Consistent interpretation of air pollution related OMICs data; A case study in environmental epidemiological research

Bjorn van Boven

Institute of Risk Assessment Science, Utrecht University, Utrecht, The Netherlands

Abstract

Environmental air pollution is observed to contribute to the prevalence and incidence of chronic diseases. The exact underlying mechanisms by which air pollutants initiate disease have not been completely elucidated. However oxidative stress and inflammation are the most common related mechanisms underlying for these health outcomes. Genetic variation in specific molecular pathways or other biomarkers can provide information on these mechanisms that are associated with a certain type of exposure. In this paper, we used an AI tool, the Euretos Knowledge Platform to identify candidate genes related to air pollution exposure and those involved in the occurrence and development of the (cardiovascular) diseases. An user case experiment was conducted based on a candidate gene set derived from literature. Next to this, an orientating experiment was performed with the use of the search engine with in the same programme. Based on those experiments, several genes and pathways related with the exposure on air pollution and associated with the diseases were identified. Genes marked for detoxification (SOD1, SOD2, GCLC, GCLM GPX1 and CAT) and metabolism of xenobiotics (GSTP1, GSTM1, CYP1B1, GCLC and GCLM) were recognised. Furthermore, genes related to inflammatory responses (IL1B, IL6, IL6R and CXCL8) are marked in this study as well. Euretos showed potential for implementation in environmental epidemiology research. However, more research and extra changes are required for data search and desirable for the user case for an direct implantation of the programme.

Introduction

The way we live and behave is impacting our own life, and the life of other people. The World Health Organisation (WHO) stated that household combustion, motor vehicles and industrial facilities are just a few but the most common sources of air pollution. Air pollution is a complex mixture include particulate matter (PM), carbon monoxide, ozone, nitrogen dioxide and sulfur dioxide. Air pollution is a growing public health concern of global significance, according to the WHO (2021).

The common consensus is that environmental (air) pollution and non-chemical stressors contribute to the prevalence and incidence of chronic diseases (Daiber et al., 2019). The authors stated that a higher risk for noncommunicable disease such as cardiovascular, metabolic and mental diseases are associated with traffic noise and air pollution. Next to this, air pollution kills an estimated seven million people worldwide every year (WHO, 2021). WHO data also shows that 99% of the global population breathe air that exceeds their WHO guideline limits containing high levels of pollutants. Air pollution is recognized as a human carcinogen, therefore being a risk factor for (lung) cancer (Xing et al., 2019). It is also a major risk factor for other acute and chronic diseases including cardiovascular disease (Lederer et al., 2021; Mannucci et al., 2019) and chronic respiratory diseases (Guan et al., 2016; Park et al., 2021). Oxidative stress and inflammation are commonly considered as presumptive mechanisms underlying for these health outcomes (Fuertes et al., 2020; Hahad et al., 2020). However the exact underlying mechanisms by which air pollutants initiate disease have not clearly been completely clarified.

Despite the fact that exposure to air pollution has been shown to induce changes in gene expression (Huang et al., 2011; Wittkopp et al., 2016), evidence for mechanisms related to different disease outcomes is scarce. Investigating genetic variation in specific molecular pathways can provide information on physiological disturbances that are associated with a certain type environmental exposure. Next to this, genetic changes and early biomarkers form various systems could help to prevent and control the damage effects of air pollution (D. Yang et al., 2017). Focusing on genes (genomics), proteins (proteomics), mRNA (transcriptomics) or metabolites (metabolomics) could help clarified the unclear pathways. Studying these different (multi-)omics data can provide biological insights for empirically observed exposure disease associations (Sun & Hu, 2016). Despite the significant research on air pollutant-associated health effects, the underlying molecular mechanisms by which air pollutants initiate disease remain unclear. Oxidative damage, inflammation and endothelial dysfunction have been suggested as potential underlying mechanisms of air-pollutant associated adverse health (Lederer et al., 2021). Recently, it has been shown that short-term exposure of air pollution could lead to adverse cardiopulmonary effects (Liu et al., 2021).

Although these studies could imply a possible impact of specific air pollutants on humans, the results are sometimes inconclusive. Therefore, research is sometimes placed in a better perspective after clutching at straws. If the data acquired from your conducted research is confirmed by another study, your obtained data will be more acceptable. Mirroring obtained data to other studies is without a doubt a great principle to check your data for significance. Unfortunately, conformation of the empirical data is in most cases possible by ‘cherry picking’ facts and statements from previous literature. Even the more debatable and rare results can be linked to previous studies. AI-driven statistic tools could provide a powerful tool in this type of data analysis. These artificial intelligence programmes can examine, observe and predict possible relations between experimental derived data. These functionality accelerates the analysis of gene- or datasets by easing the transition from data collection to biological related concepts and systems (Dennis et al., 2003).

Extensive experimental methods such as genome-wide association studies (GWAS) identify increasingly large numbers of potential interesting genes and genetic variants. These type of studies have made a significant contribution to our understanding of the genetics of complex disorders over decade (Dehghan, 2018). Next to the increasing amount of available data from these studies, both biological complexity and the significant rate of empirical data production require computational and consistent approaches to interpretate potential new associated genes and elucidate gene-disease and gene-environment relations (Hettne et al., 2016). On the account of these factors, it is nearly impossible to perform these phases of research by hand. For these cases, statistical AI tools can provide the solution.

In environmental epidemiology a variety of enrichment analysis AI tools are implemented in different types of research, despite the lack of real literature references. One example, the *Database for Annotation, Visualization and Integrated Discovery* (DAVID) is used for the classification of gene function and pathway enrichment of empirical derived gene lists in research (T.-Y. Yang et al., 2014). Whereas other studies have focus on Ingenuity Pathway Analysis (IPA) (Smit-McBride et al., 2018) to identify potential miRNA gene targets and for gene pathway analysis. An addition study combined both tools in their conducted research (Toonen et al., 2018). More recent performed research (Everson et al., 2021) enhanced their study by using EnrichR. This tool is implemented to test for enrichment of transcription factor targets from, in this type of study, ENCODE/ChEA databases.

Our research started with an alternative approach and with the use of another AI tool. This air pollution focused research was conducted with the implementation of the Euretos Knowledge Platform ([1]), an alternative AI tool. In contrast to DAVID, which was described as rapid, complete and the foundation of the integration of information-rich data analytics methods (Dennis et al., 2003), Euretos consist of different engines and tools to perform analytics. Data sets could be uploaded and created with the database and literature search. With these sets, enrichment and network analysis can be performed on the same programme. The Euretos platform could give new insights using both curated data base annotation and literature from scientific publications. This is the first time implementing this tool into environmental epidemiology to the best of our knowledge.

In contrast to environmental epidemiology, Euretos is already implemented for pharmaceutical, drug and disease research. The Euretos Knowledge Platform (EKP) is used in cancer research and other disease and drug research. The platform describes itself as recognized as on one of the leading companies in pharmaceutical research and featured as one of the major players in their field. Euretos features a large AI- integrated knowledge base containing over 275 life science databases and millions of publications. The search engine of the EKP provides ranked list in over 100 categories (f.e. genes or pathways). These list are based on a variety of different sources, therefore they should give an extensive overview of the current knowledge ([1]). Knowledge regarding to the terms created by combining synonyms of the used search terms. The data sets obtained by the search engine can be used to give new biological multi-omics insights and provide biomolecular interaction analysis. The tool provide options for gene set enrichment, gene set ranking and transcription network analysis. The search and analytics are optimised for disease and drug biology, nevertheless it possibly could give new insights in environmental epidemiology.

The Euretos Knowledge Platform consist of multiple main stages, of which the search engine for the research data, analytics of the generated data sets and formation of relation maps to identify possible new molecular pathways are used in this study. Euretos had an integrated search engine which can be used to create lists of 'concepts' associated with the input search terms. Euretos uses concepts to perform search and analytics with. The concepts can vary from genes (IL6, TNF), diseases (COPD, asthma) to simple terms (ozone) or pathways (cytokine signalling). The saved concepts lists described as a 'set' in the Euretos Platform, can be used to perform several analytics. Multiple sets can be evaluated in the program to reveal the conceptual overlap between the different sets. These concept lists could consist of genes, pathways or random associated concepts, based on the selected category list. Subsequently, after uploading own experimental data or created lists using 'Search', the next step is to evaluate the data against the platform's knowledge base. The 'Analytics' function can be used to perform enrichment or ranking analysis on single concepts lists or the overlay between different sets. Enrichment analysis can be conducted with the 'Find Related' option in the Analytics of the EKP. The ranking analysis can be performed with the 'Rank Selection' option, where (mostly) genes in a set can be ranked to determine whether a gene is differentially expressed for a disease, phenotype, or compound. This ranking is based on thousands of differential expression experiments that are available in the platform ([2]). The 'Relation map' option of the programme creates a cobweb consisting of concepts uploaded into the sets and identify potential relations between the concepts. It uses both text mined literature abstract data or data from curated data base annotations to illustrated relations and links between the different concepts. The literature based relations are resulting in possible 'abstract cooccurrences', which is one on many relation types in this option. Furthermore, the curated sources could provide more "in depth" relations types, for this study: affects; binds with; catalysis; precedes; coexists with; controls expression of; forms protein complex with; gene product is; biomarker type; gene product variant results in abnormal; inhibits; interacts with; is a; is associated with; is manifestation of; predisposes; produces and stimulates. These indicate the type of relation between the concepts, with an optional label, name tag, and arrow clarifying the direction of the relation. Next to the possibility to import data sets and lists, individual concepts can be added to improve and supplement the analysis in the relation maps. An extensive protocol and overview of Euretos can be found in Supplementary file 1 (Euretos appendix).

