January 2023

Application Of Artificial Intelligence And Multi-Omics To Understand The Effect Of Heavy Metals Exposure On Human Health

Sonalika Singhal

APPLICATION OF ARTIFICIAL INTELLIGENCE AND MULTI-OMICS TO UNDERSTAND THE EFFECT OF HEAVY METALS EXPOSURE ON HUMAN HEALTH.

by

Sonalika Singhal

Bachelor of Science, Chaudhary Charan Singh University,1999

Masters in Computer Applications, Indira Gandhi National Open University, Delhi, 2004

A Dissertation

Submitted to the Graduate Faculty

of the
University of North Dakota
in partial fulfillment of the requirements

for

the degree of

Doctor of Philosophy

Grand Forks, North Dakota

August, 2023

Name: Sonalika Singhal

Degree: Doctor of Philosophy

This document, submitted in partial fulfillment of the requirements for the degree from the University of North Dakota, has been read by the Faculty Advisory Committee under whom the work has been done and is hereby approved.

DocuSigned by:
donald sens
donald sens

DocuSigned by:
Scott Garrett
Scott Garrett

DocuSigned by:
Mary Ann Sens
Mary Ann Sens

DocuSigned by:
kouhyar Tavakolian
Kouhyar Tavakolian

DocuSigned by:
Van Doze
Van Doze

This document is being submitted by the appointed advisory committee as having met all the requirements of the School of Graduate Studies at the University of North Dakota and is hereby approved.

DocuSigned by:
Chris Nelson
Chris Nelson
Dean of the School of Graduate Studies

7/13/2023
Date

PERMISSION

Title Application of artificial intelligence and multi-omics to understand the effect of heavy metals exposure on human health.

Department Clinical Translational Science

Degree Doctor of Philosophy

In presenting this dissertation in partial fulfillment of the requirements for a graduate degree from the University of North Dakota, I agree that the library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by the professor who supervised my dissertation work or, in his absence, by the Chairperson of the department or the dean of the School of Graduate Studies. It is understood that any copying or publication or other use of this dissertation or part thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of North Dakota in any scholarly use which may be made of any material in my dissertation.

Sonalika Singhal

July, 2023

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

My passion to work with bioinformatics and molecular biology had been fulfilled by getting involved in this thesis project. It was a great opportunity for me to practice bioinformatics techniques and the wet-lab work which enabled me to gain an immense knowledge that is useful for my future research.

There are many people who have made my time as a PhD student both enjoyable and enlightening, and without whose help the work would not have been possible. First, I would like to thank Dr. Don Sens who accepted me as a graduate student in his lab and provided me with an opportunity to gain extensive knowledge and training on science and research. As my teacher and mentor, he has taught me more than I could ever give him credit for here. He has shown me, by his example, what a good scientist and person should be.

I am grateful to all of those with whom I have had the pleasure to work during this and other related projects. Thanks to my thesis advisory committee Dr. Mary Ann Sens, Dr. Scott Garrett, Dr. Kouhyar Tavakolian, and Dr. Van Doze for help and advice at various stages of this work. Each of the members of my committee has provided me with extensive personal and professional guidance and taught me a great deal about both scientific research and life in general. I considered myself extremely lucky to have amazing lab mates and colleagues.

Nobody has been more important to me in the pursuit of this project than the members of my family. I would like to thank my parents, whose love and guidance are with me in whatever I pursue. Most importantly, I wish to thank my loving and supportive husband, Sandeep and my two wonderful daughters, Sudiksha and Sonakshi, as a whole for their continuous support and understanding when undertaking my research and writing my thesis. Their belief in me has kept my spirits and motivation high during this process.

This thesis is dedicated to my lovely Daughters,

SUDIKSHA & SONAKSHI

ABSTRACT

IARC classified arsenic (iAs) as “carcinogenic to humans”, but despite the health consequences, there is no molecular signature available yet to predict when exposure may lead to the disease development.

In this study, a three-step analysis was employed: (1) the gene expression profiles obtained from diverse arsenic-exposed populations were utilized to identify differentially expressed genes associated with arsenic exposure in human subjects, (2) the gene expression profiles induced by arsenic exposure in different myeloma cancer cell lines were used to define common genes and pathways altered by arsenic exposure,(3) the genetic profiles of human bladder cancer studies were used to test the significance of the common association of genes, identified in step 1 and step 2, to develop and validate a predictive model of primary bladder cancer risk associated with arsenic exposure.

The study identified a unique set of 147 genes associated with arsenic exposure and linked to molecular mechanisms of cancer. The risk prediction model shows the highest prediction ability for recurrent bladder tumors based on a very small subset (NKIRAS2, AKTIP, and HLA-DQA1) of the 147 genes resulting in AUC of 0.94 (95% CI: 0.744-0.995) and 0.75 (95% CI: 0.343-0.933) on training and validation data, respectively.

In addition, high arsenic exposure has been associated with adverse kidney disease outcomes. Therefore, we performed a systematic analysis of the association between arsenic and various kidney disease outcomes. Because of the high prevalence of arsenic exposure worldwide, there is a need for additional well-designed epidemiologic and mechanistic studies of arsenic and kidney disease outcomes.

The human kidney is known to possess renal progenitor cells (RPCs) that can assist in the repair of acute tubular injury. The RPCs are sparsely located as single cells throughout the kidney. We recently generated an immortalized human renal progenitor cell line (HRTPT) that co-expresses PROM1/CD24 and expresses features expected on a RPCs. This included the ability to form nephrospheres, differentiate on the surface of Matrigel, and to undergo adipogenic, neurogenic, and osteogenic differentiation. These cells were used in the present study to determine how the cells

would respond when exposed to a nephrotoxin. Arsenite (iAs) was chosen as the nephrotoxin since the kidney is susceptible to this toxin and there is evidence for its involvement in renal disease. Gene expression profiles when the cells were exposed to iAs for 3, 8, 10 passages (subcultured at 1:3 ratio) identified a shift in from the control unexposed cells. The cells exposed to iAs for 8 passages were then referred with growth media containing no iAs and within 2 passages the cells returned to an epithelial morphology with strong agreement in differential gene expression between control and cells recovered from iAs exposure. Results show within 3 serial passages of the cells exposed to iAs there was a shift in morphology from an epithelial to a mesenchymal phenotype. EMT was suggested based on an increase in known mesenchymal markers. We found RPCs can undergo EMT when exposed to a nephrotoxin and undergo MET when the agent is removed from the growth media.

MOTIVATION

Environmental factors on human health are one of the greatest biological challenges because they deal with the precise foundations of mankind. And therefore, it is eventually connected with my ultimate research topic of bioinformatics. Addressing the current challenges is key to develop biomarkers to both prevent and cure diseases. Data mining, Statistical and Machine learning modeling are the tools, few among many, for addressing the challenges. A common focus of my projects has mainly been towards creating well defined biomarkers, which will help improving and optimizing diagnostics and treatment after heavy metal exposure.

Personally, this thesis has given me the passion to explore and to expand in what ways, and possibly how well, machine learning could be used to answer current challenges or problems formulations by using available genomic data.

FOCUS AND CHALLENGES

Challenges of developing human genomics disease risk predictors after exposure of heavy metals using machine learning application has been the following:

- The collection, selection, transformation and representation of genome (genomic) data in a way which enables standard machine learning algorithms to work on it.
- The adaptation of appropriate tools, implementation, within an already complex and existing framework.
- The adversity of building a tool which bridges the fields of machine learning, heavy metal toxicity and human biology.

CHAPTER 1
Data Mining, Machine Learning and Statistical Modeling

In the last couple of decades, multi-omics (especially genomic and proteomic) databases have grown exponentially, and resulted in an explosion of information and knowledge. From the beginning of the Genome Project, the numbers of published research on multi-omics experiments have grown substantially, and to utilize this information, datasets and refined computational models have been created to solve critical biological problems. As a result of the significant drop in the sequencing price, the amount of genome data is significantly increasing, as shown in Figure 1, the progress of the cumulative number of human genomes throughout the years. This amount of available genomic data enabled the establishment of large-scale genomics projects including The Cancer Genome Atlas (TCGA) [1], The Encyclopedia of DNA Elements (ENCODE) [2], and the 1000 Genomes Project Consortium [3], projects aimed to collect and store genomics data at one platform and provide it to the research community. In order to make efficient use of the collected genomics data, big data analysis techniques are essential.

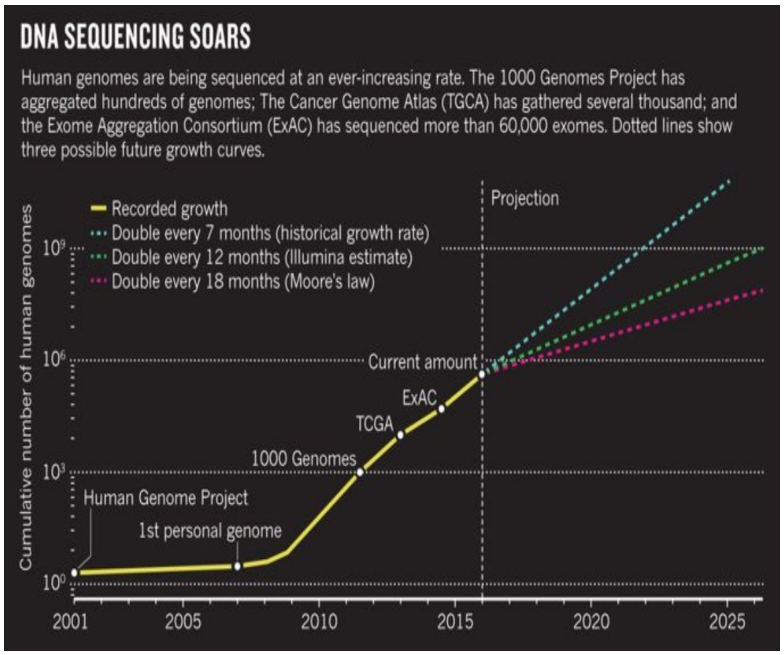


Figure 1: The number of sequenced human genomes over the years

Credit: Stephens, Z. D. et al. PLoS Biol. 13, e1002195 (2015)/CC by 4.0 http://creativecommons.org/licenses/by/4.0

Similar to any other research area, there are some challenges that arise in this field such as: integrate different data from various sources to a common schema (as witnessed in data warehousing system), computational power that suffer from the constantly evolving nature of the data and methods/algorithms to deal with different biological questions. During my research, I tried to consider all possible challenges in this area and overcome certain limitations. For data collection and selection, extensive research has been done searching through online resources and departmental collaborative institutes. To deal with data mining, machine learning and statistical modeling, a significant number of courses such as Genomic Data Science Specialization, statistical courses, programming in R and digital pathology have been completed while writing the thesis. The application development has throughout the thesis been implemented in 'RStudio' https://posit.co/download/

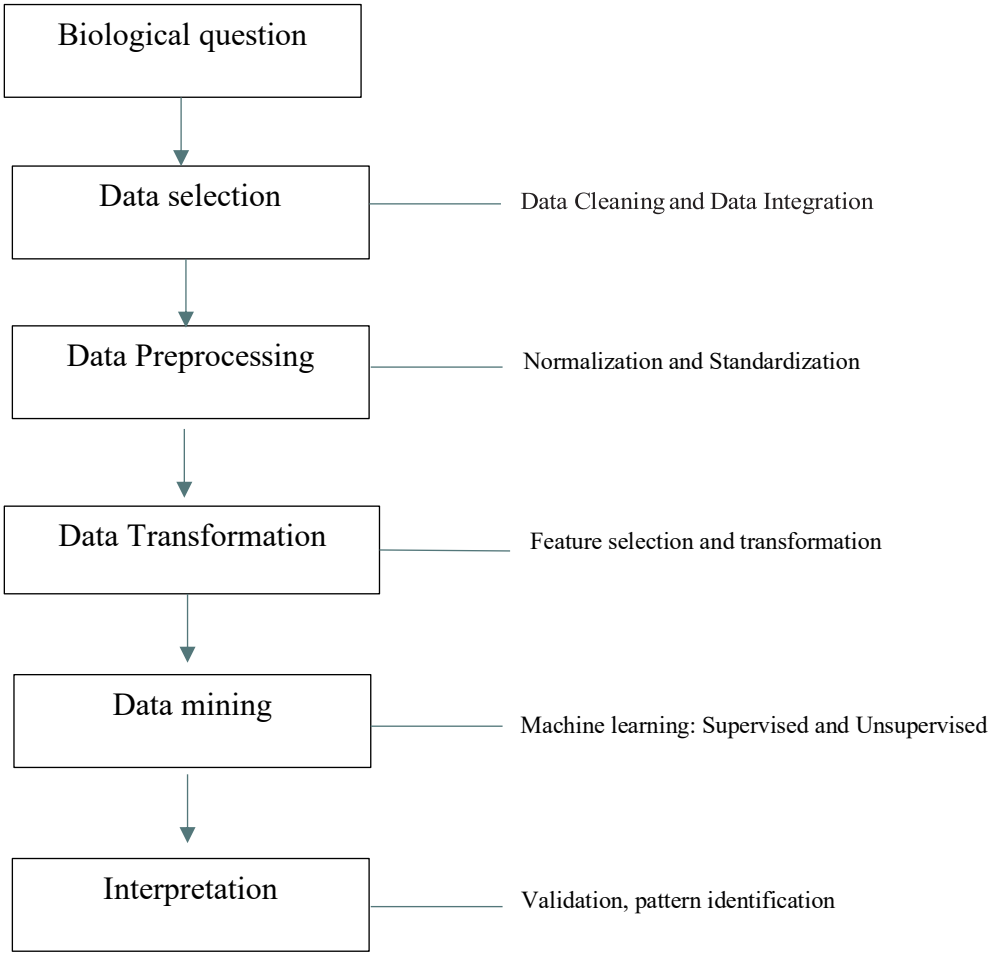


Figure 2: Flow chart of genome wide association studies

1.1 BIOLOGICAL DATA

The human genome consists of DNA containing the instruction to build and maintain cells. To carry out these instructions, DNA must be read and transcribed or copied into ribonucleic acid (RNA). In a human body, normal cells form, grow, divide as expected. Cells are also replaced when they grow old or become damaged, but if they divide and grow uncontrollably, then they are considered as disease including tumors [4]. In my research, I have used microarray and RNA-sequencing (RNA-seq) data to capture the genomic changes. Nucleotide microarrays were the first high throughput method for genomics, introduced in 1995 [5]. RNA-Seq is revolutionizing the study of the transcriptome that represents all gene readouts present in a cell.

1.2 GENE EXPRESSION OMNIBUS (GEO) DATABASE

This is the most common data platform developed and maintained by NIH to deposit and extract the all-possible omics data. Initiated by the need of a public repository for high-throughput data, the Gene Expression Omnibus (GEO) project [6] was designed to provide a flexible and open design to store, retrieve, and insert data from high-throughput experiments. It is intended to act as a central data distribution hub of gene expression data derived from coherent datasets.

I have used this resource to collect large amount of data as well as information to complete my different projects. GEO database provides access to thousands of high-quality, curated disease datasets in multiple disease areas covering over 2000 clinical measures, including disease, tissue, treatment, survival and demographics. I would like to cover the important characteristics that gene expression databases possess. The consideration is given to the major components used for analyses and segregating the characteristics of data.

Typically, all records in a database consist of five parts, raw data, process data, and annotation data (omics, sample and experimental).

1.2.1 Raw data

Raw data is a scanned image generated through the machine such as CEL files (microarray chip). All databases come with the information of the platform used to generate that particular data. A unique GEO id and GSM number is allocated to each study and sample respectively. In addition, the complete contact information of the author is provided for any additional support, question, query.

1.2.2 Process data

There are two ways to collect the process data from GEO: 1) Collect the raw files and pre-process them with an appropriate tool/package suitable, 2) each GEO ID provides the processed data with the information of the method used by the author(s) for original publication. The data from each sample is also provided with the lab/experimental protocol used to generate that data. All that information is stored separately for each sample and together as well for all the samples provided by the study. The data used for my research here is extracted from raw CEL files using the different R packages such as “affy”, “rma”, “frma” and “oligo”.

1.2.3 Gene annotation

Gene annotation information provides detailed information about each and every microarray probe. It provides all position information about a particular segment of the genome such as gene name, alternative names (if any) functions, strand and location over the chromosome etc. These annotations are collected over time and are publicly accessed from different databases. We have used the most updated database to link the probes of microarray to gene annotation.

1.2.4 Sample annotations

Annotation of sample studies is discussed in this section. We collect all the possible information about the characteristics of the microarray sample. This information can be

found for each individual sample and/or combined in a matrix file. Information pertinent to the biological sample used to extract the targets, the corresponding information such as clinical/pathological descriptions pertaining to source and sample/patient characteristics, like information that describes whether the samples are normal or cancerous, and information that describes if there are any in vitro or in vivo treatments that have been applied in addition to clinical-pathological features.

1.2.5 Experiment annotations

It contains the information regarding the protocols followed during the experiment and parameter settings used by the associated tools and software during hybridization. This section provides all the necessary information about the generation of the sample.

1.3 MICROARRAY

A microarray is a genomic tool used to detect the expression of thousands microscopic DNA spots attached to a solid surface (called probes). DNA microarrays are microscope slides that are printed with thousands of tiny spots in defined positions, with each spot containing a known DNA sequence which represents a particular gene. In a laboratory setting, a microarray is a glass or plastic slide or a bead that has a piece of an oligonucleotide or oligo (DNA or cDNA) attached to it [7]. Specific sequences are immobilized to a surface and reacted with labeled cDNA targets. A signal resulting from hybridization of the labeled target with the specific immobilized probe identifies which RNAs are present in the unknown target sample. After hybridization the microarray is washed to get all the loose labelled nucleotides off and then we put the whole thing through a scanning microscope at each label wavelength.

The primary data is a digital black and white photograph of the array. For two channel microarray arrays which have two differently labeled samples hybridized to the same probe, the data are often visualized by a picture like the one below. This is actually a composite of the black and white

photo for each label. One of the labels is represented by red and the other by green. The relative intensity of the two samples is represented by a color scale going from pure red (only the red sample has hybridized) to pure green (only the green sample has hybridized) with yellow meaning equal amounts of both samples. The intensity is represented by brightness, so that dark spots show little hybridization and bright spots have high hybridization.

