A gene regulation model reveals an ancestral adaptation response to particulate exposure triggered by nanomaterials

Giusy del Giudice
Tampere University
Angela Serra
Tampere University
Laura Saarimäki
Tampere University https://orcid.org/0000-0003-1996-286X
Konstantinos Kotsis
University College Dublin
Ian Rouse
University College Dublin
Anais Colibaba
University College Dublin
Karolina Jagiello
QSARLab Ltd.
Alicja Mikolajczyk
QSAR Lab Ltd
Anastasios Papadiamantis
University of Birmingham
Natasha Sanabria
National Institute for Occupational Health, A Division of National Health Laboratory Services
Pia Kinaret
University of Helsinki
Evangelos Voyiatzis
Novamechanics Ltd
Georgia Melagraki
Nanoinformatics Department, NovaMechanics Ltd, Nicosia
Antreas Afantitis
Nanoinformatics Department, NovaMechanics Ltd, Nicosia
Kaido Tamm
Department of Chemistry, University of Tartu, Tartu
Tomasz Puzyn

QSARLab Ltd.
Mary Gulumian
National Institute for Occupational Health, A Division of National Health Laboratory Services
Vladimir Lobaskin
University College Dublin
Iseult Lynch
University of Birmingham https://orcid.org/0000-0003-4250-4584
Antonio Federico
University of Tampere
Dario Greco (dario.greco@tuni.fi)
Tampere University https://orcid.org/0000-0001-9195-9003

Research Article

Keywords:

Posted Date: April 12th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1547187/v1

A gene regulation model reveals an ancestral adaptation response to particulate exposure triggered by nanomaterials

del Giudice, G. [1,2,3], Serra, A. [1,2,3], Saarimäki, L. A. [1,2,3], Kotsis, K .[4], Rouse, I .[4], Colibaba, S. A. [4], Jagiello, K. [5,6], Mikolajczyk, A. [5,6], Papadiamantis, A. G. [7,8], Sanabria, N. [9], Kinaret, P.A.S. [1,2,3,10], Voyiatzis, E.[8], Melagraki, G. [8], Afantitis, A. [8], Tämm, K.[11], Puzyn, T. [5,6], Gulumian, M. [9,12,13], Lobaskin, V. [4], Lynch, I. [7], Federico, A. [1,2,3] and Greco, D. [1,2,3,10]

1. *Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland*
2. *BioMediTech Institute, Tampere University, Tampere, Finland*
3. *FHAIVE, Finnish Hub for Development and Validation of Integrated Approaches, Tampere University, Tampere, Finland*
4. *School of Physics, University College Dublin, Belfield, Dublin 4, Ireland*
5. *Group of Environmental Chemoinformatics, Faculty of Chemistry, University of Gdańsk, Wita Stwosza 63, 80-308 Gdańsk, Poland*
6. *QSAR Lab Ltd, Trzy lipy 3, 80-172 Gdańsk, Poland*
7. *School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham B15 2TT, UK*
8. *Novamechanics Ltd., 1065 Nicosia, Cyprus*
9. *National Institute for Occupational Health, A Division of National Health Laboratory Services, Johannesburg, South Africa*
10. *Institute of Biotechnology, University of Helsinki, Helsinki, Finland*
11. *Institute of Chemistry, University of Tartu, Ravila 14a, Tartu 50411, Estonia*
12. *Haematology and Molecular Medicine Department, School of Pathology, University of the Witwatersrand, Johannesburg, South Africa*
13. *Water Research Group, Unit for Environmental Sciences and Management, North West University, Potchefstroom, South Africa*

* Corresponding author: dario.greco@tuni.fi

Introduction/Abstract

Toxicogenomics aims at characterising the mechanism of action (MOA) of environmental exposures, and often relies on transcriptomics to investigate the responses of exposed biological systems. However, the identification of shared toxicogenomics-derived signatures across exposed biological systems is hampered by the complexity and heterogeneity of transcriptomics data.

Given the lack of a clear transcriptomic signature of exposure, we hypothesise that common patterns of gene regulation could explain the response to engineered nanomaterials (ENM) across biological systems, disentangling the complexity of their MOA. We performed meta-analysis of a large collection of transcriptomics data from various ENM exposure studies and identified deregulation of immune functions as a prominent response across different ENM exposures. This pattern of transcriptional deregulation differed significantly from exposure to drugs. By investigating the promoter regions of genes frequently altered both *in vitro* and *in vivo* following exposure to ENM, we identified a set of binding sites for zinc finger transcription factors C_2H_2 involved in chromatin remodelling and immunomodulation. We further demonstrate that this gene regulatory model also underlies the transcriptomic MOA in non-mammal species of ecotoxicological interest exposed to ENMs, suggesting that it may be part of the innate immune system conserved by natural selection.

Main

Unravelling the potential mechanisms of toxicity of engineered nanomaterials (ENMs) is a major objective of nano-toxicogenomics, and to date a plethora of transcriptomics data have been generated for this purpose [1]. Multiple efforts aim to use mechanistic data to find commonalities among environmental exposures, hence facilitating the grouping of ENM by MOA streamlining their safety assessment[2]. However, due to the high complexity and variability of biological systems, transcriptomic profiles are often heterogeneous, further biased by the typically small-scale datasets for ENMs (low numbers of ENM, limited doses and time points). This results in a myriad of toxicogenomic signatures, which show only a moderate degree of similarity to each other (Figure S1).

Furthermore, the high heterogeneity of transcriptomic profiles prevents from finding relationships between the biological mechanisms observed *in vitro* with real life exposure scenarios. However, improving the *in vitro-in vivo* extrapolation (IVIVE) requires the definition of better models able to transpose the mechanisms of toxicity detected at shorter observations *in vitro* (hours or days) to long time scales (weeks or months) observed *in vivo*[3].

In contrast to the specificity of the transcriptional response to environmental signals, the regulatory circuits controlling gene expression are usually well conserved across species [4]. Furthermore transcriptional changes usually follow an "impulse-like" kinetic, whereas regulation of transcription is achieved by multiple layers of sustained genetic and epigenetic signals [5]. For this reason, gene regulation is associated with more stable physiopathological changes, and can be more reliable in detecting alterations induced by the exposure than the single gene expression.

Here, we leveraged this concept by searching for common patterns of gene regulation underlying the adaptive response of multiple biological systems exposed to a variety of ENM. We analysed the most comprehensive transcriptomics data collection for ENMs, in which the expression of 3,676 genes is measured across 584 experimental conditions [6–8]. This data space includes multiple human and mouse cell types and tissues, both *in vitro* and *in vivo*, exposed to ENM varying in chemistry, geometry and size.

In toxicogenomics studies, each significant variation observed between the exposed and unexposed samples is considered to be induced by the test compound, and it is commonly defined as its MOA signature. Hence, we first performed differential expression analysis for each dataset in the ENM collection, and computed the pairwise similarity among them. Our results suggest that the transcriptomic signatures of ENM exposures are substantially dissimilar from each other (Figure S1). This corroborates the hypothesis that transcriptional MOAs do not reveal common patterns of adaptation due to the high diversity of the exposures as well as the exposed biological systems [9].

Consequently, we utilised an ensemble meta-analytical approach to highlight robust patterns of molecular alteration detected in multiple ENM exposure systems *in vivo* and *in vitro*. This allowed us to rank the list of 3,676 genes according to the robustness of their expression patterns across the conditions and datasets (Table S1).