In our study, the search engine of the programme was applied to create data sets. The genes categories provides a set of genes that are associated with the input search term. Data sets consisting of genes associated to 'air pollution' and genes associated to specific concepts related to air pollution. Different types of air pollution 'concepts' were used. Particulate matter (PM), ozone, nitrogen oxides (Nitrogen dioxide), sulfur dioxide, volatile organic compounds (VOC), polycyclic aromatic hydrocarbons (PAH) and carbon monoxide were classified as pollutants of major public health concern by the WHO (2021). Next to this, we focused on different cardiometabolic pulmonary health outcomes in this study. Acute myocardial infarction (AMI), asthma, cerebrovascular accident (CVA), chronic obstructive pulmonary disease (COPD) and Diabetes Mellitus type 2 (T2DM) were selected based on their known association with air pollution (Cesaroni et al., 2014; Gehring et al., 2015; Li et al., 2021; Merid et al., 2021; Park et al., 2021; Shah et al., 2013; Zhao et al., 2021). The 'Analytics' function of the programme was used to find related pathways. Eventually, the relation map option of the EKP was implemented to identify key terms, indirect associations and concept clusters. In this case, the concepts are generally genes combined with air pollution concepts or health outcomes.

Our study is divided into two part. Firstly, a 'user case' experiment was performed. For this study we used data obtained from an existing, already performed study: Associations Between Genome-wide Gene Expression and Ambient Nitrogen Oxides (Mostafavi et al., 2017). This can be seen as a typical application for the AI tool in this type of research. In this case, the use of the EKP is the first step after obtaining the data. Secondly, we tried to use the search engine of Euretos to create our own data 'sets' consisting of genes (TOP10 test). These created lists could give new insights based on the available data in literature and databases. Which could provide with entire networks and new opportunities for additional research. The user case experiment tests the potential of the programme to compare and complement the existing obtained data. Whereas the TOP10 test provide a new approach for orientating into a new focus of (environmental epidemiological) biology. Both could be useful in their own way to strengthen research.

The overall aim of this study was to evaluate to which extent Euretos added value for future epidemiological studies within the domain of environmental research. An addition comparison was made with the DAVID pathway enrichment tool, to show if the Euretos tool did add some additional and new value to the research. The potential implementation could help the researcher to map the obtained results within an curated database. In our conducted (user) case study, we focused on the identification of air pollution associated biological pathways and genetic relations, with biological interpretation of cohort (gen)omics data from Mostafavi (2017). Our goals with that experiment, in particular, was to test the programme on performed research, with curated data and the previous described analytical functions within the programme. The programme could provide with new insights for interpretation of new empirical data which was obtained through a GWAS or other type of studies. The research is conducted with the Euretos Platform. The environmental epidemiological omics data were studied in order to have a better understanding of the underlying biological pathways that relate air pollution exposure to health outcomes. With our approach, we carefully identified to which extent this AI tool is efficient for future epidemiological studies within the domain of environmental research.

Methods

The flowchart of our approach is depicted below (Figure 1). A brief summary, more extensive explanation is described after the figure: A) Real world obtained data is uploaded in the EKP. B) Euretos' search engine results were filtered and selected for the TOP10 sets. In addition, sets obtained in step A and B were supplemented with either disease concepts or air pollution concepts were added for further analysis. C) Pathway enrichment was performed for the user case experiment, the real world observations. D) All the outputs were combined and review to create a general conclusion, that is usually visualised in several relations maps. E) Extra experts and extern analytic data could be added to contribute to and improve upon the obtained results. However, were not applied to this study. Therefore could be used for future application. F) Final results and conclusion are drawn from all the available knowledge obtained from the program and additional sources.

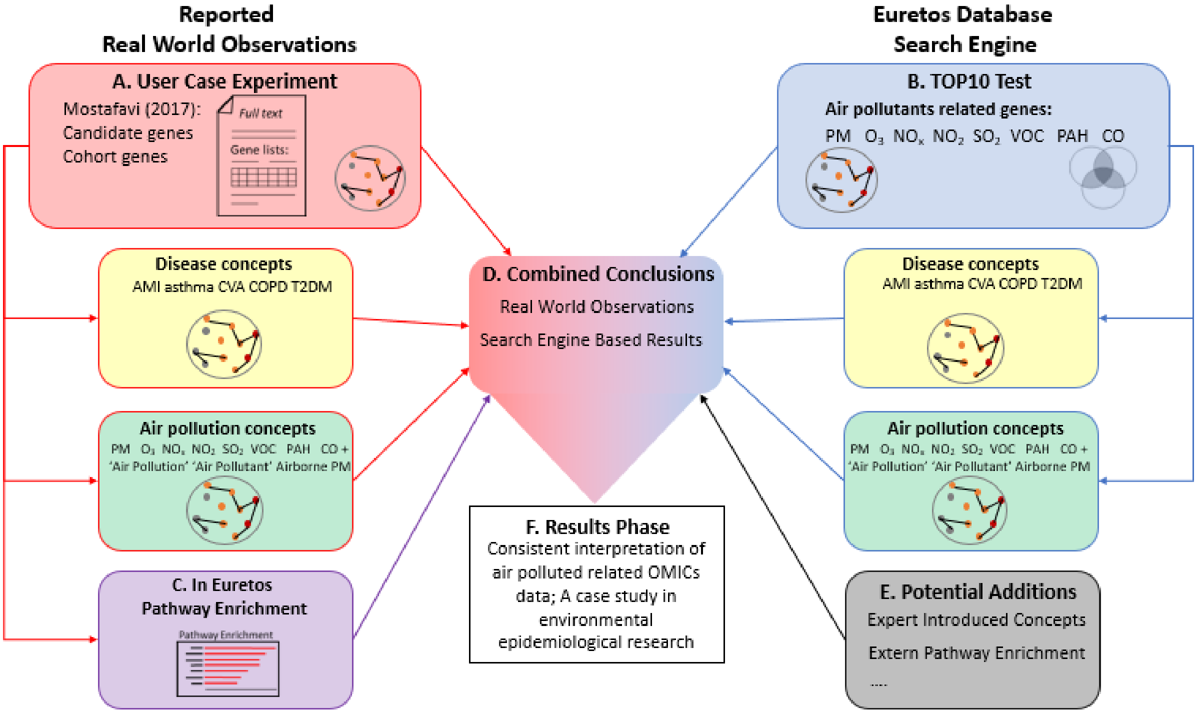

Reported
Real World Observations
Euretos Database
Search Engine
A. User Case Experiment
Mostafavi (2017):
Candidate genes
Cohort genes
Full text
Gene lists:
B. TOP10 Test
Air pollutants related genes:
PM O_3 NO_x NO_2 SO_2 VOC PAH CO
Disease concepts
AMI asthma CVA COPD T2DM
D. Combined Conclusions
Real World Observations
Search Engine Based Results
Disease concepts
AMI asthma CVA COPD T2DM
Air pollution concepts
PM O_3 NO_x NO_2 SO_2 VOC PAH CO +
'Air Pollution' 'Air Pollutant' Airborne PM
Air pollution concepts
PM O_3 NO_x NO_2 SO_2 VOC PAH CO +
'Air Pollution' 'Air Pollutant' Airborne PM
F. Results Phase
Consistent interpretation of air polluted related OMICs data; A case study in environmental epidemiological research
C. In Euretos
Pathway Enrichment
Pathway Enrichment
E. Potential Additions
Expert Introduced Concepts
Extern Pathway Enrichment
....

Figure 1. The flowchart of a potential approach of implementation of the Euretos Knowledge Platform in environmental epidemiology. Adapted from preprint Mons (2020).

Mostafavi (2017) article (A)

The obtained data was selected from articles published by the IRAS, Institute of Risk Assessment Sciences (Mostafavi, 2017). Mostafavi, et al. (2017) display a table containing genes associated with long-term exposure to NO_x and a table containing possible candidate genes, respectively table 1 and 3 in their manuscript. The experiment was done as typical application of the EKP in the field of environmental epidemiology. With the implementation of the tool, the research could be improved by providing possible new insights. Firstly, the gene list, consisting of 29 candidate genes previously associated to air pollution in the epidemiological literature was uploaded and analysed. This candidate gene list is tested as a positive control for implementation. A higher expectation applies for the candidate genes, because they are actually observed to be affected by air pollution. On the other hand, that certainty applies less to the cohort genes obtained by the conducted research.

Based on applicability, the candidate genes list should give a better idea of the potential of the EKP for environmental epidemiological studies. Because this set contained genes with relations to each other that we reported several times in literature, according to the programme. Therefore, the candidate genes data set was added to the 'Relation map' function of the EKP and used for pathway enrichment in the 'Analytic' section of the programme. (C) Pathway analysis is performed in Euretos to identify potential enrichment in the set of genes. The enrichment results are sorted by P-value. These values are adjusted for multiple testing correction using the Benjamini-Hochberg procedure. The pathway analysis was only performed on the candidate set, due to the size of the cohort gene set (consisting of 11 genes). The pathway enrichment was also conducted with DAVID. The functional annotation clustering tool of this programme was used for pathway analysis (Background: Homo Sapiens). This pathway analysis was made to compare the analysis of Euretos with the renowned pathway analysis of DAVID.