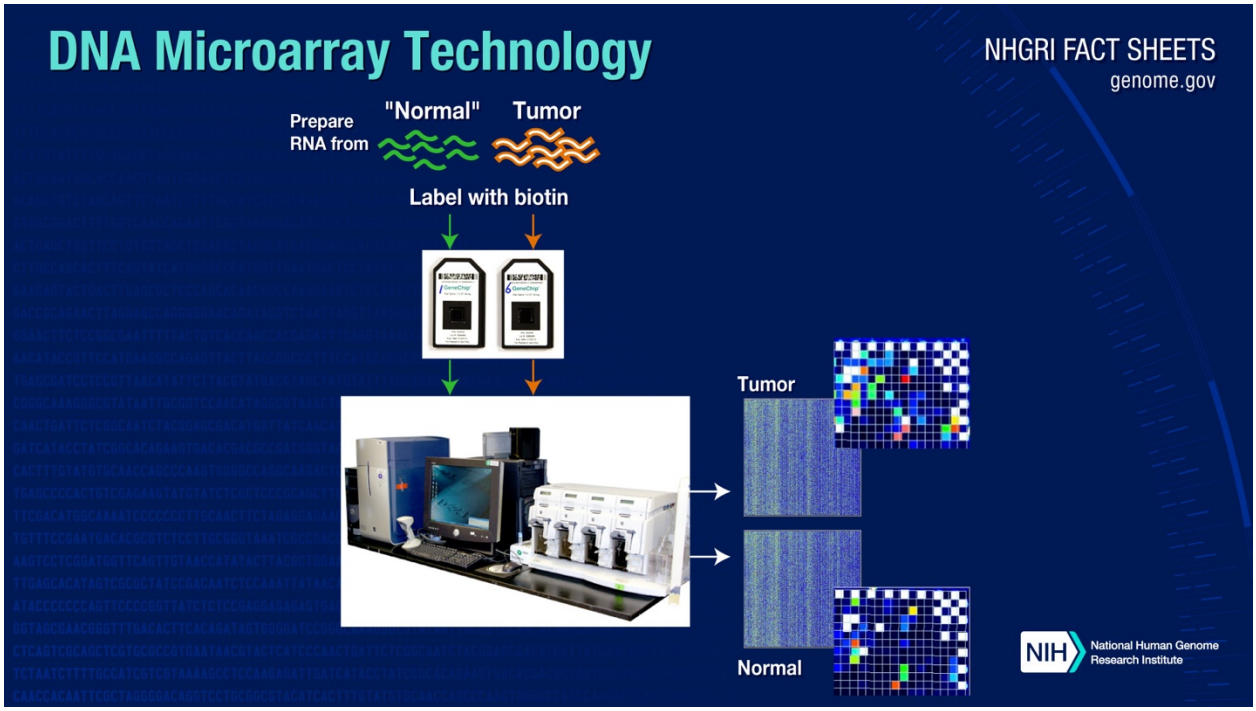


Figure 3: DNA microarray technology, Credit: DNA Microarray Technology Fact Sheet

https://www.genome.gov/about-genomics/fact-sheets/DNA-Microarray-Technology

Each probe of the microarray is a set of lots of identical strands of DNA that are supposed to be complementary to what is in the sample. The DNA is synthesized from known sequences. We may not know what features they represent, but we have the genomic sequence. To obtain a summary for the probe we need to identify

- The pixels in the probe foreground - i.e., the region which has the complementary strands.
- A probe summary such as the mean or median.

· The pixels in the background.

· A probe summary for the background.

The raw Affymetrix data is stored in a DAT file which is called a CEL file. Each CEL file has the probe id, probe location and probe intensity, as well as information which identifies the type of array. The probe identifier is used to find the annotation which links the probes to the genes.

1.3.1 Microarray Quality Control and Normalization and Summarization

Numerous studies have identified sources of inter- and intra- laboratory error and variability in results and outcomes of microarray studies. Therefore, Microarray Quality Control and Normalization of Affymetrix microarrays is an essential step before applying any statistical test on the data. It provides a comprehensive resource for ensuring quality control in every step of this complex process. From concept building, experimental design, data processing, analysis, and interpretation, we emphasis on data check at each stage of design and analysis.

Quality control of microarray data begins with the visual inspection of the scanned microarray images to make sure that there are no obvious splotches, scratches or blank areas. After feature extraction, the R packages were used to make diagnostic plots for background signal, average intensity values and percentage of genes above background to identify errors.

Normalization of data was used to control for technical variation between samples, while preserving the biological variation [8]. There are few Bioconductor package such as Affy[9], fRMA[10] that can read the files in either format, which saves us the trouble of having to identify the file format. The selection of package depends on:

· The type of array

· The design of the study/experiment

·Hypothesis of the study for example the majority of genes represented on the microarray do not change between test group to controls

· Platform used to generate the microarray data i.e., Affy, Agilent, etc.

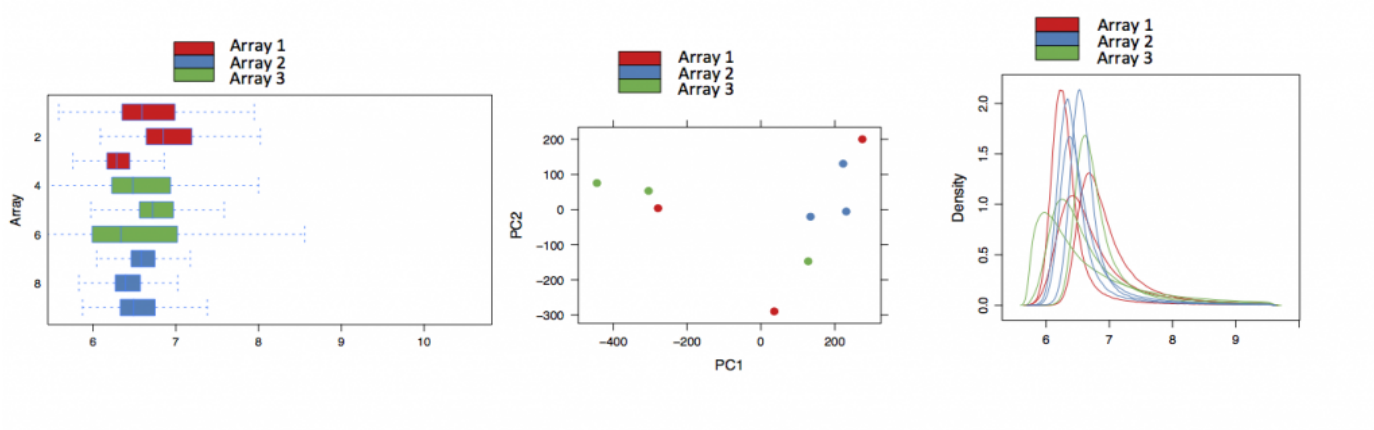


Figure 4: Visualization of sample statistics.

Box1: 1. Box plot represents the statistical summary of each sample including min, median, quartile, maximum 2. PCA component analysis showing the distribution of variation across sample groups, PC1 and PC2 explains the first two maximum variation in two-dimension space 3) Density distribution plot of gene expression value of each sample with color group corresponding

The affy software includes several of these methods and also allows the user to "mix and match" picking different background correction, probe normalization and probe-set summary methods. However, for research, it is usually best to select one of the standard methods. Of these, the most used appears to be RMA or fRMA. For Expression Atlas, Affymetrix microarray data is normalized using the 'Robust Multi-Array Average' (RMA) method within the 'oligo' package. To set the same scale across the different samples, we used fRMA, where some samples are always used together with the new dataset to get the same expression scale at the end.

Most of those packages cover all 3 steps - background correction, quantile normalization of the individual probes and then probe-set summary.

After reading into the affy tool, we do background correction, probe normalization and finally summarization of the probes into probe-sets. Background correction is a complex statistical model which supposes both additive and multiplicative noise components. After background correction to the individual probes, quantile normalization is applied. In the third and final step, the probes are summarized into probe-sets using the median polish algorithm, which is a type of robust 2-way

ANOVA, where one factor is the array and the other is the probe-set. The algorithm is robust to outlying data, so that single probes with large values are down-weighted. Because both quantile normalization and median polish use data from all the microarrays, using just a subset of the microarrays or removing a single bad array affects the normalization step for all the arrays.

Finally, we extract the probe intensities, which can be treated as continuous data. By using the log2 transformation, the data are suitable for analysis by versions of standard statistical methods.

1.4 DATA ANALYSIS

A major goal of genomics analysis is to identify genes of interest i.e., differentially expressed across the phenotypic conditions, and co-regulated genes to infer biological meaning for further studies. Source material is microarray gene expression data [11]. The significance of findings depends on appropriate study design, implementation of controls, and correct analysis. Every effort should be made to minimize data bias, because small and uncontrolled changes in an environment can result in identification of differentially expressed genes unrelated to the designed study. Sources of data bias can occur during the experiment, during the mRNA library preparation, or during the microarray run (but are not limited). Once a controlled study is designed with well-defined biological questions, a structured analytical approach is required to start to test for quality control followed by unbiased analysis of the data. In current research, we use the following approaches that include.

1.4.1 Determining Intra- and -Inter group Variability and Outliers

The first and most important analytical questions are “what is our research question/hypothesis?” and “what kind of data do we have to test this hypothesis?” in terms of selecting a statistical test. Therefore, visualization of data distribution is one of the essential parts which help us to identify the most appropriate statistics approach suitable according to the biological question/hypothesis. There are two types of statistics: Parametric and Non-parametric [12]. Parametric statistics are based on assumptions about the distribution of population from which the sample was taken should

be normally distributed. Nonparametric statistics are free from this assumption, i.e., the data can be collected from a sample that does not follow a specific distribution. The most common example which explains the selection of statistical tests to identify the differentially expressed genes: Student's t-tests (when data is normally distributed) [13], Mann-Whitney-Wilcoxon (MWW) test [13] or the Wilcoxon test (when data does not follow a normal distribution) [13].

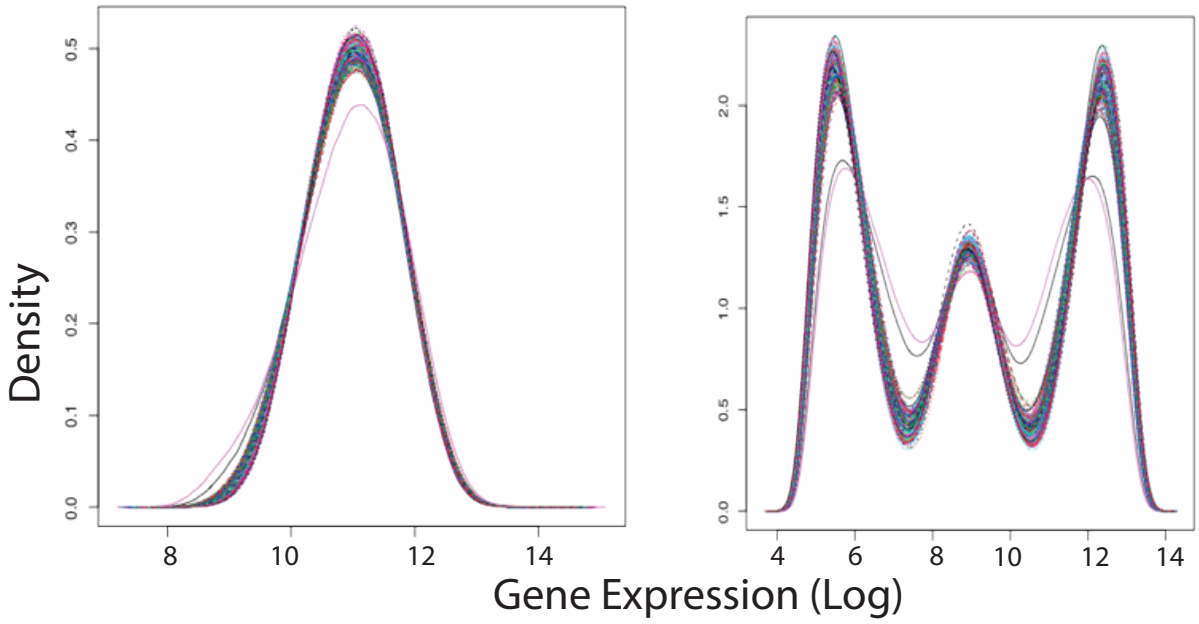


Figure 5: *Density plots of normal distribution.*

Box 2: The left-hand plot shows the normal distribution (also known as Gaussian or Bell shape curve). This data is suitable for parametric test. The right-hand plot is non-parametric (multi-model distribution) and non- parametric test works well on this kind of data.

One of the most popular methods that almost everyone knows is the histogram to see the data distribution. The histogram is a data visualization method that shows the distribution of a variable across samples. It provides the frequency of occurrence per value in the dataset, which is what shows normal vs non-normal distribution.

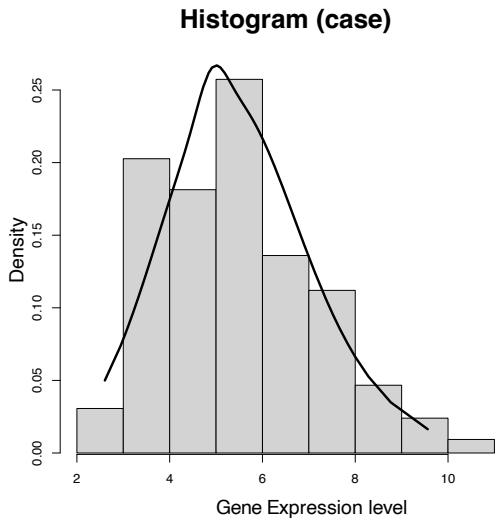


Figure 6: Histogram with the frequency distribution (gray) and line of fit (black) to provides the shape of the distribution.

Another popular method is box plot. It helps us to visualize uniformly and non-uniformly distributed samples with the help of basic statistics (including outliers) which summarizes the different variables across samples minimum. Those variables are: minimum, first quartile, median, third quartile and maximum.

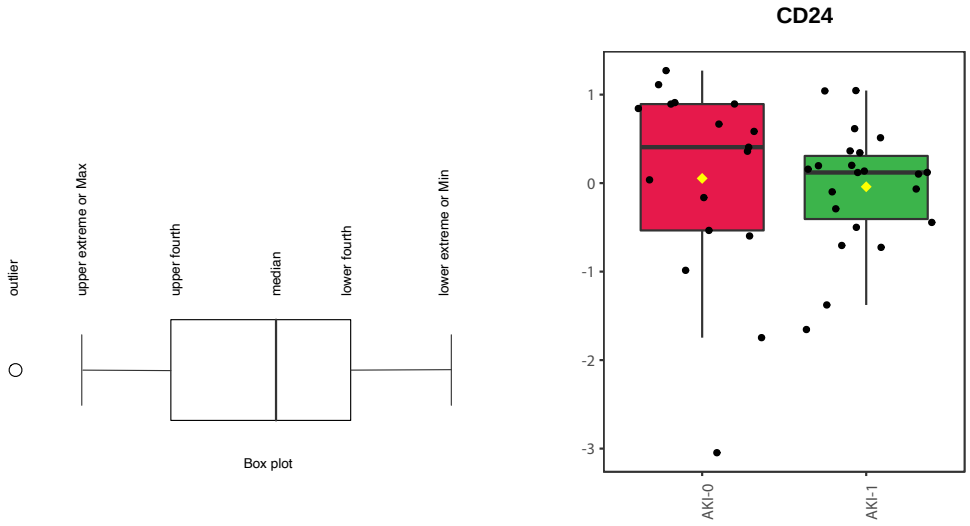


Figure 7: Different parts of a Boxplot and example of Boxplot

Box 3: Left figure shows the properties of the box plot. Right figure shows an example of box plat to see the variation of CD24 gene between two groups (AKI-0 (red) vs AKI-1 (green) in this case). Each black dot represents one sample, the black horizontal line represents the median and yellow dot represents the mean value of group.

The next stage is to understand the underlying patterns of data using the principal component analysis (PCA). PCA is an unsupervised machine learning method that helps to understand patterns present in high-dimensional data beyond the descriptive statistics and reduce the complexity of the data while retaining most of the information[13]. The method as such captures the maximum possible variance across features and projects observations onto mutually uncorrelated vectors, called components. The PCA metrics show 1. how many components capture the largest share of variance (explained variance), and 2., which features correlate with the most important components (factor loading).

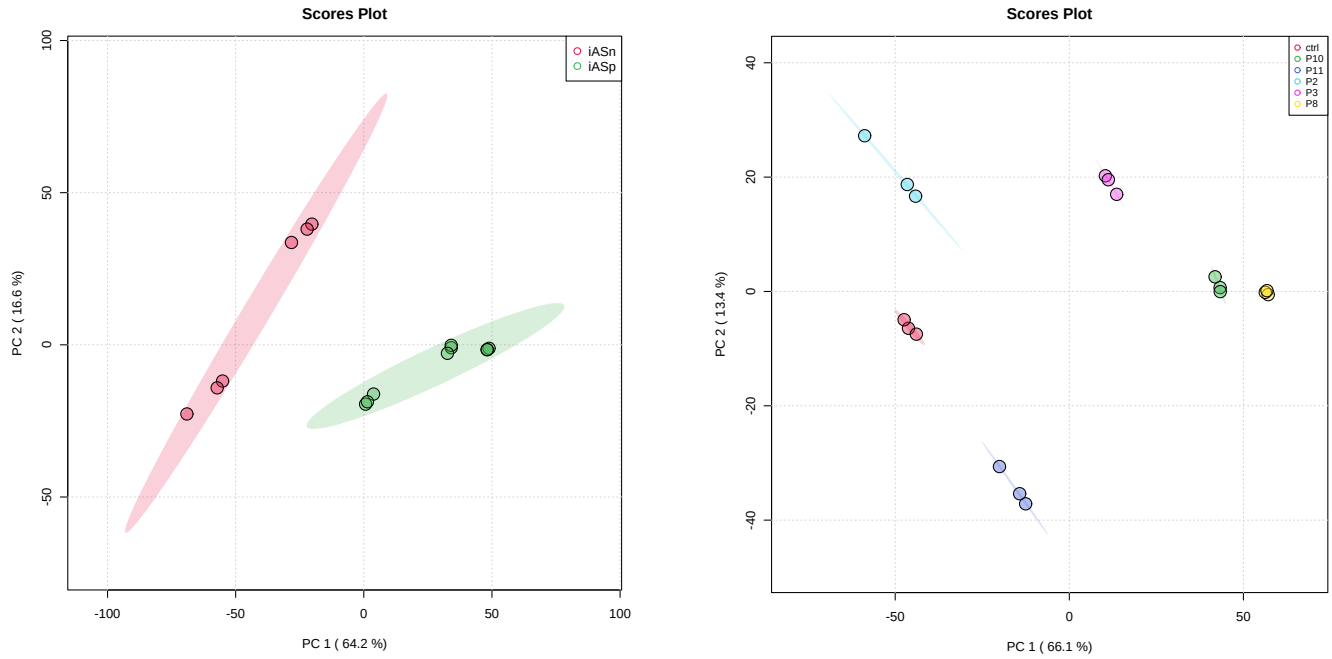


Figure 8: Example of PCA plot between two different conditions

Box 4: Left-hand side figure represents the PCA between two groups (iAS positive (red) and iAS negative (green) in this case). The first two components PC1 and PC2 explains the 64.2 and 16.6 percent of variation among the samples. The right-hand side figure represents the PCA between multiple groups (six in this case). The PC1 and PC2 represents the first two maximum variations 66.1 and 13.4 respectively

1.4.2 Sample and Gene correlation

This is an approach to determining within and in-between group variability is to calculate distance as represented by correlation between samples as well as genes. Depending upon the characteristic of data there are two commonly used tests of correlation are the Pearson's coefficient and the Spearman's rank correlation coefficient, which describe the directionality and strength of the relationship between selected variables. Pearson's correlation is a parametric test that reflects the linear relationship between two variables accounting for differences in their mean and SD, whereas the Spearman's rank correlation is a nonparametric test using the rank values of the two variables. These correlation coefficients are calculated between samples or genes and can be visualized as either a table or a heat map, allowing us to assess whether replicates (technical or biological) group together.

