Machine Learning classifier built with heavy metal signature biomarker genes as features to distinguish between heavy metal exposure from non-heavy metal exposure gene expression samples

Shradha Mukherjee

Title: Machine Learning classifier built with heavy metal signature biomarker genes as features to distinguish between heavy metal exposure from non-heavy metal exposure gene expression samples.

NAMES AND AFFILIATIONS OF EACH AUTHOR

Shradha Mukherjee, PhD
Independent Researcher.
This work is self-owned and not affiliated with any organization/university.

CORRESPONDING AUTHOR'S NAME AND EMAIL ADDRESS

Address correspondence to: Shradha Mukherjee; Independent Researcher, Calcutta, West Bengal, India
Email smukher2@yahoo.com

FINANCIAL INTERESTS OR CONFLICT OF INTEREST (IF APPLICABLE)

NA

Abstract:

There are over 350,000 registered chemicals and chemical combinations in use today globally. This is a public health concern and an active area of research, as for majority of these chemicals no scientific data is available on potential adverse effects on human health. Both U.S. Environmental Protection Agency (EPA) and Word Health Organization (WHO) have listed heavy metals, lead (Pb), mercury (Hg), cadmium (Cd) and arsenic (As) among the top chemicals of public health concern. Adverse health effects, induced by heavy metals and other non-heavy metal chemicals include neurodegenerative diseases (Alzheimer's disease, Parkinson's disease), cognitive decline, behavioral problems, kidney diseases, cancer and cardiovascular diseases. Thus, it is important to detect not only active chemical exposure but also past chemical exposures. In this paper differential gene expression (DEG) analysis and machine learning (ML) were combined to identify differentially expressed genes (DEGs) or heavy metal toxicity signature genes that were used as features in ML to classify test samples into heavy metal and non-heavy metal control groups. From NIH-GEO, RNA-seq gene expression data from a total of 827 human neuronal cell culture samples treated with 87 different chemicals were downloaded and normalized. Two groups of DEGs consisting of 80 genes (consensus of limma, edgeR and simple DEG analysis) and 879 genes (consensus of atleast 2 of the three DEG methods limma, edgeR and simple) were identified and designated as heavy metal biomarkers. The heavy metal biomarker gene sets were enriched with metal metabolism gene ontology, kidney disease and cancer diseases genes. Comparison of different ML models built with 80 DEGs and 879 DEGs showed that Logistic Regression and Support Vector Machine (SVM) were accurate (>90% success in classifying test samples into heavy metal and non-heavy metal groups) for both 80 DEG and 879 DEG features. In this paper, a combined DEG analysis and ML pipeline has been developed that can successfully detect heavy metal exposure from gene expression data. This pipeline can be applied for identification of chemical exposure, which is the first step for developing a treatment plan for patients exposed to toxic chemicals.

Introduction:

There are over 350,000 registered chemicals and chemical combinations in use today globally[1]. This is a public health concern and an active area of research, as for majority of these chemicals no scientific data is available on potential adverse effects on human health[2]. Children are especially susceptible to chemical exposure induced diseases such as autism, cerebral palsy, mental retardation, obesity and respiratory diseases[3]. Both U.S. Environmental Protection Agency (EPA) and Word Health Organization (WHO) have listed heavy metals, lead (Pb), mercury (Hg), cadmium (Cd) and arsenic (As) among the top chemicals of public health concern[4; 5]. Heavy metal refers to chemical elements with high molecular weight that cause toxicity in humans. As heavy metals accumulate in the human body overtime (bioaccumulate), exposure to even small amounts of heavy metals can cause toxicity and adverse health effects. Adverse health effects, induced by heavy metals and other non-heavy metal chemicals include neurodegenerative diseases (Alzheimer's disease, Parkinson's disease), cognitive decline, behavioral problems, kidney diseases, cancer and cardiovascular diseases[6].

Long-term adverse effects induced by chemical exposure in humans, remain even when the chemical exposure itself has stopped. Animal studies showed that chronic lead (Pb) exposure results in lead accumulation in choroid plexus and correlates with reduced production of transthyretin by choroid plexus, required for regulation of thyroid hormone critical for prenatal and early postnatal development [7; 8]. Chronic exposure to chromium (Cr) and arsenic (As) through drinking water increased incidence and size of tumors in mice [9]. Exposure to low levels of mercury (Hg) in form of methylmercury (MeHg) prenatally was associated with learning and memory deficits in children [10]. Long-term effects may occur because of persistent gene expression changes that were induced by the chemical during exposure, gene expression changes induced by release of bioaccumulated chemicals later in life or due to other hereto unknown reasons. Though the mechanism of how chemical induced gene effects persist after chemical exposure has stopped is not clearly understood, it is certain that as genes regulate all biological phenotypes, detection of these gene expression changes during the chemical exposure and after chemical exposure has stopped, is the first step towards understanding and treating the adverse disease phenotypes. Identification of chemicals by analyzing its induced gene expression profile, must also account for presence of other chemicals that work in combination with it, to produce a net chemical induced gene expression profile in humans.

Accurate toxicity predictions have the potential can reduce uncertainty and expense of clinical trials of drugs. To decipher chemical effects, artificial intelligence (AI) and machine learning (ML) techniques are now being widely used. AI/ML algorithms can learn from patterns in chemical assay results (activity) and chemical characteristics (structural) data from known chemicals, and build AI/ML models that can make predictions about unknown chemicals. Stress assay results from chemical-protein binding assays, known toxicity endpoints (hepatotoxicity, oral toxicity, cardiotoxicity, mutagenicity) and types of chemical functional groups in chemical structure are some of the training data features employed to build toxicity predictor AI/ML models. These

chemical structure and activity parameters called Quantitative Structure-Activity Relationships (QSAR) parameters are commonly used for AI/ML methods, while chemical induced gene expression data is not always used as training data features to build toxicity predictor AI/ML models. eToxPred is a ML model trained on chemical structural features and toxicity classification data for known chemicals, and can predict toxicity of new unknown chemicals [11]. DeepTox, a deep learning model normalizes and computes features to build a machine learning model that can predict toxicity of new unknown chemical compounds [12]. DTox (Deep Learning for Toxicology) is a deep learning model that takes as input chemical structure to predict its probable toxicity assay result and part of its gene expression profile underlying toxicity effects [13].

Neurotoxicity of a chemical is an underutilized toxicity endpoint relative to other toxicity endpoints hepatotoxicity, oral toxicity, cardiotoxicity and mutagenicity in AI/ML models because its difficult to maintain neural samples in culture that represent in vivo neurotoxicity. Now, its known that AI/ML models perform better when they are trained with more volume of data and more data types. However, comparison of ability of ML models to reliably predict chemical toxicity was shown to be more robust for 2D neural cell culture, than 3D neural cell culture (organoid) [14; 15]. As more drug induced gene expression data gets generated from chemical treatment of 2D and 3D neural cell culture assays, it would be a missed opportunity to not incorporate all data sources, 2D and 3D, to increase size and type of data utilized for building AI/ML models and potentially increase performance of AI/ML models. Variability in gene expression data originating from different sources, is a known bottleneck for gene expression based meta-analysis. Thus in the present study it was hypothesized that by reducing batch or source variability in gene expression data using normalization, it will be possible to use data from different sources for training a reliable and robust AI/ML model for prediction of chemical toxicity. The goal of the present study was to develop a computational pipeline that would normalize and enable utilization of different data types, 2D and 3D, and from different sources, for building a reliable AI/ML model for toxicity prediction, specifically heavy metal toxicity. Here, a pipeline was developed to normalize gene expression data from different sources and identify heavy metal signature genes (biomarkers). The biomarker genes were used as features train an AI/ML model for prediction of heavy metal toxicity. The ML algorithms, logistic regression and Support Vector Machine (SVM) were able to distinguish between heavy metal and non-heavy metal samples for biomarker gene sets of both 80 genes and 879 genes.

Methods

Code Availability

Computational code with html or pdf rendering showing input and output of code chunks is available as a git local repository at https://icedrive.net/s/h3P65RbNvf5Dh8yT1DabXxyNgWg6 with all files and as a git remote repository at https://gitlab.com/smukher2/pbothers_rnaseq_ml_feb2023 with large files ignored or removed.

Source of RNA-seq gene expression data

NCBI GEO repository https://www.ncbi.nlm.nih.gov/geo/, was searched for RNA-seq gene expression data from human pluripotent or embryonic stem cell derived neural tissue culture samples exposed to different chemicals. The results were filtered to retain those datasets, which contained untreated and/or DMSO solvent treated samples. The series numbers GSE166297, GSE128431, GSE63935 and GSE126786 were selected as they met the filtering criteria.

Normalization of RNA-seq gene expression data

RNA-seq raw counts are number of reads overlapping annotated genes in human gene annotation file Homo_sapiens.GRCh38.p13.gtf. RNA-seq raw counts generated by NCBI were obtained from NCBI GEO repository, https://www.ncbi.nlm.nih.gov/geo/info/rnaseqcounts.html#locate. A bash script with commandline tool wget was used to programmatically download the NCBI generated raw RNA-seq counts available at url of format https://www.ncbi.nlm.nih.gov/geo/download/?acc=<GSE Series ID> example, https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE166297, under 'Series RNA-seq raw counts matrix' at url of format https://www.ncbi.nlm.nih.gov/geo/download/?type=rnaseq_counts&acc=<GSE Series ID>&format=file&file=<GSE Series ID>_raw_counts_GRCh38.p13_NCBI.tsv.gz example, https://www.ncbi.nlm.nih.gov/geo/download/?type=rnaseq_counts&acc=GSE166297&format=file&file=GSE166297_raw_counts_GRCh38.p13_NCBI.tsv.gz. In R/Rstudio using package limma_3.38.3, the downloaded NCBI generated raw counts were quantile normalized and scaled to log2+1. Effect of limma quantile model normalization on batch variation was visualized by comparing by comparing pre-limma quantile normalized (before) and limma quantile normalized (after) expression data plots. Both raw counts (not scaled to log2+1) and normalized counts (scaled to log2+1), were visualized with density plots, Box-Whiskers plots, correlation plots and PCA analyses to estimate effect of normalization on variability between samples from different GSE series. Density plots and Box-Whiskers plots were made using ggplot2_3.1.1, while prcomp stats4_3.5.0 and corrplot_0.84, were used to perform PCA analyses and correlation analysis, respectively [16; 17].