For the relation map, the set was enriched with the five different health outcomes (Acute myocardial infarction (AMI), asthma, cerebrovascular accident (CVA), chronic obstructive pulmonary disease (COPD) and Diabetes Mellitus type 2 (T2DM)), for the first analysis, and enriched with the concepts related to air pollution for another. All relations among different health outcomes and among the different concepts related to air pollution were removed on forehand. The concepts were arranged to improve the interpretation of the relation maps. If necessary, the weakest associations labels were disabled to improve the overall overview and interpretation. These labels, with connection: is associated with, have been removed because these are less concrete about the content of the connection between concepts.

Secondly, the gene list, consisting of empirically obtained genes (cohort) associated with long-term exposure to NOx was uploaded manually to the programme. This corresponds to table 1 of the manuscript of Mostafavi (2017). The complete analysis was done largely the same way compared to the candidate gene list, mentioned before.

TOP10 test (B)

Initially, the sets consisting of genes associated with particulate matter were made using the Euretos Search engine. The input search term "particulate matter" was used in the search engine. The programme' search recommendation: search with synonyms was applied. The additional category button 'genes' was used to obtain genes related to particulate matter. The top 10 of most referenced genes was selected and saved as a new concept set. These single sets were created to determine the most referenced genes with the corresponding type of air pollution. The number of concepts in a set was restricted by a limitation of function of the programme. Euretos' 'relation map' function, one of the performed analytic tools, was limited to a total of 400 concepts. Therefore, these data sets are limited on number of concepts. This procedure was performed essentially identical for the remaining air pollution specific concepts, however with the corresponding input search term. This resulted in eight lists containing genes related to separately: particulate matter, ozone, nitrogen oxides, nitrogen dioxide, sulfur dioxide, volatile organic compounds, polycyclic aromatic hydrocarbons and carbon monoxide. The eight single sets, consisting of the different top 10 genes, were combined to perform analytics on the total set of air pollution associated genes. This 'TOP10 gene set' consists of 33 different genes.

The TOP10 gene sets can be used to perform relation and general analytics on relevant concepts and give more insights on the relations between the concepts. These type of analytics were performed with the Relation map option of the EKP. Therefore, as already mentioned, the concepts sets of our interest consist of genes related to the relevant type of air pollution. Despite the variety of different sources used by the programme, the genes category only used abstract text-mining to draft the references lists.

Relation maps were created to determine the relations between the individual genes. The TOP10 set was arranged in the 'Relation map' function of the programme. Followed by relation maps enriched with the different health outcomes or concepts related to air pollution. All relations among different health outcomes and among the different concepts related to air pollution were removed on forehand. This was done to create more accessible figures. The relation map option of EKP provides the user to disable and enable relation types labels. Starting with showing or hiding abstract and sentence co-occurrence relation types labels, to more in depth relation types (is associated with, promotes, binds with etc.) which can be selected individually. The relation types labels: Co-occurrence and 'is associated with' were disabled to improve interpretation. These types of relations are the most common and therefore cause an unreadable figure. Removing these relations labels with lower impact, leads to a clearer figure.

Results

Mostafavi (2017) article

Candidate genes: The relations maps revealed that a total of seven genes were not related to any of the other genes in the dataset. The complete list of relations is shown in Supplementary Table 2. Figure 2 reveals a significant position for the transcription regulator gene, NFE2L2 (homo sapiens). This gene is observed to have several interactions with other genes and their products. Therefore, it is identified as a central gene in this network. Next to this, the relation between the antioxidant genes (Fuertes et al., 2020) and among the interleukins can be observed in the same figure (Fig. 2). As extra remark, the gene name 'IL8' was not present in the EKP. This concept was changed for the concept CXCL8, another name for the same gen.

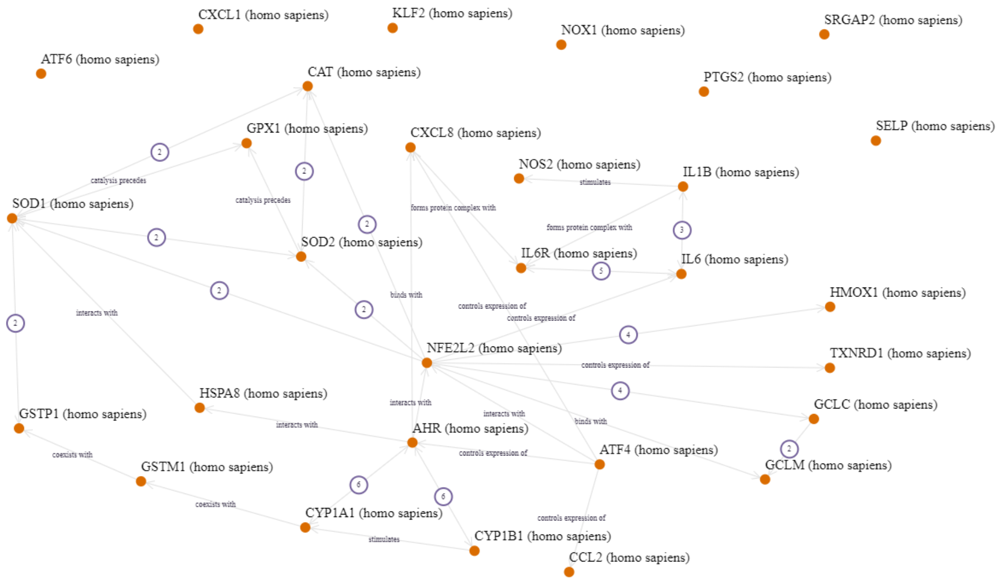
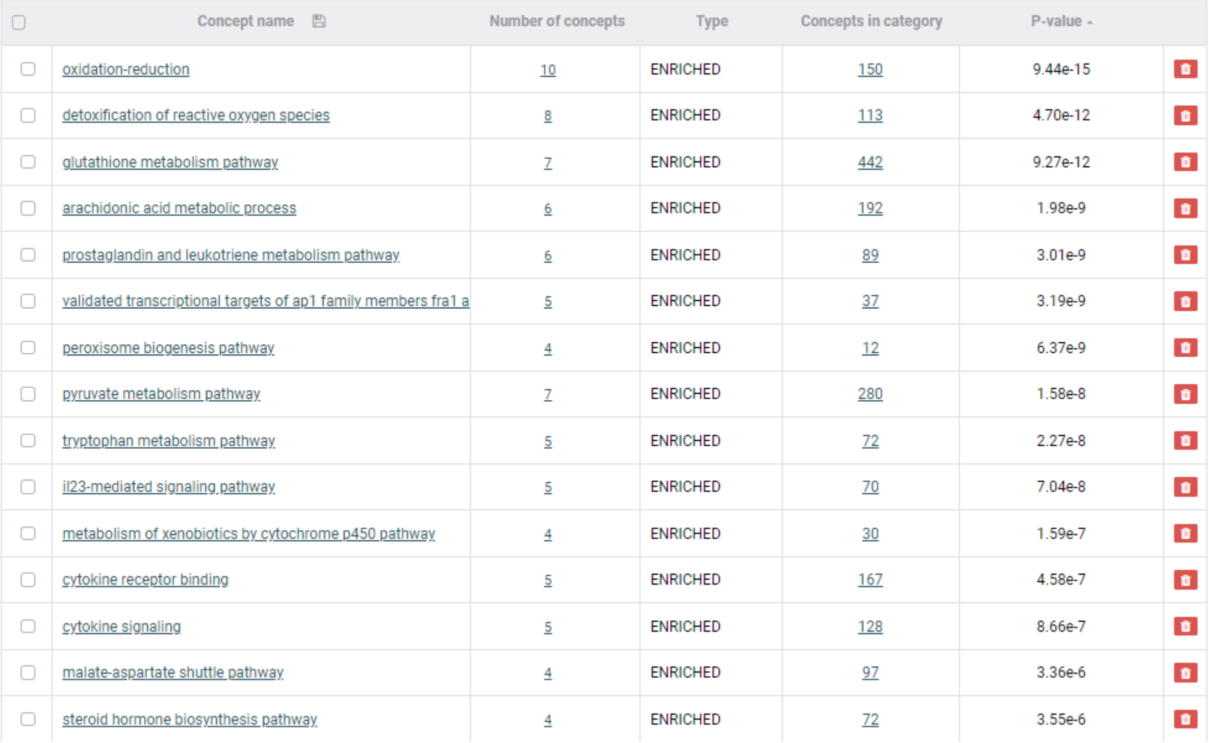


Figure 2. Relation map represents the relations between the candidate genes (Mostafavi, 2017). All relation type labels are shown. (Figure can be observed in the Supplementary File, for further identification (Supplementary Figure 2)

Supplementary File 2 Figure 3 reveals considerable relations between the candidate genes and the different health outcomes. The more in depth relations were attributed to the KLF2, SRGAP2 and the interleukins. According to the GWAS Catalog (source used by the EKP), the IL6, IL6R and Il1B gene products variants related to asthma. KLF2 and SRGAP2 possess the same relation related to this health outcome, whereas the SRGAP2 gene products also variants related to Diabetes Mellitus, type 2. The genes HMOX1 and Il6, were indicated by Supplementary File 2 Figure 3, as biomarker for respectively COPD and T2DM. As can be seen from Supplementary File 2 Figure 4, there was only one gene with a curated relation to one of the concepts related to air pollution. The HMOX1 gene product, Heme oxygenase, produces carbon monoxide, thus this CO relation is not air pollution related.

