2]. The pathogenic bacteria effect of dust inhalation can be attributed to the direct physical action of dust particles, and may be exacerbated by the toxic effects of biologically active compounds[33]. Although prediction accuracy was overall good as shown our study (Table 2), regression models such as MLR have certain limitations. For example, it is relatively difficult to reflect non-linear conditions, and multi-collinearity between independent and dependent variables usually causes MLR to be inefficient[32]. Motivated by knowledge of these limitations, we applied the CART and rule induction method to predict potential hazards of urban airborne bacteria during AD events. This CART and rule induction approach successfully evaluated the prediction performance between observed, real-time measurable atmosphere environmental parameters and airborne bacterial parameters from NGS-based screening and targeted toxin genes from qPCR results. These results could be because the training datasets fit relatively well, reflecting the relationships between airborne bacterial parameters and atmospheric environmental parameters. From these results, we suggest that the correlations between airborne bacterial parameters and atmospheric environmental parameters during AD events are an approximately good fit with the CART and rule induction method for predicting the potential bacterial hazard in urban areas. Although the 16S rRNA gene sequence has been restricted to identifying the taxonomic resolution of bacterial pathogens[13,30], combining high-throughput sequencing and qPCR results can provide relatively high resolution[34]. Because metagenomic approaches could be used to screen potential pathogens in AD samples, the identified potential pathogens subsequently could be quantified by using qPCR, which targets the potential pathogens using their biomarkers[34].

During AD events, biological concentrations significantly increase with PM_{10} concentrations, with differences in bacterial community structure. The high correlation of bacterial abundances with PM_{10} during the AD events (Table 1, Supplementary Fig. S1) and backward trajectory results (Supplementary Fig. S2) in this study indicate that desert dust might be the source of airborne bacteria. However, there were not significant changes during non-AD events. These results indicate that the high concentration of bacteria during AD events was due to the large increase of the concentration of soil-originated particles which contained higher bacterial concentration[1–3]. The airborne bacteria from AD events may have mixed with indigenous airborne bacterial communities before reaching our sampling point, having traveled through industrial, agricultural, and urban areas[5,6]. As such, the suspended particle composition (e.g., PM_{10}) may have been affected due to the addition of local pollutants and physicochemical changes in the atmospheric environment during transport; therefore, the frequency of potential pathogenic bacteria may have increased during AD events, which could affect ecosystem and human health. PM_{10} always segregated the first split of the tree, while temperature, relative humidity, and evaporation were important in predicting the airborne bacterial parameters in the rule induction (Fig. 3, Supplementary Fig. S3). PM_{10} is well established as an indicator of heavy air pollution, based on physical and chemical results and clinical evidence[35]. There is mounting evidence of the negative effects of bioaerosols associated with PM_{10} on ecosystems and human health[36,37]. However, the correlation between airborne bacterial parameters, including potential pathogens, and PM_{10} in urban areas during AD events is not well understood.

From our results, high PM_{10} concentrations were significantly correlated with potential pathogen indicators during AD events (Table 1, Supplementary Fig. S1). When the training datasets were constructed to predict bacterial abundance and diversity in the CART model, most PM_{10} concentrations were segregated into two split nodes between 65.3 and 70.8 μg/m^3 (Supplementary Fig. S3). Meanwhile, the relative abundances of potential pathogens, *B. cereus*, and the *bceT* gene were segregated into higher PM_{10} concentrations (78.4 to 92.2 μg/m^3) than bacterial abundance and diversity (Fig. 3), suggesting that the relative abundances of potential pathogens, *B. cereus*, and *bceT* gene were more significantly affected by PM_{10} concentrations and AD events than seasonal changes and local environmental effects. Our results revealed PM_{10} concentrations between 78.4 and 92.2 μg/m^3 during AD events, indicative of a relatively high risk. PM_{10} prediction has attracted special legislative and scientific attention due to its negative effects on human health[38]. Since these results could offer AD-specific bacteria or relative environmental parameters for the implementation of a robust biosurveillance network, current air pollution policy may be further improved by taking into consideration the potential of biological hazards during AD events.

Airborne bacteria growth is affected by relative humidity and temperature[39]. Temperatures above 24 °C decrease airborne bacterial survival[39], while relative humidity of 70–80% has a protective effect on aerosolized