Source of metadata for RNA-seq gene expression data

GEOquery R package was used to obtain metadata for all samples in the GSE series numbers GSE166297, GSE128431, GSE63935 and GSE126786 [18]. The metadata fields, sample id, source tissue (embryonic stem cells or pluripotent stem cells derived neural tissue), organism (homo sapiens) and chemical name (87 unique chemicals) was fetched with GEOquery for each GSE series. In the list of chemicals, lead, arsenic, cadmium and mercury were heavy metals and others were non-heavy metal chemicals.

Identification of heavy metal biomarkers

To identify heavy metal biomarker genes, normalized log2+1 expression values were used as input for estimation of DEGs (Differentially Expressed Genes). Gene

expression normalization and differential gene expression (DEG) analysis was done using previously published methods [19; 20]. Differentially expressed genes (DEGs) were identified by comparing gene expression of heavy metal and non-heavy metal exposed samples using limma_3.38.3, edgeR_3.24.3 and simple comparison expression means [21; 22; 23]. In limma and edgeR model design step variable (metadata fields) requires variables have more than one value, thus organism was not used as it contained only one value 'homo sapiens'. DEGs between heavy metal and non-heavy metal was calculated while correction for false discovery and multiple testing was done using Benjamini and Hochberf (BH) to correct for the presence of multiple genes or features in the analysis [24]. In limma and edgeR, significantly upregulated genes in heavy metal category with BH corrected adjP-values <0.05 and a fold changes >5, were designated as heavy metal specific genes or biomarker genes. In simple comparison of means method of DEG identification, gene expression from heavy metal and non-heavy metal samples was compared, and genes with P-values <0.05 and a fold changes >5, were designated as heavy metal specific genes or biomarker genes. Gene expression of DEGs for all three methods of DEG analysis, was visualized with volcano plots, Box-Whiskers plots, barplots and density plots using ggplot2_3.1.1 R package[17]. Consensus strict DEGs were those DEGs that were present in DEG lists of all three methods, limma, edgeR and simple comparison of means. Consensus relaxed DEGs were those DEGs that were present in DEG lists of atleast two of the three methods, limma, edgeR and simple comparison of means. Significance of overlap between DEG lists were calculated using GeneOverlap_1.18.0 R package and number of genes overlapping were visualized as Venn-diagrams using VennDiagram_1.6.20 R package[25; 26].

Gene Ontology (GO) analysis of heavy metal biomarker genes

To determine biological significance of the heavy metal biomarkers, EnrichR_1.0 a R package was performed to determine biological process gene ontology (GO) enriched in the heavy metal strict DEGs and relaxed DEGs [27]. GOs with adjP-values <0.05 were considered significant GOs.

Machine Learning Models

Two sets of differentially expressed genes were used as features (set of 80 DEGs consensus of all 3 methods and set of 880 DEGs consensus of atleast 2 methods) to train the machine learning algorithm with gene expression data for classifying or distinguishing between heavy metal exposure and non-heavy metal exposure human samples. To avoid class sample size bias, the non-heavy metal chemical exposure samples were not included in the ML training, as there were more non-heavy metal chemical exposure samples than heavy metal chemical exposure samples. Only 33 heavy metal chemical exposure samples, 25 untreated control samples and 15 solvent DMSO treated control samples, were included in the ML training. In Python/Spyder/Jupyter Notebook, samples were split into training (70%) and test datasets (30%) using scikit-learn train_test_split function. The models were built with scikit-learn using five ML methods Logistic Regression, Support Vector Machine (SVM), Naïve Bayes, K-means, Random Forest, and XGB.

For ML model building, supervised learning methods learn about patterns in dataset from labeled datasets, while unsupervised learning methods learn about patterns in the dataset without being given labeled datasets. Logistic Regression is a supervised learning method, which relies on odds ratio calculation to predict the probability of an event (here heavy metal and non-heavy metal) occurrence depending on value of input features (here biomarker genes). SVM is a supervised learning method, which uses labeled training dataset to create a partition or hyperplane that separates the groups (here heavy metal and non-heavy metal) using input features (here biomarker genes). Naïve Bayes, is a supervised learning method, which classifies based on probability of a feature (here biomarker genes) occurring in a group (here heavy metal and non-heavy metal). K-means, is an unsupervised leaning method, in which the datasets are split into clusters or groups (here heavy metal and non-heavy metal) based on their input features (here biomarker genes), and by an iterative process the centroid of the cluster is matched with the mean of the samples assigned into the cluster until changing assignment of samples in the cluster does not change the mean or centroid does not move anymore. Random Forest is a supervised learning method, which consists of many decision trees built using features (here biomarker genes) and prediction of class (here heavy metal and non-heavy metal) is done by calculating average prediction of each tree. XGB or XGBoost, is another decision tree based method that is fast because of parallel processing of decision trees.

Evaluation of Machine Learning (ML) heavy metal and non-heavy metal classifier models

To compare performance of these ML models, accuracy, f1-score, ROC curve and AUC were calculated and confusion matrix was plotted. To visually display quality of classification models a confusion matrix was plotted for each using scikit-learn_1.2.1 python package[28]. In confusion matrix, the columns represent model predicted labels and rows represent true labels of samples in the test dataset. Receiver Operating Characteristic (ROC) curve was plotted using scikit-learn_1.2.1 python package, where Area Under the Curve (AUC) in the ROC plot indicates resolving power of classifier models[28]. Accuracy, precision, recall and f1-score were calculated using classification_report function of the scikit-learn_1.2.1 python package[28].

Results:

Quantile normalization reduces batch effects in RNA-seq gene expression count data

As publicly available chemical exposure induced neural tissue gene expression count data was obtained from different research studies or batches, the counts were normalized to reduce cross-study variability and effect of normalization was visualized with different plots. Density and Box-Whiskers plots showed a greater overlap of samples from different batches and greater overlap of gene expression means, respectively, in limma quantile normalized count data relative to pre-normalized count data (Figure 1A, Figure 1B). Correlation plots showed greater correlation of batches after quantile normalization relative to that before quantile normalization (Figure 1C). PCA plots showed that variability or wide dispersion of samples was less in quantile normalized count data relative to pre-normalized count data (Figure 1D). Thus, limma

quantile normalization reduced variability unrelated to chemical exposure gene expression counts, making the data suitable for further analysis.

Limma identified 85 DEGs upregulated in heavy metals relative to other samples

To identify heavy metal signature genes, heavy metal samples were contrasted with samples that were treated with non-heavy metal samples (other chemical treatments and untreated controls), using limma DEG analysis. DEGs significantly upregulated or downregulated (p-value < 0.05) in heavy metals relative to other samples were visualized in volcano plot (Figure 2A). From the DEGs, 85 genes were significantly (p-value < 0.05) upregulated (fold-change > 5) in heavy metals relative to other samples were putative heavy metal signature genes. Density, Box-Whiskers and bar plots showed that the average normalized gene expression count of upregulated heavy metal DEGs was higher in heavy metal samples relative to non-heavy metal samples (Figure 2B, Figure 2C, Figure 2D).

EdgeR identified 1072 DEGs upregulated in heavy metals relative to other samples

To identify heavy metal signature genes, heavy metal samples were contrasted with samples that were treated with non-heavy metal samples (other chemical treatments and untreated controls), using edgeR DEG analysis. DEGs significantly upregulated or downregulated (p-value < 0.05) in heavy metals relative to other samples were visualized in volcano plot (Figure 3A). From the DEGs, 1072 genes were significantly (p-value < 0.05) upregulated (fold-change > 5) in heavy metals relative to other samples were putative heavy metal signature genes. Density, Box-Whiskers and bar plots showed that the average normalized gene expression count of upregulated heavy metal DEGs was higher in heavy metal samples relative to non-heavy metal samples (Figure 3B, Figure 3C, Figure 3D).

Simple method identified 2237 DEGs upregulated in heavy metals relative to other samples

To identify heavy metal signature genes, heavy metal samples were contrasted with samples that were treated with non-heavy metal samples (other chemical treatments and untreated controls), using edgeR DEG analysis. DEGs significantly upregulated or downregulated (p-value < 0.05) in heavy metals relative to other samples were visualized in volcano plot (Figure 4A). From the DEGs, 2237 genes were significantly (p-value < 0.05) upregulated (fold-change > 5) in heavy metals relative to other samples were putative heavy metal signature genes. Density, Box-Whiskers and bar plots showed that the average normalized gene expression count of upregulated heavy metal DEGs was higher in heavy metal samples relative to non-heavy metal samples (Figure 4B, Figure 4C, Figure 4D).

Heavy metal biomarker genes identified by overlaping of limma, edgeR and simple DEGs

Overlapping DEGs identified by limma, edgeR and simple comparison of means methods, identified 80 DEGs common to all 3 methods (strict overlap), and 879 DEGs common to atleast 2 methods (relaxed overlap) (Figure 5A). Jaccard index calculations

showed that edgeR and limma has significant overlap, while simple method did not significantly overlap with edgeR or limma (Figure 5B).

Gene Ontology (GO) of heavy metal biomarkers

Metal related cellular response and homeostasis were the most enriched biological processes in heavy metal GO Biological Process (Figure 5C, Figure 5D). GOs related to cellular response to metals, "cellular response to zinc ion" (GO:0071294 p-values 1.71E-10 and 1.87E-05), "response to copper ion" (GO:0046688 p-values 1.71E-10 and 1.87E-05) and "cellular response to cadmium ion" (GO:0071276 p-values 1.47E-09 and 5.11E-05), were the most significant GOs in both heavy metal biomarkers in strict group of 80 genes and relaxed group of 879 genes (Figure 5C, Figure 5D). These results are consistent with the group of genes being heavy metal biomarkers. GOs "cellular response to unfolded protein" (GO:0034620 p-value 5.29E-08) and "negative regulation of growth" (GO:0045926 p-value 4.88E-09) were significant in heavy metal biomarkers in strict group of 80 genes (Figure 5C).