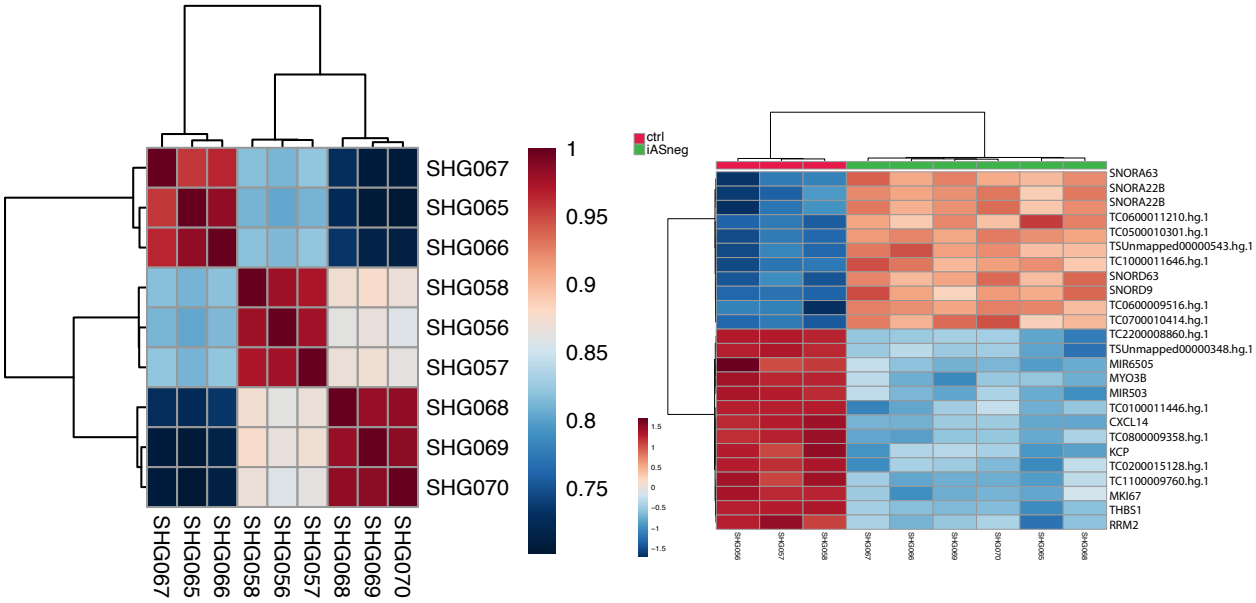


Figure 9: Example of Heatmap showing correlation among samples and differentially expressed genes.

Box 5: The heatmap of the correlation values between the sample (left) and between the top genes (right) together with the sample and linkage between the samples and genes using the hierarchal clustering.

1.4.3 Filtering Out Data Noise

After outliers are excluded and variability is assessed using the above-mentioned process, the next step is to find the distribution of expressed genes that will be helpful to determine a threshold for low expression caused by technical factors (based on sample-to-sample variation), referred to as data noise. One approach to viewing variability between samples is to generate a scatterplot comparing the normalized log2 transformed gene expression values in two different phenotypes to visualize their similarity or correlation; this provides a more detailed view of genes driving the correlation. By comparing the similarity of housekeeping genes across different samples, the user can assess the level of noise. Another approach to determining a threshold for expression above noise is to compare the number of genes expressed (up and down regulated) at different cutoffs across all samples. This can be done by using MA.plot (Figure 10).

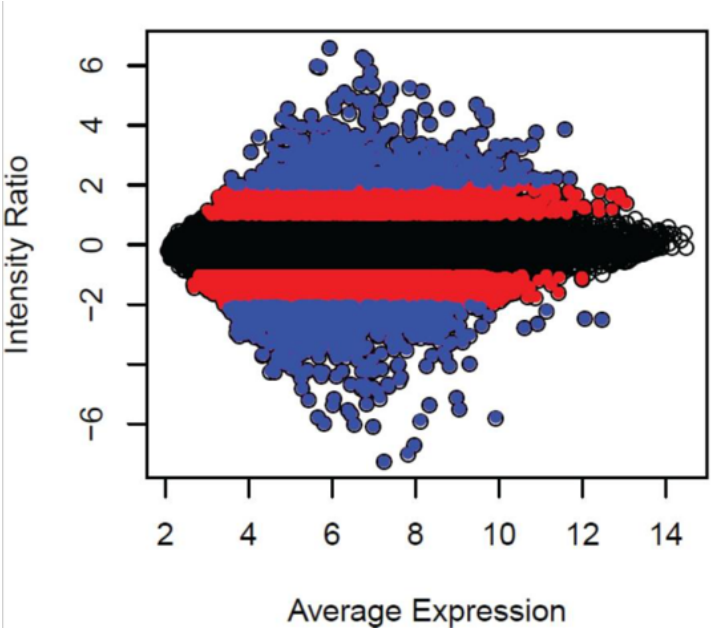


Figure 10: Gene expression MA plot.

Box 6: M is, therefore, the intensity ratio, and A is the average intensity. Each dot represents gene expression value of one probe in the plot. The three different colors represent the three ranges of gene expression intensity ([−1, 1] black, [> 1, 2] and [< −1, −2] red, [> 2] and [< −2] blue).

1.4.4 Differently Express Genes

After the quality control steps, outlier removal, and filtering, the data are ready for down stem analysis. Our next goal is to identify the differentially expressed genes (or significant genes) across different phenotypic conditions. This approach allows us to rank a long list of genes based upon the level of differentiation between the conditions. Two general approaches applied in this section: 1) pairwise comparison and 2) variance across different groups. Various online resources and software are publicly available that allow for this type of analysis but we have used an R script pipeline to perform this kind of analysis.

1.4.4.1Pairwise comparison

There are two most popular methods that identify pairwise differently express genes, 1) t-test, 2) Wilcoxon test.

The paired Student's t-test is a parametric test comparing the means of paired quantitative measurements from two groups. The t-test requires that the sample means are normally distributed. The test relies on estimations that the true difference between two groups means using the ratio of the difference in group means over the pooled standard error of both groups. After testing the hypothesis, the outcome provides the t-statistic, the t-distribution values, P-value and the degrees of freedom to determine statistical significance as an outcome of measurements. Wilcoxon is a nonparametric alternative to the t-test. It tests whether the average sum of the ranks (and thus the medians) of the two samples differ significantly from each other. This test determines if groups of comparison have the same mean on ranks. We don't use actual data values themselves, instead, a rank is assigned to each data point and those ranks are used to determine if the data in each group originates from the same distribution.

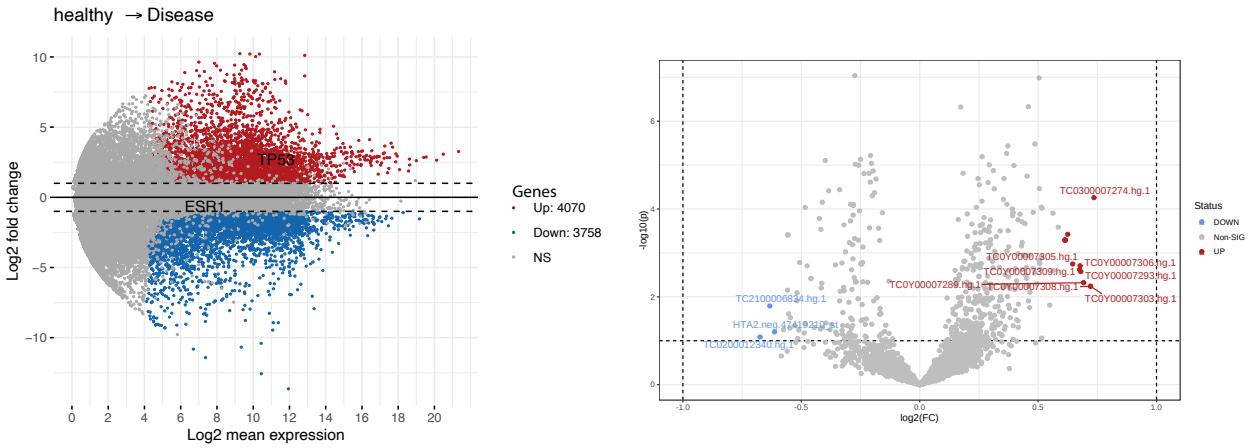


Figure 11: MA and volcano plot to demonstrate the significant gene with direction (Up or down regulated).

Box 7: MA plot to demonstrate the Up and Down regulated gene based selected threshold value (fold change greater than 2 and P<0.05 in this case.). Up regulated (Red color) and down regulated (blue color) genes. We can also add the name of genes of interest to visualize their significance level and direction such as TP53 and ESR1 genes are shown in left side MA plot and several genes are shown in volcano plot.

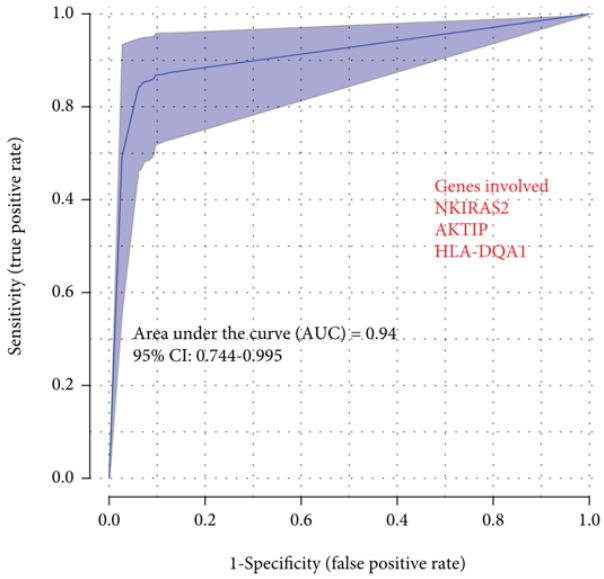
1.4.5 Modeling

In addition to the above motion comparative test, machine learning (ML) and statistical modeling approaches were used in significance testing mainly for prediction problems. These algorithms are capable of identifying important patterns/markers in large genomic data. ML falls into two main classes: unsupervised and supervised learning algorithms. Both classes are best suited to addressing distinct biological questions, and both will be required to effectively focus on characteristics of the data as well the outcome.

In a supervised learning model, the input consists of a set of training data with known labels (e.g., healthy vs patients). A supervised learning algorithm trains the system by analyzing a subset of data (called the training data) and produces an inferred function, which can be used for classifying unknown or new samples (called test/validation data). Some of the most common approaches we have used are random forest classification, linear and logistic regression. A linear regression model describes the relationship between a dependent variable, which is a continuous variable and one or more independent variables, which could be continuous and/or categorical. The dependent variable

is known as response variable and independent variables are known as explanatory or predictor variables. For example, we have developed a linear equation which explains the genomic changes with respect to patient age samples using the training dataset, based upon that linear equation we can predict the age of new samples or test samples using gene expression of genes included in that equation with degree of error. We could choose to perform univariate analysis on any of the individual variables in the dataset or multivariate analysis on a set of variables.

In an unsupervised learning model system, the input data is a set of unlabeled examples without predefined classes for example. These approaches are applied in grouping the data depending upon similar attributes (or distance), most similar patterns, or relationships amongst the dataset points or values. Distinct approaches are employed on every other algorithm in splitting up data into clusters. Various clustering algorithms are deployed in microarray analysis which is useful in clinical research in keeping track of gene expression data. For example, there are certain patterns that exist in gene expression of different cancers in a mixed dataset.



logit (P)
= 12.664 + 9.057
∗ ILMN_1677481 (NKIRAS2 – 6.497 ∗ ILMN_1665982 (AKTIP)
– 2.201 ∗ ILMN_1808405 (HLA – DQA1)

Figure 12: Example of AUROC plot with the equations for the logistic model.

Box 8: The AUROC figure (left) shows the prediction ability (AUC=0.94 with Confidence Interval (CI) = 0.74 – 0.96) of the logistic regressions model (Right). Three genes used are shown in red (left) and actual logistic regressions with weighted value of each gene in that model (right).

A principal component analysis (PCA) which is an unsupervised approach is used to identify hidden features in the data that provide the most correlated signal across the samples. The first principal component is the feature that explains most of the variability in the data. The objective of unsupervised learning is to discover hidden information within the data to detect the different groups or clusters. Some time we also used the semi-supervised algorithms which is the combination of both labeled and unlabeled data.

1.4.6 Pathway Analysis

Pathway analysis is the study of how genes and interlinkage systems of the genome contribute to different pathways. Pathway analysis uses gene lists generated from well-defined microarray study and searches for the strong association between this list and previously identified biological pathways using statistical measurements. The main objective of the pathway is to resolve how the individual segments of an organism work together to produce a particular phenotype. Pathway analysis is viewed as an intermediate step of translational research that brings biological research to be applied in clinical practice (from bench-side to bedside). Based on accomplishments of previous studies of genomics within organisms, it is inferred that the function of genes and other functional elements of the genome can be inferred more accurately only when the genome is studied in its entirety.

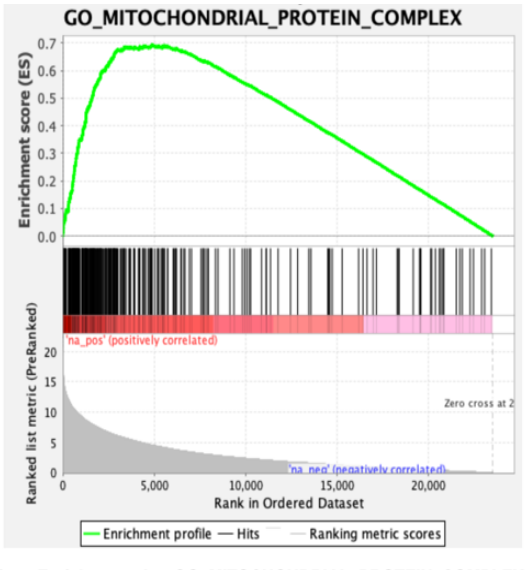
Following are some tools that used for pathway analysis:

1.4.6.1 Gene Set Enrichment Analysis (GSEA)

Gene Set Enrichment Analysis (GSEA) is a computational tool developed by Broad institute [15]. The tool has an inbuilt pathway database known as MSigDB which contains collections of gene sets, including regulatory target, oncogenic signature, and immunologic signature gene sets, among others. There ways to use the GSEA to identify the associated pathways, 1) users input an expression data set, phenotype annotation, and system creates a list of significant gene and then search for the associated pathways 2) pre-ranked GSEA, where user provides a list of gene with the

level of significance (p-value or fold change) and then system finds the significant pathways based upon the cutoff value. The gene set can be from the user's choosing or MSigDB (complete or partial set).

After running GSEA on a selected gene set and pathway database, we received two components as an outcome: GSEA statistics and GSEA reports. The GSEA statistics typically comprises enrichment score, normalized enrichment score, false discovery rate, and nominal p-value. GSEA Reports generate enrichment in phenotype, dataset details, gene set details, gene markers, and other helpful analyses to interpret gene signatures.



GeneSet	GO_MITOCHONDRIAL_PROTEIN_CO
Enrichment Score (ES)	0.69327265
Normalized Enrichment Score (NES)	1.7830322
Nominal p-value	0.0

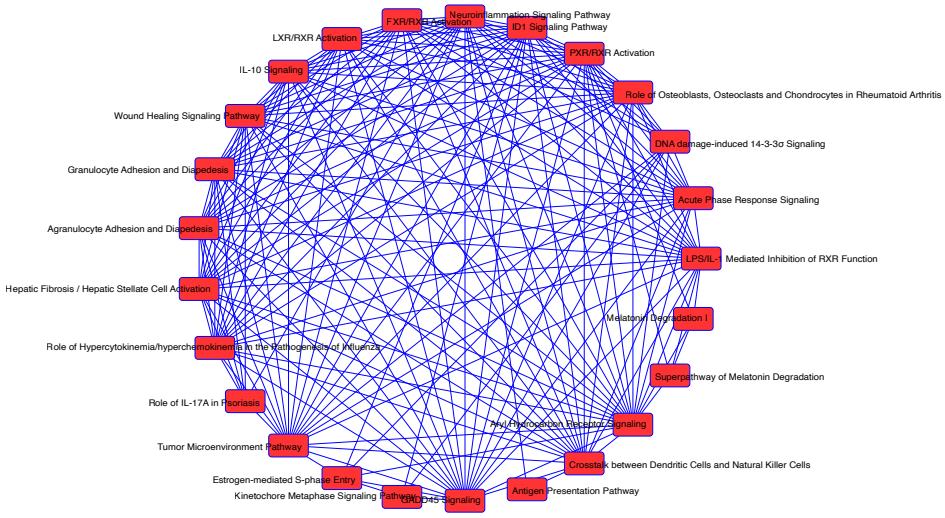
Figure 13:Example of GSEA showing the plot and statistics

Box 9: The green line shows the enrichment score across all the genes in our list (represents underneath bar lines) with the GO_Mitochondrial_protein_complex pathway from MSigDB database. Right side shows some important statistical parameter such as name of gene set, enrichment score value, Normalized enrichment score and nominal p-value.

1.4.6.2 Ingenuity Pathway Analysis (QIAGEN IPA)

Ingenuity Pathway Analysis (IPA) is a commercially available pathway analysis tool that quickly visualizes and understands complicated genomics data

(https://www.qiagenbioinformatics.com/products/ingenuity). It creates an interactive network to represent biological systems. Advanced analysis capabilities provide several options for gene selection, choice of analytical method and powerful algorithms combined with rich content to help us to identify the most critical pathways. IPA allows researchers to upload microarray data or subsets of significant genes with the platform information for pathway analysis. IPA allows an interactive network design to depict biological systems and offers a search feature for information on genes, proteins, chemicals, and medications.



Top Tox Functions

Hepatotoxicity

Name	p-value range	Molecules
Hepatocellular carcinoma	1.00E+00 - 1.76E-10	141
Liver Hyperplasia/Hyperproliferation	1.00E+00 - 1.76E-10	388
Liver Damage	3.11E-01 - 7.06E-10	35
Liver Proliferation	6.14E-01 - 3.20E-07	30
Biliary Hyperplasia	2.09E-01 - 8.52E-05	6

Nephrotoxicity

Name	p-value range	Molecules
Renal Damage	7.95E-02 - 3.30E-12	47
Renal Necrosis/Cell Death	2.95E-01 - 1.57E-07	57
Kidney Failure	6.14E-01 - 2.58E-07	35
Renal Inflammation	5.74E-01 - 2.66E-06	39
Renal Nephritis	5.74E-01 - 2.66E-06	39

Figure 14:Example of Ingenuity Pathway Analysis and top hepatotoxicity and nephrotoxicity functions from analysis.

Box 10: The close (circular) network of all the significant pathways identified using our gene list and IPA tool. All the pathways sharing some genes are connected through blue lines.

A free online web-based Database for Annotation, Visualization, and Integrated Discovery (DAVID) is a significant source for functional annotation and performing gene-annotation

enrichment analysis[16]. All tools in the DAVID resources provide functional interpretation of significant lists of genes derived from genomic (microarray in our case) studies including pathway visualization, annotation clustering, and annotation classification.