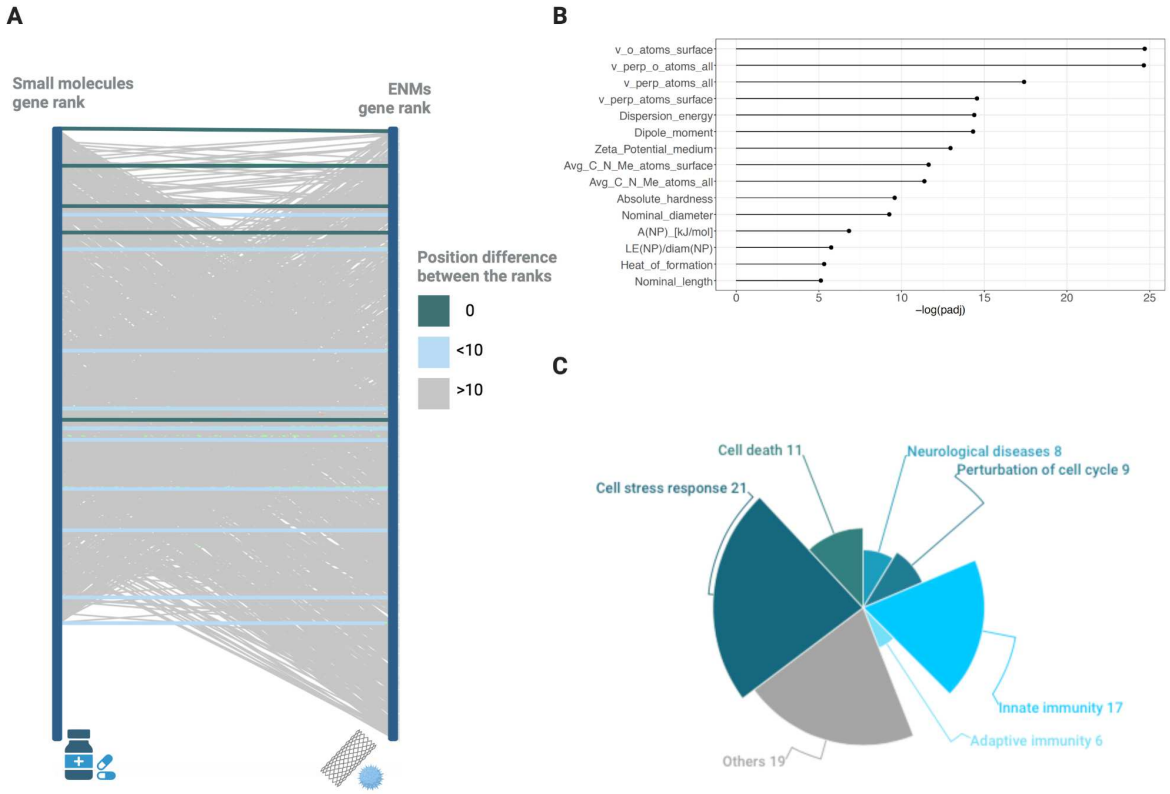

Figure 1. A Comparison of the meta-analysis ranks obtained from ENMs and drug safety derived data (TG-Gates). Genes sharing the same position in both datasets are marked in green. Genes whose position difference is smaller or bigger than 10 are marked in light blue and grey, respectively. **B** Molecular descriptors correlated with genes at the top of the ENM ranking. For each molecular descriptor, the top 10% of genes in the dataspace whose expression is correlated, is selected. A GSEA approach is used to highlight the descriptors whose correlated genes are enriched at the top of the ENM rank. For more detail, see Methods. **C** Pie chart representing the functions associated with the meta-analysis rankings, grouped into 6 main biological categories. For each section, the number of pathways falling within the category is reported. The complete list of pathways is shown in Table S3.

In order to assess the specificity of the transcriptional signature of ENM highlighted by our meta-analysis, we applied the same analytical pipeline to the Open Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System (TG-Gates) dataset, in which the transcriptomic adaptation of rat liver tissue is measured after treatment with 158 small molecules. Our analysis showed substantial difference among the two ranks (Figure 1A), suggesting that the transcriptional signature highlighted by our meta-analysis is indeed specific to ENM and is not overlapping with the effects of small molecules. ENMs differ from drugs due to their unique physicochemical properties and increased complexity, requiring more information to be described than just the structure (e.g., core chemistry, coating functionalization and biomolecule binding) [10,11]. So we next investigated whether characteristics of ENMs would be associated with the transcriptional signature. To this aim, we curated a set of 159 molecular descriptors and 46 experimental labels, and computed their correlation with gene expression across the dataset (Table S2). A subset of descriptors associated with ENM solubility, charge, structure, molecular interactions, and stability of the ENM were correlated with genes at the top of our rank (Figure 1B). These results suggest that descriptors relevant to the ENM-bio interaction or biological activity, associated with charge

and solubility, (e.g. dipole moments, dispersion energy, zeta potentials, surface energies and hydrophilicity or Hamaker constants descriptors) strongly determine the host response to ENM exposure. Similarly, structure and stability (dissolution, heat of formation and absolute hardness descriptors) seem to be key players of the induced response. Finally, the size of an ENM and its surface appear to significantly affect the genes at the top of the meta-analysis rank. Interestingly, some of these descriptors have been previously successfully used to predict ENM toxicity, highlighting the significance of the ENM surface reactivity [12].

The complete meta-analysis rank enriched biological pathways belonging to six main categories: cell stress response, innate immunity, cell death, perturbation of cell cycle, neurological diseases, and adaptive immunity (Figure 1C). Oxidative stress is a prevalent mechanism of ENM toxicity: the reactive surface of ENMs can promote the generation of intracellular reactive hydroxyl radicals, which in turn results in lipid peroxidation, interference with proteins, and DNA damage [13]. Similarly, the overlap of the ENMs energy band gap, i.e., the energy levels of the valence and conduction band with the cellular redox potential have been correlated with oxidative stress [14] [12]. Cytotoxicity is a routinely tested effect of ENMs, which is often modulated by the particle physicochemical characteristics [15]. Pathways associated with neurological diseases were also found enriched in the ENM toxicogenomics dataset. Indeed, many ENMs have been previously associated with possible neuronal toxicity [2], and protein unfolding at ENM surfaces may also influence protein fibrillation [16].

Finally, pathways related to both innate and adaptive immunity were strongly enriched. Among the innate immunity related pathways, IL-12, IL4, IL6, Jak-STAT signalling pathway and interferon responses clearly emerged (Table S3). Interferons are known to mediate the early innate and ancestral defence mechanisms, especially upon viral infections [17,18], and have also been reported as being activated by exposure to some ENMs [19–22] Interestingly, immune-related pathways appear to be frequently altered across multiple cell types and tissues, even those which do not have a primary immune function, such as hepatocytes, respiratory epithelial cells or human fibroblasts, in response to ENMs with diverse intrinsic properties.