Performance of ML models with heavy metal 80 and 879 biomarker as features

To build and test ML models, randomly picked 70% of samples (25 heavy metal and 26 control DMSO or untreated samples) were used for training and 30% of samples (8 heavy metal and 14 control DMSO or untreated samples) were used as testing datasets (Figure 6A, Figure 7A). A total of seven models, Logistic Regression, K-means, Naïve-Bayes, SVM, Random Forest, XGB with grid and XGB without grid were compared to determine their ability to distinguish or classify heavy metal and control non-heavy metal samples. The performance of the models was compared using confusion matrix, accuracy, precision, f1-score, recall, ROC and AUC (Figure 6, Figure 7). For 80 heavy metal biomarker genes as features, both Logistic Regression and SVM outperformed other models, with 95.4% accuracy, 100% precision, 0.94 AUC, 0.875 recall and 0.933 f1-score (Figure 6I). Random Forest also showed good performance, with 90.9% accuracy, 100% precision, 0.88 AUC, 0.75 recall and 0.857 f1-score (Figure 6I). Other models, K-Means, Naïve-Bayes, XGB with grid and XGB without grid showed an accuracy between 27.27% to 81.81% (Figure 6I). For 879 heavy metal biomarker genes as features, both Logistic Regression and SVM outperformed other models, with 95.4% accuracy, 100% precision, 0.94 AUC, 0.875 recall and 0.933 f1-score (Figure 7I). XGB without grid also showed good performance, with 90.9% accuracy, 100% precision, 0.88 AUC, 0.75 recall and 0.857 f1-score (Figure 7I). Other models, K-Means, Naïve-Bayes, Random Forest and XGB with grid showed an accuracy between 45.45% to 86.36% (Figure 7I). For both 80 genes and 879 gene features (heavy metal biomarkers), confusion matrix constructed from the different ML models showed visually the true and predicted, heavy metal and control non-heavy metal labels consistent with the precision, AUC, recall and f1-score calculations (Figure 6B-H, Figure 7B-H). Taken together, these results show that Logistic Regression and SVM outperform other models with 879 gene features, as well as 80 gene features.

Discussion:

Neurotoxicity cell culture models

The nervous system is protected by the blood-brain-barrier form many infectious agents and harmful chemicals. However, this protection weakens making the nervous system susceptible to chemicals such as neurotoxic heavy metals, when the blood-brain-barrier shows dysfunction in adults with aging and disease [29; 30]. Thus, it is important to study neurotoxicity to enable detection of neurotoxicity, understand its molecular mechanisms and discover therapeutic targets for intervention. However, this has been challenging as there are no standard neural cell culture protocols, so gene expression data originating from different 2D or 3D neural culture conditions is highly variable. In iPSC derived organoids developed to study Alzheimer's disease, large variability was found in morphology and electrophysiological activity of neurons inside organoids even when they were prepared from the same cell line [31]. This variability could result from inherent stochastic nature of in vitro self-organization which makes neural differentiation process and neuronal cell type characteristics inside organoids variable. In the present study, a pipeline was developed to reduce variability in gene expression with normalization to make it usable for detection of heavy metal neurotoxicity with ML using heavy metal molecular biomarkers as features.

Meta-analysis and normalization of RNA-seq

Most gene expression RNA-seq raw reads from published studies are stored in National Centre for Biotechnology Information Sequence Read Archive (NCBI SRA) or National Centre for Biotechnology Information Gene Expression Omnibus (NCBI GEO) public repository that can be used for meta-analysis. Meta-analysis involves integration of samples from different studies related to a research topic, to increase sample size which improves robustness of results. Different studies have processing variability as they use different methods to convert raw RNA-seq reads to annotated gene expression [32]. Thus, for uniformity in meta-analysis combining gene expression data from different studies requires extensive pre-processing where all the raw reads from different studies are processed with same pipeline to annotated gene expression. To overcome this caveat of having to re-run RNA-seq data from different studies available in the public NCBI GEO repository, NCBI GEO made annotated gene expression available for RNA-seq studies in their repository. Annotated gene expression has variability due to different source of tissue origin that can be normalized using several methods such as TPM, FPKM and quantile normalization. Comparison of these normalization methods in a study of human tumour xenograft showed normalization of counts had lower median coefficient of variation than FPKM and TPM normalization [33]. Thus, to reduce variation in annotated gene expression downloaded from NCBI GEO, quantile normalization method was applied. Quantile normalization of annotated gene expression for meta-analysis of combined GSE166297, GSE128431, GSE63935 and GSE126786 batches, significantly reduced batch variability (Figure 1).

Feature or biomarker gene selection for ML

Feature selection is process of selecting most informative features, here genes, which are most likely to be relevant distinguishing characteristics of the sample labels, here heavy metal toxicity and non-heavy metal toxicity. More features a ML model has, more

time it takes to run and more challenging it becomes to understand effects of features on classification prediction made by ML model. Feature selection algorithms can be broadly classified into filter methods, wrapper methods, embedded methods and hybrid methods [34]. However, these feature selection methods are based on 'trial and error' where features or genes are put through several rounds of iteration, features are pruned after each iteration and model is re-build with only those features that are most relevant to ML model's prediction accuracy. In the present study to make feature selection for more explainable, instead of prevalent computational feature selection methods, bioinformatics DEG selection methods were used. It was hypothesized that DEG genes that are significantly upregulated in heavy metals toxicity, designated heavy metal toxicity biomarkers, by virtue of their biological relevance will help create better ML models. Heavy metal DEGs or biomarkers that were upregulated significantly by atleast 5 fold, were identified with three commonly used DEG methods, limma, edgeR and simple comparison of means (Figure 2, Figure 3, Figure 4). Overlap of all 3 DEG methods resulted in 80 genes (strict overlap) and overlap of atleast 2 DEG methods resulted in 879 (relaxed overlap) genes (Figure 5A, Figure 5B). ML models were built with both 80 genes and 879 genes, so that robustness of the ML models could be tested.

Insights into selected biomarker features from GO analysis

GO analysis of 80 genes and 879 genes was done to understand molecular mechanism of heavy metal neurotoxicity and make the ML model more explainable, by determination of major biological groups to which the selected features belonged. All the GOs related to cellular response to metals, contained metallothionein (MT) genes MT2A, MT1M, MT1F, MT1G, MT1X, MT1H and MT1E (Figure 5C, Figure 5D). Metallothionein proteins are rich in amino acid cysteine (formula HOOC−CH(−NH 2)−CH 2 −SH) and the sulphur (S) gives them ability to bind with metal ions. MT-1 and MT2 family of MTs are induced by presence of metals such as zinc (Zn), cadmium (Cd), copper (Cu) and lead (Pb) [35; 36; 37]. On the DNA, MT genes contain metal response elements (MREs) on their promoters, that get activated by binding of metal regulatory transcription factor (MTFs). The binding of MTFs to MREs and consequent activation of MT genes is regulated by present of heavy metals [36; 38]. MTs are required for zinc (Zn) and copper (Cu) homeostasis, oxidative stress response and detoxification of heavy metals from the body [39; 40].

Robustness of ML models for heavy metal toxicity classification

AI and ML models are being widely investigated, and applied in biomedical field. Artificial intelligence (AI) used in monitoring cancer disease state by analysis of radiographic images from patients can provide quantitative assessment, while physicians can only provide a qualitative assessment [41]. Machine Learning (ML) analysis revealed cancer biomarkers and therapeutic targets in soft tissue sarcoma datasets [42]. Research into ML models for classification is focused on development of new ML models (algorithms), reducing errors in ML models and finding avenues where ML models can be applied. Thus, here existing ML models were applied and tested in the context of heavy metal exposure detection. As the dataset was imbalanced, with more samples for other types of chemicals than heavy metals, to avoid overfitting issue,

for training and testing of ML models, only heavy metals and controls were used as their sample size was comparable (Figure 6A, Figure 7A). This comparable number of samples in each group is conducive to Logistic Regression modelling and makes it well suited for comparison with other models. Comparison of ML models to determine their ability to distinguish or classify heavy metal and control non-heavy metal samples, showed that Logistic Regression and Support Vector Machine (SVM) outperformed other models (Figure 6, Figure 7). Logistic Regression and SVM performances were consistent for both 80 feature genes, and 879 feature genes, with 95.4% accuracy, 100% precision, 0.94 AUC, 0.875 recall and 0.933 f1-score (Figure 6B, Figure 6E, Figure 6I, Figure 7B, Figure 7E, Figure 7I). Random Forest does not work well with missing values, while XGB can automatically fill missing values. Now as there were no missing values of gene expression for features of 80 genes and 879 genes, expectedly performance of Random Forest and XGB with/without gird were comparable. The performance of Random Forest and XGB with/without gird, were moderate for both 80 feature genes, and 879 feature genes, with range 72 to 90% accuracy, range 66 to 100% precision, range 0.68 to 0.81 AUC, range 0.5 to 0.75 recall and range 0.57 to 0.85 f1-score (Figure 6F, Figure 6G, Figure 6H, Figure 6I, Figure 7F, Figure 7G, Figure 7H, Figure 7I). The performance of K-Means got worse with increase in number of features (accuracy 27.2% for 80 genes and 68.1% for 879 genes), while the performance of Naïve Bayes got better with increase in number of features (accuracy 63.6% for 80 genes and 45.4% for 879 genes) (Figure 6C, Figure 6D, Figure 6I, Figure 7C, Figure 7D, Figure 7I). This could be because Naïve Bayes model works better with high dimension data, so with more features its performance got better.