CHAPTER 2
Association between arsenic level, gene expression in Asian population and in vitro carcinogenic bladder tumor

2.1 INTRODUCTION

Arsenic (As) is a ubiquitous element in the environment, ranked the 20th most abundant element on earth. The toxic impact of arsenic on human health has been documented in numerous studies leading to arsenic identification as a known carcinogen by the International Agency Research on Carcinogens (IARC), the National Toxicity Program (NTP), and the United States Environmental Protection Agency (EPA) [1, 2]. In addition to cancer, long-term exposure to arsenic has been associated with developmental effects, cardiovascular disease, neurotoxicity, and diabetes (WHO, https://www.who.int/news-room/fact-sheets/detail/As). Typically, arsenic would only be found in background levels in soil and groundwater. However, high levels of arsenic accumulates in these medians from anthropogenic activities such as indiscriminate waste disposal from mining, milling, and smelting of ores [3], raw and spent oil shale [4], and coal fly ash amendments [5]. The usage pattern in the 1960s for arsenic compounds in the United States was 77% pesticides, 18% as glass, and 4% industrial chemicals. The past use of arsenic as a pesticide in agriculture is exemplified by New Jersey, where between 1900 and 1960, it is estimated that approximately 15 million pounds of arsenic were applied to New Jersey soils alone [6]. Leaching of arsenic from soils into the water supply has now resulted in the significant contamination of drinking water in many areas of the United States and the World. This past usage of arsenic in anthropogenic activities has now resulted in exposure to arsenic being a global public health problem [7-9]. This is illustrated by the fact that over 120 million people are affected by arsenic exposure, many of which reside in Bangladesh and India [8, 10]. A recent study has modeled the role of atmospheric exposure to arsenic as being additive to overall exposure levels [11]. Despite the health consequences of arsenic exposure, there is no molecular signature that might predict the risk of developing cancer or other diseases following exposure to arsenic.

On the other hand, the use of arsenicals as therapeutic agents in medicine is very well known dating back more than 2400 years to ancient Greece and Rome[12]. In the 19th century, potassium arsenite was used to treat different types of disease[13] including diabetes, psoriasis, syphilis, skin ulcers and joint diseases. More recently, phase I/II trials have been conducted in heavily pretreated patients with relapsed or refractory multiple myeloma shows Arsenic trioxide (ATO) is the most active, single agent in acute promyelocytic leukemia (multiple myeloma: types of blood cancers)[14]. Another study suggested that ATO can be used as an effective alternative therapeutic for the treatment of retinoblastoma which is the most common intraocular cancer in children[15]. The study shows an antitumor activity of arsenic which mainly targets multiple pathways in malignant cells, resulting in the promotion of differentiation or in the induction of apoptosis, which would be very helpful to understand the molecular mechanism of arsenic-exposed cancer biology as a reverse engineering approach.

Biomarkers are classified based on exposure, effect, and susceptibility[16]. For arsenic, biomarkers of exposure have received the greatest attention and success in defining individual exposures[17]. Human susceptibility to arsenic, especially as it applies to predicting disease states, is probably the least studied area of biomarkers. A few biomarkers of interest attracting study include clastogenicity in peripheral lymphocytes, micronuclei in oral mucosa and bladder cells, and induction of heme oxygenase[16, 18, 19]. Though years of research have been done, clinical implementation remained unsuccessful due to lack of risk–assessment which should be based on mechanistic detailing of individual risk of toxicity and developing strategies to counter arsenic toxicity at the molecular, social-economical, geographical, and environmental perspectives. For example, the total concentration of inorganic arsenic (iAs) and its metabolites, monomethylarsonic acid (MMA) and subsequently to dimethylarsinic acid (DMA) in humans urine has been recommended for the biological monitoring of occupational iAs exposure by the American Conference of Governmental Industrial Hygienists (ACGIH)[16] but as described by Buchet et al.[17], certain types of seafood can contain small quantities of DMA than the urine sample should abstain from eating seafood for 3–4 days prior to urine collection. In such cases where diet cannot be controlled, such biomarkers will not perform well. Additionally, the short half-life of inorganic and organic arsenic species in blood and invasive collection limits the utility of arsenic biomarkers in blood and urine. The goal of the present study was to identify differentially expressed genes in arsenic exposed humans and determine if a molecular signature could be developed that would

stratify and predict the risk of urothelial cancer for those with known exposure to arsenic. Urothelial cancer, which is the most common type of bladder cancer, was chosen as an initial proof of principle since epidemiological and other evidence is strong for the link between arsenic and the development of urothelial cancer, and there are publicly available databases for data mining [7, 20-24]. A theme of such studies shows a strong association at more extreme levels (>150 μg/L) whereas there is uncertainty of health effects that may develop below this threshold. Suggested mechanisms for arsenic carcinogenesis include oxidative damage, epigenetic effects, and interference with DNA repair. In addition, the development of bladder cancer is known to have a strong association with environmental exposures from mentioned anthropogenic activities [25]. Overproduction of reactive oxygen species (ROS) due to arsenic exposure primarily follows direct toxicity or the metabolic processes of arsenic products. Inhibiting succinic dehydrogenase activity in mitochondrial complexes I and III in electron transport chain produces superoxide radical anion, while monomethylarsonic acid (MMA) and dimethylarsinic acid (DMA) will form radicals in the cell and specifically the enoplasmic reticulum [26] [27]. Since inorganic arsenic compounds tends to be more toxic than organic, ATO is of interest for its global concern along with its involvement in oxidative and nitrosative stress properties. Translational damage from reactive species can regulate MAPK family or induce extended states of inflammation, genetic and epigenetic mechanisms such as these are indicative of oxidative/nitrosative damage and well associated with the development of bladder cancer [28], [29], [30]. ILK signaling and Neuroinflammation Signaling Pathway were the most frequent pathways affected by the exposure of arsenic and both of them are highly associated with oxidative stress. Oxidative stress and neuroinflammation could potentiate each other to promote progression of mental disorders[31], whereas, ILK plays a complex roles in the modulation of oxidant species production[32]

The strategy used in this study involved three steps. The first step was a blood cell gene expression analysis of two diverse human populations with known levels of exposure to arsenic. One population was stratified to low and high exposure, and the second population to low, medium, and high exposure with correlation to human global gene expression. After identifying statistically significant genes unique to the mentioned test conditions, we found cancer was the most significant disease and lipid metabolism (which is considered as a major metabolic pathway involved in the progression of cancer) were most significant molecular and cellular functions associated with genes differentially expressed due to different levels of arsenic. Therefore, the next stage was to

compare it with data from four independent myeloma cell lines that had been treated with iAs trioxide (ATO) to understand the molecular mechanism of cancer. Many of the genes that were up- and down-regulated due to arsenic-exposure are associated with cancer biology. There genes lists were then subjected to enrichment analysis to identify statistically significant pathways and further scrutinized for functional relevance. The third step was to develop a model by examining the ability of the most significant genes to predict the progression, and possible development of bladder cancer using publicly available patient biopsy samples. Using this approach, we developed a robust regression model of three significant probes and corresponding genes results with AUC of 0.94 (95% CI: 0.744-0.995) and 0.75 AUC (95% CI: 0.343-0.933) on the training and validation data respectively. The most significant pathway identified is integrin-linked kinase (ILK) which plays a key role in eliciting a protective response to oxidative damage in epidermal cells.[32].

2.2 MATERIALS AND METHODS

2.2.1 Data

Two publicly available gene expression datasets of previously conducted experiments were accessed from two independent populations. The set from Bangladesh (Gene Expression Omnibus GEO ID: GSE57711) had 29 individuals, 16 were males and 13 females. The second dataset was from Pakistan (GSE110852 ID) had 57 individuals composed of 31 males and 26 females. In this report, the set from Bangladesh is denoted as Data1 and that from Pakistan is Data2 and remain unchanged from their original, respective studies. Data1 samples were part of a clinical trial in June 2011 [33]. For these samples, ‘low’ exposure levels correlate to a range of 50-200µg/L, whereas ‘high’ levels correlate to a range from 232-1000µg/L (there were no samples collected from patients exposed in the range of 201-231µg/L). Data2 samples were from two main districts of rural Pakistan, Lahore and Kasur. The study aimed to investigate the blood transcriptome profile among the exposed samples to correlated gene expression to exposure levels of iAs [34]. Urine sampling was used to define levels of arsenic exposure, with ‘low’ being 0-50 µg/g creatinine, ‘medium’ as 51-100 µg/g creatinine, and ‘high’ as >101 µg/g creatinine. The general characteristics of both data sets are detailed in Table 1.

Table 1 Characteristics of arsenic exposed gene expression data.

	Total Samples	Gender		Low Exposure	Medium Exposure	High Exposure
Data1				Water As 50–200(µg/L)	-	Water As 232–1000(µg/L)
GSE57711	29	16	13	15	-	14
Data2		Males	Females	Water As 122.22 ± 86.13(µg/L)	Water As 130 ± 128.9(µg/L)	Water As 148 ± 105.01(µg/L)
GSE110852	57	31	26	18	19	20

The results from 4 multiple myeloma cell lines treated with ATO was obtained from the GEO database, series GSE14519, [35]. These cell lines; U266, MM1S, KMS11, and 8226S were exposed to ATO for 6hr, 28hr, and 48hr before analysis. Gene expression profiling was used to determine differences in cell line response to ATO. This study was used as a reference point in the present study since it documents the effects of arsenic compounds on gene expression at different exposure levels.

The two databases of previously conducted experiments containing biopsies of bladder cancer were obtained from GEO, GSE13507 [36, 37] and GSE3167[38]. The GSE13507 contained 165 samples for primary bladder cancer, 23 recurrent non-muscle invasive tumor tissues, 58 normal-looking bladder mucosa surrounding cancer, and 10 normal bladder mucosa. This dataset was originally used in microarray analysis for the identification of genes with prognostic significance. GSE3167 contained 28 samples of superficial bladder tumors, 13 samples of muscle-invasive carcinomas, and 9 normal samples. This dataset was previously used for gene expression signatures among various stages of carcinomas. These 2 datasets were used in the present study to obtain a prognostic gene-based prediction for bladder cancer.

The QC report of the datasets was examined and only qualified samples were included. Since the data was generated using different platforms (such as Affymetrix, Agilent etc.), no single approach would work on those datasets. Therefore, the raw datasets were pre-processed to extract expression using the same approach provided within publication of study, for example the Affymetrix package in R used for GSE57711 while an in-house QC pipeline (github.com/BiGCAT-

UM/arrayQC_Module) for data GSE110852. Before applying any statistical test, the distribution of each data tested and transformed into a normal distribution using logarithm and pareto scaling (mean-centered and divided by the square root of the standard deviation of each variable) transformation. Entire gene expression analyses were performed using R Bioconductor (https://www.r-project.org). The analysis was performed using Bioconductor packages such as Stat, Dplyr (https://cran.r-project.org/web/packages/dplyr/index.html), ggplot2 (https://cran.r-project.org/web/packages/ggplot2/index.html), randomForest (https://cran.rproject.org/web/packages/randomForest/), e1071 (https://cran.rproject.org/web/packages/e1071/index .html), and pvclust (https://cran.rproject.org/web/packages/pvclust/index.html).

Sample Size calculation to find the power of detection with available number of samples. The number of samples with categories of arsenic exposure are 1) GSE57711 – (low, n= 15/high, n=14) 2) GSE110852 – (low, n= 18/medium, n=19 /high, n=20). The power analysis of GSE57711 data shows that a minimum of 14 samples are required to achieve the 80% power with minimum genes 11626 (per-sample), acceptable number of false positives is 5, fold change differences desired of 2, standard deviation of 0.6, and alpha (per-gene) 0.00043. These sample size computations assume that the expression of each gene is normally distributed on the log scale and believe that gene expression measurements are independent[39].

2.2.2 Machine Learning (ML) methods

The machine learning based classification approach was used to understand the population characteristics; partial least squares discriminant analysis (PLS-DA) was used in which the properties of PLS regression (PLS-R) is combined with the discrimination power of the classification technique [40]. The goal here is to determine the distribution of samples and visualize how the global gene expression profile scattered in different groups (sex and arsenic exposure) and which features best describe the differences between them.

A Random Forest (RF) was implemented to understand further which genes correlate to classification between sex and categories of arsenic concentration [41], specifically, if the gene expressions of a combination of genes can correctly differentiate between categories. This

classification provides insight into which genes are expressed differentially depending on an individual's condition, such as sex and/or exposure.

The correlation between the samples was calculated using the Pearson's coefficient [42, 43] and the heatmap method[44]. These were used to plot the correlation coefficient values to find the most correlated samples. Hierarchical clustering, an unsupervised learning approach, was then employed to calculate a dendrogram to determine the closest related samples. A hierarchical clustering an unsupervised learning approach was then employed to calculate a dendrogram to determine the closest related samples [45, 46].

2.2.3 Statistical methods

The statistical significance of each gene within each dataset was calculated by running t-tests [47] between the categories for conditions (male/female, low/high, low/medium, medium/high As exposure levels). Along with t-tests for pairwise comparison, One-way ANOVA[48] (Analysis of Variance) with post-hoc Tukey HSD (Honestly Significant Difference) [49] tests were performed for comparing multiple groups, i.e., level of As together with sex effect. The gene is filtered based on p-value with threshold 0.05 without statistical methods that control the false discovery rate (fdr) to avoid the loss of a large number of genes at initial level without further evaluation.

2.2.4 Pathway enrichment analysis

The Ingenuity Pathway Analysis (IPA, version 2020; Ingenuity Systems; QIAGEN)[50] was used for pathway enrichment, and functional analysis of the significant genes among the Human arsenic exposed samples. The KEGG pathways[51], PFAM protein domains[52], Uniport keywords[53], biological processes, molecular functions, cellular components, and Reactome pathways[54] were used to find associated pathways with statistically significant genes. In addition, we further built a network of gene-gene associations using STRING[55] that leads us to the gene subsets corresponding to a part of a particular function or pathway.

2.2.5 Prediction model

Classical univariate AUROC [56] analysis was performed to find out the prediction ability of each gene independently using logistic regression [57] method with 10-fold cross validation approach; next, all the genes were ranked according to this AUC value, and all possible combination of genes were tested by adding one gene at a time to the logistics model of top gene in a multivariate. The final model was selected based on the highest AUC (with 95% confidence intervals CI) among all possible combinations of the selected genes and performance was tested using the Monte Carlo cross-validation (MCCV).

The flow of the study is demonstrated by two charts (Supplementary Figure 1A and 1B). The first chart shows the flow of the analytical approaches used in parallel to analyze the Data1 and Data2 to find differently expressed genes, those specific to sex, those specific to arsenic exposure, and specific to both sex and arsenic exposure. It shows the differentially expressed genes and pathways with overlapping significance following statistical methods described and provided for clarity. The second supplementary flow chart describes and shows the flow of the next stage of this study in taking the statistically significant genes to find commonality with the four cancer cell lines and creating a bladder cancer risk predictor with high accuracy.

2.3 RESULTS

2.3.1 Global Gene Expression analysis of two As Exposed Sets of Human Data

PLS-DA and Pearson correlation were performed together with patient sex and arsenic exposed level to determine the similarities and differences in global gene expression patterns between the arsenic exposed cohorts.

For the Data1 samples, PLS-DA analysis demonstrated a clear separation between high arsenic exposed females compared to low arsenic exposed males (Figure 15A). The global gene expression profile distribution moves from low too high for samples of high exposure of arsenic in females (HF), low exposure of arsenic in females (LF), high exposure of iAs in males (HM), and low exposure of iAs in male (LM). The analysis did not demonstrate an explicit separation between low

exposure females and high exposed males. The heatmap of Pearson correlation coefficients and the dendrogram among the samples show variability within Data1 (Figure 15C) with correlation value 0.92 to 1. There was no clear separation based solely upon arsenic exposure or sex.

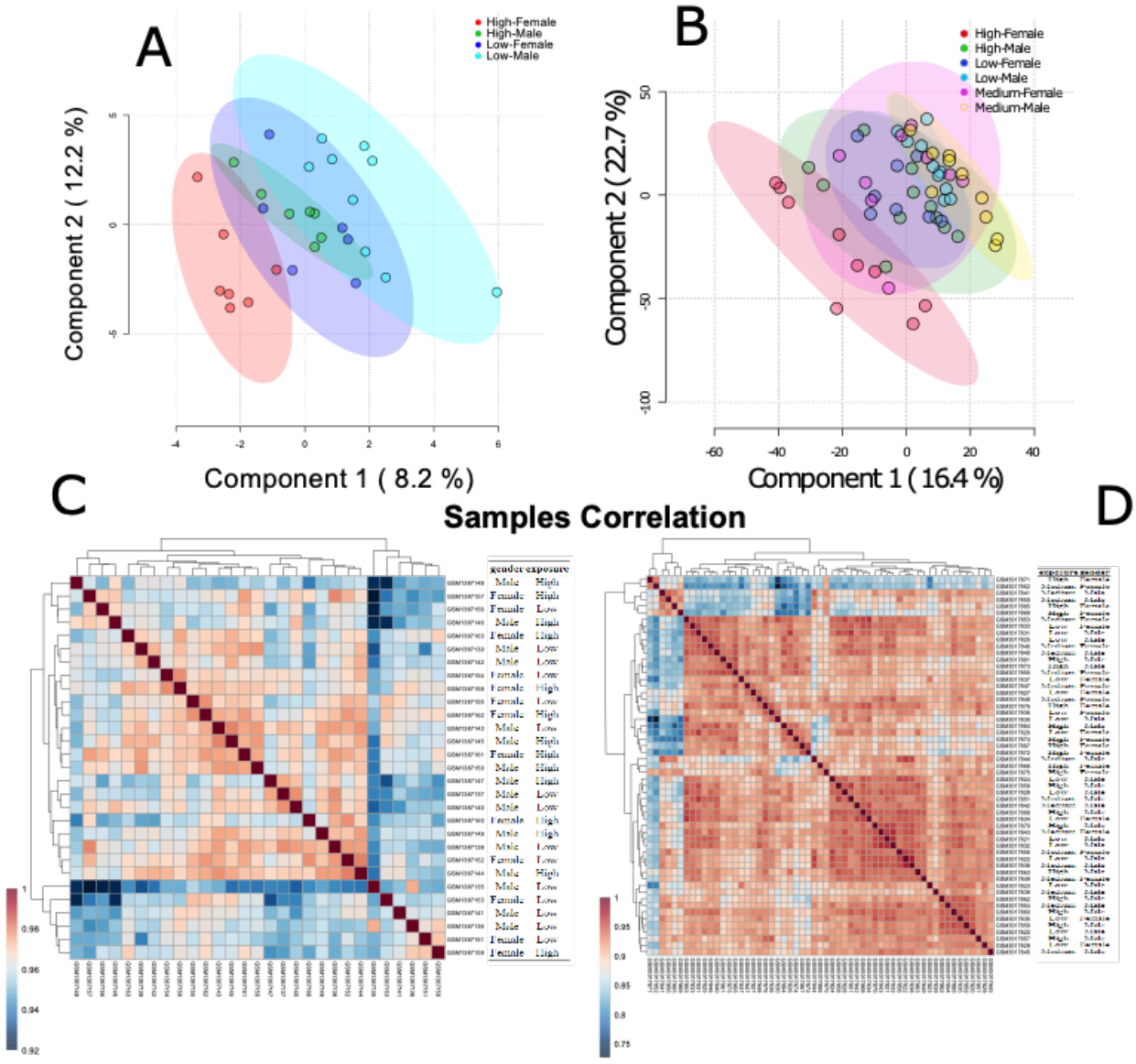


Figure 15: Sample distribution of gene expression profiles.