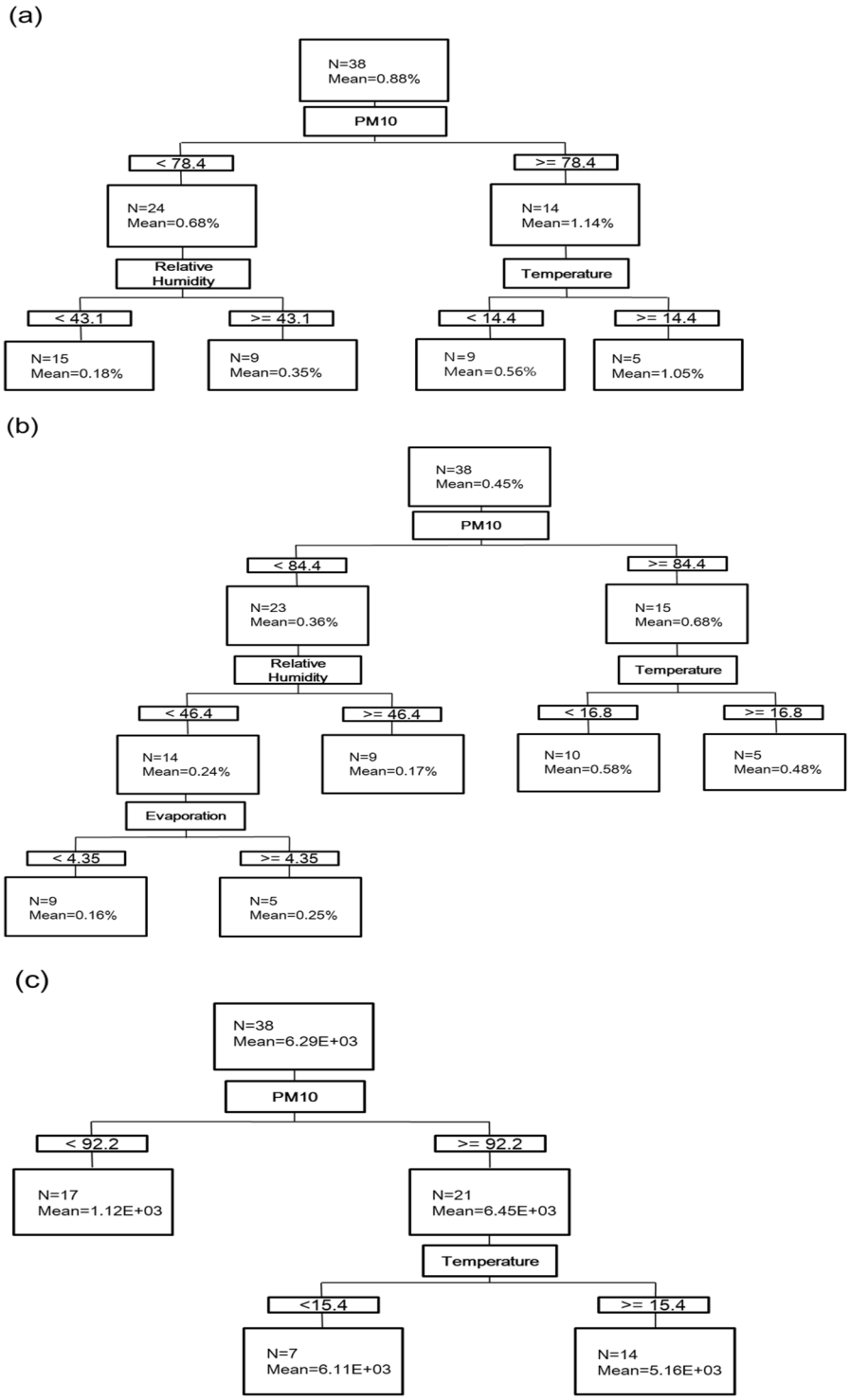

Figure 3. Determination of the relative importance of the predictor variables in the CART model for prediction of relative abundance of potential pathogens (**a**) and *B. cereus* (**b**), and *bceT* gene abundance (**c**) by binary regression tree analysis.

bacteria[40,41]. The temperature during most AD events (13–17 °C) may have supported airborne bacteria survival; however, the relative humidity (40–50%) may have adversely affected survival. The CART approach reflected the characteristics of these heterogeneous atmospheric conditions during AD events better than descriptive statistics, and successfully identified key atmospheric parameters associated with AD events and airborne bacteria. Thus, although aerosol bacterial populations are variable, the airborne bacteria community during AD events might be associated with specific atmospheric conditions.

Endospore-forming bacteria (e.g., *Bacillus*) have been isolated from inter-continentally transported dust[2,42,43]. These high-tolerance bacteria could survive during long-range dispersal and be efficiently transported by atmospheric dust[1,2], shielded from inactivation by ultraviolet light and low relative humidity by attaching to crevasses within coarse particles. The trajectories pathway (Supplementary Fig. S2) is also considered to represent a protective mode that allows for the survival of *B. cereus* in hostile environments. Numerous fungal, bacterial, and viral species have been found in desert dust samples[2,42]. Endotoxins and other biologic compounds in $PM_{10\text{-}2.5}$ from dust storms can activate inflammatory responses[44,45]. For example, in North Carolina ambient $PM_{10\text{-}2.5}$ exacerbated allergic response to airborne bacteria[44], and in six European cities the $PM_{10\text{-}2.5}$ fraction triggered the highest inflammatory effect[45].