Summary

This pipeline combines normalization, DEG analysis, GO analysis, and ML modelling that is a reusable in silico method that can be adapted for assay of various potentially toxic chemicals. This pipeline can be re-used for other datasets to study effects of chemical exposure by detection of neurotoxicity, understand molecular mechanism and discover therapeutic targets for toxic chemicals. Neurodegenerative diseases, bone diseases and cancers are some of the adverse effects of heavy metal toxicity. Detection of active and past transient heavy-metal chemical exposure is critical to device a treatment plan and plan lifestyle changes to safeguard the patient from adverse short-term and long-term effects of heavy metal toxicity. Robust classification of patients into putative heavy metal and non-heavy metal exposure classes, based patient's gene expression profile, will help detect cases of heavy metal toxicity. These results can then guide a healthcare provider to take necessary actions to treat the patient for heavy metal toxicity. The methods developed in this paper can also be applied and extended to distinguish between any other toxic chemicals or chemical combinations and untreated controls. Chemicals for which toxicity scientific data is not available but chemical induced gene expression profile is available in patients, ML methods developed in this paper can be used to determine if the chemical induced gene expression profile is more like toxic chemicals or untreated controls. For example, gene expression profile from a patient could be run through the ML model and scored for similarity with available toxic chemical and non-toxic chemical gene expression profiles.

Bibiography:

[1] Z. Wang, G.W. Walker, D.C.G. Muir, and K. Nagatani-Yoshida, Toward a Global Understanding of Chemical Pollution: A First Comprehensive Analysis of National and Regional Chemical Inventories. Environmental science & technology 54 (2020) 2575-2584.

[2] E.D. Pellizzari, T.J. Woodruff, R.R. Boyles, K. Kannan, P.I. Beamer, J.P. Buckley, A. Wang, Y. Zhu, and D.H. Bennett, Identifying and Prioritizing Chemicals with Uncertain Burden of Exposure: Opportunities for Biomonitoring and Health-Related Research. Environ Health Perspect 127 (2019) 126001.

[3] L.R. Goldman, and S. Koduru, Chemicals in the environment and developmental toxicity to children: a public health and policy perspective. Environ Health Perspect 108 Suppl 3 (2000) 443-8.

[4] WHO, 10 chemicals of public health concern. https://www.who.int/news-room/photo-story/photo-story-detail/10-chemicals-of-public-health-concern (2020).

[5] U.S.E.P.A. (EPA), Chemicals and Toxics Topics _ US EPA. https://www.epa.gov/environmental-topics/chemicals-and-toxics-topics (2022).

[6] M. Balali-Mood, K. Naseri, Z. Tahergorabi, M.R. Khazdair, and M. Sadeghi, Toxic Mechanisms of Five Heavy Metals: Mercury, Lead, Chromium, Cadmium, and Arsenic. Front Pharmacol 12 (2021) 643972.

[7] W. Zheng, W.S. Blaner, and Q. Zhao, Inhibition by lead of production and secretion of transthyretin in the choroid plexus: its relation to thyroxine transport at blood-CSF barrier. Toxicol Appl Pharmacol 155 (1999) 24-31.

[8] W. Zheng, H. Shen, W.S. Blaner, Q. Zhao, X. Ren, and J.H. Graziano, Chronic lead exposure alters transthyretin concentration in rat cerebrospinal fluid: the role of the choroid plexus. Toxicol Appl Pharmacol 139 (1996) 445-50.

[9] X. Wang, A.K. Mandal, H. Saito, J.F. Pulliam, E.Y. Lee, Z.J. Ke, J. Lu, S. Ding, L. Li, B.J. Shelton, T. Tucker, B.M. Evers, Z. Zhang, and X. Shi, Arsenic and chromium in drinking water promote tumorigenesis in a mouse colitis-associated colorectal cancer model and the potential mechanism is ROS-mediated Wnt/beta-catenin signaling pathway. Toxicol Appl Pharmacol 262 (2012) 11-21.

[10] S.T. Orenstein, S.W. Thurston, D.C. Bellinger, J.D. Schwartz, C.J. Amarasiriwardena, L.M. Altshul, and S.A. Korrick, Prenatal organochlorine and methylmercury exposure and memory and learning in school-age children in communities near the New Bedford Harbor Superfund site, Massachusetts. Environ Health Perspect 122 (2014) 1253-9.

[11] L. Pu, M. Naderi, T. Liu, H.C. Wu, S. Mukhopadhyay, and M. Brylinski, eToxPred: a machine learning-based approach to estimate the toxicity of drug candidates. BMC Pharmacol Toxicol 20 (2019) 2.

[12] A. Mayr, G. Klambauer, T. Unterthiner, and S. Hochreiter, DeepTox: Toxicity Prediction using Deep Learning. Frontiers in Environmental Science 3 (2016).

[13] Y. Hao, J.D. Romano, and J.H. Moore, Knowledge-guided deep learning models of drug toxicity improve interpretation. Patterns (N Y) 3 (2022) 100565.

[14] F. Kuusisto, V.S. Costa, Z. Hou, J. Thomson, D. Page, and R. Stewart, Machine learning to predict developmental neurotoxicity with high-throughput data from 2D bio-engineered tissues. Proc Int Conf Mach Learn Appl 2019 (2019) 293-298.

[15] M.P. Schwartz, Z. Hou, N.E. Propson, J. Zhang, C.J. Engstrom, V. Santos Costa, P. Jiang, B.K. Nguyen, J.M. Bolin, W. Daly, Y. Wang, R. Stewart, C.D. Page, W.L. Murphy, and J.A. Thomson, Human pluripotent stem cell-derived neural constructs for predicting neural toxicity. Proc Natl Acad Sci U S A 112 (2015) 12516-21.

[16] T. Wei, and V. Simko, R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). (2017).

[17] H. Wickham, ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York (2016).

[18] S. Davis, and P.S. Meltzer, GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics 23 (2007) 1846-7.
[19] S. Mukherjee, Immune gene network of neurological diseases: Multiple sclerosis (MS), Alzheimer's disease (AD), Parkinson's disease (PD) and Huntington's disease (HD). Heliyon 7 (2021) e08518.
[20] S. Mukherjee, Quiescent stem cell marker genes in glioma gene networks are sufficient to distinguish between normal and glioblastoma (GBM) samples. Scientific reports 10 (2020) 10937.
[21] M.E. Ritchie, B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi, and G.K. Smyth, limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research 43 (2015) e47.
[22] M.D. Robinson, D.J. McCarthy, and G.K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26 (2010) 139-40.
[23] Z. Zhao, F. Meng, W. Wang, Z. Wang, C. Zhang, and T. Jiang, Comprehensive RNA-seq transcriptomic profiling in the malignant progression of gliomas. Scientific data 4 (2017) 170024.
[24] A. Reiner, D. Yekutieli, and Y. Benjamini, Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 19 (2003) 368-75.
[25] L. Shen, and M. Sinai, GeneOverlap: Test and visualize gene overlaps. R package version 1.22.0 (2019).
[26] H. Chen, and P.C. Boutros, VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC bioinformatics 12 (2011) 35.
[27] E.Y. Chen, C.M. Tan, Y. Kou, Q. Duan, Z. Wang, G.V. Meirelles, N.R. Clark, and A. Ma'ayan, Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC bioinformatics 14 (2013) 128.
[28] Pedregosa F, G.e. Varoquaux, Gramfort A, Michel V, Thirion B, Grisel O, and e. al., Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12 (2011) 2825-2830.
[29] E.G. Knox, M.R. Aburto, G. Clarke, J.F. Cryan, and C.M. O'Driscoll, The blood-brain barrier in aging and neurodegeneration. Mol Psychiatry 27 (2022) 2659-2673.
[30] W. Zheng, M. Aschner, and J.F. Ghersi-Egea, Brain barrier systems: a new frontier in metal neurotoxicological research. Toxicol Appl Pharmacol 192 (2003) 1-11.
[31] J.H. Lee, G. Yoo, J. Choi, S.H. Park, H. Shin, R. Prasad, Y. Lee, M.R. Ahn, I.J. Cho, and W. Sun, Cell-line dependency in cerebral organoid induction: cautionary observations in Alzheimer's disease patient-derived induced pluripotent stem cells. Mol Brain 15 (2022) 46.
[32] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M.W. Szczesniak, D.J. Gaffney, L.L. Elo, X. Zhang, and A. Mortazavi, A survey of best practices for RNA-seq data analysis. Genome Biol 17 (2016) 13.
[33] Y. Zhao, M.C. Li, M.M. Konate, L. Chen, B. Das, C. Karlovich, P.M. Williams, Y.A. Evrard, J.H. Doroshow, and L.M. McShane, TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. J Transl Med 19 (2021) 269.
[34] N. Pudjihartono, T. Fadason, A.W. Kempa-Liehr, and J.M. O'Sullivan, A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Front Bioinform 2 (2022) 927312.
[35] E. Cobb, J. Hall, and D.L. Palazzolo, Induction of Metallothionein Expression After Exposure to Conventional Cigarette Smoke but Not Electronic Cigarette (ECIG)-Generated Aerosol in Caenorhabditis elegans. Front Physiol 9 (2018) 426.
[36] R.D. Palmiter, Regulation of metallothionein genes by heavy metals appears to be mediated by a zinc-sensitive inhibitor that interacts with a constitutively active transcription factor, MTF-1. Proc Natl Acad Sci U S A 91 (1994) 1219-23.