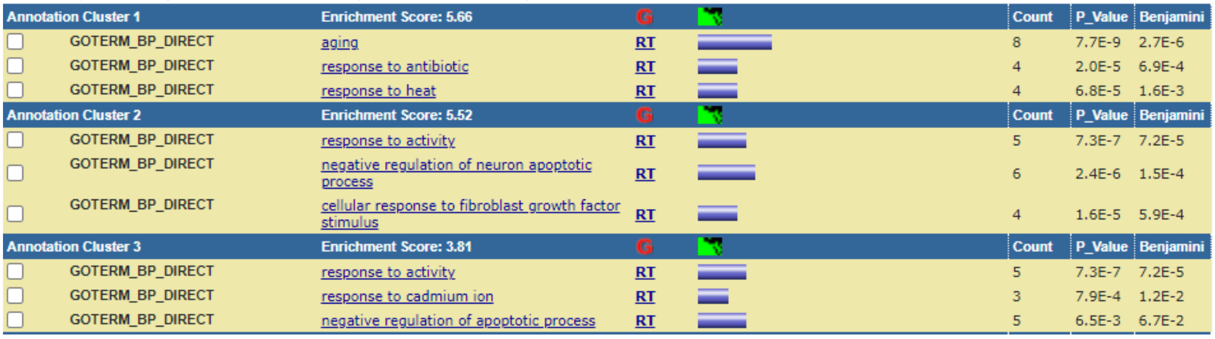
The pathway analysis identified 203 significant pathways, with an extra 20 enriched pathways that did not pass the false discovery rate of 0.05. The top of the chart can be found in Table 1, the complete table is found in Supplementary Table 1. Most pathways that were identified were oxidation related, complemented with signalling and specific metabolic pathways. Based on the pathway analysis, we can compare the previously mentioned DAVID annotation tool with the Euretos AI tool. The pathway analysis of the candidate gene set in both Euretos and DAVID resulted in pathways concerning oxidation, signalling and metabolic pathways. Only exception was the pathways regarding aging, which were indicated by DAVID (Table 2). Besides, this was the pathway resulting in the highest enrichment score on the DAVID clustering tool. DAVID provides the user with this extra tool to look at the internal relationships of clustered terms and makes the biological interpretation more focused at a group level.

Table 1. Top of the Euretos pathway analysis table. Listing the concept (pathway) name, number of concepts of the selected set represented in the relevant pathway, column for type indication (Enriched or Depleted), number of concepts in the category (pathway) and the corresponding Fisher's exact test p-values.



□	Concept name	Number of concepts	Type	Concepts in category	P-value ▴	
□	oxidation-reduction	10	ENRICHED	150	9.44e-15	🗑
□	detoxification of reactive oxygen species	8	ENRICHED	113	4.70e-12	🗑
□	glutathione metabolism pathway	7	ENRICHED	442	9.27e-12	🗑
□	arachidonic acid metabolic process	6	ENRICHED	192	1.98e-9	🗑
□	prostaglandin and leukotriene metabolism pathway	6	ENRICHED	89	3.01e-9	🗑
□	validated transcriptional targets of ap1 family members fra1 a	5	ENRICHED	37	3.19e-9	🗑
□	peroxisome biogenesis pathway	4	ENRICHED	12	6.37e-9	🗑
□	pyruvate metabolism pathway	7	ENRICHED	280	1.58e-8	🗑
□	tryptophan metabolism pathway	5	ENRICHED	72	2.27e-8	🗑
□	il23-mediated signaling pathway	5	ENRICHED	70	7.04e-8	🗑
□	metabolism of xenobiotics by cytochrome p450 pathway	4	ENRICHED	30	1.59e-7	🗑
□	cytokine receptor binding	5	ENRICHED	167	4.58e-7	🗑
□	cytokine signaling	5	ENRICHED	128	8.66e-7	🗑
□	malate-aspartate shuttle pathway	4	ENRICHED	97	3.36e-6	🗑
□	steroid hormone biosynthesis pathway	4	ENRICHED	72	3.55e-6	🗑

Table 2. Top of the DAVID pathway analysis table. Listing the system of classification, the GO term of the biological process, potential related terms (RT), number of concepts of the selected set represented in the relevant pathway, number of concepts in the category (pathway) and the corresponding P-value and adjusted for Benjamini test score.



				Count	P_Value	Benjamini
Annotation Cluster 1		Enrichment Score: 5.66	G	Count	P_Value	Benjamini
□	GOTERM_BP_DIRECT	aging	RT	8	7.7E-9	2.7E-6
□	GOTERM_BP_DIRECT	response to antibiotic	RT	4	2.0E-5	6.9E-4
□	GOTERM_BP_DIRECT	response to heat	RT	4	6.8E-5	1.6E-3
Annotation Cluster 2		Enrichment Score: 5.52	G	Count	P_Value	Benjamini
□	GOTERM_BP_DIRECT	response to activity	RT	5	7.3E-7	7.2E-5
□	GOTERM_BP_DIRECT	negative regulation of neuron apoptotic process	RT	6	2.4E-6	1.5E-4
□	GOTERM_BP_DIRECT	cellular response to fibroblast growth factor stimulus	RT	4	1.6E-5	5.9E-4
Annotation Cluster 3		Enrichment Score: 3.81	G	Count	P_Value	Benjamini
□	GOTERM_BP_DIRECT	response to activity	RT	5	7.3E-7	7.2E-5
□	GOTERM_BP_DIRECT	response to cadmium ion	RT	3	7.9E-4	1.2E-2
□	GOTERM_BP_DIRECT	negative regulation of apoptotic process	RT	5	6.5E-3	6.7E-2

Cohort genes: Apart from the variant results from LARP1B and GNA15, in respectively T2DM and asthma, the other relations shown in Supplementary File 2 Figure. 5 were based on abstract co-occurrence or the simple ‘is associated with’ relation type. Next to this, Supplementary File 2 Figure 5 reveals only a single abstract co-occurrence between the cohort acquired genes. This co-occurrence could be assigned to the abstract of the chosen article, Mostafavi (2017).

<u>TOP10 test</u>

The set consisting of the top 10 most referred genes per type of pollutant was used for the first analysis. The program identified 33 individual genes in the set. However, several genes observed to be a false positive result. For example, the PAH (homo sapiens) gen was incorrectly present in all the pollutant sets according to Euretos. The association of all the air pollutants with this gene raised questions. In the end it turned out that the PAH gene was referred to due to the abbreviation of polycyclic aromatic hydrocarbon, PAH. The text mining did not made any distinguish between the gene or air pollutants. The complete set with genes was gone through. More false positive results will be discussed later. Next to this, the ten most referred genes associated with VOC different from 8 to 3 references. Well under the number of references of the other pollutants. These consist of genes with at least ten references for the tenth most referenced gene. The set of 33 genes were displayed in a relation map (Supplementary File 2 Figure 6), to illustrate possible associations between the genes from the same set. The relations in Supplementary File 2 Figure 6 revealed that a total of nine genes exhibited no direct link with at least one other gene from this gene dataset. Next, a relation map was created to visualize the relations between the obtained genes and the five health outcomes of interest (AMI, asthma, CVA, COPD and T2DM) (Supplementary File 2 Figure 7 and 8). The following genes failed to show any relation with at least one of the different health outcomes: DBP, IMPACT and GSTT1 (Supplementary File 2 Figure 7). These will be discussed later. As can be seen in Supplementary File 2 Figure 8a, the figure is packed with mainly lines representing the ‘is associated with’ relation. To give a better overview of the pathways, the ‘is associated with’ relation is removed. This relation type is the least defined type, it represents a possible association without further depth. There is no clear evidence for a possible interaction (f.e. the relation type: forms protein complex with) nor an indication for a positive or negative indirect relation (f.e. the relation type: gene product is biomarker). Therefore, with the removal of this relation type, the better defined relation will be more outstanding. The more in-depth relations can be identified from Supplementary File 2 Figure 8b. Indicating some genes with a considerable impact. The TNF gen, for example, is linked to four out of five health outcomes. The complete list of the relations between the genes and the various health outcomes are displayed in Supplementary Table 3.

Essentially the same relations maps were created to visualize the relations between the TOP10 genes and a list of different types and synonyms of air pollution. In addition to the previous selection of air pollutants, the concepts: air pollution, air pollutant and ambient particulate matter where added to the concept set. According to this relation map, the majority of the complete set of air pollution related concepts were not associated with the TOP10 genes. Data in Supplementary File 2 Figure 9 indicates the associations of the MPO, HMOX1 and HMOX2 gens with the production of nitrogen dioxide for MPO and the production of carbon monoxide for HMOX1 and 2.

Discussion

This study provides some evidence for a possible implementation of the Euretos Knowledge Platform in environmental epidemiology research. We observed an useful application of the enrichment and relation analysis functions of the programme. These functions operate optimal when using data sets consisting of both reasonable quality and quantity. In addition to this, more environmental epidemiology, curated database annotated information is necessary for Euretos to work.