The correlation between bacterial abundance and particulate matter in the air is likely a result of the dependence of bacteria on coarse particles (e.g., PM_{10}) rather than on fine particles (e.g., $PM_{2.5}$)[46]. Thus, molecular airborne bacteria community data with PM_{10} characteristics is rational to investigate the distribution and changes in airborne bacterial communities during AD events by resolving genetic diversity and populations. There are two reasons for excluding the possibility of a correlation between airborne bacterial communities and $PM_{2.5}$. First, a large amount of $PM_{2.5}$ are basically produced via homogeneous processes in the atmosphere, with no direct association with pre-existing particles[47]. Second, the suggested correlation is potentially wrong, since coarse and fine particles are not significantly correlated, according to the *Murata and Zhang*[46] study. There are usually primary particles among $PM_{2.5}$ such as fine particles, and the increase of coarse particles such as PM_{10} is commonly accompanied with an increase in fine particles in East Asia. This is supported by the dependence of airborne bacteria on dust particles[5,43].

This study quantified the independent effects of different PM_{10} fractions, included a large distribution of complete differences among PM_{10} concentrations on case and control days, which provided acceptable statistical significance to detect relative high or low significant effects, with minimizing misclassification. Although machine learning and rule induction from small data sets makes the modeling procedure difficult and prone to overfitting, there are many situations in which organizations must work with small data sets in environmental analysis[48]. Thus, it is worthwhile to start developing appropriate forecasting models with smaller variance of forecasting error and good accuracy based on small data sets. To avoid overfitting due to the use of the small data set, k-fold cross-validation and random sampling alternatively can be used in the CART model[23,49]. Previous studies reported that k-fold cross-validation and random sampling are useful when no test sample is available and the learning sample is too small to have the test sample removed from it[49,50]. Although we tried to decrease error and biased predictors, relatively small-sized training and test data still can result in overfitting or misclassifications in this study. Therefore, further validation of our results is needed. Because recent studies have suggested that resampling and virtual data generation significantly improved predictive accuracy[48,51], resampling and virtual data generation can be considered as an alternative method to improve problems inherent within small data sets. Additionally, if a sufficiently large dataset were obtained to further test the feasibility of this approach, the concepts outlined in this study could have potentially broad applications in real-time forecasts. Our concept can be potentially useful for further designing the spatial distribution of monitoring networks to protect public health during AD events. In addition, it could provide a scientific reference for the policy maker in developing future policies.

Material and Methods

Bioaerosol Sample Collection. We collected 55 air samples from 2011 to 2013 in Seodaemun-gu of Seoul, Korea, of which 16 were from the rooftop of the Seoul Air Monitoring Station in Bulgwang (37°61′31″N, 126°93′01″E) in 2011, and 39 were from the rooftop of the 3rd Engineering building of Yonsei University in Shinchon (37°33′42″N, 126°56′07″E) in 2012 and 2013. These sites are located about 10 km from each other in an urban area characterized by human activities without industrial complexes. All air samples were collected 20–30 m above the ground. Ten AD events occurred in Seoul, Korea in 2011 and 2013. All data were separated into AD (ten samples) and non-AD (45 samples) events based on the "Asian Dust Occurrence Reports" from the National Institute of Environmental Research (NIER), Korea.

Bioaerosol samples were collected with a high-volume air sampler (Thermo Scientific, MA, USA). Samples were collected for 24 h at air flow rates of 300–500 L/min on 8 × 10-in. track-etched polycarbonate membrane filters (0.2 μm pore size; Whatman, GE, USA). The filters were autoclaved before sampling, and the filter holder in sampling apparatus was cleaned with 70% ethanol before each sampling event to avoid microbial contamination. After sampling, each filter was stored at −20 °C before DNA extraction.

DNA Extraction from Bioaerosol Samples. Genomic DNA was extracted using a Fast DNA spin for Soil Kit (MPBiomedicals, OH, USA) following a previous method[52], with slight modifications[15]. A negative control was included with every set of DNA extractions. These negative controls were treated exactly the same as all the samples through the entire experiment process, including amplification and sequencing. The extracted DNA samples were stored at −20 °C until use.

Total Bacterial and bceT gene Quantification in Bioaerosol Samples. The total numbers of bacterial 16S rRNA genes copied from each bioaerosol sample were measured using qPCR with an iQ5 Real-Time PCR Detection System (Bio-Rad, CA, USA). The total reaction volume was 20 μL, containing 1× SYBR Master Mix (Bio-Rad), primer sets (300 nM each), and 10-fold-diluted template DNA. The primers targeting bacterial 16S rRNA gene and *bceT* gene have been described previously[53,54]. Because *bceT* is the pathogenic gene in *B. cereus*, and usually causes illness through the production of enterotoxin[55], we used it to quantitatively examine the presence of potential pathogenic bacteria. A total of 1×10^1 to 1×10^7 copies/reaction of PCR products of *Escherichia coli* W3110 and *Bacillus cereus* strain KACC 11240 were used as the standard DNA template to

generate a standard curve to quantify the 16S rRNA and *bceT* genes. For 16S rRNA gene, the thermal cycling conditions were followed as: 94 °C for 10 min, followed by 40 cycles at 94 °C for 15 s and 60 °C for 60 s. For *bceT* gene, the thermal cycling conditions were followed as: 95 °C for 5 min, followed by 37 cycles of 95 °C for 10 s and 60 °C for 45 s. Gene copy numbers (per m^3) were calculated as described previously[56]. For the qPCR run of each sample, triplicate reactions were performed with positive and negative controls. Melting curve analysis (Tm) was performed for 1 cycle of 95 °C for 15 s, 1 cycle of 60 °C for 20 s and 1 cycle from 60 °C to 95 °C for 20 min.