[37] H. Ikebuchi, R. Teshima, K. Suzuki, T. Terao, and Y. Yamane, Simultaneous induction of Pb-metallothionein-like protein and Zn-thionein in the liver of rats given lead acetate. Biochem J 233 (1986) 541-6.
[38] M.K. Yagle, and R.D. Palmiter, Coordinate regulation of mouse metallothionein I and II genes by heavy metals and glucocorticoids. Mol Cell Biol 5 (1985) 291-4.
[39] R. Dallinger, B. Berger, P. Hunziker, and J.H. Kagi, Metallothionein in snail Cd and Cu metabolism. Nature 388 (1997) 237-8.
[40] D.H. Hamer, Metallothionein. Annu Rev Biochem 55 (1986) 913-51.
[41] A. Hosny, C. Parmar, J. Quackenbush, L.H. Schwartz, and H. Aerts, Artificial intelligence in radiology. Nat Rev Cancer 18 (2018) 500-510.
[42] I.D.G.P. van, K. Szuhai, I.H. Briaire-de Bruijn, M. Kostine, M.L. Kuijjer, and J. Bovee, Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas. PLoS Comput Biol 15 (2019) e1006826.

Figure Legend

Figure 1. Effect of quantile normalization on chemical exposure and control RNA-seq datasets (GSE166297, GSE128431, GSE63935 and GSE126786). **A:** Density plot representation of gene expression per study before and after quantile normalization. **B:** Box and whisker plot representation of gene expression per study before and after quantile normalization. **C:** Pearson correlation plot of studies before and after quantile normalization. **D:** Principal Component Analysis (PCA) of datasets before and after quantile normalization. For this figure the following R packages and built-in R functions were used: prcomp from stats4_3.5.0, corrplot_0.84, ggplot2_3.1.1 and limma_3.38.3.

Figure 2. Heavy metal biomarkers or DEGs identified with limma. **A:** Volcano plot of differentially expressed genes showing logFC (log fold change) and its p-value in -log10(adjusted.p-value). Red dots have significant p-value <0.05 and black boxed genes have fold-change > 5. **B:** Box and whisker plot representation of gene expression per group, here heavy metal and non-heavy metal group. **C:** Density plot representation of gene expression per group, here heavy metal and non-heavy metal group. **D:** Bar plot of logFC (log fold change) of top 100 genes (here 85 genes) upregulated in heavy metal relative to non-heavy metal group. For this figure the following R packages and built-in R functions were used: ggplot2_3.1.1 and limma_3.38.3.

Figure 3. Heavy metal biomarkers or DEGs identified with edgeR. **A:** Volcano plot of differentially expressed genes showing logFC (log fold change) and its p-value in -log10(adjusted.p-value). Red dots have significant p-value <0.05 and black boxed genes have fold-change > 5. **B:** Box and whisker plot representation of gene expression per group, here heavy metal and non-heavy metal group. **C:** Density plot representation of gene expression per group, here heavy metal and non-heavy metal group. **D:** Bar plot of logFC (log fold change) of top 100 genes upregulated in heavy metal relative to non-heavy metal group. For this figure the following R packages and built-in R functions were used: ggplot2_3.1.1 and edgeR_3.24.3.

Figure 4. Heavy metal biomarkers or DEGs identified with simple comparison of means. **A:** Volcano plot of differentially expressed genes showing logFC (log fold change) and its p-value in -log10(adjusted.p-value). Red dots have significant p-value <0.05 and black boxed genes have fold-change > 5. **B:** Box and whisker plot representation of gene expression per group, here heavy metal and non-heavy metal group. **C:** Density plot representation of gene expression per group, here heavy metal and non-heavy metal group. **D:** Bar plot of logFC (log fold change) of top 100 genes upregulated in heavy metal relative to non-heavy metal group. For this figure the following R packages and built-in R functions were used: ggplot2_3.1.1.

Figure 5. Heavy metal biomarker 80 genes from overlap of all 3 DEG methods and 879 biomarker genes from overlap of atleast 2 methods, and their GO analysis. **A:** Venn Diagram for overlapping DEGs significantly upregulated by > 5 FC (fold change). **B:** Heat-map for overlap of significance for genes significantly upregulated by > 5 FC (fold change) in limma, edgeR and simple comparison of means. **C:** GO Analysis (Gene Ontology Analysis) Biological Process of heavy metal biomarker 80 genes from overlap of all 3 DEG methods. **D:** GO Analysis (Gene Ontology Analysis) Biological Process of heavy metal biomarker 879 genes from overlap of atleast 2 methods. For this figure the following R packages and built-in R functions were used: GeneOverlap_1.18.0, VennDiagram_1.6.20, EnrichR_1.0 and ggplot2_3.1.1.

Figure 6. Evaluation of machine learning (ML) models built with heavy metal biomarker 80 genes (strict overlap of DEG methods) as features. **A:** Number of datasets from heavy metal and non-heavy metal groups used to train and test ML models. Confusion matrix and ROC curve for **B:** Logistic Regression AUC=0.94 **C:** K-Means AUC=0.38 **D:** Naïve Bayes AUC=0.58 **E:** Support Vector Machine (SVM) AUC=0.94 **F:** Random Forest AUC=0.88 **G:** XGB Grid AUC=0.68 **H:** XGB No Grid AUC=0.75 **I:** Summary of accuracy, precision, recall and f1-score for all models. For this figure the scikit-learn_1.2.1 python package was used.

Figure 7. Evaluation of machine learning (ML) models built with heavy metal biomarker 879 genes (relaxed overlap of DEG methods) as features. **A:** Number of datasets from heavy metal and non-heavy metal groups used to train and test ML models. Confusion matrix and ROC curve for **B:** Logistic Regression AUC=0.94 **C:** K-Means AUC=0.56 **D:** Naïve Bayes AUC=0.57 **E:** Support Vector Machine (SVM) AUC=0.94 **F:** Random Forest AUC=0.81 **G:** XGB Grid AUC=0.68 **H:** XGB No Grid AUC=0.88 **I:** Summary of accuracy, precision, recall and f1-score for all models. For this figure the scikit-learn_1.2.1 python package was used.

Access to Codes:
https://icedrive.net/s/h3P65RbNvf5Dh8yT1DabXxyNgWg6 and
https://gitlab.com/smukher2/pbothers_rnaseq_ml_feb2023.
Please cite this paper if you use these codes. Thank you.