The cohort and candidate gene sets of Mostafavi (2017) were used as positive control. These gene sets, especially the candidate genes, showed us a fraction of the possibilities and opportunities of the EKP. The pathway enrichment of the genes from the candidate gene set resulted in pathways related to oxidation/reduction, detoxification and metabolism of anti-oxidants (glutathione). Genes clusters of this pathways can be identified from later analysis of Supplementary File 2 Figure 2 and 6. However, no significant relations were observed between the candidate set genes and the air pollution concepts. Which indicate that the programme does not association the genes with air pollution. Doubts can be drawn about this, because these genes were selected by Mostafavi (2017) because of their known relationship with air pollution. The pathway analysis was not performed on the cohort gene set, because the set consist of 11 concepts. Which did not resulted in a significant analysis. The quality of the analytical functions of the programme seems related with the quality and quantity of the data sets, to a certain degree. More concepts leads to a better analyses.

Relations in Fig. 2 suggest that a few clusters on genes are present in the candidate gene lists. According to this figure, NFE2L2, AFT4 and AHR play a major role in the control of expression the genes. The candidate gene set contains clusters for both detoxification (SOD1, SOD2, GCLC, GCLM GPX1 and CAT) (Klusek et al., 2018) and metabolism of xenobiotics (GSTP1, GSTM1, CYP1B1, GCLC and GCLM) (Saghaeian Jazi et al., 2021). Genes related to inflammatory responses are marked as candidate gene as well (IL1B, IL6, IL6R and CXCL8). This indicates the possible mechanisms of air pollution exposure. Which is in line with performed studies (Fuertes et al., 2020; Hahad et al., 2020). The same candidate genes did not show many relations towards the different health outcomes. Besides the interleukins, remaining genes with a possible relation were not included in any of the possible clusters in the candidate gene list. The KLF2 (Kruppel like factor) is associated with cardiovascular diseases (Hu et al., 2018) and according to Euretos is the variant gene product of KLF2 related to asthma. SRGAP2 showed a gene product variant for asthma as well. Next to this, was the SRGAP2 gene in the same way related to T2DM. The exposure to air pollution is related to cardiovascular diseases (Fiordelisi et al., 2017; Lederer et al., 2021; Mannucci et al., 2019), despite the absence of relations from the clusters of the candidate lists towards the different health outcomes. The reason why the associations between the exposure to air pollution and cardiovascular diseases are missing is unclear. The cohort set from Mostafavi (2017) consist of only 11 concepts. This cause some standalone, individual genes without relations to other genes and disease or air pollution concepts. More research is needed to fill in the potential missing links and underlying processes of this effect. Mostafavi’s cohort findings are currently not supported by the information already present in the system of Euretos.

The relation maps representing relations among the TOP10 genes illustrated complex networks between several genes and concepts. In common application of the programme, in drugs and disease research, the use of database annotation is favoured over the text-mining. Unfortunately, there is no curated database for environmental epidemiological research data implemented in the system. Therefore, the relations between pollutants and genes are purely based on simple text-mining. After formation of the relations maps (Fig. 6-9), the 'pollutant to gene' relations were looked into in more depth. A significant number of genes were linked incorrectly to the different types of pollutants. The thorough analysis of the genes were performed on the genes with at least 3 references to maximise the efficient use of time. However, it is reasonable to assume that there were more false positive results. These incorrect links, caused by text-mining, can be explained differently per gene. Firstly, some of the genes were abbreviations of other terms applied in the research. One example of these false positive abbreviations was the PAH gene. This gene was marked as related to the pollutants because of the abbreviation for polycyclic aromatic hydrocarbons, PAH. The gene PAH was not mentioned in the full article. Secondly, the name of the gene is commonly used in general writing (the gene IMPACT or TANK) or a part of a longer word (the gene SRI, used in Sri Lanka). Lastly, some of the genes have various aliases, another name for the same gene. These aliases sometimes resulted in a false positive result due to one of the previously mentioned reasons. The complete list of (partly) false positive genes are mentioned in Supplementary Table 4. Despite some false-positive results in the TOP10 gene sets, the remaining true-positive genes displays a useful network consisting of systems related to anti-oxidation, detoxification genes (GSTP1, GSTT1, GSTM1), xenobiotic-metabolizing genes (CYP1A1, AHR, CYP1A2, CYP1B1), different interleukins (IL6, IL1B, CXCL8) and other curated genes with multiple interactions (TNF, NFHB1 and EGFR). In combination with the relations found in Figure 2, several key players in detoxification, oxidation (Fuertes et al., 2020) and other genes related to redox and regulation are observed. This shows the potential application to enhance all different types of gene related research. As already mentioned, sets containing more concepts can expose more systems and pathways. Both the candidate genes and the TOP10 genes illustrated considerable associations between the genes and health outcomes. This includes the less descriptive 'is associated with' relation and more in-depth relations. Firstly, Asthma is related to several interleukins and SRGAP2 (homo sapiens), where variants of the gene product are resulting in abnormal phenotypes. Secondly, some gene products from IL6 and HMOX1 are biomarkers for respectively DMT2 and COPD. Last, genetic changes in the MB (homo sapiens) gen can predispose AMI, whereas CVA could be the result of a specific changes in the TNF gen. The relation map function of the programme can provide new and fast insights with these different relations.

On the other hand, the relations regarding the concepts associated with air pollution are less impactful. Unfortunately, these exist of purely relations where the air pollution concept is a product of a gene (product). HMOX1 and HMOX2 produces CO (by-product) and nitrogen dioxide is formed after nitrite oxidation by myeloperoxidase, the gene product of MPO (homo sapiens). Euretos does not distinguish physiological and environmental substances. Which could lead to problems for both the search engine as well as analytic or relation related functions. The NO_x compounds are physiological signalling molecules, on the other hand, the NO_x molecules are a significantly type of air pollution. Therefore the association of the air pollution NO_x with a type of gene is doubtful in some cases. These compounds with a physiological function do cause issues with the analytics as well. For example, the CO related to the physiological carbon monoxide instead of the air pollution variant. In addition, Supplementary File 2 Figure 5c identified a few relations, the abstract co-occurrence between AHCYL2, MTMR2 and 'air pollutants' and 'nitrogen oxides'. These co-occurrences could be tracked back to the Mostafavi (2017) article. To prevent this type of circular reasoning, relation analytics could be performed before publication.

The EKP was implemented to give new insights into the pathway and genes involved in air pollution exposure and different health outcomes. The AI tool can provide insights into systems biology to some extent. The data set creation based on curated databases allows the user to compare and analysis the empirical derived data. Next to this, the program provided additional evidence for the potential therapeutic targets for diseases such as Autosomal Dominant Polycystic Kidney Disease (ADPKD) (Malas et al., 2020). The (pathway) analysis option can provide new insights on related pathways and regulations. In addition, this option can rank the users selection of concepts based on expression and interaction. The relation map option provides the user with a visual representation of the relations and association of the concepts of interest. This function of Euretos is used in other studies (Malas et al., 2019) to integrate semantic information within a knowledge graph, which describes known relationships between biomedical concepts (e.g drugs, diseases or genes). However, the programme needs to be more optimized for environmental epidemiological research, to provide a first stap in consistence interpretation of empirical (gen)omics data. More in-depth evaluation of the technical aspects of the programme is discussed later.

Classification of the concepts in Euretos is resulting in varied possibilities in the programme. The concept 'air pollution' on itself is classified as phenomena. This classification offers fewer additional filter options compared to the classification 'genes' or 'molecules'. Where asthma, and other health outcomes, feature significantly, beneficial available filters related to the search option and dataset creation. For instance, regulations of proteins (Dys-, up- and downregulated) or gene cell type expression (Transcripts per million). Most important, the category gene is not available for the concept 'air pollution', which caused a limit for further search and creation options. Another variant on the concept 'air pollution' was used to search for associated genes. The concept 'air pollutants' was used for further set creation. With the use of only the concept 'air pollutants' in the EKP, we were doubtful about the fact that this concept is not related to majority of the known genes. The list with genes related to 'air pollutants' was missing for example several interleukins, CAT and NOX1. Genes that were present in the candidate genes from Mostafavi (2017) and were observed to show relations with other genes in the same set (Fig. 2). Therefore we had to add detail to this concept to conduct a more significant experiment. These detailed concepts provided a broader perspective on the genes associated with air pollution in general, compared to the narrow approach of the single concept 'air pollution'. With the broader approach on air pollution, less significant information should be missed out by the programme. Therefore a broad approach was the optimal way to describe the exposure side of the experiment. The health outcome concepts (AMI, asthma, CVA, COPD and T2DM) were sufficiently elaborate to function properly in the programme.

The relations among the genes and between the health outcomes were removed on forehand. In our approach, the advantage of improved interpretability was chosen over the completeness of the figures. Preserving all the relations, lead to figures that were almost impossible to interpretate due to the low legibility. However, network statistics and system analysis could be used for the interpretation for the more complex figure. Existing studies in both humans and model organisms highlighted the complexity of genomic information flow, together with the interactive networks in biological mechanisms and the onset and development of diseases (Sun & Hu, 2016). Insurmountable, this lead to the loss of biological clusters between the genes and among the different diseases. Next to the removal of the relations among the concepts of the same set, the 'is associated with' relation was determined as least defined. This assumption was made because this relation type was less specific compared to other relation types (e.g. forms protein complex with or gene product variant results in abnormal). The relegation of the relation type 'is associated with', made it possible to hide this type of relation to improve the interpretation of the figures. This assumption limited this study, due to the fact that numerous environmental epidemiological studies cannot define specific relations. These studies can provide new possible association, however these signals are too mild and uncertain to provide a more in depth relation.