NGS Targeting Bacterial 16S rRNA Gene in Bioaerosol Microbial Communities. In this study, 454 FLX pyrosequencing was used to characterize microbial communities between AD and non-AD events. To provide PCR amplicons for the pyrosequencing, 563 F/16 (5′-AYTGGGYDTAAAGNG-3′) and BSR926/20 (5′-CCGTCAATTYYTTTRAGTTT-3′) targeting V4-V5 regions of 16S rRNA gene were amplified as described previously[57]. Forward primers included pyrosequencing adapter sequences and 8-bp barcode to distinguish each sample in the pool of amplicons[15]. PCR was conducted with a C1000TM Thermal Cycler (Bio-Rad) as follows: 3 min for 94 °C, followed by 35 cycles of 94 °C for 1 min, 55 °C for 30 s, 72 °C for 1 min, and a final extension at 72 °C for 5 min[15]. Negative controls consisting of the same process were included in each PCR run. Amplicons were pooled at equal concentrations using a NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA), and PCR purification was performed using the MinElute PCR Purification Kit (Qiagen, CA, USA). Pyrosequencing was performed on a 454 GS-FLX Titanium Instrument (Roche, NJ, USA) at Macrogen (Seoul, Korea).

Quality control and taxonomic analysis of the 16S rRNA gene sequence reads were performed with Mothur package v.1.30 according to Schloss' SOP[58]. All sequencing analysis process was performed following our previous work[15]. The obtained sequences were separated according to the barcodes, and quality filtering was performed using the Flowgram filtering method. Low-quality sequences with more than one mismatch to the barcode, two mismatches to the primer, or ambiguous nucleotides, negative controls were discarded. Sequences were removed if the homopolymers were longer than 8 bps and/or sequences were shorter than 300 bps[59]. UCHIME was used to remove expected chimeras derived from PCR using chimera.uchime from Mothur[60]. To remove or reduce PCR amplification and sequencing errors, sequences were denoised using the shhh.seqs command in AmpliconNoise in Mothur[61]. After quality filtering, sequences were aligned with the SILVA reference database using the NAST algorithm[58,62], and similar sequences ($\geq$97% similarity) were clustered into operational taxonomic units (OTUs). Sequences were assigned to phylotypes using the RDP classifier[63]. Non-metric multidimensional scaling (NMDS) was performed using the vegan package in R to visualize the taxonomic structure differences between AD and non-AD samples. The data were based on the Bray–Curtis dissimilarity measure of the binary matrix information of 55 air samples.

To screen for human pathogenic bacteria sequence candidates, representative 16S rRNA gene sequences of the bacterial genera OTUs were matched with the reference list of 16S rRNA gene sequences for known human pathogenic bacteria (Supplementary Table S1) from existing databases and studies[11,13,64] using BLAST (blastn, cut-off identity $\geq$97%)[65], and the first-cut screened sequences were matched again (identity >97%) using EzTaxon[66] to identify bacterial 16S rRNA gene sequences similar to those of known pathogenic isolates.

Characteristics of Atmosphere Environmental Parameters. Daily atmospheric environmental parameter measurements were obtained from the NIER, Korea (http://www.airkorea.or.kr/) using fully automated and daily measurements of atmospheric environmental parameters (e.g., PM_{10}, temperature, relative humidity, wind speed, duration of sunshine, evaporation, and surface temperature). Available atmospheric environmental parameter data were extracted from the NIER daily, and averaged over the sampling time. Where data were missing for particular atmospheric environmental parameters on a given day, the values from the remaining data were used to compute the average. Daily information was provided by the Korea Meteorological Administration (KMA) (http://web.kma.go.kr/eng/index.jsp). Descriptive statistics were calculated for each parameter using SAS v.9.2 (SAS Institute Inc., USA).

Data Processing of Multiple Linear Regression and CART. Multiple linear regression (MLR) is one of the most widely used methodologies for modeling the dependence of a dependent variable on several independent variables[17]. In general, a linear regression model assumes that (a) the error term has a normal distribution with a mean of 0, (b) the variance of the error term is constant across cases and independent of the variables in the model and (c) the value of the error term for a given case is independent of the values of the variable in the model and of the values of the error term for other cases.

MLR is one of the modeling techniques to investigate the relationship between a dependent variable and several independent variables[17,18]. In the MLR model, the error term denoted by ε is assumed to be normally distributed with mean 0 and variance σ^2 (which is a constant). ε is also assumed to be uncorrelated. Thus, the regression model can be written as[17]:

$$y = b_0 + \sum_{i=1}^{n} b_i x_i + \varepsilon \quad (1)$$

where b_i are the regression coefficients, x_i are independent variables and ε is stochastic error associated with the regression. To estimate the value of the parameters, the least squares method was used.