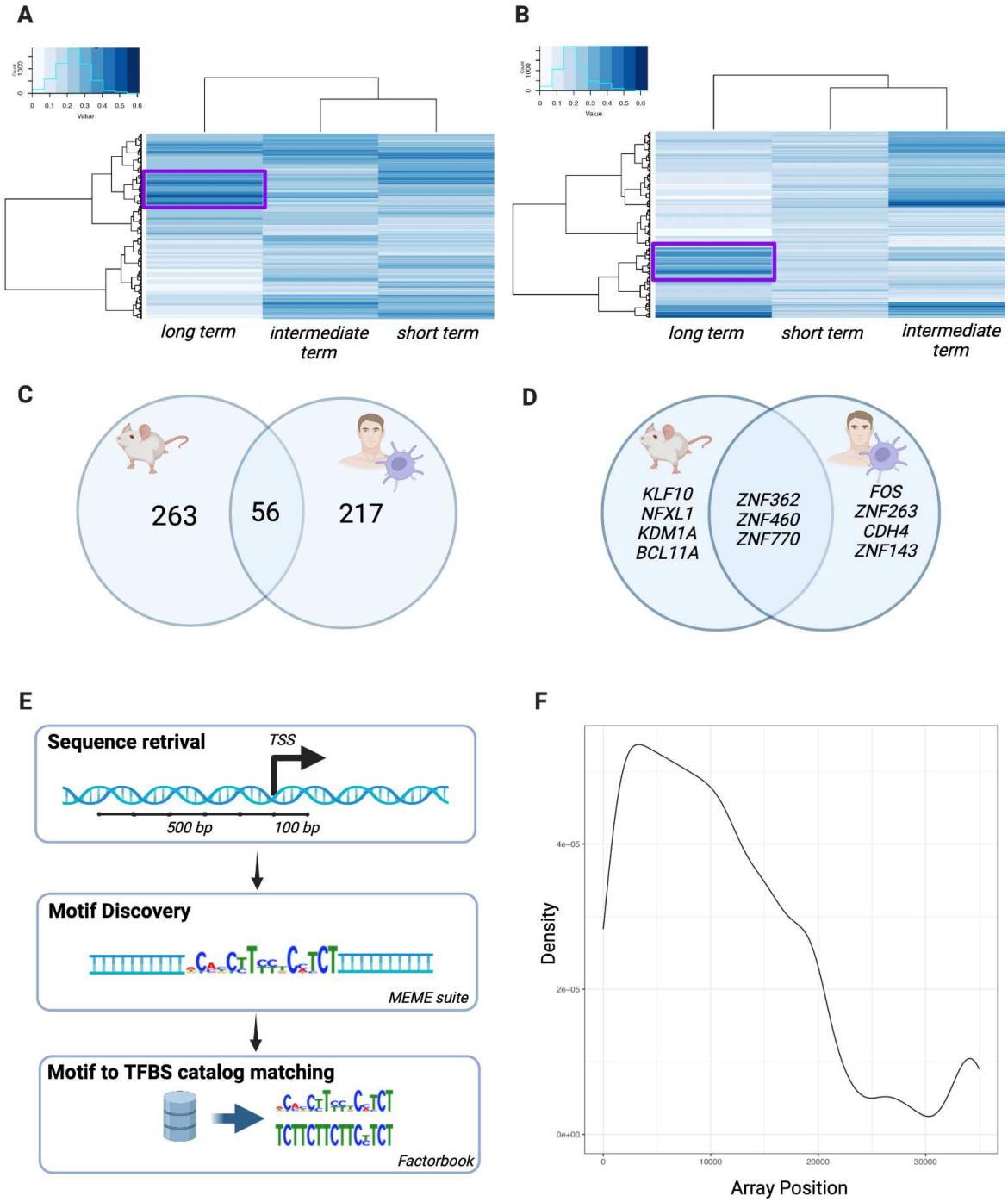

Figure 2. A-B Genes prioritised through functional analysis *in vitro* (A) and *in vivo* (B) are clustered according to the deregulation frequency following short, intermediate and long term post-exposure monitoring. The cluster containing genes altered in at least 40% of the long-term monitoring samples (highlighted by the purple boxes), are selected for further investigation. **C** Venn diagram showing the overlap between the two clusters of genes (*in vitro* and *in vivo*). **D.** Venn diagram showing the overlap of the regulatory motifs whose expression was statistically significant in the promoter regions of the two clusters of genes. **E.** Conceptual pipeline of the promoter analysis performed in this study. First, the DNA sequence of the [-500,+100] region around the transcriptional start sites (TSS) is retrieved. Motif discovery is performed through the MEME suite, finding all the DNA motifs between 6 and 15 base pairs which would satisfy

a p-value threshold of 0.05. Finally, all the motifs are matched to the closest transcription factor binding site (TFBS) returned by the Factorbook database. **F.** Density plot of the position of genes regulated by the C_2H_2 family of transcription factors in the original complete ENM dataset.

We hypothesised that the meta-analysis of the transcriptomics data sets could allow the identification of specific patterns of transcriptional alterations common to both *in vitro-in vivo* ENM exposures monitored over the long term. For this, we focused our attention on the subset of the top 1,872 genes of the meta-analysis rank, which showed the most significant enrichment in the functional analysis (see Methods).

By comparing patterns of gene expression alteration in long term post-exposure monitoring *in vitro* and *in vivo* (cfr. Methods, Figure 2A-B), we identified two distinct clusters of genes altered in at least 40% of the samples (Figure 2C and Table S4-S5). The *in vivo* and *in vitro* clusters contained 319 and 273 genes, respectively, sharing a set of 56 genes. Functional annotation of these genes highlighted unfolded protein response (UPR) and apoptosis. UPR is induced after accumulation of misfolded proteins in the endoplasmic reticulum, activating multiple biological processes to ease the reticulum stress, and ultimately inducing apoptosis if it cannot be reverted [23]. ENMs are known to induce endoplasmic reticulum (ER) stress through various mechanisms, including reactive oxygen species (ROS) production, which can eventually lead to UPR in many cell types [24,25]. ENMs are also known to contribute to unfolding of proteins and protein fibrillation, and to shed proteins from extracellular locations intracellularly as they are transported into and around cells, potentially disrupting cellular proteostasis [16]. The acquisition of complement proteins, acute phase, and tissue leakage involved proteins into the ENM corona have also been shown to act as "thorns" inducing a proinflammatory response [26] .

Long term physiopathological changes are well captured by epigenetic modifications, while transient patterns of molecular response are usually associated with rapid transcriptional alteration [27]. Therefore, we investigated whether these two clusters of genes that remained up-regulated for extended periods, shared a regulatory mechanism underlying the transcriptomic response.

We retrieved the promoter sequences of each gene both *in vitro* and *in vivo*, and searched for conserved motifs around the transcription starting site (TSS, Figure 2E). We identified ten and eight DNA motifs significantly (p-value < 0.05) associated with the *in vivo* and the *in vitro* sets of genes, respectively (Figure 2D, Table S6). Each of these motifs was then matched to a transcription factor binding site (Figure 2E).

Of note, when the regulatory layer of the transcriptome is investigated, the similarity between *in vitro* and *in vivo* increases from 20% to more than 40% (Figure 2 C-D). More interestingly, the set of discovered motifs mainly binds members of the C_2H_2 zinc finger subfamily (C_2H_2-ZNF), possibly suggesting a central role of this family of transcription factors in the conserved response to ENMs [28].

C_2H_2-ZNF are a large family of transcription factors in the eukaryotic genomes with a key role in development and differentiation of several tissues [29]. C_2H_2-ZNF are conserved throughout evolution and bind to repeated and contiguous motifs on target chromatin sites, often associated with transposable elements [30]. It is estimated that more than 700 C_2H_2 ZNF genes exist in humans accounting for more than 2 percent of the total human genes [31].

Some members of this family have been directly linked to immunomodulation and inflammation, including the transcription factors CTCF, Ikaros and ZAS [32–34]. Disruption of the ZNF structures has been associated with metal-ion toxicity, as a secondary effect of the metal concentration changes, causing interference with gene expression and DNA repair [35].

Importantly, while they have already been studied in plants as abiotic stress master regulators [36], their driving role in human toxicological responses remains to be clarified to date.

Interestingly, C_2H_2-ZNF transcription factors play an important role in chromatin plasticity and recruiting of repressor complexes, such as polycomb [37]. Modulation of chromatin structure is a commonly observed response of hosts to exogenous stimuli, e.g., reduction of chromatin accessibility to avoid possible viral insertion in the genome [38]. However, epigenetic mechanisms are involved in generic stress responses, and can also alter the chromatin structure to produce transient translocation and nuclear reorganisation [39,40]. This effect primes the cells, providing an epigenetic memory of the environmental stimulus [41]. Notably, this mechanism has been observed already in plants, and in more simple eukaryotic organisms [39,40]. Furthermore, topological associated domains (TADs) borders are enriched in the C_2H_2-ZNF member CTCF, whose binding determines chromatin domains [42]. These results suggest that the epigenetic mechanisms found here can significantly aid the reconstruction of ENM MOA, but have remained largely unexplored to date.

Next, we tested whether this gene regulatory model would result in information loss, when compared to the complete transcriptional profile in all the transcriptomics datasets previously curated (Table S7). Indeed, our results showed that the C_2H_2-ZNF regulatory model explains the most relevant transcriptional alterations in all the individual transcriptomic profiles (Figure 2F and Figure S2). Taken together, these findings suggest that the C_2H_2-ZNF regulatory model resolves the complexity and disparity of individual transcriptomic responses to ENM exposures, allowing key epigenetic changes and longer-term, persistent regulatory gene expression responses to be unravelled. Moreover, analysis of DNA regulatory elements provides a higher and more stable level of similarity across biological systems, as compared to the gene-level toxicogenomic signatures.