Uploading of sets requires some manual assist of the user. The programme is very sensitive with regard to extra notations on the genes or very specific in the types of aliases per genes. In our case, this Il8 was substituted for CXCL8, an alias for the interleukin-8. The programme does not make use of all the aliases of genes and therefore decisions have to be made by the user. In our case, the Il8 was mentioned in the Mostafavi (2017) article, however is not recognised by the programme as gene. Therefore during uploading, an extra clarification was needed to be able to proceed with the data set. The CXCL8 (homo sapiens) gene was present in the programme and thus the substitution for the Il8 gene. Which is one off the aliases of Il8. Other manual adaptations could be regarding extra notations, symbols or punctuation marks.

The search engine of the platform was used to create data sets containing genes. The TOP10 test was done to run the analytics of the programme. A maximum of 10 referred genes per type of air pollution was selected due to a limit of concepts for some analytic options in the tool. The restricted maximum of 400 concepts in a relation map caused the selection criteria was set on the ten most referred genes per pollutant. An elevation of the numerical limit could enlarge the potential set size used in the relation map option of Euretos. However, in our study, this numerical limit could be replaced by a relative limit based on the numbers of references in list. This limitation caused for some debatable results. Some air pollution species resulted in significantly more results compared to others. Where 'particulate matter' resulted in over 500 references for some genes, 'volatile organic compounds', on the other hand, resulted in less than 10 references for the most referred genes. This could cause that the CYP1A1 (homo sapiens) gen, with 72 unique references (number 11 of PM references), was excluded from the set, while the SARS1 (homo sapiens) gen, with merely four references (number 6 of VOC references), was included. Perhaps, a relative limit based on the number of reference fitted the data set creation better. To perform analyses on the derived gene lists, the lists were combined into one set. The main drawback of this approach is the loss of magnitude. Genes that were referenced in six out of the eight species, were on the same level as genes that were associated for one type of air pollution. Genes that could have a larger impact, compared to substantially less impactful genes lost their magnitude due to this approach. Nevertheless, this method of set creation by making use of the search engine of the programme was rejected. During the process it was found out that the text-mining based data sets were containing several false-positive results. Clearly, the obtained search derived data sets should be validated before future analysis to overcome this problem. If these problematic concepts are fixed, literature derived gene sets can be used to test the analytical functions in the programme.

The major advantage from Euretos over DAVID (^{3}) is the refresh on literature based data. This provides the user with recent data published on different life science and biomedical search engines, alongside curated data bases. While the databases of DAVID are criticized for being outdated (Zhou et al., 2019). Euretos integrates over 200 biomedical knowledge sources, which can be distinguished in: life-science databases, textual and publication sources, and semantic and ontological sources. The programme allows the user to search within the programme in which the analysis and pathway enrichment can be performed. The analysis and enrichment can be conducted on the obtained search results. Contrary, DAVID consist only of the option to upload gene sets. Based on the pathway analysis, we compared DAVID annotation tool with the Euretos Knowledge Platform. Besides the pathway regarding 'aging', both enrichment analysis resulted in pathways concerning oxidation, signalling and metabolic pathways. As mentioned before, DAVID provides the user with this extra tool to look at the internal relationships of clustered terms and makes the biological interpretation more focused at a group level. In Euretos, this function can be imitated by saving the pathways as a set and exporting them to the relation map option of the AI tool. The different strategy of gene set analysis with the Euretos Knowledge Platform resulted in more visual appealing approach, compared to the alternative DAVID. Toonen *et al.* used both Euretos and IPA for pathway analysis and exploration of metabolite-phenotype links. They suggest that the Euretos platform performed a successful analysis. Taken together, both strategies and studies suggest that the pathway analysis of the EKP is performing well, on the condition that the gene sets needs a certain quality. At the end, the pathway analysis cluster tool of DAVID causes the programme to transcend Euretos. However, this tool can be mimic by the relation map option. Which by itself, provides the user with a more visual approach to look into genetic relations, on pathway networks and individual gene levels.

The main drawback is the lack of environmental epidemiological data available in the programme. In addition, the programme operate with abstract scanning instead of using the full text for analysis. The way the epidemiologic articles are organized, did lead to some difficulties. Normally, environmental epidemiology papers do not mention all the findings in the abstract. The abstract is used for the results with the best evidence. Some major findings, which were left out of the abstract, are not referred to by the programme due to this abstract text-mining. Other major limits of the programme are the potential circular reasoning and possible false positive annotations. The circular reasoning with text co-occurrences caused the condition in with own results can be used to reinforce the same results through Euretos. The obtained false positive annotations are extensively described already. After the completing this study, potential hand-operated solutions are applied by the employees of Euretos.

During the process, other serious challenges and limitations were faced. In the search engine of the EKP, the terminology and incorrect relations could cause some problems, resulting in incorrect sets. First of all, the synonym lists of the concepts are not always complete. This results in concepts that are not related according to the programme, concepts that are related while using other search engines such as PubMed. The selection of articles obtained from the EKP differs from the selection obtained from PubMed, the latter consist of more articles corresponding to the same search terms. This is principally caused by the fact that abstract text-mining is resulting in less search results compared to full test scanning. Besides, the abbreviation for particulate matter, PM, is not recognised by Euretos as abbreviation for the corresponding type of air pollutant. These search issues combined resulting in incomplete search results in the EKP.

It is reasonable to assume that if the curated epidemiological data bases are expended or the signals of empirical epidemiolocal data are amplified, the analysis of the EKP can provide a framework for environmental epidemiolocal research. Euretos is a powerful tool for drugs and disease research, however is lagging behind on the field of environmental epidemiology. The EKP is hindered by the lack of curated epidemiological data, frequent present of type 1 errors and potential circular reasoning.

Conclusion

The Euretos Knowledge Platform showed serious potential for implementation in environmental epidemiology research. In addition to the useability for enrichment analysis, the AI tool provide the user with additional other analytic features. Which can be performed on the same set of data. The user case showed the potential to broader the performed research with pathway enrichment and network analysis options. The data search experiment, TOP10 test, showed less encouraging results. Circular reasoning and false positive annotation are to major limits in the programme.

The analytic options of the programme function optimal after uploading experimental data to avoid false positive results in the starting datasets. Next to this, the implementation of curated environmental epidemiological data bases will provide better search results compared to the literature search alone. Further work is planned, using research data of the epidemiological research group (population health sciences) of the Institute of Research Assessment Science (IRAS).

More research and extra changes are required for data search and desirable for the user case to implement Euretos directly. Changes are necessary for the programme to create an AI tool suitable for epidemiological applications. Minor changes could make the EKP feasible for widespread use in different fields of research.

Sources

[1] Euretos platform. 2021. https://www.euretos.com/home

[2] Euretos help-center. 2021. https://knowledge.euretos.com/help-center

[3] Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.8. 2021. https://david.ncifcrf.gov/

Literature

Cesaroni, G., Forastiere, F., Stafoggia, M., Andersen, Z. J., Badaloni, C., Beelen, R., Caracciolo, B., de Faire, U., Erbel, R., Eriksen, K. T., Fratiglioni, L., Galassi, C., Hampel, R., Heier, M., Hennig, F., Hilding, A., Hoffmann, B., Houthuijs, D., Jockel, K.-H., … Peters, A. (2014). Long term exposure to ambient air pollution and incidence of acute coronary events: prospective cohort study and meta-analysis in 11 European cohorts from the ESCAPE Project. *BMJ*, *348*(jan21 3), f7412–f7412. https://doi.org/10.1136/bmj.f7412

Daiber, A., Lelieveld, J., Steven, S., Oelze, M., Kröller-Schön, S., Sørensen, M., & Münzel, T. (2019). The “exposome” concept – how environmental risk factors influence cardiovascular health. *Acta Biochimica Polonica*. https://doi.org/10.18388/abp.2019_2853

Dehghan, A. (2018). *Genome-Wide Association Studies* (pp. 37–49). https://doi.org/10.1007/978-1-4939-7868-7_4

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., & Lempicki, R. A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology*, *4*(5), P3.