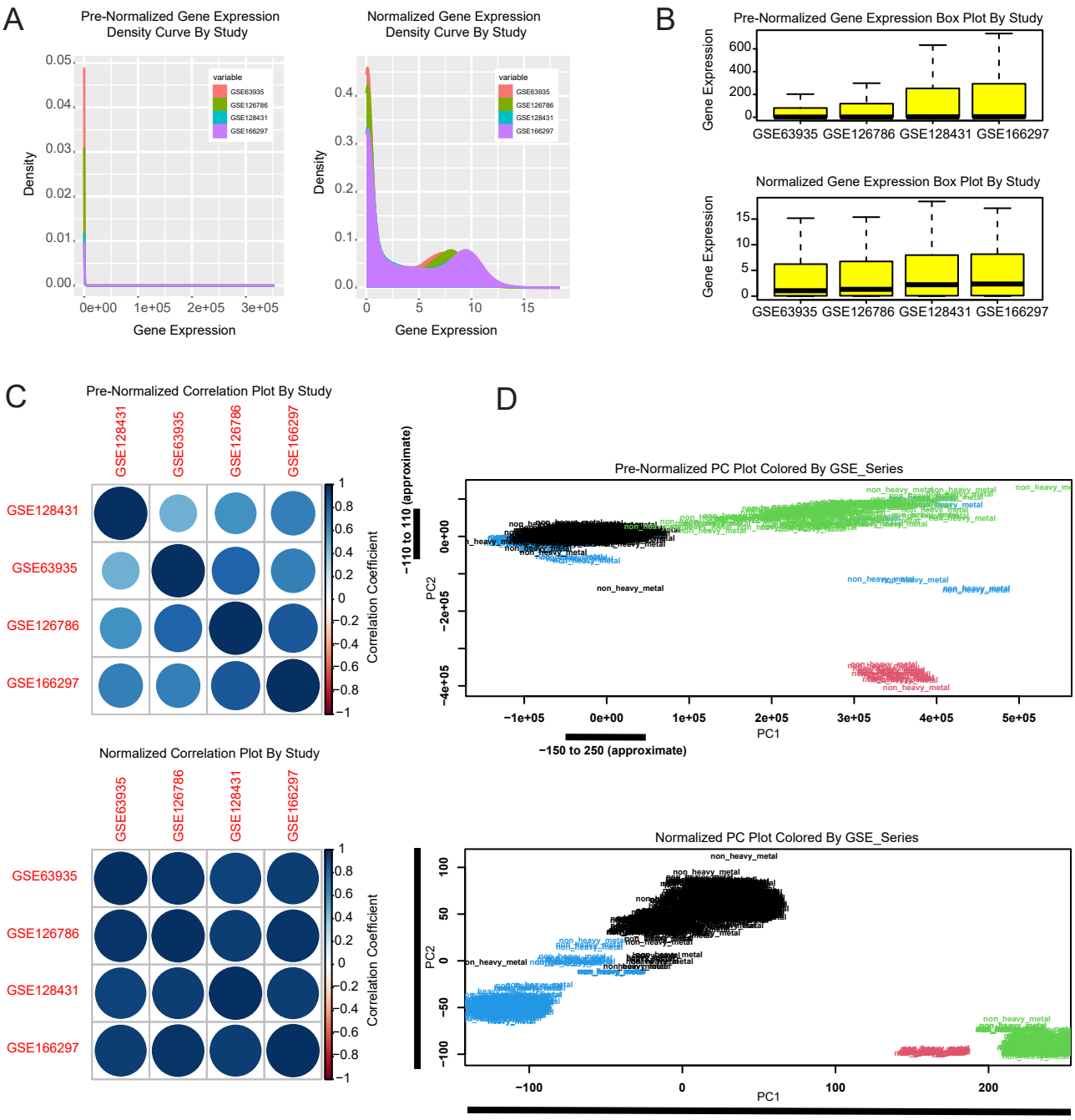


Figure 1 Mukherjee

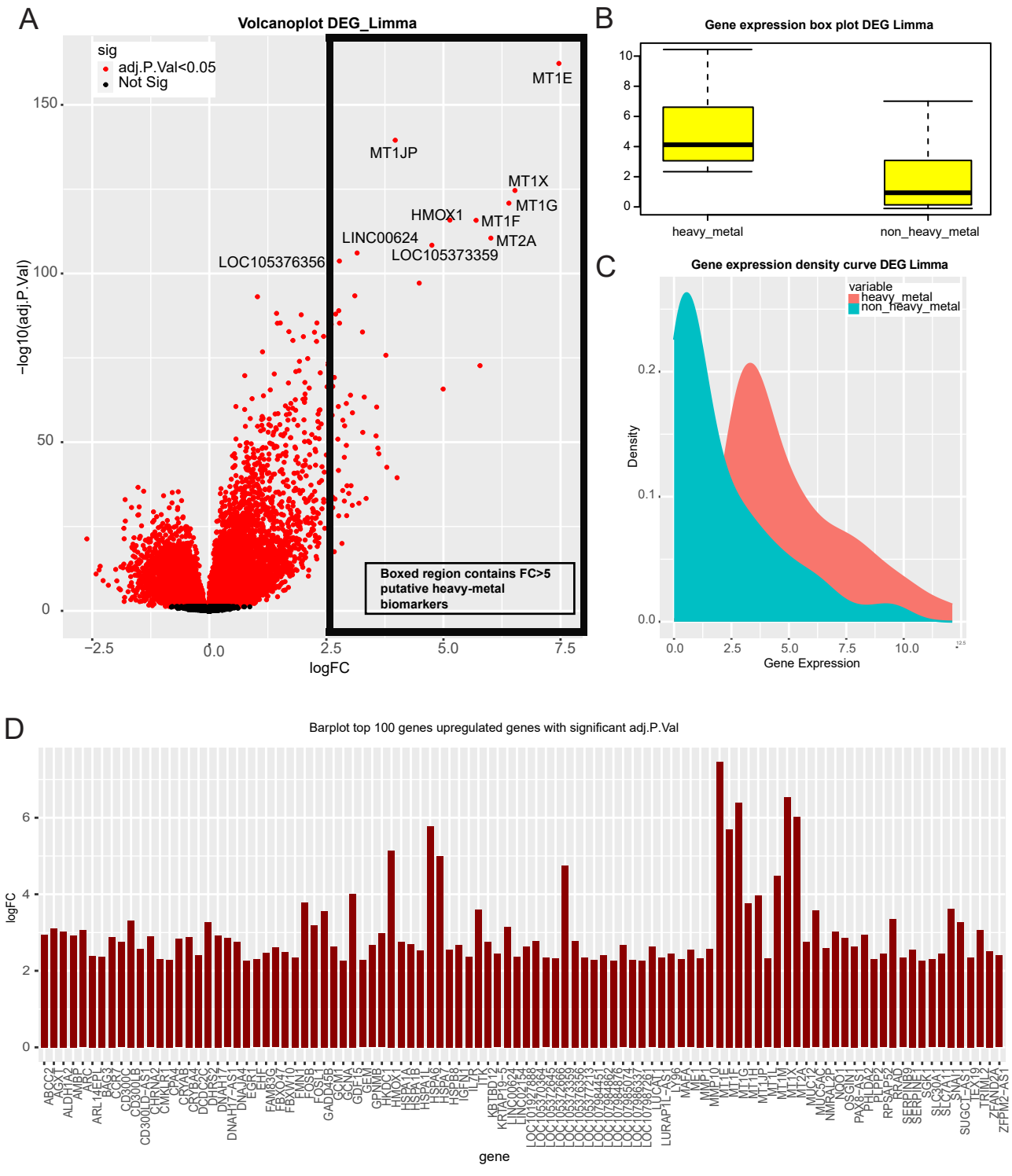


Figure 2 Mukherjee

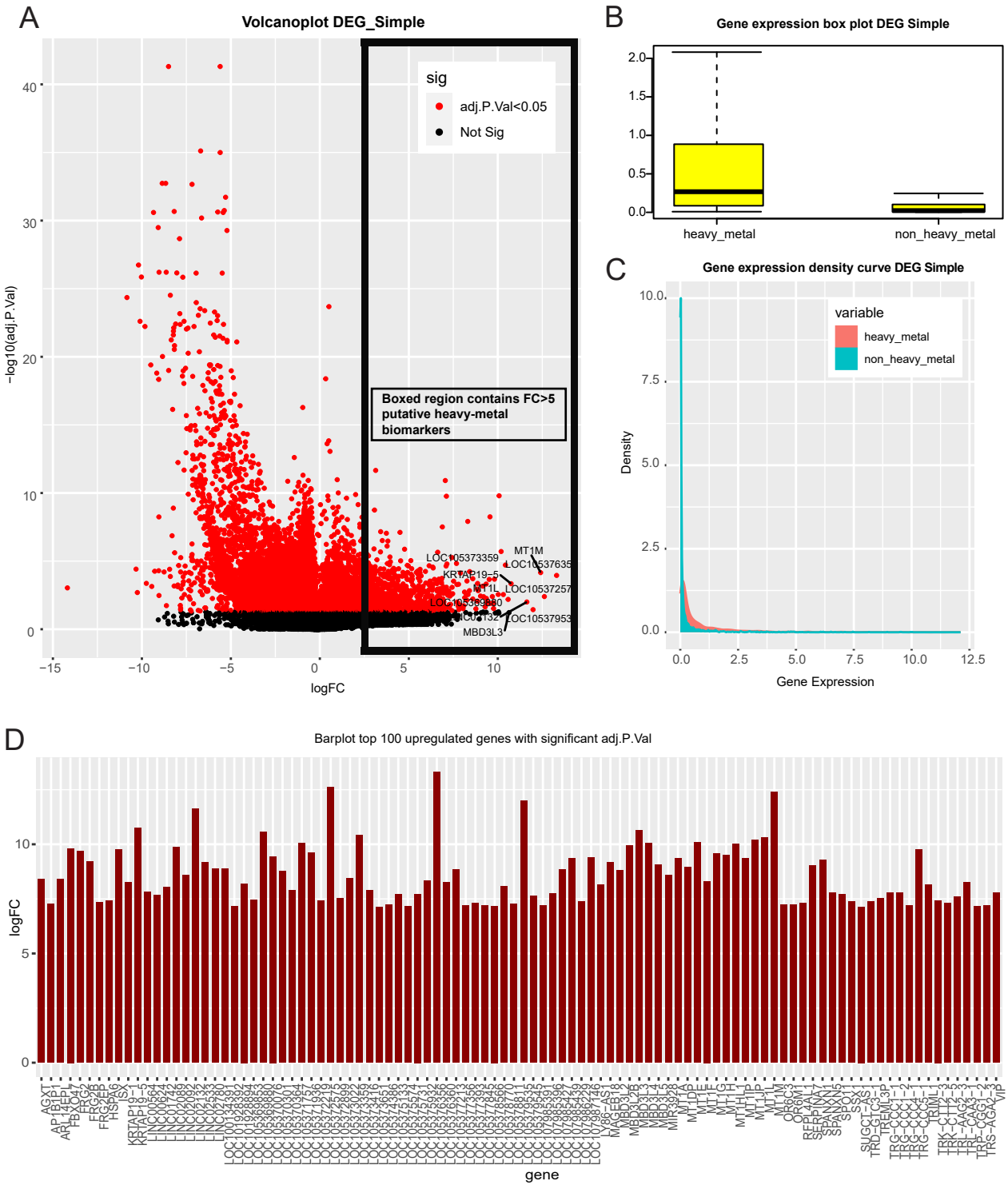


Figure 3 Mukherjee

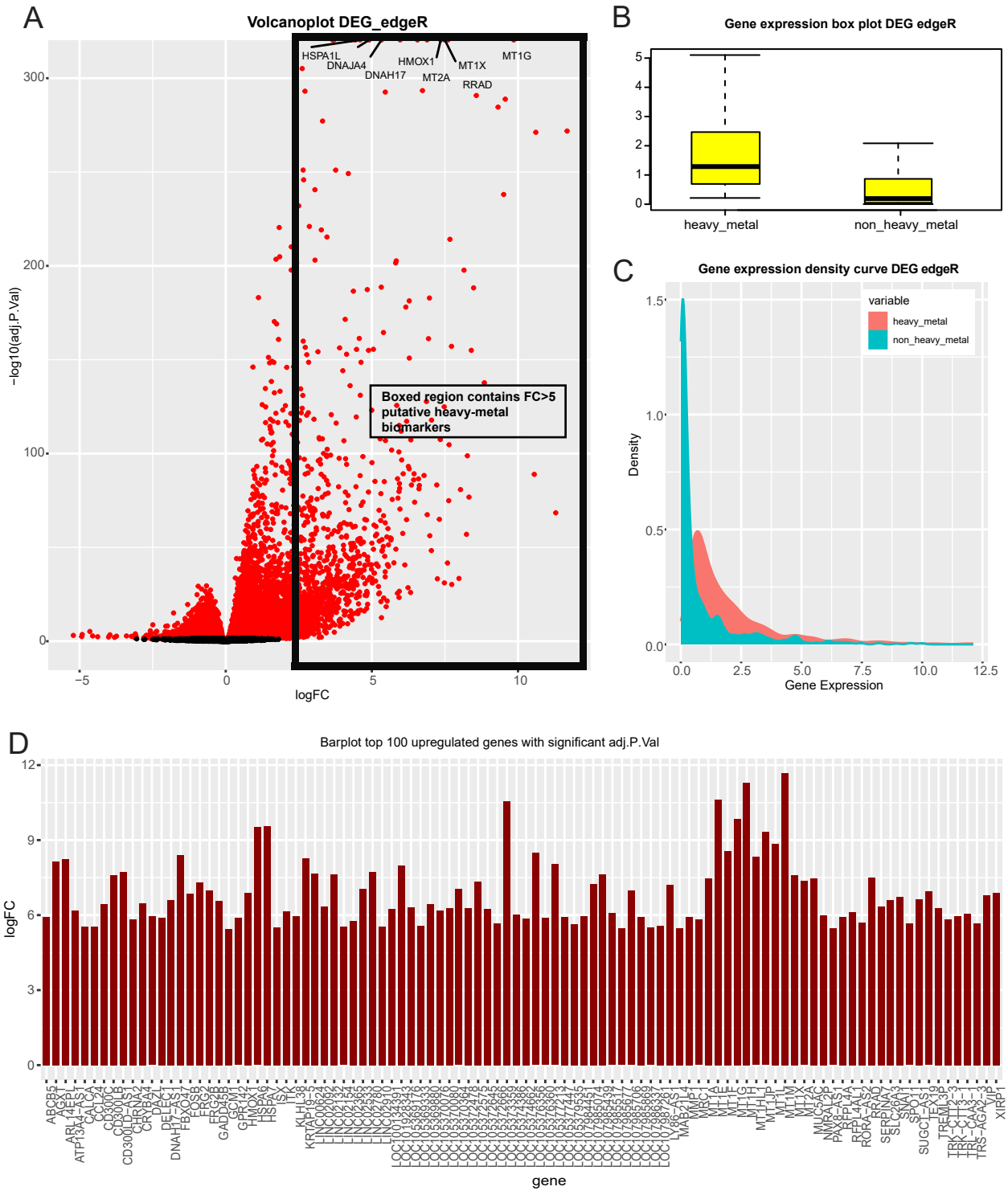
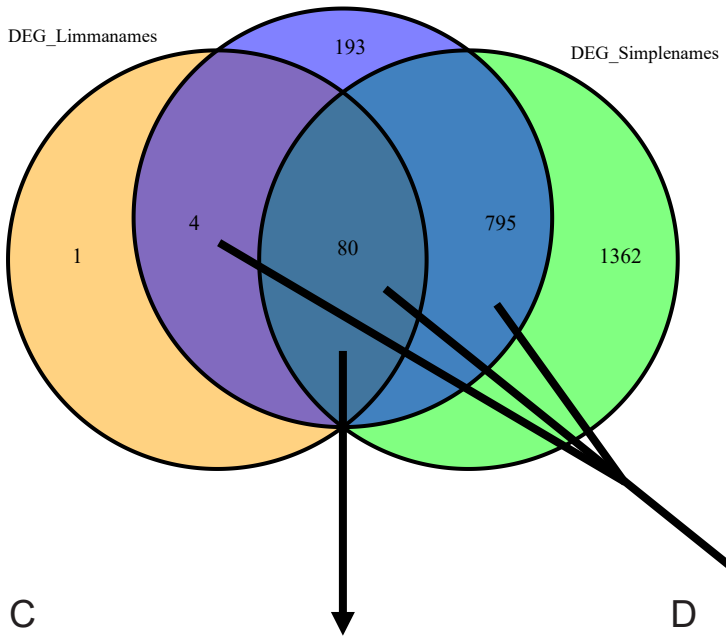
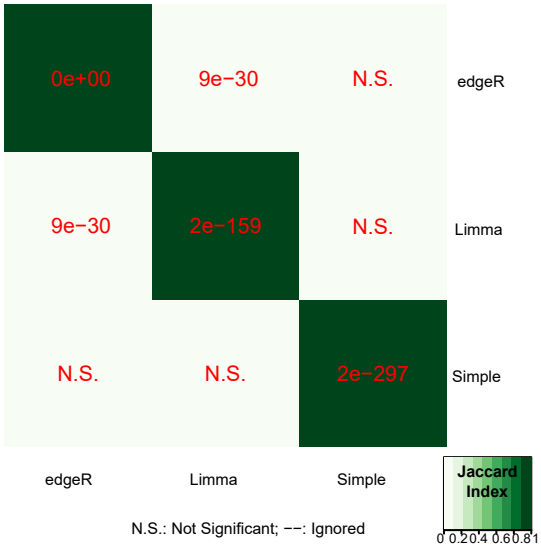


Figure 4 Mukherjee

A



B



C

Overlap of atleast 3 DEG methods (80 heavy-metal biomarker genes): GO Biological Process

	Term	Adjusted.P. value	Genes
1	cellular response to zinc ion (GO:0071294)	1.71E-10	MT2A;MT1M;MT1F;MT1G; MT1X;MT1H;MT1E
2	cellular response to copper ion (GO:0071280)	1.71E-10	MT2A;MT1M;MT1F;MT1G; MT1X;MT1H;MT1E
3	response to copper ion (GO:0046688)	5.49E-10	MT2A;MT1M;MT1F;MT1G; MT1X;MT1H;MT1E
4	cellular response to cadmium ion (GO:0071276)	1.47E-09	MT2A;MT1M;MT1F;MT1G; MT1X;MT1H;MT1E
5	cellular zinc ion homeostasis (GO:0006882)	1.47E-09	MT2A;MT1M;MT1F;MT1G; MT1X;MT1H;MT1E
6	response to cadmium ion (GO:0046686)	1.47E-09	MT2A;MT1M;MT1F;MT1G; MT1X;MT1H;MT1E
7	response to zinc ion (GO:0010043)	1.47E-09	MT2A;MT1M;MT1F;MT1G; MT1X;MT1H;MT1E
8	zinc ion homeostasis (GO:0055069)	1.67E-09	MT2A;MT1M;MT1F;MT1G; MT1X;MT1H;MT1E
9	negative regulation of growth (GO:0045926)	4.88E-09	MT2A;MT1M;MT1F; OSGIN1;MT1G;MT1X; MT1H;HSPA1B;MT1E;HSPA1A
10	cellular response to unfolded protein (GO:0034620)	5.29E-08	HSPA1L;BAG3;HSPB8; HSPA6;HSPA1B;HSPA1A

D