CART is a nonparametric statistical technique developed by *Breiman et al.*[23] that can solve classification and regression problems for categorical and continuous dependent variables. One notable advantage is that the models are scalable to large problems and small datasets[23]. CART is constructed by subsets of a dataset using all predictor variables to repeatedly create two child nodes beginning with the entire dataset[23], and uses a stepwise

method to establish splitting rules[23]. Although there are seven single variable splitting criteria, the Gini index is the default method, and it usually performs best[23].

We included seven properties (PM_{10}, temperature, relative humidity, wind speed, duration of sunshine, evaporation, and surface temperature) as independent variables and five properties (bacterial abundance, bacterial diversity, relative abundance of potential pathogenic bacteria, *B. cereus*, and *bceT* gene) as dependent variables in MLR and CART model. In CART, the Gini index was used to determine the dataset. To evaluate model performance, we partitioned the data into training (70% of the dataset for each class) and testing (remaining 30% of the entire dataset) datasets. The training dataset was used to find an optimal value from one or more predictors during the CART model construction. The testing dataset was used to evaluate the optimal value by verifying the prediction accuracy of the dependent variables. We used the SAS for the MLR model learning and SAS Enterprise Miner v.9.2 (SAS Inc.) for the CART model learning. Ten-fold cross-validation was used to avoid model over-fitting[23,67]. In this study, the data randomly broke into ten different parts. We used nine of these parts to train the model and the remaining part to test the model performance. We repeated these nine more times, using each of the ten parts as testing data. Then, we averaged the accuracy of the model in classifying the testing samples over each of the ten datasets to obtain a measure for the accuracy of MLR and CART.

Model Performance Criteria. We evaluated the performance of the constructed MLR and CART model statistically, using the root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2)[18] to evaluate the MLR and CART model performance between the dependent variables and predicted values of the response. Each performance criteria term indicates specific information regarding the predictive performance efficiency[18]. RMSE is a quadratic scoring rule that measures the average magnitude of the error. It gives a relatively high weight to large errors; hence, it is most useful when large errors are undesirable[18]. MAE measures the average magnitude of the error in a set of predictions without considering their direction. It is a linear score, implying that all individual differences between predictions and corresponding observed values are weighted equally in the average[18]. R^2 is the best single measure of how well the predicted values match the observed values[18]. RMSE, MAE, and R^2 are defined by the equations:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(Q_{pre} - Q_{obs})^2}{n}} \tag{2}$$

$$\mathrm{MAE} = \left[\frac{\sum_{i=1}^{n}|Q_{pre} - Q_{obs}|}{n}\right] \tag{3}$$

$$R^2 = \left[1 - \sum_{i}\frac{(Q_{obs} - Q_{pre})^2}{(Q_{obs} - \overline{Q}_{obs})^2}\right] \tag{4}$$

where Q_{obs} = observed value; $\overline{Q}_{obs}$ = the mean of the observed data; Q_{pre} = predicted value; i = number of observations; and n = number of points in the dataset. The best score for RMSE and MAE is defined as minimizing the training error; the measure is 1 for R^2 and 0 for the other measures.

References

1. Goudie, A. S. Desert dust and human health disorders. *Environment International* **63**, 101–113, https://doi.org/10.1016/j.envint.2013.10.011 (2014).
2. Griffin, D. W. Atmospheric movement of microorganisms in clouds of desert dust and implications for human health. *Clinical Microbiology Reviews* **20**, 459–477, https://doi.org/10.1128/cmr.00039-06 (2007).
3. Uno, I. *et al.* Asian dust transported one full circuit around the globe. *Nature Geoscience* **2**, 557–560, https://doi.org/10.1038/ngeo583 (2009).
4. Griffin, D., Kellogg, C. & Shinn, E. Dust in the wind: long range transport of dust in the atmosphere and its implications for public and ecosystem health. *Global Change and Human Health* **2**, 20–33, https://doi.org/10.1023/A:1011910224374 (2001).
5. Maki, T. *et al.* Variations in the structure of airborne bacterial communities in a downwind area during an Asian dust (Kosa) event. *Science of the Total Environment* **488**, 75–84, https://doi.org/10.1016/j.scitotenv.2014.04.044 (2014).
6. Jeon, E. M. *et al.* Impact of Asian dust events on airborne bacterial community assessed by molecular analyses. *Atmospheric Environment* **45**, 4313–4321, https://doi.org/10.1016/j.atmosenv.2010.11.054 (2011).
7. Yoo, K. *et al.* Molecular approaches for the detection and monitoring of microbial communities in bioaerosols: A review. *Journal of Environmental Sciences* **51**, 234–247, https://doi.org/10.1016/j.jes.2016.07.002 (2017).
8. Akhlaq, M., Sheltami, T. R. & Mouftah, H. T. A review of techniques and technologies for sand and dust storm detection. *Reviews in Environmental Science and Bio-Technology* **11**, 305–322, https://doi.org/10.1007/s11157-012-9282-y (2012).
9. Peccia, J., Milton, D. K., Reponen, T. & Hill, J. A role for environmental engineering and science in preventing bioaerosol-related disease. *Environmental Science & Technology* **42**, 4631–4637 (2008).
10. DeLeon-Rodriguez, N. *et al.* Microbiome of the upper troposphere: Species composition and prevalence, effects of tropical storms, and atmospheric implications. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 2575–2580, https://doi.org/10.1073/pnas.1212089110 (2013).
11. Bibby, K., Viau, E. & Peccia, J. Pyrosequencing of the 16S rRNA gene to reveal bacterial pathogen diversity in biosolids. *Water Research* **44**, 4252–4260, https://doi.org/10.1016/j.watres.2010.05.039 (2010).
12. Huang, K. L., Zhang, X. X., Shi, P., Wu, B. & Ren, H. Q. A comprehensive insight into bacterial virulence in drinking water using 454 pyrosequencing and Illumina high-throughput sequencing. *Ecotoxicology and Environmental Safety* **109**, 15–21, https://doi.org/10.1016/j.ecoenv.2014.07.029 (2014).
13. Ye, L. & Zhang, T. Pathogenic Bacteria in Sewage Treatment Plants as Revealed by 454 Pyrosequencing. *Environmental Science & Technology* **45**, 7173–7179, https://doi.org/10.1021/es201045e (2011).