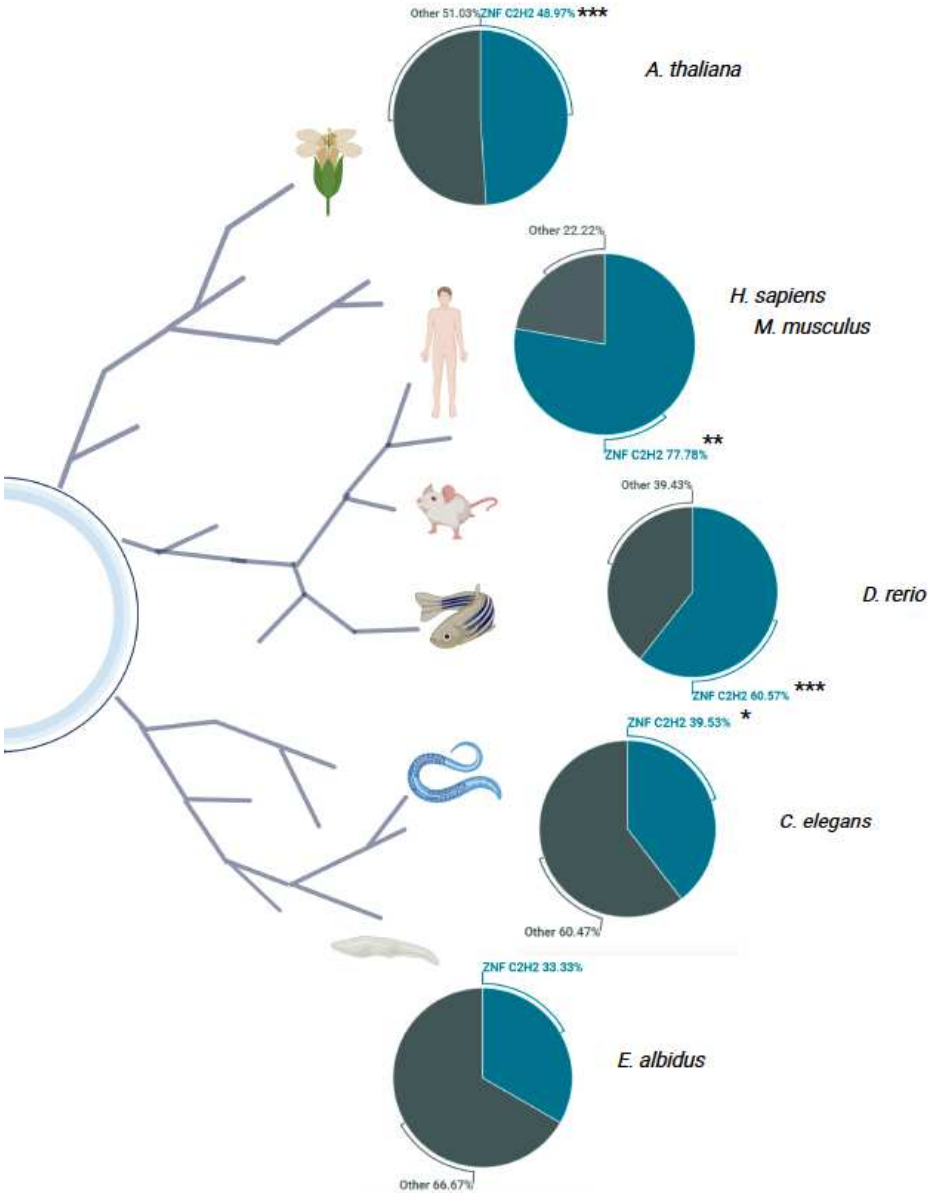

Figure 3. Overview of the species included in our study, and relative proportion of C_2H_2 ZNF transcription factors binding sites in the promoter region of the genes involved in their adaptation response. The statistical significance is indicated by the number of stars (*). * p-value < 0.005 **p-value = 1.4 e-08 and *** p-value < 2.2 e-16

We further hypothesised that C_2H_2-ZNF could mediate the ENM transcriptomic responses in other species of eco-toxicological interest. To this end, we analysed seventeen datasets recently curated, including ENM exposures to *D. rerio*, *C. elegans*, *E. albidus* and *A. thaliana* [9,43] . Our results indicate that also in non-mammal organisms the adaptive response to ENM is regulated by the C_2H_2-ZNF transcription factors family (Table S8). Interestingly, the relative proportion of TFBS matching C_2H_2-ZNF members decreases down the phylogenetic tree, suggesting a possible association with organismal complexity [31] (Figure 3). The organisms included in our analysis have variable levels of organisation of the adaptive and innate immunity. Plants, such as *A. thaliana*, display host-immunity that is controlled by polymorphic host genes, where resistance protein-mediated activation of defence is based on an 'altered-self' mechanism of recognition [44]. These mechanisms also commonly use epigenetic modification and chromatin remodelling to establish memory of the infection, achieving immunity even in absence of

specialised immune cells [45]. Furthermore, in a recent study, Curtis et al. described immune-deregulation and stress response as a shared feature to lithium cobalt oxide (LCO) exposure across taxonomic groups, including *D. rerio* [46].

In conclusion, the response to exogenous stimuli is a primitive function of the cell. Response to the environment shapes the genome and influences evolution, developing epigenetic mechanisms to minimise potentially harmful events, and eventually increase the fitness of the host. Although ENMs are a result of industrialisation, all organisms have been exposed to nanoscale particles of natural origin across geological eras, generated during volcanic eruptions, wildfires, and other natural processes [47–49]. This, in turn, suggests that the responses and mechanisms of adaptation are the result of a long evolutionary history of exposure to nano-sized matter. Indeed, our results suggest that a very ancestral regulatory model may be conserved across the tree of life and can explain the adaptation of biological entities to ENM exposures. This proposed model solves the apparent heterogeneity of the transcriptomic MOA landscape and points to a shared immunomodulatory response to ENMs, even in cells whose primary function is not immunological.

The stability of the regulatory layer across species paves the way to the design of new toxicological tests that can bridge ecotoxicology and human toxicology [50] , and eventually facilitate the development of more effective *in vitro-in vivo* extrapolation testing strategies.

It is an established paradigm that the exposome impacts on the epigenome, often causing disease susceptibility [51]. For this reason, the association between specific DNA modifications (e.g. DNA methylation) and exposures to environmental factors have been widely investigated [52,53]. Our results further suggest that the role of other layers of genomic regulation on priming future responses to ENM could be investigated.

Finally, this study shed light on an unexplored layer of the adaptation response, in which ENMs induce a chromatin dynamic response and potentially point to epigenetic marks of the exposures.

Methods

Data collection and preprocessing of nanomaterial datasets

The meta-analysis of multiple toxicogenomics studies of different classes of ENMs *in vitro* and *in vivo* can help to identify common molecular mechanisms of action independent from the genotype of the biological system under evaluation. To this end, we implemented a meta-analysis of 66 transcriptomic datasets derived from the public data collection curated by Saarimäki et al.[6] (https://zenodo.org/record/3949890#.YlPUri0RqH0), supplemented with additional data previously published in Kinaret et al. [8] (GEO accession number GSE157266) and Gallud et al. [7] (GEO accession number GSE148705) (Table S1). From the original collection we excluded datasets derived from old microarray platforms, as they shared a small number of probes with the more recent versions (Table S1).

The datasets GSE148705 and GSE157266 were pre-processed using the eUTOPIA software, as described in Saarimäki et al [6,54] . Briefly, for the probe filtering step we selected only probes with a value higher than the 0.8 quantile against the negative control in at least 75% of the samples. Data were normalised between arrays using quantile normalisation [55]. No batch correction was needed for the dataset GSE148705, while, for dataset GSE157266 we corrected for technical variation associated with variables "dye" and "slide" using the ComBat method [56]. Finally, we used the limma package to compute the gene expression difference between each exposure in the dataset and the corresponding controls, further correcting the p-value using the Benjamini & Hochberg procedure. The aggregated, normalised and corrected expression matrix was then exported, applying no threshold on the log fold change value.

In this study, we selected all pairwise comparisons between time/dose exposures and their respective controls. The final dataset comprised multiple human and mouse tissues and cell types exposed to various nanomaterials. From this selection, we were able to define a data space of 584 specific exposure conditions (treatment, exposure time and dose, and biological system) and 3,676 genes.