Everson, T. M., Vives-Usano, M., Seyve, E., Cardenas, A., Lacasaña, M., Craig, J. M., Lesseur, C., Baker, E. R., Fernandez-Jimenez, N., Heude, B., Perron, P., Gónzalez-Alzaga, B., Halliday, J., Deyssenroth, M. A., Karagas, M. R., Íñiguez, C., Bouchard, L., Carmona-Sáez, P., Loke, Y. J., … Bustamante, M. (2021). Placental DNA methylation signatures of maternal smoking during pregnancy and potential impacts on fetal growth. *Nature Communications*, *12*(1), 5095. https://doi.org/10.1038/s41467-021-24558-y

Fiordelisi, A., Piscitelli, P., Trimarco, B., Coscioni, E., Iaccarino, G., & Sorriento, D. (2017). The mechanisms of air pollution and particulate matter in cardiovascular diseases. *Heart Failure Reviews*, *22*(3), 337–347. https://doi.org/10.1007/s10741-017-9606-7

Fuertes, E., van der Plaat, D. A., & Minelli, C. (2020). Antioxidant genes and susceptibility to air pollution for respiratory and cardiovascular health. *Free Radical Biology and Medicine*, *151*, 88–98. https://doi.org/10.1016/j.freeradbiomed.2020.01.181

Gehring, U., Wijga, A. H., Hoek, G., Bellander, T., Berdel, D., Brüske, I., Fuertes, E., Gruzieva, O., Heinrich, J., Hoffmann, B., de Jongste, J. C., Klümper, C., Koppelman, G. H., Korek, M., Krämer, U., Maier, D., Melén, E., Pershagen, G., Postma, D. S., … Brunekreef, B. (2015). Exposure to air pollution and development of asthma and rhinoconjunctivitis throughout childhood and adolescence: a population-based birth cohort study. *The Lancet Respiratory Medicine*, *3*(12), 933–942. https://doi.org/10.1016/S2213-2600(15)00426-9

Guan, W.-J., Zheng, X.-Y., Chung, K. F., & Zhong, N.-S. (2016). Impact of air pollution on the burden of chronic respiratory diseases in China: time for urgent action. *The Lancet*, *388*(10054), 1939–1951. https://doi.org/10.1016/S0140-6736(16)31597-5

Hahad, O., Lelieveld, J., Birklein, F., Lieb, K., Daiber, A., & Münzel, T. (2020). Ambient Air Pollution Increases the Risk of Cerebrovascular and Neuropsychiatric Disorders through Induction of Inflammation and Oxidative Stress. *International Journal of Molecular Sciences*, *21*(12), 4306. https://doi.org/10.3390/ijms21124306

Hettne, K. M., Thompson, M., van Haagen, H. H. H. B. M., van der Horst, E., Kaliyaperumal, R., Mina, E., Tatum, Z., Laros, J. F. J., van Mulligen, E. M., Schuemie, M., Aten, E., Li, T. S., Bruskiewich, R., Good, B. M., Su, A. I., Kors, J. A., den Dunnen, J., van Ommen, G.-J. B., Roos, M., … Schultes, E. A. (2016). The Implicitome: A Resource for Rationalizing Gene-Disease Associations. *PLOS ONE*, *11*(2), e0149621. https://doi.org/10.1371/journal.pone.0149621

Hu, W., Lu, H., Zhang, J., Fan, Y., Chang, Z., Liang, W., Wang, H., Zhu, T., Garcia-Barrio, M. T., Peng, D., Chen, Y. E., & Guo, Y. (2018). Krüppel-like factor 14, a coronary artery disease associated transcription factor, inhibits endothelial inflammation via NF-κB signaling pathway. *Atherosclerosis*, *278*, 39–48. https://doi.org/10.1016/j.atherosclerosis.2018.09.018

Huang, Y.-C. T., Karoly, E. D., Dailey, L. A., Schmitt, M. T., Silbajoris, R., Graff, D. W., & Devlin, R. B. (2011). Comparison of Gene Expression Profiles Induced By Coarse, Fine, and Ultrafine Particulate Matter. *Journal of Toxicology and Environmental Health, Part A*, *74*(5), 296–312. https://doi.org/10.1080/15287394.2010.516238

Klusek, J., Nasierowska-Guttmejer, A., Kowalik, A., Wawrzycka, I., Lewitowicz, P., Chrapek, M., & Głuszek, S. (2018). GSTM1, GSTT1, and GSTP1 polymorphisms and colorectal cancer risk in Polish nonsmokers. *Oncotarget*, *9*(30), 21224–21230. https://doi.org/10.18632/oncotarget.25031

Lederer, A. M., Fredriksen, P. M., Nkeh-Chungag, B. N., Everson, F., Strijdom, H., de Boever, P., & Goswami, N. (2021). Cardiovascular effects of air pollution: current evidence from animal and human studies. *American Journal of Physiology-Heart and Circulatory Physiology*, *320*(4), H1417–H1439. https://doi.org/10.1152/ajpheart.00706.2020

Li, Y.-L., Chuang, T.-W., Chang, P., Lin, L.-Y., Su, C.-T., Chien, L.-N., & Chiou, H.-Y. (2021). Long-term exposure to ozone and sulfur dioxide increases the incidence of type 2 diabetes mellitus among aged 30 to 50 adult population. *Environmental Research*, *194*, 110624. https://doi.org/10.1016/j.envres.2020.110624

Liu, S., Huang, Q., Zhang, X., Dong, W., Zhang, W., Wu, S., Yang, D., Nan, B., Zhang, J., Shen, H., Guo, X., & Deng, F. (2021). Cardiorespiratory Effects of Indoor Ozone Exposure Associated with Changes in Metabolic Profiles among Children: A Repeated-Measure Panel Study. *The Innovation*, *2*(1), 100087. https://doi.org/10.1016/j.xinn.2021.100087

Malas, T. B., Leonhard, W. N., Bange, H., Granchi, Z., Hettne, K. M., van Westen, G. J. P., Price, L. S., ’t Hoen, P. A. C., & Peters, D. J. M. (2020). Prioritization of novel ADPKD drug candidates from disease-stage specific gene expression profiles. *EBioMedicine*, *51*, 102585. https://doi.org/10.1016/j.ebiom.2019.11.046

Malas, T. B., Vlietstra, W. J., Kudrin, R., Starikov, S., Charrout, M., Roos, M., Peters, D. J. M., Kors, J. A., Vos, R., ‘t Hoen, P. A. C., van Mulligen, E. M., & Hettne, K. M. (2019). Drug prioritization using the semantic properties of a knowledge graph. *Scientific Reports*, *9*(1), 6281. https://doi.org/10.1038/s41598-019-42806-6

Mannucci, P. M., Harari, S., & Franchini, M. (2019). Novel evidence for a greater burden of ambient air pollution on cardiovascular disease. *Haematologica*, *104*(12), 2349–2357. https://doi.org/10.3324/haematol.2019.225086

Merid, S. K., Bustamante, M., Standl, M., Sunyer, J., Heinrich, J., Lemonnier, N., Aguilar, D., Antó, J. M., Bousquet, J., Santa-Marina, L., Lertxundi, A., Bergström, A., Kull, I., Wheelock, Å. M., Koppelman, G. H., Melén, E., & Gruzieva, O. (2021). Integration of gene expression and DNA methylation identifies epigenetically controlled modules related to PM2.5 exposure. *Environment International*, *146*, 106248. https://doi.org/10.1016/j.envint.2020.106248

Mostafavi, N., Vlaanderen, J., Portengen, L., Chadeau-Hyam, M., Modig, L., Palli, D., Bergdahl, I. A., Brunekreef, B., Vineis, P., Hebels, D. G. A. J., Kleinjans, J. C. S., Krogh, V., Hoek, G., Georgiadis, P., Kyrtopoulos, S. ., & Vermeulen, R. (2017). Associations Between Genome-wide Gene Expression and Ambient Nitrogen Oxides. *Epidemiology*, *28*(3), 320–328. https://doi.org/10.1097/EDE.0000000000000628

Park, J., Kim, H.-J., Lee, C.-H., Lee, C. H., & Lee, H. W. (2021). Impact of long-term exposure to ambient air pollution on the incidence of chronic obstructive pulmonary disease: A systematic review and meta-analysis. *Environmental Research*, *194*, 110703. https://doi.org/10.1016/j.envres.2020.110703

Saghaeian Jazi, M., Mohammadi, S., Zare Ebrahimabad, M., Sedighi, S., Abdolahi, N., Tabarraei, A., & Yazdani, Y. (2021). Genetic variation in CYP1A1 and AHRR genes increase the risk of *systemic lupus erythematosus* and exacerbate disease severity in smoker patients. *Journal of Biochemical and Molecular Toxicology*, *35*(12). https://doi.org/10.1002/jbt.22916

Shah, A. S., Langrish, J. P., Nair, H., McAllister, D. A., Hunter, A. L., Donaldson, K., Newby, D. E., & Mills, N. L. (2013). Global association of air pollution and heart failure: a systematic review and meta-analysis. *The Lancet*, *382*(9897), 1039–1048. https://doi.org/10.1016/S0140-6736(13)60898-3

Smit-McBride, Z., Nguyen, J., Elliott, G. W., Wang, Z., McBride, R. A., Nguyen, A. T., Oltjen, S. L., Yiu, G., Thomasy, S. M., Pinkerton, K. E., Lee, E. S., Cunefare, D., Farsiu, S., & Morse, L. S. (2018). Effects of aging and environmental tobacco smoke exposure on ocular and plasma circulatory microRNAs in the Rhesus macaque. *Molecular Vision*, *24*, 633–646.