14. Chen, Q. L. *et al*. Long-term field application of sewage sludge increases the abundance of antibiotic resistance genes in soil. *Environment International* **92–93**, 1–10, https://doi.org/10.1016/j.envint.2016.03.026 (2016).
15. Yoo, K. *Decision Tree-based Data Mining and Rule Induction for Environmental Impact Assessment* Ph.D thesis, Yonsei University (2015).
16. Al-Alawi, S. M., Abdul-Wahab, S. A. & Bakheit, C. S. Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone. *Environmental Modelling & Software* **23**, 396–403, https://doi.org/10.1016/j.envsoft.2006.08.007 (2008).
17. Kovač-Andrić, E., Brana, J. & Gvozdić, V. Impact of meteorological factors on ozone concentrations modelled by time series analysis and multivariate statistical methods. *Ecological Informatics* **4**, 117–122, https://doi.org/10.1016/j.ecoinf.2009.01.002 (2009).
18. Singh, K. P., Gupta, S. & Rai, P. Identifying pollution sources and predicting urban air quality using ensemble learning methods. *Atmospheric Environment* **80**, 426–437, https://doi.org/10.1016/j.atmosenv.2013.08.023 (2013).
19. Koo, Y. S., Choi, D. R., Kwon, H. Y., Jang, Y. K. & Han, J. S. Improvement of PM10 prediction in East Asia using inverse modeling. *Atmospheric Environment* **106**, 318–328, https://doi.org/10.1016/j.atmosenv.2015.02.004 (2015).
20. Shao, Y. & Dong, C. H. A review on East Asian dust storm climate, modelling and monitoring. *Global and Planetary Change* **52**, 1–22, https://doi.org/10.1016/j.gloplacha.2006.02.011 (2006).
21. Livingstone, D. J. & Salt, D. W. Judging the significance of multiple linear regression models. *Journal of Medicinal Chemistry* **48**, 661–663, https://doi.org/10.1021/jm049111p (2005).
22. Berry, M. & Linoff, G. *Data Mining Techniques*. (Indianapolis, 2004).
23. Breiman, L., Friedman, J., Olshen, R. & Stone, C. In *Classification and Regression Tree* (Champman and Hall, New York, 1984).
24. Yoo, K., Shukla, S. K., Ahn, J. J., Oh, K. & Park, J. Decision tree-based data mining and rule induction for identifying hydrogeological parameters that influence groundwater pollution sensitivity. *Journal of Cleaner Production* **122**, 277–286, https://doi.org/10.1016/j.jclepro.2016.01.075 (2016).
25. Lee, J., Im, J., Kim, U. & Loffler, F. E. A Data Mining Approach to Predict *In Situ* Detoxification Potential of Chlorinated Ethenes. *Environmental Science & Technology* **50**, 5181–5188, https://doi.org/10.1021/acs.est.5b05090 (2016).
26. Smeti, E. M., Thanasoulias, N. C., Lytras, E. S., Tzoumerkas, P. C. & Golfinopoulos, S. K. Treated water quality assurance and description of distribution networks by multivariate chemometrics. *Water Research* **43**, 4676–4684, https://doi.org/10.1016/j.watres.2009.07.023 (2009).
27. Price, B., Venso, E., Frana, M., Greenberg, J. & Ware, A. A comparison of ARA and DNA data for microbial source tracking based on source-classification models developed using classification trees. *Water Research* **41**, 3575–3584, https://doi.org/10.1016/j.watres.2007.05.026 (2007).
28. Vega, F. A., Matias, J. M., Andrade, M. L., Reigosa, M. J. & Covelo, E. F. Classification and regression trees (CARTs) for modelling the sorption and retention of heavy metals by soil. *Journal of Hazardous Materials* **167**, 615–624, https://doi.org/10.1016/j.jhazmat.2009.01.016 (2009).
29. Burrows, S. M., Elbert, W., Lawrence, M. G. & Poschl, U. Bacteria in the global atmosphere - Part 1: Review and synthesis of literature data for different ecosystems. *Atmospheric Chemistry and Physics* **9**, 9263–9280 (2009).
30. Zhou, Y. J. *et al*. Metagenomic Approach for Identification of the Pathogens Associated with Diarrhea in Stool Specimens. *Journal of Clinical Microbiology* **54**, 368–375, https://doi.org/10.1128/jcm.01965-15 (2016).
31. Boldo, E. *et al*. Health impact assessment of a reduction in ambient PM2.5 levels in Spain. *Environment International* **37**, 342–348, https://doi.org/10.1016/j.envint.2010.10.004 (2011).