The pool of 3,676 genes represents the intersection of the genes present in all the selected human and mouse experiments, limiting the mouse-human conversion to 1:1 orthology relationships (i.e., where both genes in the pair have only one ortholog in the other species, where an ortholog is a gene in a different species that catalyses the same reaction (i.e., has the same function)). The ortholog genes were converted using the getLDS function of the biomaRt R package [57].

Collection and preprocessing of transcriptomics data for drug exposure

In order to assess the specificity of the meta-analysis rank, we performed a similar analysis on a dataset of small molecules. To this aim, raw microarray data for 158 drugs was downloaded from the Open Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System (TG-GATEs) database [58]. The microarray data comes from the *in vivo* exposure of rats to three dose levels for each drug. For each dose level, the rats received a single exposure of the drugs and subsequently liver samples were harvested at four different time points (3, 6, 9 and 24 hours post-exposure).

Raw data was imported into R by using the justRMA function from the R library Affy [59]. Probe annotation to Ensembl genes was performed by the custom annotation files rat2302rnensgcdf (v. 22.0.0), downloaded from the brain array website, leading to 12,153 genes. The expression values were quantile normalised by means of the normalizeQuantile function from the R limma library [55].

Lastly, differential expression analysis was performed by means of the R limma library [55]. Differential expression analysis was performed for each drug, for each combination of dose level and time point available, leading to 1,839 comparisons. The analyses were performed comparing the treated samples to the matched control samples of the same time point. As a result, log2-fold-changes, p-values, and adjusted p-values (by means of the FDR correction) were retrieved for all genes for each comparison. In order to make a comparable evaluation, from the complete set of genes we selected only ones that were also included in our original meta-analysis set.

Characterization of the 584 experimental conditions

In our ENM meta-analysis dataspace we were able to identify 584 different experimental conditions, which are unique because of the biological system, nanomaterial or experimental condition used (e.g. time, dose). For each experimental condition, we manually curated the information provided in the original publications and annotated each instance into several categories (Table S9). First, we grouped samples according to the biological system properties and exposure setting (organism, biological system, biological system grouping, and whether the experiment was *in vivo* or *in vitro*). We also included information on the exposure duration (Time, Time_period, Time_h) by grouping samples into short, intermediate, and long exposures. Different thresholds were defined for *in vivo* and *in vitro* settings. In detail, for *in vitro* experiments, exposures have been considered short, intermediate and long at 24 hours, between 24 and 72 hours and after 72 hours, respectively. For *in vivo* post-exposure periods, thresholds were posed at up to 3 days for short exposures, between 3 days and 1 month for intermediate, and more than 1 month for long term ones.
The nanomaterials used in the experiment were classified according to the generic chemical characteristics of the material (core Material, specific Material, chemistry, metal, oxide, geometry, material), and the presence or absence of functionalized groups.
A panel of further information was extracted from the original publications (when possible), covering nanomaterial characteristics (crystal phase, purity, whether the absence of endotoxins was confirmed (or not), coating, stabiliser and supplier information) as well as protocol information (dispersant and description of dispersion).
Finally, in the studies in which the ENM had been characterised, data were reported regarding: nominal diameter, length and specific surface area; TEM diameter, width, and length; BET surface area; number of walls; DLS mean diameter and PDI in water and medium; Zeta Potential in water and medium.
Furthermore, a list of 159 molecular descriptors (MD) covering both molecular and electronic structure properties was computed, when possible, for each ENM (see next section).

Computation of molecular descriptors

A set of 159 descriptors have been computed for the ENMs, including liquid drop model (LDM) descriptors, electronic structure descriptors, interaction descriptors, image (shape) descriptors, periodic table descriptors. LDM molecular attributes (commonly referred to as "nanodescriptors") [60] are calculated based on the assumption that ENM can be represented as a spherical drop, in which elementary molecules are tightly packed, while the density of clusters is equal to the particle mass density [60,61]. The LDM-based descriptors can be applied both in the case of analysing a single ENMs and agglomerates of ENMs [60]. Among the calculated ionic parameters is: the Wigner-Seitz radius (r_w), the number of ENMs in the analysed agglomerate (n), the number of surface elements (*S*), the surface – volume ratio (*SV*), the aggregation parameter (*AP*) [60]. The Wigner-Seitz radius characterises the minimum radius of interactions

between individual molecules (or ENMs in agglomerates) and is represented by the following formula (1):

$$r_w = \left(\frac{3M}{4\pi\rho N_A}\right)^{\frac{1}{3}}$$

where: M - molecular weight; - mass density, N_A - Avogadro constant.
The number of ENMs in the analysed agglomerate (n) is represented using the following formula (2):

$$n = \left(\frac{r_0}{r_w}\right)^{3}$$

where r_0 the radius of each ENMs.
The number of surface elements (S) is represented by the following formula (3):

$$S = 4n^{-\frac{1}{3}}$$

where: *S* describes the ratio of surface molecules to molecules in the volume (or surface EMs in agglomerates).
The surface – volume ratio (*SV*) is represented using the following formula (4):

$$SV = \frac{S}{1-S}$$

where: *SV* is the feature that describes the ratio of surface molecules to molecules in volume (or surface ENMs in agglomerates).
Size-dependent interfacial thickness (*h*) was calculated with the following formula (5)

$$h = 0.01 \cdot (T - 273) \cdot r^{0.35}$$

where r – is the nominal size of the ENM, T – temperature [62].
The electronic structure descriptors of ENMs were computed by density functional theory (DFT) and semi-empirical quantum chemical methods, while the Hamaker constants are evaluated from atomistic force fields and a continuum method [63,64]. All results are obtained with a good resolution in a fast manner and the proposed methodologies can be applied on a broad range of ENMs. The bulk of the ENMs interacts via the long-range van der Waals interaction, which is a major contribution in the calculation of the adsorption energies of biomolecules in water. Therefore, Hamaker constants are evaluated to describe bionano interactions in water through an atomistic force field approach and via Lifschitz theory [64]. In the Lifschitz theory [64] two materials are interacting through a medium; the Hamaker constant for the ENM and a biomolecule in water is calculated from optical parameters that are experimentally determined (Table S10), while in the force field approach long-range dispersion interaction is calculated using the Lorentz-Berthelot rules for sigma (atom size) and epsilon (atom-atom interaction amplitude) [65,66], i.e. combining rules that provide the interaction energy between two non-bonded atoms. In the case of metal ENMs, we use CHARMM force field parameters [67], and for metal oxides and carbon ENMs as well as amino acids, lipids and sugars the force fields as reported in [68]. All the force field parameters have been applied in molecular dynamics simulations for many ENM properties, including potentials of mean force. Considering all atom-atom interactions between two molecular entities, the Hamaker constant

is derived by a simple approximation of the combined sigma and epsilon dispersion parameters[69]. In this work we considered the Hamaker constants of ENMs interacting with biomolecules (amino acids, lipids, hydrocarbons) in water or just the interaction between two ENM pieces in water. The geometric structures of the bulk ENMs were optimised with DFT and the PBE functional [70] employing the SIESTA code [71]. The band gaps were also calculated by PBE [70], while the heat of formation, electronegativity, absolute hardness, dispersion energy per atom, dipole moment and static polarizability descriptors [63] were obtained on the self-consistent field (SCF) level through the semi-empirical code MOPAC (HTTP://OpenMOPAC.net) using the PM6-D3 [72] parametrization. Finally, ionisation potentials, electron affinities and the global electrophilicity index were computed through self-consistent charge calculations (ΔSCC calculation) for the electronic states of the neutral and ion ENMs via the GFN1-xTB parametrization of the GFN-xTB code [73–77].