Box 11: *A) Data1 and B) Data2: The Partial least squares discriminant analysis (PLS-DA) plot showing clusters of samples based on similarity. The first two components of PLS-DA (PC1 and PC2) of gene expression profile and overall variance between the groups are displayed. Each dot represents a sample color coded by both gender and level of arsenic exposure level. Pearson correlations were calculated between each sample of total population and correlation coefficient values were shown by heatmap of Data1 (C) and Data2 (D). The color-coding bar proves the value of correlation-coefficient. The dendrogram represents the relation between the samples created by using hierarchical clustering approach.*

This provides evidence that sex and arsenic exposure has no bias impact on gene expression profile of Data1. An identical analysis of the samples in Data2 demonstrates that the distribution of global gene expression is almost the same for high and low exposure between males and females, where females with high arsenic exposure have the lowest expression profile when compared to low iAs exposed males (Figure 15B). The global gene expression profile moves low to high starting lowest in high exposure of arsenic in female (HF), to high exposure of iAs in male (HM), then low exposure of iAs in female (LF), and low exposure of iAs in male (LM), however, the medium exposure of iAs is mixed with low iAs exposure. The heatmap of Pearson correlation coefficients and the dendrogram among the samples show variability within Data2 (Figure 15D) with correlation value 0.75 to 1, and there is no clear separation between either based upon arsenic exposure or sex. This indicates that this data has no bias impact for both the factors.

As described in our methodology and shown via supplementary figures, the most significant, common genes were screened across the two datasets to distinguish difference among all four scenarios of sex and arsenic exposure in Data1 (Figure 16A) and all six scenarios in Data2 (Figure 16B).

This analysis showed several probes for the genes XIST (X inactive specific transcript), MALAT1 (metastasis-associated lung adenocarcinoma transcript 1), XLOC_008276 (long intergenic non-protein coding RNA 278), USP9Y (Ubiquitin Specific Peptidase 9 Y -Linked), SEPTIN6 (Septin 6), DDX3X (DEAD-Box Helicase 3 X-Linked), KDM6A (Lysine Demethylase 6A) and ZFX (Zinc Finger Protein X-Linked) as most significant in the RF as well as in the hierarchical clustering approach (Figure 16 A-D).

In addition to the advanced machine learning approach, the ANOVA test with post-hoc test was used to compare different arsenic exposure levels together with sex. This identified 476 probes (corresponding to 476 unique genes symbols) (Supp. Table 2) that were differently expressed in Data1 and 529 probes (corresponding to 439 unique genes symbols) (Supp. Table 2) that were differently expressed in Data2 (p-value < 0.05).

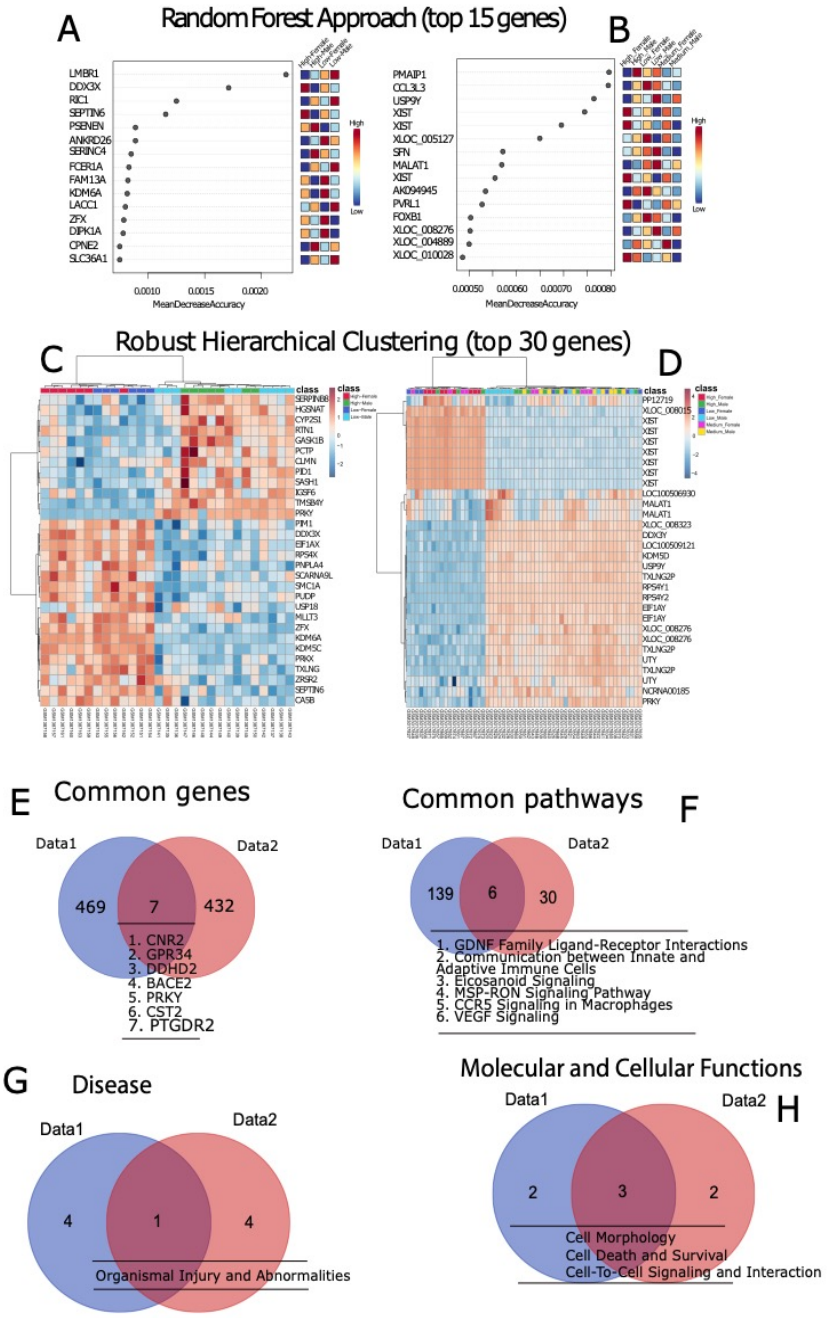


Figure 16: Global gene expression profile analysis.

Box 12: A-B) A list of top 15 genes is displayed with Mean decrease in accuracy value (X-axis) calculated using Random Forest Approach for Data1 and Data2 respectively. The small box (right side) and color-coding bar represents the expression value (from low to high) of each gene in different conditions. C-D) Top 30 genes identified by robust hierarchical clustering approach for Data1 and Data2 respectively. The heatmap represents the gene expression value across different samples. The top line on the x-axis, each box represents one sample. Two color-bar codes provide the gene expression value and condition of the sample (ultra-right). E-F) Venn diagram, overlap of most significantly differentially expressed genes and significantly associated pathways (p-value≤ 0.05) between both cohorts Data1(blue) and Data2 (red) respectively. The complete lists of significant genes are provided in supplementary table -2 and significant pathways are in supplementary table-3.

An IPA analysis was performed to identify the functional relevance of these genes in terms of pathway association, resulting in identifying a total of 145 and 36 significant pathways for Data1 and Data2, respectively (Supp. Table 3). A total of 7 common genes and 6 common pathways were found to overlap for the two populations (Figure 16 E, F). The common genes identified were; CNR2 (Cannabinoid Receptor 2), GPR34 (G Protein-Coupled Receptor 34), DDHD2 (DDHD Domain Containing 2), BACE2 (Beta-Secretase 2), PRKY (Protein Kinase Y-linked Pseudogene), CST2 (Cystatin SA), and PTGDR2 (Prostaglandin D2 Receptor 2) (Figure 16 E-F).

Organismal Injury and Abnormalities was the only disease common between two datasets due to combined change of arsenic level and sex and Cell Morphology, Cell Death and Survival and Cell-To-Cell Signaling and Interaction were common Molecular and Cellular Functions (Figure 16 G-H, supplementary table: 1B).

2.3.2 Sex-Specific Gene Expression

PLS-DA analysis was performed on Data1 and Data2 to determine only sex-based changes in overall gene expression profiles. The PLS-DA plot for Data1 demonstrated a clear separation among the 17 male and 13 female samples, with females having an overall lower gene expression profile (Figure 17A). The analysis of Data2 showed the same pattern among the 31 males and 26 females, with a relatively low separation because of the high variability of female gene expression profiles (Figure 17B). A t-test was used to identify the significant differentially expressed genes between the sexes ($p < 0.05$). This identified 532 and 373 genes for Data1 and Data2, respectively (Supp. Table 4). Volcano plots were generated for both data sets to demonstrate high statistical significance as determined by p-value together with a fold change difference of 2 (Figure 17 C, D). This analysis identified 3 biologically significant genes for Data1, and no significant genes for Data2. Of the 3 genes identified, two were down-regulated, PRKY (protein kinase Y-linked – pseudogene), and TMSB4Y (thymosin beta 4 Y-linked), while the one up-regulated gene was KI67 (a marker of proliferation, Ki-67).

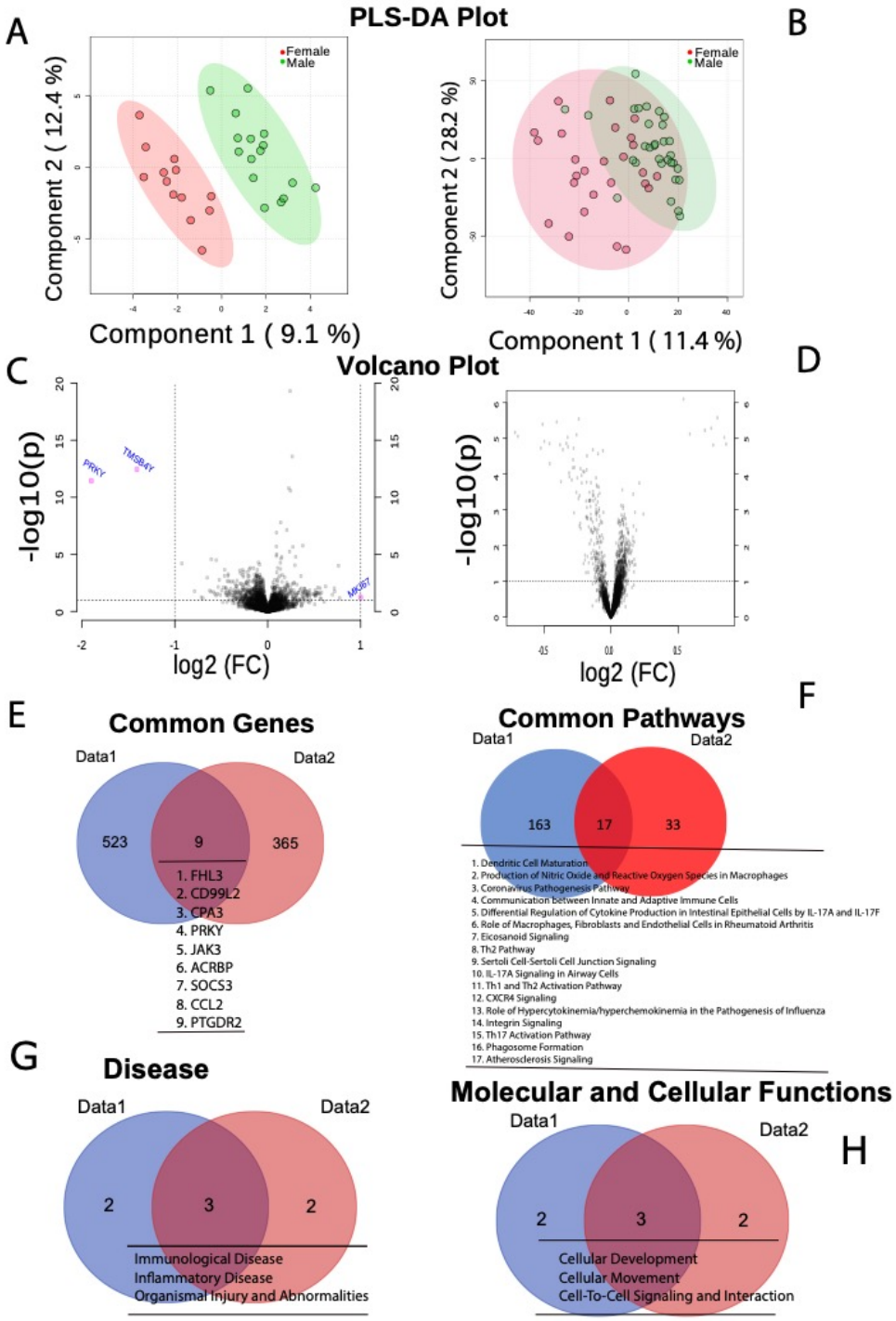


Figure 17: Sex-dependent genetic variations

Box 13: A-B) PLS-DA plot showing the gene expression profile distribution of each sample between females (red) and males (green) for Data1 and Data2 respectively. The first two components of PLS-DA (PC1 and PC2) of gene expression profile and overall variance between the groups are displayed. Each dot represents a sample color coded by gender. C-D) Volcano plot displays the log2 fold change and -log10(p-value) of gene expression differentiating due to gender effect for Data1 and Data2 respectively. Genes with higher than two-fold (p-value ≤ 0.05) are highlighted in red. E-F) Venn diagram, overlap of most significantly differentially expressed genes and significantly associated pathways (p-value≤ 0.05) between both cohorts Data1(blue) and Data2 (red) respectively. The names of common genes are provided in a table (underneath). The complete lists of significant genes are provided in supplementary table -4 and significant pathways are in supplementary table-5.

The determining overlapping significant genes (p-value<0.05) identified 9 genes that were common among the 532 differentially expressed genes of Data1 and the 373 genes of Data2 (Figure 17E). These 9 genes were: FHL3 (Four and A Half LIM Domains 3); CD99L2 (CD99 Molecule Like 2); CPA3 (Carboxypeptidase A3); PRKY (protein kinase Y-linked – pseudogene); JAK3 (Janus Kinase 3); ACRBP (Adrenoceptor Beta 3); SOCS3 (Suppressor of Cytokine Signaling 3); CCL2 (C-C Motif Chemokine Ligand 2); and PTGDR2 (prostaglandin D2 receptor 2). The genes for PRKY and PTGDR2 also appeared in the comparisons for unique overlapping genes in the global population analysis performed in the previous section. An IPA pathways analysis demonstrated that 180 pathways were significantly associated in Data1 as compared to 50 pathways in Data2 (Figure 17C-D, Supp. Table 5, Figure 17F). A comparison of these pathways demonstrated that there were 17 overlapping common pathways with respect to sex in the 2 populations.

Immunological Disease, Inflammatory Disease, Organismal Injury and Abnormalities were the common disease and Cellular Development, Cellular Movement, Cell-To-Cell Signaling and Interaction were the common Molecular and Cellular Functions (supplementary table: 1B) associated due the sex difference in both the datasets.

2.3.3 As-Specific Human Gene Expression

An identical analysis used above for sex was employed to compare the differences in gene expression due to arsenic exposure for Data1 and Data2. PLS-DA demonstrated a prominent division of high and low arsenic exposure for those in Data1, where high exposure showed an overall low gene expression profile (Figure 18A). The analysis of Data2 showed a substantial division between the high versus medium levels of arsenic exposure, but no separation between medium and low level of exposure (Figure 18B). The differentially expressed genes were identified for Data1 using the t-test with significance at $p<0.05$. For Data2, the differentially expressed genes were identified using paired t-test and ANOVA between all three groups with a threshold level of $p < 0.05$.

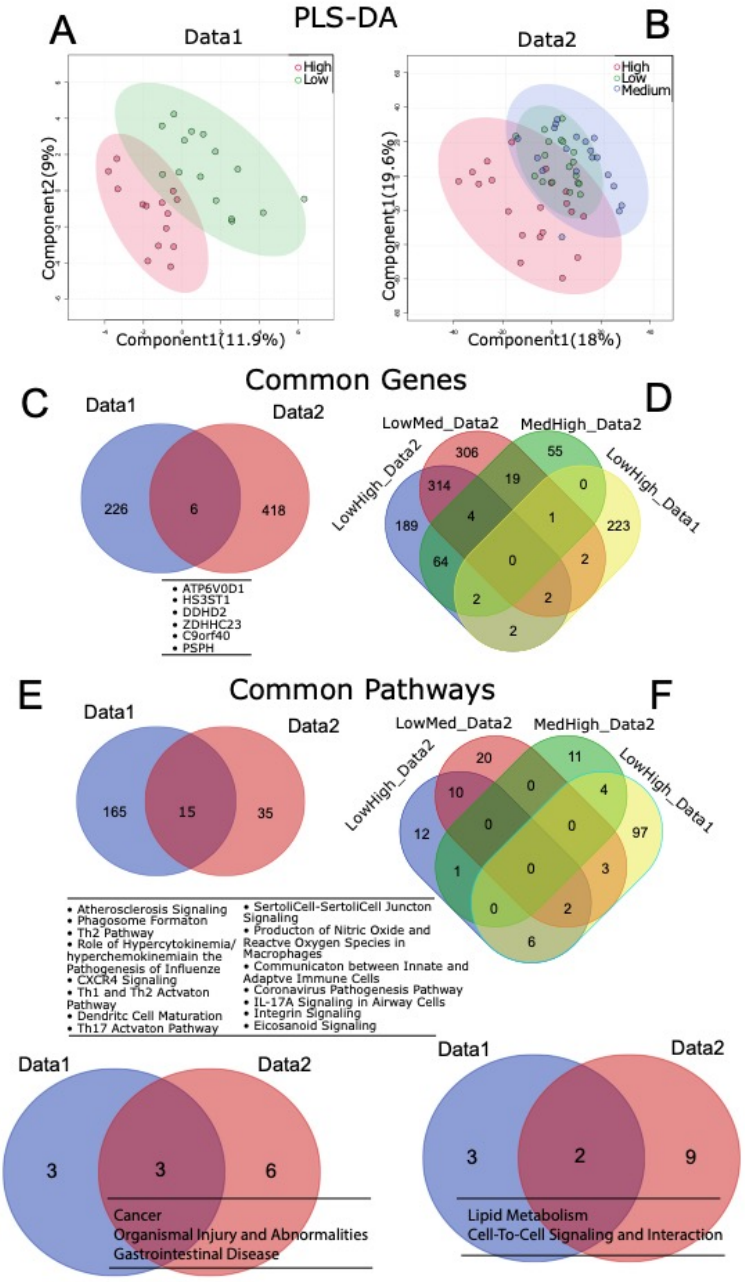


Figure 18: Arsenic-level dependent genetic variations.

Box 14: A-B) PLS-DA plot showing the gene expression profile distribution of each sample between different levels of arsenic exposure (red: High, purple: Medium and green: Low) for Data1 and Data2 respectively. The first two components of PLS-DA (PC1 and PC2) of gene expression profile and overall variance between the groups are displayed. Each dot represents a sample color coded by As-level. C) Venn diagram, overlap of most significantly differentially expressed genes when comparing the Low vs High for Data1 and Low-Medium-High (p-value≤ 0.05) for Data2 and represented by blue (Data1) and red (Data2) respectively. The names of common genes are provided in a table (underneath). D) Venn diagram, overlap of most significantly differentially expressed gene when compared the low vs High for Data1 and pairwise comparison between Low-Medium-High with (p-value≤ 0.05) and represented as Data1(yellow) and Data2 (blue: Low vs High, red: low vs Medium, green: medium vs High) respectively. E-F) Venn diagram, associated pathways for the genes identified in Figure C-D with same classification and color coding described.