32. de Andres, J. M., Borge, R., de la Paz, D., Lumbreras, J. & Rodriguez, E. Implementation of a module for risk of ozone impacts assessment to vegetation in the Integrated Assessment Modelling system for the Iberian Peninsula. Evaluation for wheat and Holm oak. *Environmental Pollution* **165**, 25–37, https://doi.org/10.1016/j.envpol.2012.01.048 (2012).
33. Leski, T. A., Malanoski, A. P., Gregory, M. J., Lin, B. C. & Stenger, D. A. Application of a Broad-Range Resequencing Array for Detection of Pathogens in Desert Dust Samples from Kuwait and Iraq. *Applied and Environmental Microbiology* **77**, 4285–4292, https://doi.org/10.1128/aem.00021-11 (2011).
34. Aw, T. G. & Rose, J. B. Detection of pathogens in water: from phylochips to qPCR to pyrosequencing. *Curr. Opin. Biotechnol.* **23**, 422–430 (2012).
35. Tao, Y., An, X. Q., Sun, Z. B., Hou, Q. & Wang, Y. Association between dust weather and number of admissions for patients with respiratory diseases in spring in Lanzhou. *Science of the Total Environment* **423**, 8–11, https://doi.org/10.1016/j.scitotenv.2012.01.064 (2012).
36. Camatini, M., Corvaja, V., Pezzolato, E., Mantecca, P. & Gualtieri, M. PM10-biogenic fraction drives the seasonal variation of proinflammatory response in A549 cells. *Environmental Toxicology* **27**, 63–73, https://doi.org/10.1002/tox.20611 (2012).
37. Wiseman, C. L. S. & Zereini, F. Airborne particulate matter, platinum group elements and human health: A review of recent evidence. *Science of the Total Environment* **407**, 2493–2500, https://doi.org/10.1016/j.scitotenv.2008.12.057 (2009).
38. de Longueville, F. *et al*. Saharan Dust Impacts on Air Quality: What Are the Potential Health Risks in West Africa? *Human and Ecological Risk Assessment* **19**, 1595–1617, https://doi.org/10.1080/10807039.2012.716684 (2013).
39. Cox, C. S. *The microbiology of air. In Topley & Wilson's microbiology and microbial infections*. 9th ed. edn, (Oxford University Press, 1998).
40. Marthi, B., Fieland, V. P., Walter, M. & Seidler, R. J. Survival Of Bacteria During Aerosolization. *Applied and Environmental Microbiology* **56**, 3463–3467 (1990).
41. Shaman, J. & Kohn, M. Absolute humidity modulates influenza survival, transmission, and seasonality. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 3243–3248, https://doi.org/10.1073/pnas.0806852106 (2009).
42. de la Campa, A. S., Garcia-Salamanca, A., Solano, J., de la Rosa, J. & Ramos, J. L. Chemical and Microbiological Characterization of Atmospheric Particulate Matter during an Intense African Dust Event in Southern Spain. *Environmental Science & Technology* **47**, 3630–3638, https://doi.org/10.1021/es3051235 (2013).
43. Yamaguchi, N., Ichijo, T., Sakotani, A., Baba, T. & Nasu, M. Global dispersion of bacterial cells on Asian dust. *Scientific Reports* **2**, https://doi.org/10.1038/srep00525 (2012).
44. Alexis, N. E. *et al*. Biological material on inhaled coarse fraction particulate matter activates airway phagocytes *in vivo* in healthy volunteers. *Journal of Allergy and Clinical Immunology* **117**, 1396–1403, https://doi.org/10.1016/j.jaci.2006.02.030 (2006).
45. Happo, M. S. *et al*. Dose and time dependency of inflammatory responses in the mouse lung to urban air coarse, fine, and ultrafine particles from six European cities. *Inhalation Toxicology* **19**, 227–246, https://doi.org/10.1080/08958370601067897 (2007).
46. Murata, K. & Zhang, D. Z. Transport of bacterial cells toward the Pacific in Northern Hemisphere westerly winds. *Atmospheric Environment* **87**, 138–145, https://doi.org/10.1016/j.atmosenv.2013.12.038 (2014).
47. Seinfeld, J. H. & Pandis, S. N. *Dynamics of aerosol populations, Atmospheric Chemistry and Physics: from air pollution to climate change* (John Wiley, 1998).
48. Khayyam, H., Golkarnarenji, G. & Jazar, R. N. *Nonlinear Approaches in Engineering Applications: Energy, Vibrations, and Modern Applications* (ed. Dai, L & Jazar, R. N.) 345–379 (Springer, 2018).
49. Refaeilzadeh, P., Tang, L. & Liu, H. *Cross Validation* in Encyclopedia of Database Systems (ed. Liu, L & Özsu, M. T.) 532–538 (Springer, 2009).
50. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal* **13**, 8–17, https://doi.org/10.1016/j.csbj.2014.11.005 (2015).