Sun, Y. v., & Hu, Y.-J. (2016). *Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of Complex Human Diseases* (pp. 147–190). https://doi.org/10.1016/bs.adgen.2015.11.004

Toonen, L. J. A., Overzier, M., Evers, M. M., Leon, L. G., van der Zeeuw, S. A. J., Mei, H., Kielbasa, S. M., Goeman, J. J., Hettne, K. M., Magnusson, O. Th., Poirel, M., Seyer, A., 't Hoen, P. A. C., & van Roon-Mom, W. M. C. (2018). Transcriptional profiling and biomarker identification reveal tissue specific effects of expanded ataxin-3 in a spinocerebellar ataxia type 3 mouse model. *Molecular Neurodegeneration*, *13*(1), 31. https://doi.org/10.1186/s13024-018-0261-9

Wittkopp, S., Staimer, N., Tjoa, T., Stinchcombe, T., Daher, N., Schauer, J. J., Shafer, M. M., Sioutas, C., Gillen, D. L., & Delfino, R. J. (2016). Nrf2-related gene expression and exposure to traffic-related air pollution in elderly subjects with cardiovascular disease: An exploratory panel study. *Journal of Exposure Science & Environmental Epidemiology*, *26*(2), 141–149. https://doi.org/10.1038/jes.2014.84

Xing, D. F., Xu, C. D., Liao, X. Y., Xing, T. Y., Cheng, S. P., Hu, M. G., & Wang, J. X. (2019). Spatial association between outdoor air pollution and lung cancer incidence in China. *BMC Public Health*, *19*(1), 1377. https://doi.org/10.1186/s12889-019-7740-y

Yang, D., Yang, X., Deng, F., & Guo, X. (2017). *Ambient Air Pollution and Biomarkers of Health Effect* (pp. 59–102). https://doi.org/10.1007/978-981-10-5657-4_4

Yang, T.-Y., Hsu, L.-I., Chiu, A. W., Pu, Y.-S., Wang, S.-H., Liao, Y.-T., Wu, M.-M., Wang, Y.-H., Chang, C.-H., Lee, T.-C., & Chen, C.-J. (2014). Comparison of genome-wide DNA methylation in urothelial carcinomas of patients with and without arsenic exposure. *Environmental Research*, *128*, 57–63. https://doi.org/10.1016/j.envres.2013.10.006

Zhao, K., Li, J., Du, C., Zhang, Q., Guo, Y., & Yang, M. (2021). Ambient fine particulate matter of diameter ≤ 2.5 μm and risk of hemorrhagic stroke: a systemic review and meta-analysis of cohort studies. *Environmental Science and Pollution Research*, *28*(17), 20970–20980. https://doi.org/10.1007/s11356-021-13074-7

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., Benner, C., & Chanda, S. K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communications*, *10*(1), 1523. https://doi.org/10.1038/s41467-019-09234-6

Supplementary File 1 (Euretos appendix): *See additional file*

Supplementary File 2 (Figures): *See additional file*

Supplementary Table 1. Pathway analysis of the Mostafavi (2017) candidate genes. Listing the concept (pathway) name, number of concepts of the selected set represented in the relevant pathway, column for type indication (Enriched or Depleted), number of concepts in the category (pathway) and the corresponding Fisher's exact test p-values.

See additional file

Supplementary Table 2. Overview of different relation types of Supplementary File 2 Figure 2.

AFT4	controls expression of	AHR
AFT4	interacts with	NFE2L2
AFT4	controls expression of	CXCL8
AFT4	controls expression of	CCL2
AHR	interacts with	HSPA8
AHR	controls expression of	CYP1A1
AHR	binds with	CYP1A1
AHR	stimulates	CYP1A1
AHR	interacts with	CYP1A1
AHR	interacts with	CYP1B1
AHR	controls expression of	CYP1B1
AHR	stimulates	CYP1B1
AHR	binds with	CYP1B1
AHR	interacts with	NFE2L2
AHR	binds with	CXCL8
CAT	interacts with	SOD2
CAT	interacts with	SOD1
CXCL8	forms protein complex with	IL6R
CYP1A1	stimulates	AHR
CYP1A1	inhibits	AHR
CYP1A1	coexists with	GSTM1
CYP1B1	stimulates	AHR
CYP1B1	inhibits	AHR
CYP1B1	stimulates	CYP1A1
GCLC	interacts with	GCLM
GCLC	forms protein complex with	GCLM
GSTM1	coexists with	GSTP1
GSTP1	interacts with	SOD1
HSPA8	interacts with	SOD1
IL1B	controls expression of	IL6
IL1B	stimulates	IL6
IL1B	forms protein complex with	IL6R
IL1B	stimulates	NOS2
IL6	binds with	IL6R
IL6	forms protein complex with	IL6R
IL6	interacts with	IL6R
IL6	stimulates	IL6R
IL6	interacts with	IL1B
IL6R	binds with	IL6
NFE2L2	controls expression of	SOD2
NFE2L2	stimulates	SOD2
NFE2L2	controls expression of	SOD1

NFE2L2	stimulates	SOD1
NFE2L2	controls expression of	IL6
NFE2L2	controls expression of	CAT
NFE2L2	stimulates	CAT
NFE2L2	binds with	GCLM
NFE2L2	binds with	GCLC
NFE2L2	controls expression of	GCLC
NFE2L2	inhibits	GCLC
NFE2L2	stimulates	GCLC
NFE2L2	controls expression of	TXNRD1
NFE2L2	controls expression of	HMOX1
NFE2L2	inhibits	HMOX1
NFE2L2	stimulates	HMOX1
NFE2L2	binds with	HMOX1
SOD1	catalysis precedes	GPX1
SOD1	interacts with	SOD2
SOD1	catalysis precedes	GSTP1
SOD1	catalysis precedes	CAT
SOD2	catalysis precedes	GPX1
SOD2	is a	SOD1
SOD2	catalysis precedes	CAT

Supplementary Table 3. Overview of different relation types (-is associated with) of Supplementary File 2 Figure 8b.

Asthma	produces	TNF
CRP	affects	Asthma
CRP	affects	COPD
CVA	is manifestation of	CBS
HMOX1	gene product is biomarker type	COPD
IL1B	gene product variant results in abnormal	Asthma
IL6	gene product variant results in abnormal	Asthma
IL6	gene product is biomarker type	T2DM
MB	predisposes	AMI
MPO	gene product variant results in abnormal	COPD
NFKB1	gene product variant results in abnormal	Asthma
NFKB1	gene product is biomarker type	COPD
T2DM	produces	TNF
TNF	gene product is biomarker type	Asthma
TNF	causes	Asthma
TNF	affects	Asthma
TNF	predisposes	Asthma
TNF	predisposes	CVA
TNF	affects	COPD
TNF	causes	COPD
TNF	causes	T2DM
TNF	augments	T2DM
VEGFA	gene product variant results in abnormal	T2DM

Supplementary Table 4. False positive results obtained during the data set creation (Search→ Particulate matter → Genes) (02-06-2021)

Gene		Reason of false positive result
ABO	*	ABO Blood group
ADM	*	Atmospheric dispersion models
AFM		Atomic force microscopy
AHR	*	Adjusted hazard ratio
APCS		Part of PCA-APCS (principal component analysis with absolute principal component scores)
ATR		Part of ATR- FT-IR (attenuated total reflection Fourier transform infrared)
BCR		Bureau of reference (Road dust (trace element))
BMF		Biomass fuel
CAMP		cAMP (cyclic AMP), camp (place with tents etc)
CAPS		Concentrated ambient particles, Cooking and Pneumonia Study
CAST		Cast-iron, cast (urinary formed secretion), part of open-cast mining (type of mining)
CCK		Cell counting kit
CCS		Congestion Charging scheme (Daily charge if you drive, London)
CFD		Computational fluid dynamics, Chinese fine dust
COIL		Mosquito coil (mosquito repellent)
CPM		Condensable particulate matter
DAP		Dissolved aqueous phase
DBP		Diastolic blood pressure
EBP		Electrostatic-bag precipitator (removes particles from a gas)
ECD		Exceedance concentrations days
ERAS		Environmental risk assessment
FEV		Forced expiratory volume
GAA		Greater Athens Area
HBM		Hierarchical Bayesian Model
HGF		Human gingival fibroblasts
HPD		Heavily polluted days, high pollution district
IMPACT		Impact (influence)
LPA	*	Low polluted area, low-PB-accumulation
MBP	*	Mean blood pressure
MMD		Mass median diameter
MMUT	*	Aliases: MCM→ (Bayesian) multicity multi-outcome, part of Al-MCM-41(member of mesoporous molecular sieves family (material))
MOS		Model outcome statistic, margin of safety
MPI		Term in air quality monitoring, multidimensional poverty index, part of MPI/TOF-MS (multiphoton ionization time-of-flight mass spectrometry)
MSSD		Something to do with heart rate
MTTP		Aliases: ABL→ Atmospheric boundary layer
NHS		Nurses' Health study
NMB		Normalized median bias

NPS		Nanoparticles (NPs)
NPY	*	2-nitrogenpyrene (nitro-PAH isomer)
NRAP		Near roadway air pollution
PAH		Polycyclic aromatic hydrocarbons, pulmonary arterial hypertension
PAM		Personal air quality monitor
PFAS		Per- and polyfluoroalkyl substances
PGF*	*	8 iso prostane (8 iso-PGF-(2a))
POR		Prevalence odds ratio
PRAME		Aliases: MAPE → Mean absolute percentage error
PSD		Particle size distribution, passive sampling device
RPE		Retinal pigment epithelial
SRI		Part of Sri Lanka
SRM		Standard reference materials (dust/diesel numbers)
TANK		Storage chamber
TERT		Part of special type of cell line, part of some chemical names (e.g. tert-butyl ether)
TPR		Temperature programmed reduction, total peripheral resistances, true prediction rate
TRAP		Traffic related air pollution
*These gene relations are partially incorrect. For these genes there were true positive results, next to the false positive relations.		