This analysis identified 232 genes from Data1 and 424 genes from Data2 that were significant with 6 genes being common between the datasets (Supp. Table 6, Figure 18C). These 6 genes were: ATP6V0D1 (ATPase H+ Transporting V0 Subunit D1); HS3ST1 (Heparan Sulfate-Glucosamine 3-Sulfotransferase 1); DDHD2 (DDHD Domain Containing 2); ZDHHC23 (Zinc Finger DHHC-Type Palmitoyltransferase 23); C9orf40 (chromosome 9 Open Reading Frame 4); and PSPH (Phosphoserine Phosphatase). All Common genes were between all possible pairwise combinations of different arsenic levels of Data2 together with those of Data1 (Supp. Table 8, Figure 18D). The total number of significant pathways were 180 and 50, for Data1 and Data2, with 15 pathways common between them (Supp. Table 7, Figure 18E). The interaction of significant pathways identified by the paired analysis was also determined (Supp. Table 9, Figure 18F). ILK signaling and Neuroinflammation Signaling Pathway are the most frequent pathways identified through the comparative analysis of pathways (Supp. Table 9).

Cancer, Organismal Injury and Abnormalities and Gastrointestinal Disease were the common disease and Lipid Metabolism, Cell-To-Cell Signaling and Interaction were the common Molecular and Cellular Functions (Figure 18E-F, supplementary table: 1B) associated due the sex difference in both the datasets.

2.3.4 Myeloma Cancer Cell Lines Exposed to As Trioxide (ATO)

The “methods” section provided time course and ATO exposure details for the U266, MM.1s, KMS11, and 8226/S multiple myeloma cell lines used to generate genomic data obtained from GEO. The gene expression results from Data1 and Data2 were compared with the global gene expression results from the 4 myeloma cancer cell lines exposed to ATO. The results of this comparison demonstrated that 58, 78, 59, and 38 genes were found to be commonly expressed in the arsenic exposed population and the 4 cell lines (Figure 19 A-D). An examination of this data demonstrated that there was a total of 147 unique genes (Supp. Table 10) that appeared common in the 4 cell lines and the arsenic exposed populations (Data1 and Data2). This set of 147 unique genes was used to predict urothelial cancer development and progression in the next section of results.

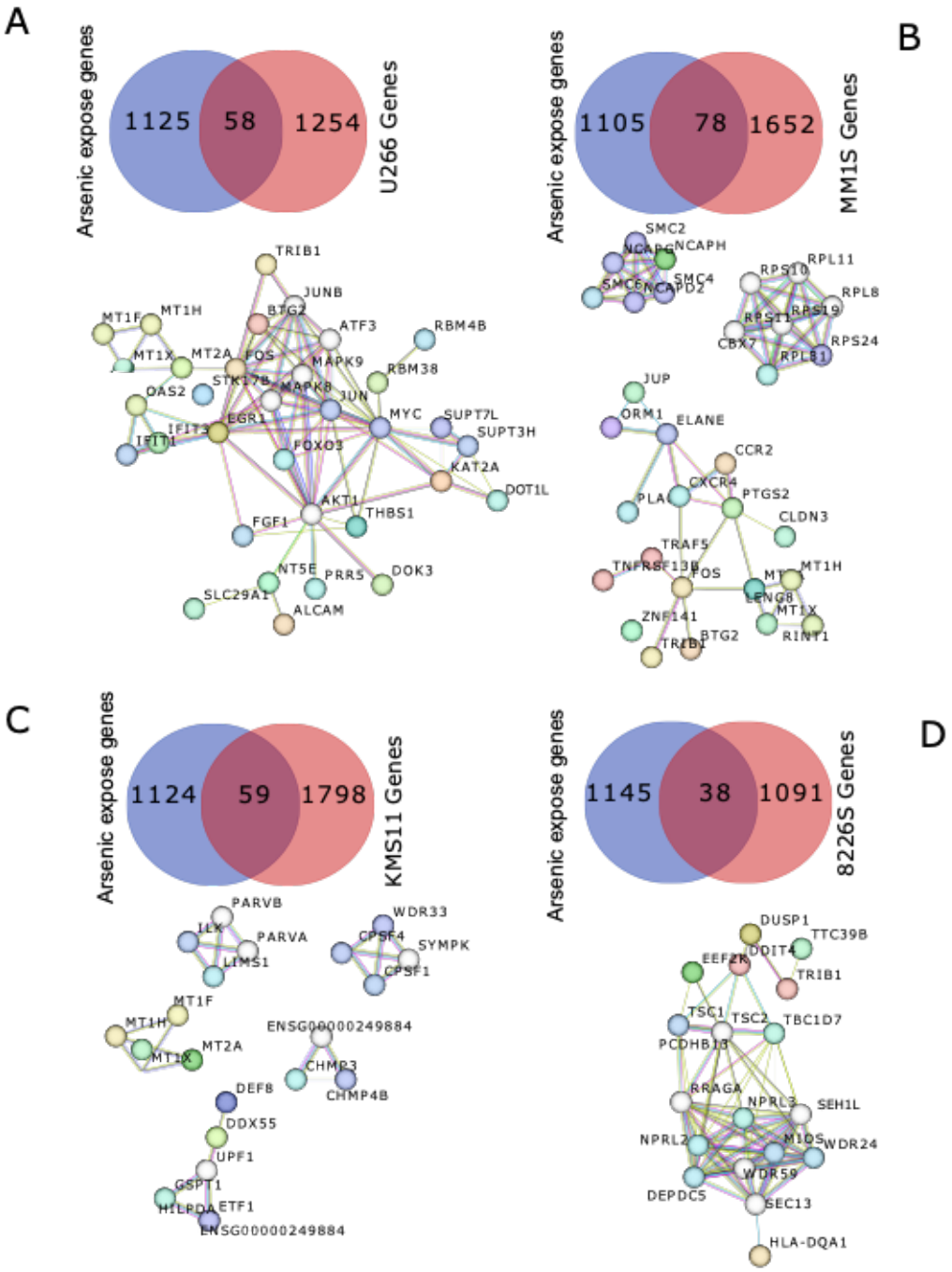


Figure 19: Identification of previously known arsenic exposed genes association with cancer progression.

Box 15: *A-D) The common arsenic exposed gene-set from Data1 and Data2 compared with differentially expressed genes within four Arsenic trioxide (ATO) cell lines. Venn diagram, showing total number of common genes, (A) U266, (B) MM1S, (C) KMS11, and (D) 8226S. Interaction networks functional enrichment analysis plots using STRING were demonstrated (underneath). The plots were generated with common genes identified between each cell-line and arsenic exposed gene-list, the connected lines represent the degree of interconnectivity and enrichment in characteristic molecular functions.*

The functional association of common genes and interaction networks were determined by functional enrichment analysis using the STRING for each cell line (Figure 19 A-D). The gene interaction networks for each myeloma cell line identified 6 genes that were central to the interaction of the networks. These 6 genes were important transcription factors or second messengers. The EGR1 gene encodes a zinc finger protein that is a transcriptional regulator that plays a major role in cell survival, proliferation, and cell death. Its activation of p53/TP53 and TGFB1 suppresses tumor formation. MAPK8 and 9 genes are integration points for multiple biochemical signals and can influence a wide variety of cellular processes such as proliferation, differentiation, transcription regulation, and development. The FOXO3 gene functions as a transcriptional activator that regulates apoptosis and autophagy. The MYC gene is a proto-oncogene that plays a major role in cell cycle progression, apoptosis, and cellular transformation. The AKT1 gene is activated by platelet-derived growth factor and is looked upon as a survival factor that can inhibit apoptosis. The STK17B gene is a kinase involved in the regulation of apoptosis and autophagy.

The gene set enrichment shows various functional aspects in all four cell lines (Figure 20). The gene functions significant for U266 were associated with cellular response to the metal ion cadmium, external stimulus, and cytokine. The gene sets were also a part of pathways related to colorectal cancer, choline metabolism in cancer, HTLV-1 infection, TNF (Tumor necrosis factor) signaling factor, and prolactin signaling pathway. The gene functions significant in MM1S were associated with mitotic/meiotic chromosome condensation, cellular response to Zn ion, nuclear-transcribed mRNA catabolic process, and SRP-dependent co-translational protein targeting to the membrane. Among these genes, the pathways associated with this comparison are mineral absorption and the ribosome pathway. The biological functions related to KMS11 that are significant are a cellular response to Zn ion, mRNA polyadenylation, termination of RNA polymerase II transcription, positive regulation of viral life cycle, and viral release from the host cell. And the pathways related to these gene sets are a component of mineral absorption and mRNA surveillance. While looking at the gene sets from 8226S, the significant biological functions include TOR (target of rapamycin) signaling, response to amino acid starvation, and nutrient levels. The pathways that were related in these gene sets were mTOR signaling and autophagy pathways.

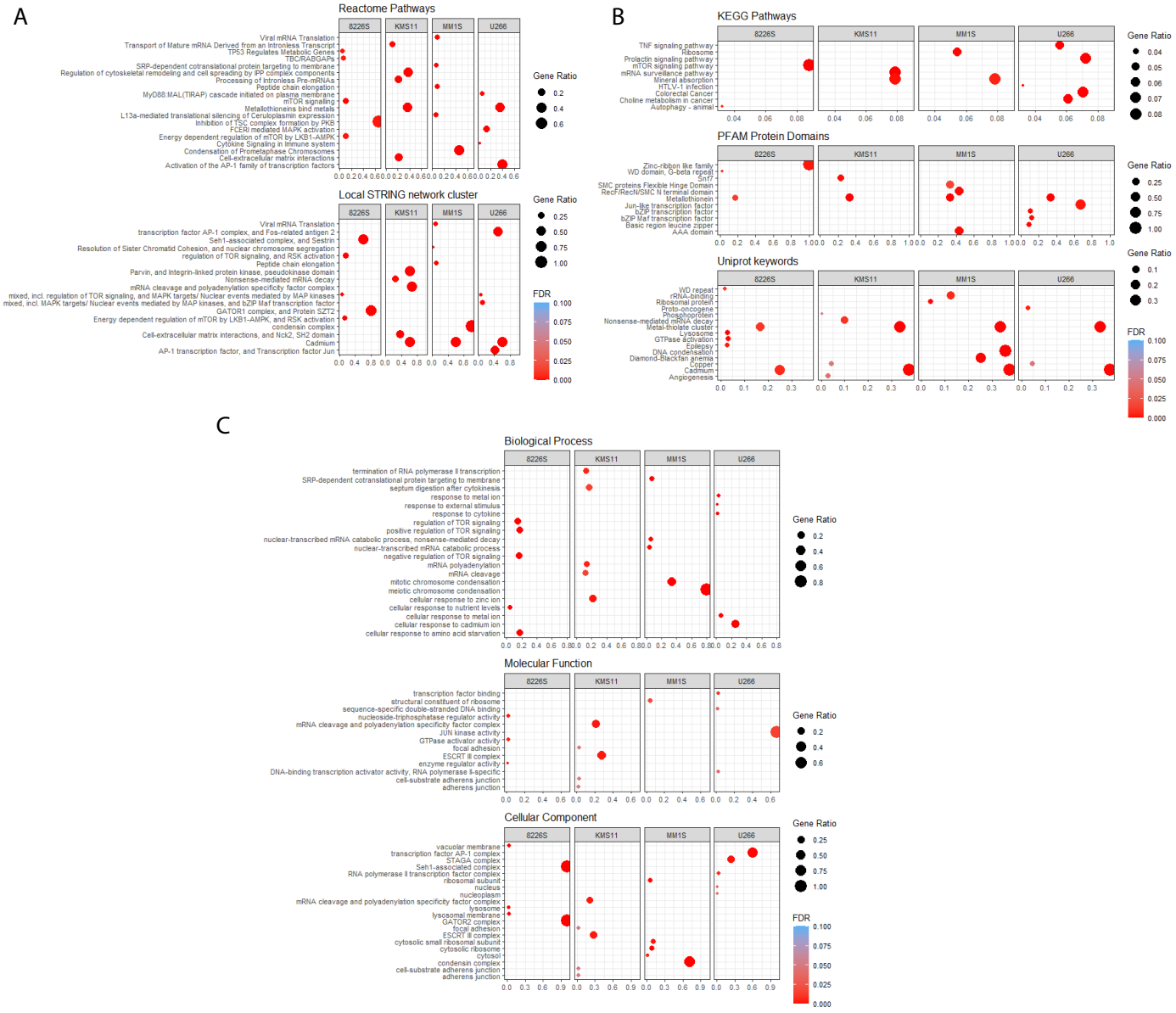


Figure 20: Functional analysis of arsenic exposed and cancer associated genes.

Box 16: To visualize the enriched terms, dot plots are generated using significantly associated pathways with arsenic exposed and cancer progression. It depicts the enrichment scores (p values), gene ratio as bar height and color. The pathway databases used for significance are: A) Reactome, local STRING network clusters, B) KEGG pathways, PFAM protein domains, Uniprot keywords, C) Biological processes, Molecular functions, and cellular components.

When including sex factor to identify the arsenic exposed sex specific gene association with ATO, we find an overlap of total 18 genes (BTG2, CXCR4, BACE2, EGR1, PHACTR1, CRIM1, TRIB1, TNFRSF12A, TSPAN5, RGS1, CD24, DDIT4, OLFM4, DDX3Y, PMAIP1, SLC29A1, SMAD5 and MYB) between the differentially expressed genes in arsenic exposed male/female (from Sex-Specific Gene Expression section) and differently expressed genes within ATO (Supp. Table 11)

2.3.5 Bladder Cancer Prediction Model

The 147 genes generated from the previous results section (section: Myeloma Cancer Cell Lines Exposed to As Trioxide) were utilized to develop a bladder cancer prediction model for the purpose of early diagnosis and prevention. Two publicly available human datasets were used as a training and validation data to test the prediction ability of those genes using the prediction model approach described in method section. The first (GSE13507[36, 37]) was used as a training dataset. The logistic model shows that primary tumor with three genes NKIRAS2, AKTIP, and HLA-DQA1, out of 147 with AUC 0.96 (0.82-0.99), (Figure 21A). The equations for the logistic model are given below with probe id together with gene name in brackets.

GSE13507 data modeling:

A) Normal Vs Primary tumor (GSE13507):

*logit(P) = 12.664 + 9.057 * ILMN_1677481 (NKIRAS2 – 6.497 * ILMN_1665982 (AKTIP) – 2.201*

** ILMN_1808405 (HLA-DQA1)*

Outcome Area under the curve (AUC) = 0.94(95% CI: 0.744-0.995) (Figure 21A)

The same three genes were used with another set of bladder cancer data (GSE3167 [38]) to validate the primary bladder tumor predictor. It was seen that the genes (NKIRAS2, AKTIP, HLA-DQA1), shows the prediction ability of AUC: 0.75 (95% CI: 0.34-0.93) on this dataset (Figure 21B).

GSE3167 data modeling:

B) Normal Vs primary bladder tumor model:

*logit(P) = 2.265 + 13.24 * 276 218240_at (NKIRAS2) – 4.20 * 218373_at (AKTIP) – 1.55 * 203290_at (HLA-DQA1)*

Outcome Area under the curve (AUC) = 0.75 (95% CI: 0.343-0.933) (Figure 21B).

Best expression cut off: Based on the FPKM value of each gene, patients were classified into two groups and association between prognosis (survival) and gene expression (FPKM) was examined. The best expression cut-off refers the FPKM value that yields maximal difference with regard to survival between the two groups at the lowest log-rank P-value. Best expression cut-off was selected based on survival analysis .

To measure the effect of the sex on this model, we wanted to include this parameter to the model but the sex information of normal samples was not provided with data GSE13507 and therefore, we have used the intersection of arsenic exposed sex differentiated genes to find if any above gene is significantly different between male and female. We found none of those three gene were the part of 18 gene common between the sex specific arsenic exposed cancer gene. The human protein atlas data[58] shows two out of three genes i.e., NKIRAS2 (unfavorable), and AKTIP (favorable), are prognostic marker in renal cancer (Figure 21C).

Another study reported 19 ATO target genes associated with multiple cancer types (the most common association being pancreatic cancer)[60]. Six of these genes (AKT1, CCND1, CDKN2A, IKBKB, MAPK1, and MAPK3) were strongly associated and were used to find further mutation information. In addition, 20 ATO interacting genes were also related to other diseases such as hepatitis B, leukemia, and prostate cancer. And finally, CCND1 and MAPK1 were found to be prognostic factors in patients with pancreatic cancer. The genes responsible for metabolizing arsenic (AS3MT, GSTOs, and PNP) are of interest due to their variation in populations across different regions. More recently, phase I/II trials have been conducted in heavily pretreated patients with relapsed or refractory multiple myeloma shows Arsenic trioxide (ATO) is the most active, single agent in acute promyelocytic leukemia (multiple myeloma: types of blood cancers)[59].

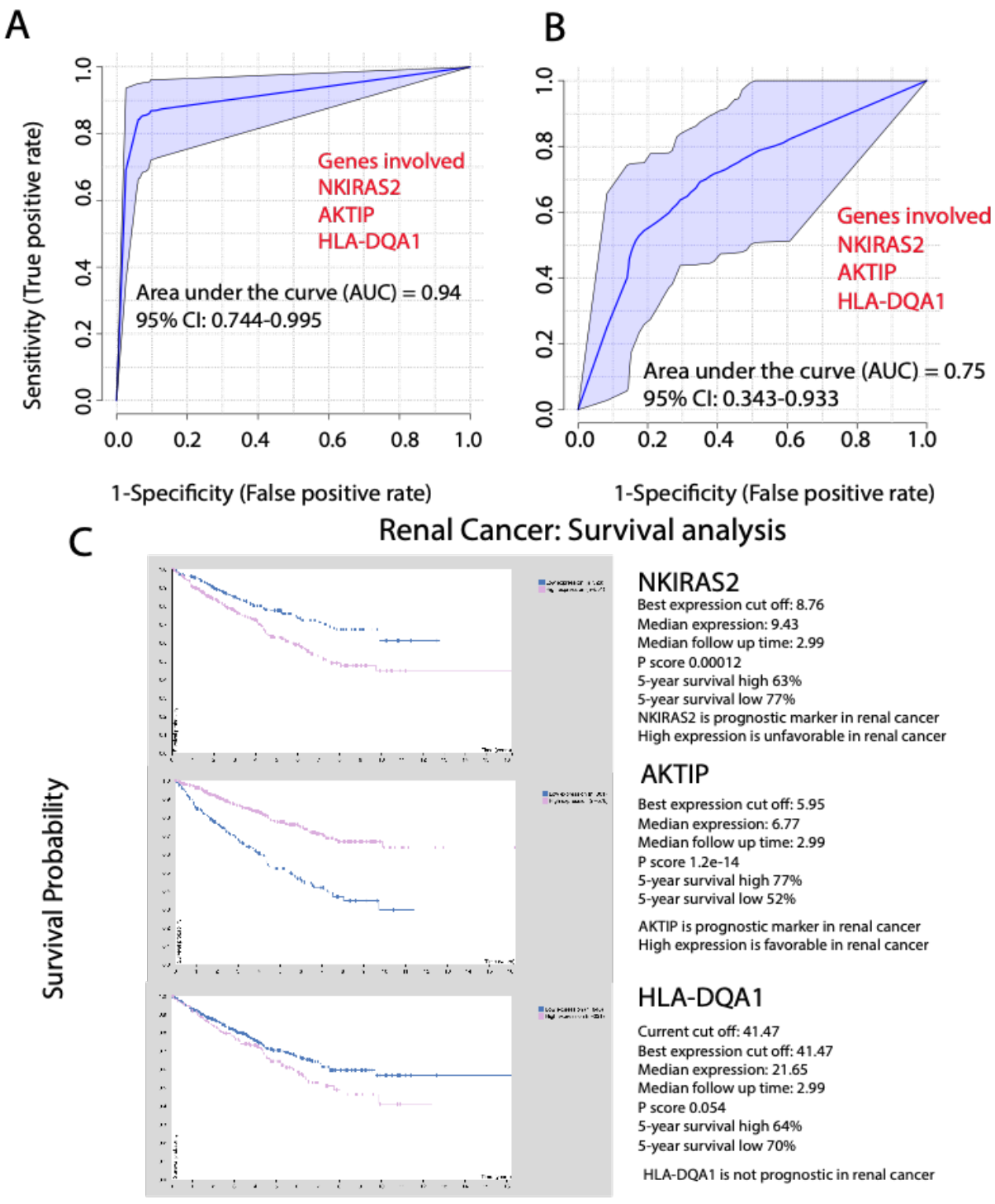


Figure 21: Bladder cancer prediction model.