51. Li, D. C., Lin, W. K., Lin, L. S., Chen, C. C. & Huang, W. T. The attribute-trend-similarity method to improve learning performance for small datasets. *International Journal of Production Research* **55**, 1898–1913, https://doi.org/10.1080/00207543.2016.1213447 (2017).
52. Radosevich, J. L., Wilson, W. J., Shinn, J. H., DeSantis, T. Z. & Andersen, G. L. Development of a high-volume aerosol collection system for the identification of air-borne micro-organisms. *Letters in Applied Microbiology* **34**, 162–167, https://doi.org/10.1046/j.1472-765x.2002.01048.x (2002).
53. Harms, G. *et al*. Real-time PCR quantification of nitrifying bacteria in a municipal wastewater treatment plant. *Environmental Science & Technology* **37**, 343–351, https://doi.org/10.1021/es0257164 (2003).
54. Shannon, K. E., Lee, D. Y., Trevors, J. T. & Beaudette, L. A. Application of real-time quantitative PCR for the detection of selected bacterial pathogens during municipal wastewater treatment. *Science of the Total Environment* **382**, 121–129, https://doi.org/10.1016/j.scitotenv.2007.02.039 (2007).
55. Priest, F. G., Barker, M., Baillie, L. W. J., Holmes, E. C. & Maiden, M. C. J. Population structure and evolution of the Bacillus cereus group. *Journal of Bacteriology* **186**, 7959–7970, https://doi.org/10.1128/jb.186.23.7959-7970.2004 (2004).
56. He, J. Z., Ritalahti, K. M., Yang, K. L., Koenigsberg, S. S. & Loffler, F. E. Detoxification of vinyl chloride to ethene coupled to growth of an anaerobic bacterium. *Nature* **424**, 62–65, https://doi.org/10.1038/nature01717 (2003).
57. Claesson, M. J. *et al*. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Research* **38**, https://doi.org/10.1093/nar/gkq873 (2010).
58. Schloss, P. D., Gevers, D. & Westcott, S. L. Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. *Plos One* **6** https://doi.org/10.1371/journal.pone.0027310 (2011).
59. Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L. & Mark Welch, D. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology* **8**, https://doi.org/10.1186/gb-2007-8-7-r143 (2007).
60. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194–2200, https://doi.org/10.1093/bioinformatics/btr381 (2011).
61. Quince, C., Lanzen, A., Davenport, R. J. & Turnbaugh, P. J. Removing Noise From Pyrosequenced Amplicons. *Bmc Bioinformatics* **12**, https://doi.org/10.1186/1471-2105-12-38 (2011).
62. Pruesse, E. *et al*. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* **35**, 7188–7196, https://doi.org/10.1093/nar/gkm864 (2007).
63. Cole, J. R. *et al*. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research* **37**, D141–D145, https://doi.org/10.1093/nar/gkn879 (2009).
64. EPA. *Microbial Risk Assessment Guideline: Pathogenic microorganisms with focus on food and water*. (U.S. Environmental Protection Agency, 2012).
65. Feazel, L. M. *et al*. Opportunistic pathogens enriched in showerhead biofilms. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 16393–16398, https://doi.org/10.1073/pnas.0908446106 (2009).
66. Chun, J. *et al*. EzTaxon: a web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences. *International Journal of Systematic and Evolutionary Microbiology* **57**, 2259–2261, https://doi.org/10.1099/ijs.0.64915-0 (2007).
67. Cawley, G. C. & Talbot, N. L. C. Efficient leave-one-out cross-validation of kernel Fisher discriminant classifiers. *Pattern Recognition* **36**, 2585–2592, https://doi.org/10.1016/s0031-3203(03)00136-5 (2003).

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No. 2011-0030040). In addition, this research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by Ministry of Education (N0. 2018R1A6A1A08025348).

Author Contributions

K.Y. and H.Y. collected bioaerosol samples. K.Y. and J.P. designed experiments and drafted the manuscript. K.Y. performed the modeling analyses. H.Y., J.L. and S.S. helped in generating and interpreting the sequencing and statistical data. J.L. and S.S. prepared figures and edited the manuscript. All authors discussed the results and contributed to the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-29796-7.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.