We further included a set of atomistic descriptors that are based on the chemical composition, potential energy, lattice energy , topology, size, and force vectors [78,79]. Constitutional descriptors are the counts of atoms of different identity and/or location. Potential energy descriptors are derived from the force-field calculations, corresponding to the arithmetic means of the potential energies for specific atom types and/or locations in the ENM. Lattice energies are based on the same potential energies but presented as per metal oxide nominal units (MxOy) and describe the energy needed to rip away said unit from the ENM surface. All potential energy-related descriptors are presented in eV units. The coordination number of atoms is defined as the count of the neighbouring atoms which lie inside the radius,

$$R = 1.2 \times (R_M \; and \; R_O)$$

where R_M and R_O are the ionic radii of metal and oxygen ions, respectively. A low coordination number indicates that some atoms have missing neighbours and thus makes the ENM more unstable. The representatives of the size related descriptors were derived from the actual calculated ENM diameter, obtained as the maximum separation (distance) between any two atoms in an ENM. The last group of descriptors represent different force vector lengths that have been derived from the structure optimization. For example to derive the average length (V) of the surface normal component of the force vector for a shell region atom, its coordinates (x, y, z), force vector components (f_x,f_y,f_z) and distance from the centre-of-mass (d) are used:

$$V = \frac{(xf_x + yf_y + zf_z)}{d}$$

List of developed atomistic descriptors with sample values for 10 nm TiO_2 ENM are presented in TS11. The described set of atomistic descriptors is also applicable to amorphous ENMs. Contrary to crystal ENMs, the generation of representative amorphous structures is a multistep procedure. It requires the simulation of bulk metal or metal oxide materials above their melting temperature, the extraction of ENMs with the desired size and shape and their subsequent cooling at the temperature of interest with a prescribed rate. Such a procedure has been applied in an automated way to build Au/Pt/Cu/Cr/Mo/SiO_2 spherical amorphous ENMs with the aid of the Enalos Demokritos KNIME nodes. All the data are hosted at the NanoPharos database (db.nanopharos.eu) and were converted into a ready-for-modelling format for the development of *in silico* alternative testing strategies and integrated approaches to testing and assessment. The nanoPharos database provides structured, curated and harmonised datasets. All datasets are made publicly available and offered in a ready-for-modelling format so that users can directly import them into computational workflows.

Meta analysis implementation

In order to identify genes associated with every ENM exposure, we implemented a consensus of 3 algorithms for meta-analysis, which produced a prioritisation of the 3,676 genes.
As proposed by Kan et al. [80,81] our analysis pipeline is based on the effect size, p-value based and rank product methods. Usually, the most common meta-analysis framework is the Effect size methods, assessing within- and between-study variation across multiple studies. This method outperforms others when there is large between-study variation and small sample sizes. To implement the effect size-based method, the "effect_sizes" function from the esc R package was used to compute effect sizes using the p-value argument and the "chi_esc" function [82]. The Fisher's sum of logs method combines individual p-values, and is considered the most straightforward method, in that it does not require additional analysis. Fisher's sum of logs method was implemented by using the "sumlog" function of the R package metap, giving as input the p-values of each gene [83]. Finally, the rank product is a non-parametric statistical method that can be used to combine the results of the differential gene expression analysis from individual studies based on the within-study ranks of the genes. To this end, the genes in each experiment were ranked based on the relevance of their associated p-values, and the "RP.advance" function of the RankProd R package was used to merge them by means of the one-class analysis of the rank product method [84,85]. This function allows combining of data coming from different studies, such as in the case of datasets generated by different laboratories. For each method, a rank based on the meta-analysis p-values was generated. Finally, all the ranks were combined through the Borda function of the TopKlists R package [86]. The final mean rank is reported in Table S1.

Gene set enrichment analysis and feature selection step

In order to select the most biologically relevant features on which to carry out further analyses, we performed a Gene Set Enrichment Analysis (GSEA) of the final meta-analysis rank on 5 different databases: Wikipathways [87], Gene Ontology [88], Reactome [89], KEGG [90] and MsigDB [91]. In each case, the "fgsea" function from the fgsea R package was used [92]. For each test, we identified the position of the rank having the highest peak value of cumulative enrichment statistics. By doing so, we created a reduced representation of the meta-analysis gene rank by setting as a threshold the top 10^{th} percentile of such values. In this way, 1,873 genes were selected for further analyses.

Computation of the frequency score and hierarchical clustering

In order to find the genes associated in vitro and in vivo in long term exposures, we split the samples into *in vivo* and *in vitro* biological systems and grouped them according to the time of post-exposure. For each gene of the reduced meta-analysis rank, we calculated a frequency score as the percentage of samples in which the gene is statistically significant.
Finally, the genes were clustered according to the euclidean distance of their frequency scores. The hierarchical clustering algorithm, with Ward linkage method, implemented into the hclust function of the R package "stats" was used.
For each type of exposure system, we selected the cluster with the most frequently deregulated genes. In order to functionally annotate the set of genes frequently altered in long term exposures, we performed a pathway enrichment analysis through the EnrichR online tool, using the MsigDB and Reactome databases [89,91,93–95].

Computation of the molecular descriptors-gene expression correlation

To identify associations between chemical properties of the ENMs and the molecular alterations induced in cells / organisms by their exposure, the Pearson correlation coefficient was computed for each pair of gene and molecular descriptor. Before computing the correlations, the two data layers were pre-processed. Particularly, a Winsorize function of the DescTools R package [96] was used to replace extreme values of log2-fold-changes with less extreme ones. Moreover, a cube root transformation was applied to the molecular descriptor values.
Since the molecular descriptor data layer contains missing data, for each descriptor the Pearson correlation was computed across the subset of samples for which its values are computed.
For each descriptor the top 10% of the most correlated genes were selected and further investigated. To this end, two different analyses were performed. First the gene sets were enriched against the KEGG pathways by means of the FunMappOne tool [97]. Only pathways with FDR corrected p-values < 0.05 were considered significantly enriched. The molecular descriptors were further clustered in 9 groups based on the Jaccard Index similarity of the shared enriched pathways.
Lastly, the fgsea R package [92] was used to perform a GSEA analysis and identify the molecular descriptors whose set of associated genes is enriched on the top of the ranked list of genes identified with the meta-analysis approach. Only molecular-descriptors with an adjusted p-value < 0.01 were considered for further characterization.

Promoter analysis

For each gene in the subset of interest, the promoter region [-500 bp, +100 bp] around the TSS. The sequence was downloaded using the biomart package and the getSequence function in 'coding_gene_flank' mode [57]. In this modality the function returns the flanking region of the gene including the UTRs, this must be accompanied with a given value for the upstream or downstream attribute.
Motif discovery was conducted with the MEME suite [98]. The motif site distribution was set as any number of repetitions, and the search was restricted to motifs ranging between 6 and 15 bases. We selected a p-value threshold of 0.05.
For each result, the Factorbook database was interrogated to explore the transcription factor binding sites in their genomic and epigenetic contexts [https://www.factorbook.org/]. Factorbook is a transcription factor (TF)-centric web-based repository associated with ENCODE ChIP-seq data, as well as multiple databases of transcription factors binding sites. The transcription factor binding site that best matches the query according to the tool is reported.
For each organism we annotated whether the transcription factor binding site would be recognised by a member of the C2H2 ZNF or not. In order to evaluate the statistical significance of C2H2 ZNF overrepresentation, we performed a fisher test by using the fisher.test function in the stats R package. For the statistical test, we used as a background the set of non-redundant transcription factor (TF) binding profiles provided in JASPAR [99]. The contingency matrix was built by using the set of TF binding sites for members of the C2H2 family, and all the others, respectively.

Acknowledgement

This research was funded by EU H2020 NanoSolveIT project (grant number 814572) and by the Academy of Finland (Grant No. 322761).

References

1. Kinaret, P. A. S. *et al.* Transcriptomics in toxicogenomics, part I: experimental design, technologies, publicly available data, and regulatory aspects. *Nanomaterials (Basel)* **10**, (2020).
2. Serra, A. *et al.* INSIdE NANO: a systems biology framework to contextualize the mechanism-of-action of engineered nanomaterials. *Sci. Rep.* **9**, 179 (2019).
3. Krewski, D. *et al.* Toxicity testing in the 21st century: a vision and a strategy. *J. Toxicol. Environ. Health B Crit. Rev.* **13**, 51–138 (2010).
4. ENCODE Project Consortium. The ENCODE (encyclopedia of DNA elements) project. *Science* **306**, 636–640 (2004).
5. Yosef, N. & Regev, A. Impulse control: temporal dynamics in gene transcription. *Cell* **144**, 886–896 (2011).
6. Saarimäki, L. A. *et al.* Manually curated transcriptomics data collection for toxicogenomic assessment of engineered nanomaterials. *Sci. Data* **8**, 49 (2021).
7. Gallud, A. *et al.* Multiparametric profiling of engineered nanomaterials: unmasking the surface coating effect. *Adv Sci (Weinh)* **7**, 2002221 (2020).
8. Kinaret, P. A. S. *et al.* Toxicogenomic profiling of 28 nanomaterials in mouse airways. *Adv Sci (Weinh)* **8**, 2004588 (2021).
9. Burkard, M., Betz, A., Schirmer, K. & Zupanic, A. Common Gene Expression Patterns in Environmental Model Organisms Exposed to Engineered Nanomaterials: A Meta-Analysis. *Environ. Sci. Technol.* **54**, 335–344 (2020).
10. Papadiamantis, A. G. *et al.* Metadata Stewardship in Nanosafety Research: Community-Driven Organisation of Metadata Schemas to Support FAIR Nanoscience Data.