Overlap of atleast 2 DEG methods (879 heavy-metal biomarker genes): GO Biological Process

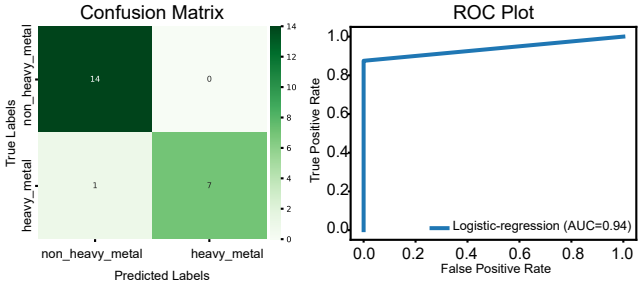
	Term	Adjusted.P. value	Genes
1	response to copper ion (GO:0046688)	1.87E-05	IL1A;MT2A;MT1A;MT1M;MT1F; MT1G;MT1X;MT1H;MT1E; MT1HL1
2	cellular response to zinc ion (GO:0071294)	1.87E-05	MT2A;MT1A;MT1M;MT1F; MT1G;MT1X;MT1H;MT1E; MT1HL1
3	cellular response to copper ion (GO:0071280)	2.28E-05	MT2A;MT1A;MT1M;MT1F; MT1G;MT1X;MT1H;MT1E; MT1HL1
4	cellular response to cadmium ion (GO:0071276)	5.11E-05	MT2A;MT1A;MT1M;MT1F; MT1G;MT1X;MT1H;FOS; MT1HL1;MT1E
5	cellular zinc ion homeostasis (GO:0006882)	6.10E-05	MT2A;MT1A;MT1M;MT1F; MT1G;MT1X;SLC30A1; MT1H;MT1HL1;MT1E
6	response to cadmium ion (GO:0046686)	7.46E-05	MT2A;MT1A;MT1M;MT1F; MT1G;MT1X;MT1H;FOS; MT1HL1;MT1E
7	zinc ion homeostasis (GO:0055069)	9.21E-05	MT2A;MT1A;MT1M;MT1F; MT1G;MT1X;SLC30A1; MT1H;MT1HL1;MT1E
8	response to zinc ion (GO:0010043)	0.000622445	MT2A;MT1A;MT1M;MT1F; MT1G;MT1X;MT1H;MT1HL1; MT1E
9	negative regulation of inclusion body assembly (GO:0090084)	0.130986911	DNAJB1;DNAJA4;HSPA1B; HSPA1A
10	cellular response to unfolded protein (GO:0034620)	0.162569664	HSPA1L;BAG3;HSPB8; HSPA6; HSPA1B;HSPA1A

Figure 5 Mukherjee

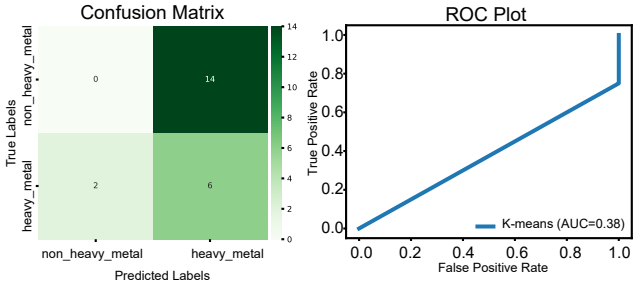
A Number of test and train datasets

Train Datasets	
26	non_heavy_metal
25	heavy_metal
Test Datasets	
14	non_heavy_metal
8	heavy_metal