Box 17: A-B) Plot of the ROC curve as an outcome of logistic regression prediction model in multivariate fashion, by AUC ROCs. The 95% confidence intervals (CI) are shown. A) Prediction outcome of primary tumor on GSE13507 dataset. B) Re-evaluation of prediction ability of 3 genes previously identified on GSE3167 dataset. C) Survival outcome of three genes using The Human Protein Atlas Data.

Another study suggested that ATO can be used as an effective alternative therapeutic for the treatment of retinoblastoma which is the most common intraocular cancer in children[60]. The study shows an antitumor activity of arsenic which mainly targets multiple pathways in malignant cells, resulting in the promotion of differentiation or in the induction of apoptosis, which would be very helpful to understand the molecular mechanism of arsenic-exposed cancer biology as a reverse engineering approach. Another study targeted gene associated with lung cancer and found four key genes that may affect lung cancer prognosis: MTIF2, ACOX1, CAV1, and MRPL17 [61]. This study also predicted Quinostatin as a reversal to As-induced lung cell malignancy. For urothelial cancer, WNT7B, SFRP1, DNAJB2, and ATF3 were reported as target genes with cantharidin predicted as a reversal drug[62]. Some genes captured in this study have been previously identified for an association with cancer to include CNR2 that is associated with bladder cancer cell growth and motility which is linked to the cannabinoid 2 receptor-mediated modifications[63], GRR34 knockdown was shown to impair proliferation and migration of HGC-27 gastric cancer cells [64], DDHD2 as a potential cancer marker in human urine [65], and BACE2 as a prognostic marker in cervical cancer [66]. The significance of MAPK signaling, Integrin-linked kinase, growth inhibitor family member 2, and NRF2-mediated oxidative stress response pathways provide an important linkage of involvement of oxidative stress and DNA damage after arsenic exposure in human which lead to carcinogenesis through dysregulation of these signaling pathway.

The key difference between these closely related studies and the current study is the process followed for capturing significant genes, where independent population data was used to generate a gene set, which was then compared to a reference dataset. In the initial analysis, two Asian populations exposed to arsenite were used to determine the common genes and pathways between the two populations based on sex and level of arsenic exposure among the 1,183 As-exposed genes. The 1,183 As-exposed genes were then correlated with the gene expression profiles of 4 multiple myeloma cell lines exposed over time to varying exposures of arsenic to generate common set of arsenic associated genes involved in cancer biology, which resulted in a set of overlapping genes and relevant pathways. These genes were then examined on the patients with bladder cancer to test the cancer association with the help of developing risk prediction model. For the first time, we developed a risk prediction model for bladder cancer using an innovative new method by combining genetic data of bladder cancer risk with genetic data of arsenic exposed cancer risk factors. Importantly, we validated our model in an independent group of patients to ensure the reliability of our risk

prediction, a vital step for clinical implementation.

The above process resulted in identifying 3 genes: NKIRAS2 (NFKB Inhibitor Interacting Ras Like 2); AKTIP (AKT-interacting protein); and HLA-DQA1 (Major histocompatibility complex, class II, DQ alpha 1), able to distinguish between normal urothelium and the primary urothelial carcinoma with a predictive ability of 94% using a pre-existing public patient dataset. The three genes have seen only limited study as regards arsenic exposure and urothelial cancer, with the majority of information available from literature searches with bladder cancer and urothelial cancer as key words, and from web-based resources such as the Human Protein Atlas (HPA), Gene Cards (GC), NCBI, and My Cancer Genome (MCG). In most cases, the Human Protein Atlas was an excellent source of information. None of the three genes were found to be prognostic for bladder cancer (HPA). The expression of the 3 genes in urothelial cancer range from moderate for NKIRAS2 (NCBI), variable for AKTIP (HPA), and variable for HLA-DQA1 as determined by an immune transcriptome analysis in bladder cancer[67]. Moreover, the same genes (NKIRAS2, AKTIP, and HLA-DQA1) were also found to make a prediction ability of 75% using a validation dataset. The predictive nature of these genes clearly supports additional study to define their roles in urothelial cancer independent of sex in general, and with exposure to arsenic in particular.

The studies leading up to the above prediction model also identified several interesting genes and pathways in the two populations exposed to arsenic. Three genes were identified that distinguished differences among all four scenarios of sex and arsenic exposure for the Data1 population and all six scenarios for the Data2 population. Two of these genes were noteworthy due to reports of their involvement in important biological processes. The XIST gene (X inactive specific transcript) is a non-coding RNA on the X chromosome that transcriptionally silences one of the pairs of X chromosomes for dosage equivalence between sexes. This gene is reported to be associated with several cancer types[68, 69] and has potential prognosis capabilities[70]. The expression of MALAT1 (metastasis-associated lung adenocarcinoma transcript 1) has also been associated with carcinogenesis and is a prognostic marker for lung cancer metastasis[71]. The XLOC_008276 (long intergenic non-protein coding RNA 278) is not strongly linked to any biological process. Previous research has also found several of these genes to be significance genes in cancer progression, such as, the high expression of XIST association with tumor progression and poor prognosis in bladder cancer patients [72], and high expression of MALAT1 as a possible

independent prognostic factor for overall survival in patients with bladder cancer [73]. Seven common genes and six common pathways were found to overlap between the 2 populations. The CNR2 protein, while not prognostic for urothelial cancer (HPA), has been shown to modify growth and motility of human urothelial cancer cell lines[63]. The gene and protein are expressed in approximately 33% of urothelial tumors. Four of the genes, BACE2, PRKY, CST2, PTGDR2 were reported to have no expression in urothelial cancer (HPA, GC). The remaining 2 genes, GPR34 and DDHD2, were expressed in urothelial cancer at 50% and 15%, respectively (HPA).

Nine genes were found to be overlapping when the two populations were assessed for sex-based changes. The data shows some cellular response patterns, such as cellular development, cellular movement, cell to cell signaling and interaction function significantly changed between male and female after arsenic exposure. Our interactive pathways show organismal injury, inflammatory and autoimmune diseases are significantly different between male and female. Several inflammatory diseases are associated with deregulated intracellular signal transduction pathways. This process results in pathogenic interactions between immune and stromal cells, which could induce a change in cell activation, proliferation, migratory capacity, and cell survival. Two genes, PRKY and PTGDR2, were also found overlap between the two populations where level of arsenic is also considered together with sex. PRKY biological functions are not yet well discovered although it is speculated to encode a ubiquitously expressed protein kinase that may have important signaling functions [74]. Whereas, PTGDR2, is upregulated in male lungs compared to females, is believed to be essential for the pro-inflammatory cytokines induction as well as asthma pathogenesis [75].

Two additional genes, JAK3 and ACRBP, were reported to have no expression in urothelial cancer (HPA, NCBI, MCC). The CPA3 gene is expressed in 90% of urothelial cancers (PHA) and can induce urothelial injury, but otherwise has not been studied in urothelial cancer. The remaining genes were of substantial interest for urothelial cancer and arsenic exposure. The FHL gene has been studied in a variety of cancers[74] and is prognostic for breast (favorable), renal (unfavorable) and liver (unfavorable) (HPA). The gene and protein have not been studied in urothelial cancer. The CD99L2 gene has been reported to be prognostic for urothelial cancer (unfavorable), pancreatic cancer (favorable) and lung cancer (favorable). The gene is expressed in 40% of urothelial cancers. The gene is reported to be active in a variety of tumors[75]. The SOCS3 gene is prognostic for renal cancer (unfavorable) and breast cancer (favorable) and reported to not be

expressed in urothelial cancer (HPA). However, studies have reported its expression in the T24 urothelial cancer cell line[76]. The CCL2 gene is expressed in 50% of urothelial cancers and has been implicated in the growth and metastasis of urothelial cancer[77, 78]. Additionally, the 18 genes are important sex dependent genomic markers which were differently express between male and female exposed to As and associated with likelihood of cancer. Most of the genes are prognostic marker of renal cancer, whereas some of them such as BTG2, CD24, OLFM4 and BACE2 are specific to females i.e., breast cancer, cervical cancer and MYB specific to males i.e., prostate cancer (Supp. Table 11).

Six genes were found to be overlapping when the two populations were assessed for level of arsenic exposure. The DDHD2 gene was also present in the above analysis of common genes between the populations in Data1 and Data2. Searching the Human Protein Atlas, ATP6V0D1 gene was prognostic for renal cancer (favorable) and pancreatic cancer (favorable), the PSPH gene was prognostic for liver cancer (unfavorable), breast cancer (unfavorable) and pancreatic cancer (favorable), and ZDHHC23 was prognostic for renal cancer (favorable), endometrial cancer (unfavorable) and thyroid cancer (unfavorable). Only the ZDHHC23 gene had confirmed expression in urothelial cancer (20%). Detailed studies in the literature for these genes in urothelial cancer were not found. The HS3ST1 gene was reported as a favorable prognostic marker for urothelial cancer, renal cancer, and endometrial cancer. Literature-based studies of these genes in urothelial cancer were not found.

2.5 LIMITATION

A major limitation in the current study was the lack of patient-level clinical-pathological information such as age, smoking status, disease history, etc. on the two populations exposed to As. Since datasets were developed on different platforms, not all the genes were present on different datasets. Therefore, to find the highest possible number of genes between those dataset, multiple testing correction was not performed which controls the Type I and Type II errors. However, the robustness of outcome was tested using different machine learning approaches such as the bladder cancer model was tested using MCCV. We did not find any genomic dataset which could provide the direct relationship between arsenic exposure and cancer

in humans, therefore, we utilized the best possible option to combine the iAs exposed humans and cell-line to establish the relationship.

2.6 CONCLUSION

This study identified significant genes and pathways of interest associated with arsenic-exposure in humans as well as their linkage with Myeloma cancer cell lines. Oxidative stress in terms of identified genes and associated pathways shows as one of the major components associated with disease development after exposure of arsenic. To test the prediction power of those genes, we developed a regression model for urothelial carcinoma that defined a set of 3 genes: NKIRAS2; AKTIP; and HLA-DQA1; which provides the likelihood of development of primary urothelial carcinoma with same estimation for male and female.

CHAPTER 3
Arsenite Exposure to Human RPCs (HRTPT) Produces a Reversible Epithelial Mesenchymal Transition (EMT): In-vitro and In-silico study

3.1 ABBREVIATIONS

PCA	Principal Component Analysis
DEG	Differentially Expressed Gene
iAs	inorganic Arsenite
P0	Passage Zero Control
P3	Passage 3 Arsenic Exposed
P8	Passage 8 Arsenic Exposed
P10	Passage 10 Arsenic Exposed
P3vsCtrl	Describes a statistical comparison of each gene tested between P3 and P0 conditions samples
P3vsCtrl_Up	Only up-regulated genes identified between P3 versus Control
P3vsCtrl_Dn	Only down-regulated genes identified between P3 versus Control
P8vsCtrl	Describes a statistical comparison of each gene tested between P8 and P0 conditions samples
P8vsCtrl_Up	Only up-regulated genes identified between P8 versus Control
P8vsCtrl_Dn	Only down-regulated genes identified between P8 versus Control
P10vsCtrl	Describes a statistical comparison of each gene tested between P10 and P0 conditions samples
P10vsCtrl_Up	Only up-regulated genes identified between P10 versus Control
P10vsCtrl_Dn	Only down-regulated genes identified between P10 versus Control
P2	Passage 2 Arsenic Recovered

P11	Passage 11 Arsenic Recovered
P2vsCtrl	Describes a statistical comparison of each gene tested between P2 and P0 conditions samples
P2vsCtrl_Up	Only up-regulated genes identified between P2 versus Control
P2vsCtrl_Dn	Only down-regulated genes identified between P2 versus Control
P11vsCtrl	Describes a statistical comparison of each gene tested between P11 and P0 conditions samples
P11vsCtrl_Up	Only up-regulated genes identified between P11 versus Control
P11vsCtrl_Dn	Only down-regulated genes identified between P11 versus Control
iAs+	Pooled samples of exposed to iAs (i.e., combination of P3, P8, P10)
iAs-	Pooled samples of after recovery of iAs (i.e., combination of P2, P11)
iAs+vs iAs-	Describes a statistical comparison of each gene tested between iAs+ and iAs- samples

3.2 INTRODUCTION

The tubular epithelium of the human kidney has the capacity to regenerate, repair, and re-epithelialize in response to injury by various insults. In the human kidney, a population of resident cells with progenitor characteristics, identified by the PROM1 stem cell marker, were localized to the Bowman's capsule, proximal tubules, and the inner medullary papilla.[1-3] The number of cortical PROM1-expressing tubular cells increased in patients with acute renal injury.[4] Further studies have shown renal epithelial cells co-expressing PROM1 and CD24 have the capacity to participate in the regeneration of renal tubule cells.[5-9] Cultures of human renal epithelial cells that co-express PROM1 and CD24 also display features expected of RPCs; such as, spheroid formation, ability to undergo adipogenic, neurogenic, osteogenic differentiation, and form tubule-like structures on Matrigel. These cells provided a potential model to define the mechanisms underlying the progenitor cell's ability to participate in renal epithelial cell regeneration. However, these cultures were shown to possess two cell types, one co-expressing PROM1 and CD24 and another expressing only CD24.[10] Subsequently, our laboratory identified an immortalized human

renal proximal tubule epithelial cell line, RPTEC/TERT1, that also display the two cell populations, one that cell sorting was used to isolate two new immortalized cell lines, one HRTPT that co-expresses PROM1 and CD24, and another, HRECT24T that expresses CD24 and no PROM1.[11] The HRTPT cells expressed the features defined for RPCs while the HRECT24T cells displayed no features of RPCs.[11-13] The HRTPT cells provide a human cell culture model to determine if PROM1/CD24 co-expressing RPCs are susceptible to nephrotoxic agents. To the author's knowledge, this is an unexplored area as regards RPCs.

Exposure of the HRTPR cells to inorganic arsenic (iAs) was chosen to test this hypothesis. Inorganic arsenic has an extensive distribution in the environment[14-16]. The kidney is the most susceptible of all organ systems to iAs exposure.[17, 18] There is evidence that exposure to iAs is associated with renal disease. A study of 6,093 participants from arseniasis-endemic areas in northeastern Taiwan showed a temporal relationship between arsenic concentrations ≥ 10 mg/L in drinking water and CKD (chronic kidney disease).[19] The study also demonstrated a dose-dependent association between well-water arsenic concentration and kidney diseases. Other studies have also shown an association of iAs exposure with alterations of renal function and disease.[20-22] Thus, there is evidence from population studies that exposure to iAs is association with renal disease, however studies defining the concentration of iAs within the human kidney and specific cells of the nephron are rare. Accumulation is possible due to the presence of metallothionein (MT), a small molecular weight protein that is known to bind and sequesters iAs within cells.[23-25]

3.3 MATERIALS AND METHODS

3.3.1 Study Design

A flowchart of study design is shown in visual abstract (Figure 22).

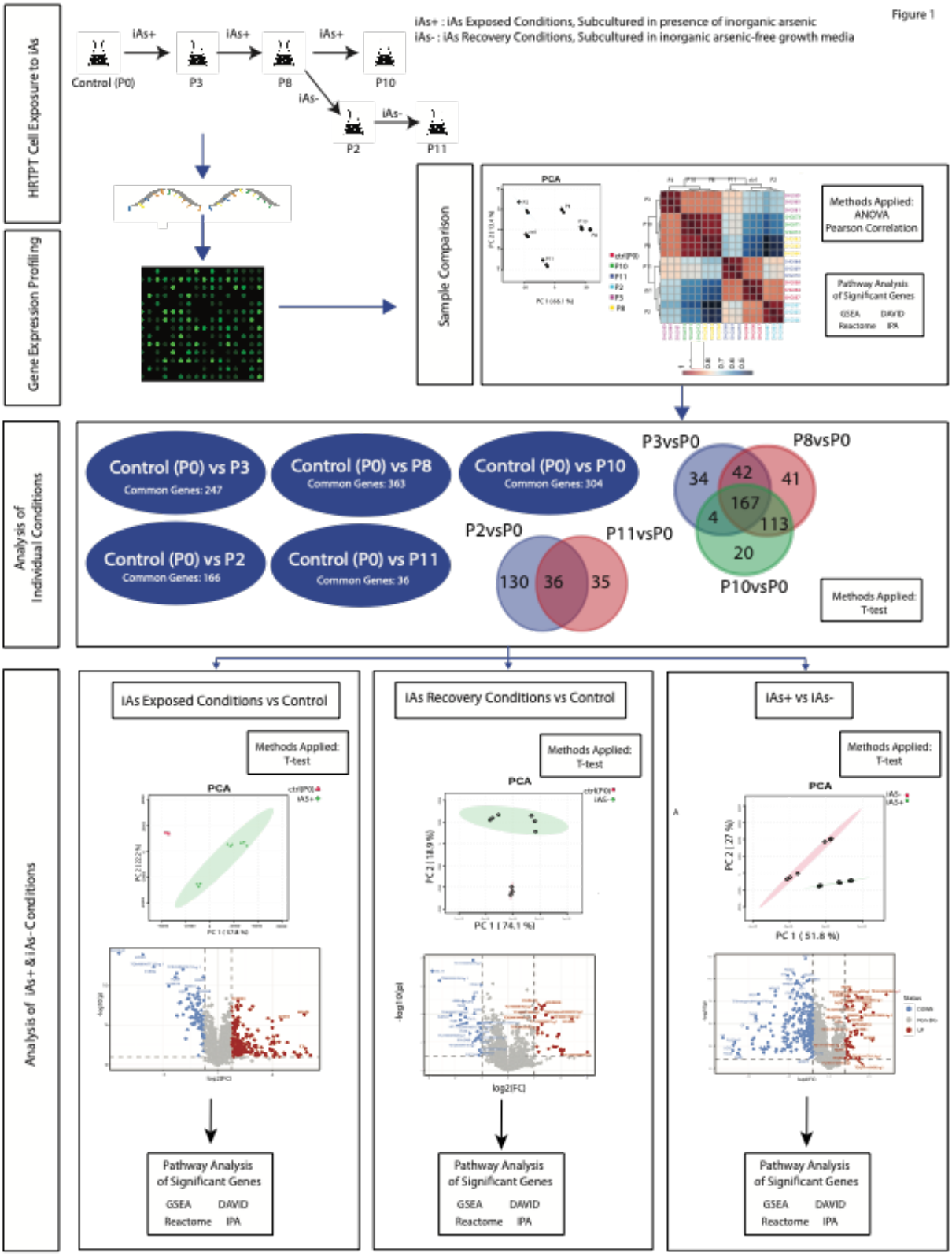


Figure 22 Flowchart of study design

3.3.2 Cell Culture

The isolation and serum-free culture conditions for the HRTPT cells has been previously described.[10, 11] Confluent cultures of HRTPT cells were exposed to 4.5 μM iAs for 24 hrs and then sub-cultured at a 1:3 ratio in the continued presence of iAs until confluent. Following confluence, the cells were serially sub cultured again in the presence of iAs until confluent. This was repeated for 10 serial passages. Additional cultures of iAs exposed cells at passage 8 were sub-cultured into iAs free growth media and continued in iAs free media for 11 additional passages.