586 *Nanomaterials (Basel)* **10**, (2020).
587 11. Walczyk, D., Bombelli, F. B., Monopoli, M. P., Lynch, I. & Dawson, K. A. What the
588 cell "sees" in bionanoscience. *J. Am. Chem. Soc.* **132**, 5761–5768 (2010).
589 12. Papadiamantis, A. G. *et al.* Predicting Cytotoxicity of Metal Oxide Nanoparticles using
590 Isalos Analytics Platform. *Nanomaterials (Basel)* **10**, (2020).
591 13. Manke, A., Wang, L. & Rojanasakul, Y. Mechanisms of nanoparticle-induced oxidative
592 stress and toxicity. *Biomed Res. Int.* **2013**, 942916 (2013).
593 14. Zhang, H. *et al.* Use of metal oxide nanoparticle band gap to develop a predictive
594 paradigm for oxidative stress and acute pulmonary inflammation. *ACS Nano* **6**, 4349–
595 4368 (2012).
596 15. Sun, H., Jiang, C., Wu, L., Bai, X. & Zhai, S. Cytotoxicity-Related Bioeffects Induced
597 by Nanoparticles: The Role of Surface Chemistry. *Front. Bioeng. Biotechnol.* **7**, 414
598 (2019).
599 16. Mahmoudi, M., Kalhor, H. R., Laurent, S. & Lynch, I. Protein fibrillation and
600 nanoparticle interactions: opportunities and challenges. *Nanoscale* **5**, 2570–2588 (2013).
601 17. Sadler, A. J. & Williams, B. R. G. Interferon-inducible antiviral effectors. *Nat. Rev.*
602 *Immunol.* **8**, 559–568 (2008).
603 18. Shaw, A. E. *et al.* Fundamental properties of the mammalian innate immune system
604 revealed by multispecies comparison of type I interferon responses. *PLoS Biol.* **15**,
605 e2004086 (2017).
606 19. Chen, H. *et al.* In vivo study of spherical gold nanoparticles: inflammatory effects and
607 distribution in mice. *PLoS ONE* **8**, e58208 (2013).
608 20. Khatri, M. *et al.* Evaluation of cytotoxic, genotoxic and inflammatory responses of
609 nanoparticles from photocopiers in three human cell lines. *Part. Fibre Toxicol.* **10**, 42
610 (2013).

21. Perkins, T. N. *et al.* Differences in gene expression and cytokine production by crystalline vs. amorphous silica in human lung epithelial cells. *Part. Fibre Toxicol.* **9**, 6 (2012).
22. Ye, S.-F., Wu, Y.-H., Hou, Z.-Q. & Zhang, Q.-Q. ROS and NF-kappaB are involved in upregulation of IL-8 in A549 cells exposed to multi-walled carbon nanotubes. *Biochem. Biophys. Res. Commun.* **379**, 643–648 (2009).
23. Adams, C. J., Kopp, M. C., Larburu, N., Nowak, P. R. & Ali, M. M. U. Structure and molecular mechanism of ER stress signaling by the unfolded protein response signal activator IRE1. *Front. Mol. Biosci.* **6**, 11 (2019).
24. Simard, J.-C., Durocher, I. & Girard, D. Silver nanoparticles induce irremediable endoplasmic reticulum stress leading to unfolded protein response dependent apoptosis in breast cancer cells. *Apoptosis* **21**, 1279–1290 (2016).
25. Christen, V., Camenzind, M. & Fent, K. Silica nanoparticles induce endoplasmic reticulum stress response, oxidative stress and activate the mitogen-activated protein kinase (MAPK) signaling pathway. *Toxicol. Rep.* **1**, 1143–1151 (2014).
26. Cai, R. *et al.* Corona of Thorns: The Surface Chemistry-Mediated Protein Corona Perturbs the Recognition and Immune Response of Macrophages. *ACS Appl. Mater. Interfaces* **12**, 1997–2008 (2020).
27. Scala, G. *et al.* Multi-omics analysis of ten carbon nanomaterials effects highlights cell type specific patterns of molecular regulation and adaptation. *NanoImpact* **11**, 99–108 (2018).
28. Kurowska, E. *et al.* The C2H2 zinc finger transcription factors are likely targets for Ni(II) toxicity. *Metallomics* **3**, 1227–1231 (2011).
29. Cassandri, M. *et al.* Zinc-finger proteins in health and disease. *Cell Death Discov.* **3**, 17071 (2017).

30. Najafabadi, H. S. *et al.* C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat. Biotechnol.* **33**, 555–562 (2015).
31. Seetharam, A. & Stuart, G. W. A study on the distribution of 37 well conserved families of C2H2 zinc finger genes in eukaryotes. *BMC Genomics* **14**, 420 (2013).
32. Stik, G. *et al.* CTCF is dispensable for immune cell transdifferentiation but facilitates an acute inflammatory response. *Nat. Genet.* **52**, 655–661 (2020).
33. Wu, L.-C. ZAS: C2H2 zinc finger proteins involved in growth and development. *Gene Expr.* **10**, 137–152 (2002).
34. Gounari, F. & Kee, B. L. Fingerprinting ikaros. *Nat. Immunol.* **14**, 1034–1035 (2013).
35. Hartwig, A. Zinc finger proteins as potential targets for toxic metal ions: differential effects on structure and function. *Antioxid. Redox Signal.* **3**, 625–634 (2001).
36. Han, G. *et al.* C2H2 zinc finger proteins: master regulators of abiotic stress responses in plants. *Front. Plant Sci.* **11**, 115 (2020).
37. Schmitges, F. W. *et al.* Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome Res.* **26**, 1742–1752 (2016).
38. Liu, X. *et al.* Human virus transcriptional regulators. *Cell* **182**, 24–37 (2020).
39. Pascual-Ahuir, A. *et al.* Dose dependent gene expression is dynamically modulated by the history, physiology and age of yeast cells. *Biochim. Biophys. Acta Gene Regul. Mech.* **1862**, 457–471 (2019).
40. Ahmed, S. *et al.* DNA zip codes control an ancient mechanism for gene targeting to the nuclear periphery. *Nat. Cell Biol.* **12**, 111–118 (2010).
41. McLeod, D. V., Wild, G. & Úbeda, F. Epigenetic memories and the evolution of infectious diseases. *Nat. Commun.* **12**, 4273 (2021).
42. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin domains: the unit of chromosome organization. *Mol. Cell* **62**, 668–680 (2016).