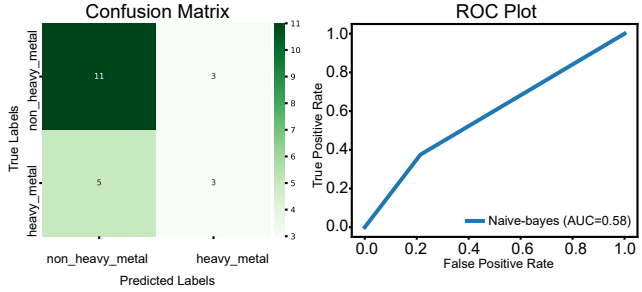
B Logistic Regression 80 genes



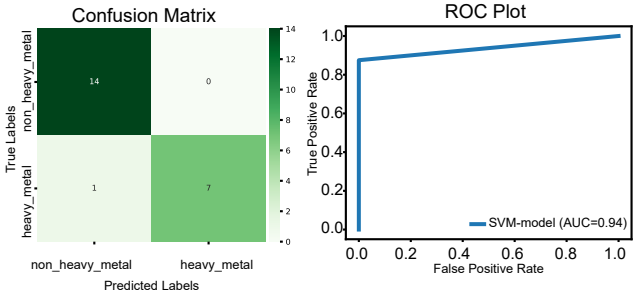
C K-Means 80 genes



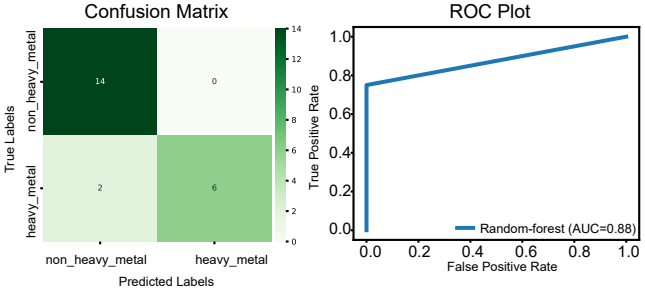
D Naive Bayes 80 genes



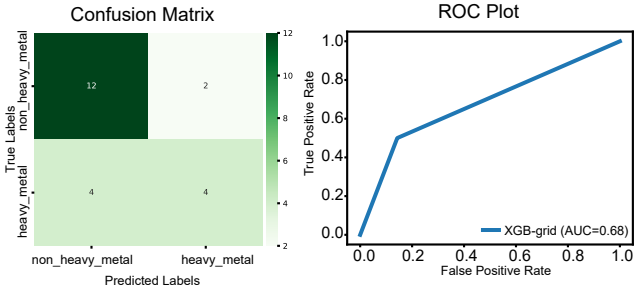
E SVM Model 80 genes



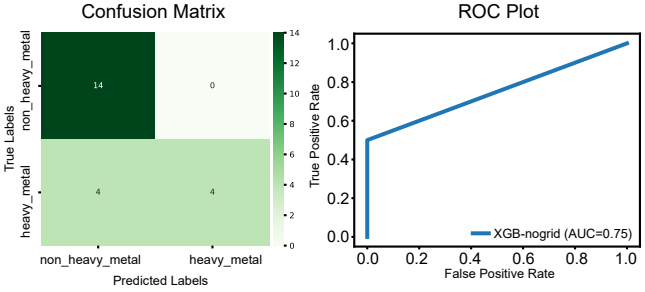
F Random Forest 80 genes



G XGB Grid Model 80 genes



H XGB No Grid Model 80 genes



I Summary of Models

Logistic Regression	precision	recall	f1-score
non_heavy_metal	0.9333333	1	0.9655172
heavy_metal	1	0.875	0.9333333
accuracy	0.9545455		

K-Means	precision	recall	f1-score
non_heavy_metal	0	0	0
heavy_metal	0.3	0.75	0.4285714
accuracy	0.2727273		

Naïve-Bayes	precision	recall	f1-score
non_heavy_metal	0.6875	0.7857143	0.7333333
heavy_metal	0.5	0.375	0.4285714
accuracy	0.6363636		

SVM Model	precision	recall	f1-score
non_heavy_metal	0.9333333	1	0.9655172
heavy_metal	1	0.875	0.9333333
accuracy	0.9545455		

Random-Forest	precision	recall	f1-score
non_heavy_metal	0.875	1	0.9333333
heavy_metal	1	0.75	0.8571429
accuracy	0.9090909		

XGB-Grid	precision	recall	f1-score
non_heavy_metal	0.75	0.8571429	0.8
heavy_metal	0.6666667	0.5	0.5714286
accuracy	0.7272727		

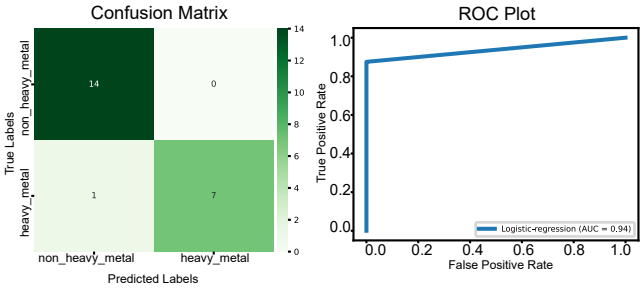
XGB-NoGrid	precision	recall	f1-score
non_heavy_metal	0.7777778	1	0.875
heavy_metal	1	0.5	0.6666667
accuracy	0.8181818		

Figure 6 Mukherjee

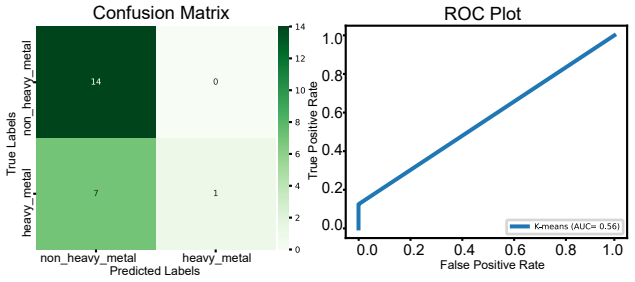
A Number of test and train datasets

Train Datasets	
26	non_heavy_metal
25	heavy_metal
Test Datasets	
14	non_heavy_metal
8	heavy_metal

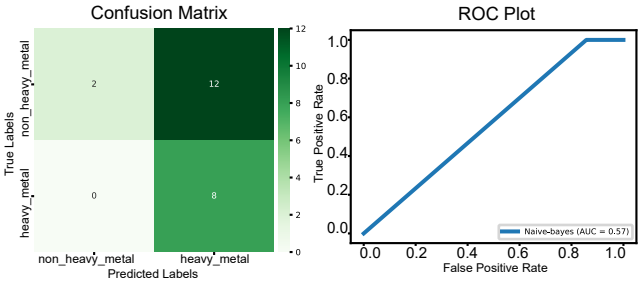
B Logistic Regression 879 genes



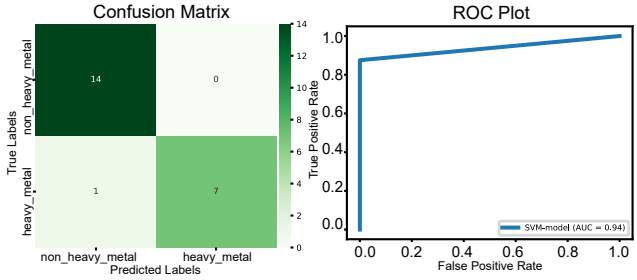
C K-Means 879 genes



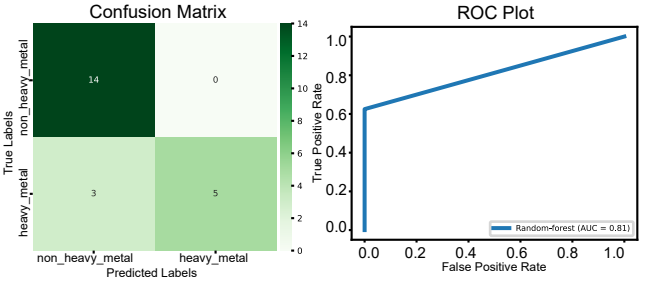
D Naive Bayes 879 genes



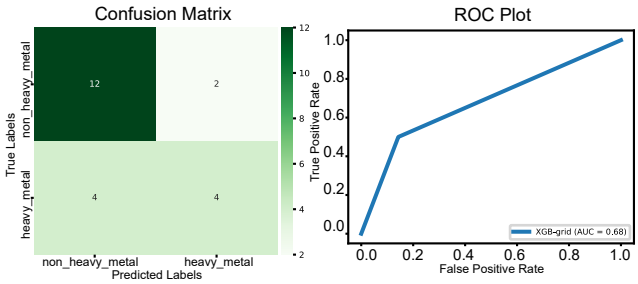
E SVM Model 879 genes



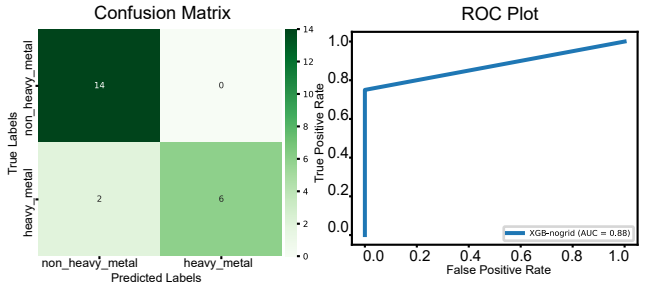
F Random Forest 879 genes



G XGB Grid Model 879 genes



H XGB No Grid Model 879 genes



I Summary of Models

Logistic Regression	precision	recall	f1-score
0	0.93333333	1	0.96551724
1	1	0.875	0.93333333
accuracy	0.95454545		

K-means	precision	recall	f1-score
0	0.66666667	1	0.8
1	1	0.125	0.22222222
accuracy	0.68181818		

Naive-bayes	precision	recall	f1-score
0	1	0.142857143	0.25
1	0.4	1	0.571428571
accuracy	0.454545455		

SVM-model	precision	recall	f1-score
0	0.93333333	1	0.96551724
1	1	0.875	0.93333333
accuracy	0.95454545		

Random-forest	precision	recall	f1-score
0	0.82352941	1	0.90322581
1	1	0.625	0.76923077
accuracy	0.86363636		

XGB-grid	precision	recall	f1-score
0	0.75	0.85714	0.8
1	0.66666667	0.5	0.57142857
accuracy	0.72727273		

XGB-NoGrid	precision	recall	f1-score
0	0.875	1	0.93333333
1	1	0.75	0.85714286
accuracy	0.90909091		

Figure 7 Mukherjee