661 43. Gomes, S. I. L., Scott-Fordsmand, J. J. & Amorim, M. J. B. Profiling transcriptomic
662 response of Enchytraeus albidus to Cu and Ni: comparison with Cd and Zn. *Environ.*
663 *Pollut.* **186**, 75–82 (2014).
664 44. Sanabria, N. M., Huang, J.-C. & Dubery, I. A. Self/nonself perception in plants in innate
665 immunity and defense. *Self Nonself* **1**, 40–54 (2010).
666 45. Spoel, S. H. & Dong, X. How do plants achieve immunity? Defence without specialized
667 immune cells. *Nat. Rev. Immunol.* **12**, 89–100 (2012).
668 46. Curtis, B. J. *et al.* Cross-species transcriptomic signatures identify mechanisms related to
669 species sensitivity and common responses to nanomaterials. *Nat. Nanotechnol.* (2022)
670 doi:10.1038/s41565-022-01096-2.
671 47. Sharma, V. K., Filip, J., Zboril, R. & Varma, R. S. Natural inorganic nanoparticles--
672 formation, fate, and toxicity in the environment. *Chem. Soc. Rev.* **44**, 8410–8423 (2015).
673 48. Gislason, S. R. *et al.* Characterization of Eyjafjallajokull volcanic ash particles and a
674 protocol for rapid risk assessment. *Proc Natl Acad Sci USA* **108**, 7307–7312 (2011).
675 49. Lespes, G., Faucher, S. & Slaveykova, V. I. Natural nanoparticles, anthropogenic
676 nanoparticles, where is the frontier? *Front. Environ. Sci.* **8**, (2020).
677 50. Rivetti, C. *et al.* Vision of a near future: Bridging the human health-environment divide.
678 Toward an integrated strategy to understand mechanisms across species for chemical
679 safety assessment. *Toxicol In Vitro* **62**, 104692 (2020).
680 51. Jirtle, R. L. & Skinner, M. K. Environmental epigenomics and disease susceptibility.
681 *Nat. Rev. Genet.* **8**, 253–262 (2007).
682 52. Agier, L. *et al.* Early-life exposome and lung function in children in Europe: an analysis
683 of data from the longitudinal, population-based HELIX cohort. *Lancet Planet. Health* **3**,
684 e81–e92 (2019).
685 53. Silver, M. J. *et al.* Environmentally sensitive hotspots in the methylome of the early

686 human embryo. *eLife* **11**, (2022).

687

688 **Methods only References**

689

690 54. Marwah, V. S. *et al.* eUTOPIA: solUTion for Omics data PreprocessIng and Analysis.
691 *Source Code Biol. Med.* **14**, 1 (2019).
692 55. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing
693 and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
694 56. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression
695 data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
696 57. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the
697 integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.*
698 **4**, 1184–1191 (2009).
699 58. Igarashi, Y. *et al.* Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic*
700 *Acids Res.* **43**, D921-7 (2015).
701 59. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy--analysis of Affymetrix
702 GeneChip data at the probe level. *Bioinformatics* **20**, 307–315 (2004).
703 60. Sizochenko, N. *et al.* From basic physics to mechanisms of toxicity: the “liquid drop”
704 approach applied to develop predictive classification models for toxicity of metal oxide
705 nanoparticles. *Nanoscale* **6**, 13986–13993 (2014).
706 61. Sizochenko, N. *et al.* How the toxicity of nanomaterials towards different species could
707 be simultaneously evaluated: a novel multi-nano-read-across approach. *Nanoscale* **10**,
708 582–591 (2018).
709 62. Sizochenko, N., Syzochenko, M., Gajewicz, A., Leszczynski, J. & Puzyn, T. Predicting
710 physical properties of nanofluids by computational modeling. *J. Phys. Chem. C* **121**,
711 1910–1917 (2017).

63. Rouse, I. *et al.* First principles characterisation of bio-nano interface. *Phys. Chem. Chem. Phys.* **23**, 13473–13482 (2021).
64. *Intermolecular and surface forces*. (Elsevier, 2011). doi:10.1016/C2011-0-05119-0.
65. Lorentz, H. A. Ueber die Anwendung des Satzes vom Virial in der kinetischen Theorie der Gase. *Ann. Phys.* **248**, 127–136 (1881).
66. Berthelot, D. Sur le mélange des gaz. in *Comptes rendus hebdomadaires des séances de l'Académie des Sciences* vol. 126 1703–1855 (1966).
67. Martin, L., Bilek, M. M., Weiss, A. S. & Kuyucak, S. Force fields for simulating the interaction of surfaces with biological molecules. *Interface Focus* **6**, 20150045 (2016).
68. Brandt, E. G. & Lyubartsev, A. P. Molecular dynamics simulations of adsorption of amino acid side chain analogues and a titanium binding peptide on the tio_2 (100) surface. *J. Phys. Chem. C* **119**, 18126–18139 (2015).
69. Munaò, G., Correa, A., Pizzirusso, A. & Milano, G. On the calculation of the potential of mean force between atomistic nanoparticles. *Eur. Phys. J. E Soft Matter* **41**, 38 (2018).
70. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
71. Soler, J. M. *et al.* The SIESTA method for ab initio order-N materials simulation. *Phys.: Condens. Matter* **14**, 2745 (2002).
72. Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **132**, 154104 (2010).
73. Grimme, S., Bannwarth, C. & Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1-86). *J. Chem. Theory Comput.* **13**, 1989–2009 (2017).

74. Bannwarth, C. *et al.* Extended tight-binding quantum chemistry methods. *WIREs Comput Mol Sci* (2020) doi:10.1002/wcms.1493.
75. Pracht, P., Caldeweyher, E., Ehlert, S. & Grimme, S. A Robust Non-Self-Consistent Tight-Binding Quantum Chemistry Method for large Molecules. (2019) doi:10.26434/chemrxiv.8326202.v1.
76. Bannwarth, C., Ehlert, S. & Grimme, S. GFN2-xTB-An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **15**, 1652–1671 (2019).
77. Spicher, S. & Grimme, S. Robust atomistic modeling of materials, organometallic, and biochemical systems. *Angew Chem Int Ed Engl* **59**, 15665–15673 (2020).
78. Burk, J. *et al.* Fe-Doped ZnO nanoparticle toxicity: assessment by a new generation of nanodescriptors. *Nanoscale* **10**, 21985–21993 (2018).
79. Tämm, K. *et al.* Parametrization of nanoparticles: development of full-particle nanodescriptors. *Nanoscale* **8**, 16243–16250 (2016).
80. Kan, M., Shumyatcher, M., Diwadkar, A., Soliman, G. & Himes, B. E. Integration of Transcriptomic Data Identifies Global and Cell-Specific Asthma-Related Gene Expression Signatures. *AMIA Annu. Symp. Proc.* **2018**, 1338–1347 (2018).
81. Hong, F. & Breitling, R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics* **24**, 374–382 (2008).
82. Lüdecke, D. Esc: Effect Size Computation for Meta Analysis. *CRAN* (2019).
83. Dewey, M. metap: meta-analysis of significance values. *CRAN* (2022).
84. Del Carratore, F. *et al.* RankProd 2.0: a refactored bioconductor package for detecting differentially expressed features in molecular profiling datasets. *Bioinformatics* **33**,

762 2774–2775 (2017).
763 85. Hong, F. *et al.* RankProd: a bioconductor package for detecting differentially expressed
764 genes in meta-analysis. *Bioinformatics* **22**, 2825–2827 (2006).
765 86. Lin, S. Space oriented rank-based data integration. *Stat. Appl. Genet. Mol. Biol.* **9**,
766 Article20 (2010).
767 87. Martens, M. *et al.* WikiPathways: connecting communities. *Nucleic Acids Res.* **49**,
768 D613–D621 (2021).
769 88. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**,
770 25–29 (2000).
771 89. Jassal, B. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **48**, D498–
772 D503 (2020).
773 90. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG:
774 integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**, D545–D551 (2021).
775 91. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**,
776 1739–1740 (2011).
777 92. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *BioRxiv* (2016)
778 doi:10.1101/060012.
779 93. Xie, Z. *et al.* Gene Set Knowledge Discovery with Enrichr. *Curr. Protoc.* **1**, e90 (2021).
780 94. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server
781 2016 update. *Nucleic Acids Res.* **44**, W90-7 (2016).
782 95. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment
783 analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
784 96. Signorell, A. DescTools: Tools for Descriptive Statistics. *R package version 0.99.43*
785 (2021).
786 97. Scala, G., Serra, A., Marwah, V. S., Saarimäki, L. A. & Greco, D. FunMappOne: a tool

to hierarchically organize and visually navigate functional gene annotations in multiple experiments. *BMC Bioinformatics* **20**, 79 (2019).

98. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).
99. Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87–D92 (2020).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigureS1.pdf
- FigureS2.pdf
- TableS1.pdf
- TableS2.pdf
- TableS3.pdf
- TableS4.pdf
- TableS5.pdf
- TableS6.pdf
- TableS7.pdf
- TableS8.pdf
- TableS9.pdf
- TableS10.pdf
- TableS11.pdf