3.3.3 Microarray Gene Expression

The gene expression profile was determined using the Clariom D Human Microarray (platform ID: GPL23126) on triplicate samples of control HRTPT cells (P0) and HRTPT cells exposed to 4.5 μM iAs for 3, 8, and 10 serial passages (named as P3, P8, P10) and after recovery (named as P2, P11) (GSE215904). Each sample gone through the quality control processing before downstream analysis. The total 138745 probes were analyzed for each sample in different conditions. Confluent cultures were used for the isolation of RNA.

3.3.4 Individual Gene mRNA and Protein Expression

The mRNA and protein expression of individual genes was determined using RT qPCR, western blotting, and flow cytometry as described previously.[10, 11]

3.3.5 Statistical Analysis

Statistical significance of genes was calculated by running t-tests[26] between pair of groups such as all subset of passages with respect to P0, iAs+ (combination of P3, P8 and P10 passages) with respect to P0, iAs- (combination of P2, P11 passages) with respect to P0 and iAs+ versus iAs-. When compared more than two groups, one-way ANOVA (Analysis of Variance) was performed. Scattered volcano plots were used to show the statistically significance genes with p-value <0.05 and fold-change (FC) greater than two in both direction (up or down regulation). The foldchange for each gene was calculated based upon antilog-expression value between two phenotypic

conditions. Most of the genes provided in different tables were selected based upon p-value < 0.05 with or without fold change (FC < 0.5 or FC > 2). A principal component analysis (PCA) were performed to test the distribution of replicates of passage samples and Pearson correlation were used to find the relationship between the different passage conditions as well as genes[27, 28]. The first two components of PCA i.e., PC1 and PC2 were used to explain the amount of the variance in the data according to phenotypic condition(s). Venn diagrams were used to demonstrate the union and intersection of genes in different conditions. Entire analysis was performed using R/Bioconductor.

3.3.6 Pathway Analysis

Different significant gene list were examined using the commercially available pathway tools such as QIAGEN Ingenuity Pathway Analysis (IPA),[29] as well as freely available Gene Set Enrichment Analysis (GSEA),[30] Reactome,[31] Panther,[32] and DAVID[33, 34] software databases.

3.3.7 Gene Set Enrichment Analysis

Gene set enrichment analysis was performed on different gene sets identified through t-tests or ANOVA using p < 0.05 +/- FC < 0.5 or > 2. In some cases, some genes with fold change > 4-fold were also included, regardless of p-value to sets the relevance at functional level. Probes were ranked according to their P-value and/or log2 fold change and all the probes without gene symbols were excluded as they cannot map with the different pathway databases. Ranked lists were used in the Gene Set Enrichment Analysis pre-ranked software for minimum gene size of 5 with a maximum gene set size of 500[30]. The MSigDB pathway database was used to identify enriched gene sets including Hallmark, C2, and C3[35].

3.4 RESULTS

3.4.1 EMT as a Function of Exposure of HRTPT Cells to iAs

Examination of the HRTPT cells exposed to iAs by light microscopy demonstrated a change in cell morphology at passage 3 when compared to cells unexposed to iAs (Figure 23A, B). When compared to control, the iAs exposed cells were less closely packed, more disorganized, and had lost the ability to form “domes”. Domes are raised areas of the monolayer due to fluid accumulation and are a manifestation of vectorial active ion transport.[10] This change in morphology was evident for at least 7 more serial passages (Figure 23C).

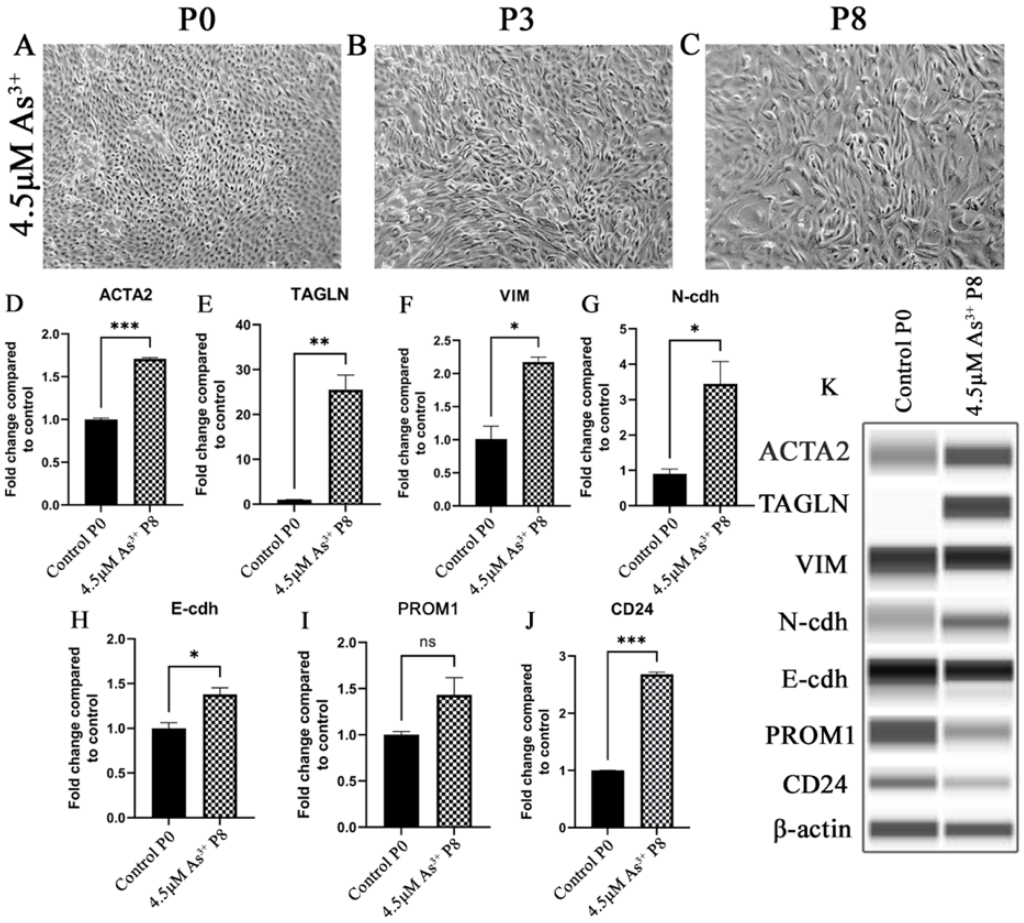


Figure 23: HRTPT cells exposed to iAs under light microscopy

Box 18: HRTPT cells exposed to iAs under light microscopy for A) P0 B) P3 C) P8. Expression of P8 for epithelial-to-mesenchymal transition genes D) ACTA2 E) TAGLN F) VIM G) N-cdh H) E-cdh and expression of I) CD133 J) CD24. K) Western blot results confirmed protein level expression. Scale bar = 50 µm and Magnification x10

The change in morphology suggested that the cells could have undergone an epithelial-to-mesenchymal transition (EMT). Further evidence for the possibility of EMT was provided at passage 8 by increased expression of ACTA2, TAGLN, VIM, and CDH2 and a modest decrease in expression of CDH1, (Figure 23D, E, F, G, H). The co-expression of PROM1 and CD24 mRNA was retained by the cells exposed to iAs but the expression was clearly reduced (Figure 23I, J, K).

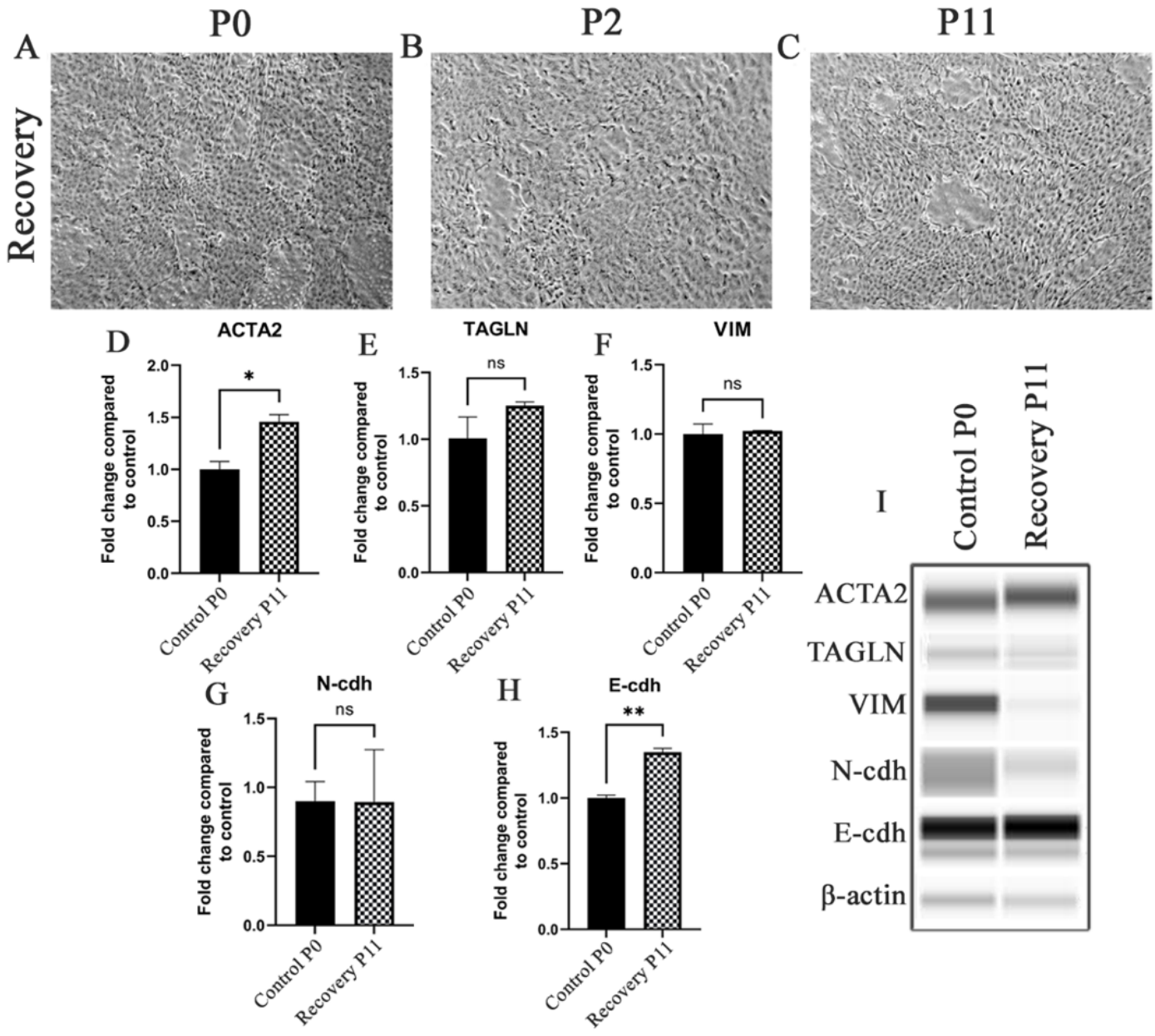


Figure 24: HRTPT cells exposed to iAs in recovery under light microscopy

Box 19: HRTPT cells exposed to iAs in recovery under light microscopy including A) P0 B) P2 C) P11. Expression of D) ACTA2 E) TAGLN F) VIM G) N-cdh H) E-cdh for recovery P2 and recovery P11. I) Western blot showed protein levels in P2 and P11. Scale bar = 50 μm and Magnification x10

The change in morphology and gene expression by the iAs-exposed cells suggested global gene expression technology might assist in providing additional information if iAs was inducing EMT or a related mesenchymal alteration in the HRTPT cells. Lower concentrations of iAs (1.0 and 2.0 μM) elicited a similar shift in morphology and increased expression, but at an extended number of serial passages (Supplementary Figure S1).

The HRTPT cells exposed to iAs were assessed for their morphology and expression of the above genes when iAs was removed from the growth media. Light microscopic examination showed that by the 2nd passage the iAs- cells displayed a morphology similar to the HRTPT controls (Figure 24A, B) and by the 11th passage they were indistinguishable from the control (Figure 24C). The iAs- cells regained dome formation at both P2 and P11 following iAs removal from the growth media. The expression of the ACTA2, TAGLN, VIM, N-cdh and E-cdh genes were also assessed and all except VIM, which was absent from the iAs-cells, showed a trend to return to control values (Figure 24D-H). The change in morphology and gene expression after removal of iAs suggested that the cells might have undergone a mesenchymal-to-epithelial transition (MET). These results presented the opportunity to examine the global gene expression profile of a toxin exposed renal progenitor cell that shows evidence of undergoing EMT and, upon toxin removal, the ability to undergo MET and return to an epithelial morphology. The ability of the iAs- HRTPT cells to dome is strong evidence of epithelial differentiation.

3.4.2 Global gene expression and Impacted Pathway Analysis

The above morphology and gene expression changes suggested that exposure of HRTPT cells to iAs induced EMT, and when iAs was removed, MET back to the morphology and gene expression of the control HRTPT cells. Global gene expression was employed to further explore the ability of HRTPT cells to undergo EMT and MET as a function of exposure to iAs. In this section, triplicates of all the samples i.e., control cells (P0), iAs exposed cells at P3, P8 and P10 and P8 cells and iAs recovered P2, P11 included to determine the global distribution and to identify all possible relation among different conditions. The first two components of the PCA plot, PC1 and PC2, carry 66.1% and 13.4% of the variance of the data and the P0 is far removed from P10 and P8 as compared to P2 and P11 (Figure 25A). Correlation analysis supported this relationship and demonstrated that P2 samples were most closely related to the P0 (Figure 25C).

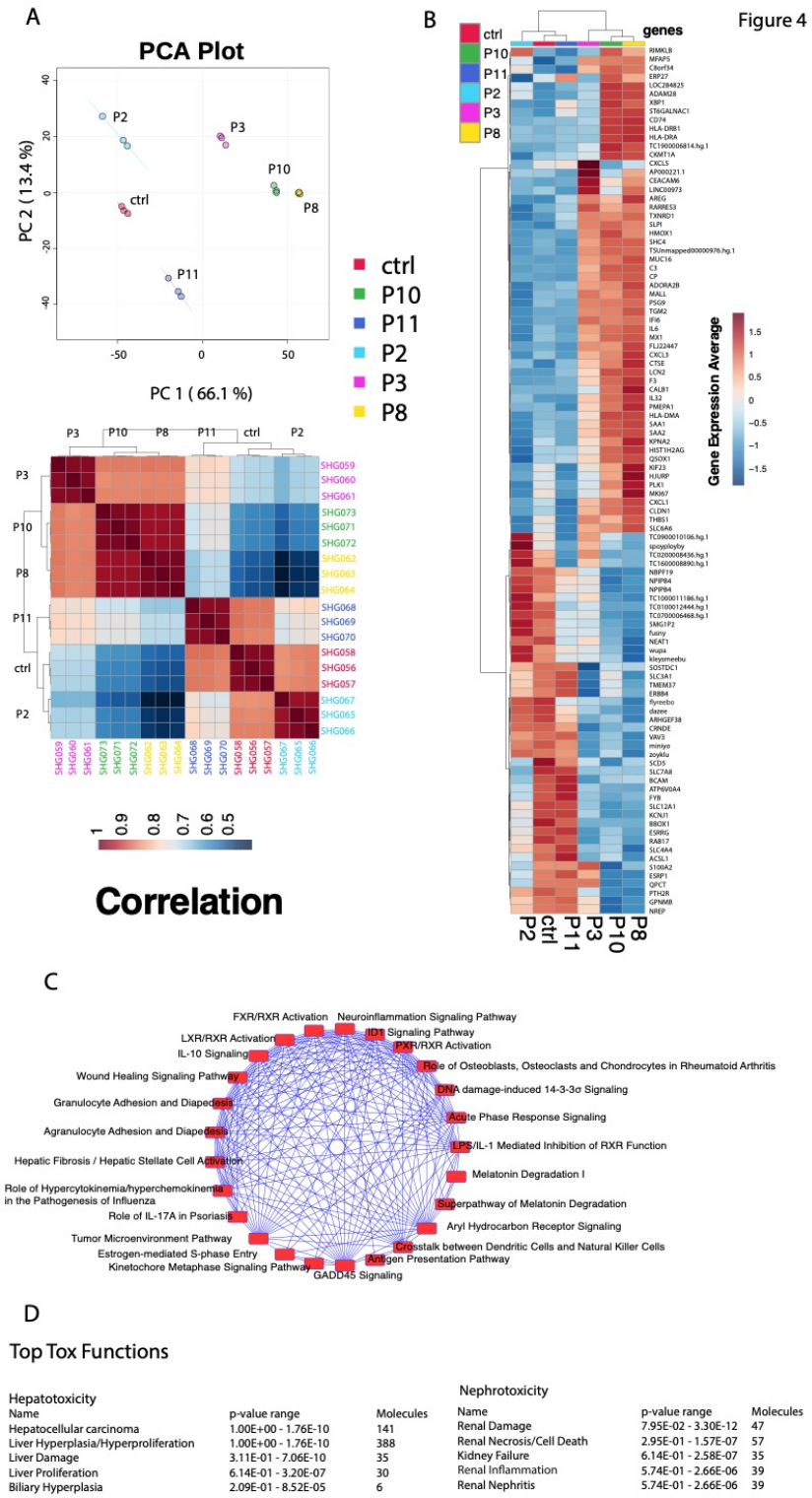


Figure 4

Figure 25: Sample Comaprision and Pathway Analysis

Box 20: A) Principal component analysis between the different passage conditions. B) Hierarchical cluster analysis and heatmap of correlation. C) Heatmap of gene expression averages for the top 100 differentially expressed genes identified through ANOVA analysis for the different passage conditions. D) Top hepatotoxicity and nephrotoxicity functions from Ingenuity Pathway Analysis.

Differential gene expression analysis using Post hoc test with ANOVA identified 2478 probes varieties across all possible conditions (Supplementary Table 1). The hierarchical clustering of the top 100 differentially expressed genes was determined from these 2,478 probes (Figure 25B). IPA was performed using the 2478 probe varieties across all possible conditions which identified hepatocellular carcinoma as the top hepatotoxicity function and renal damage as the top nephrotoxicity function (Figure 25D). GSEA analysis on 2478 probes identified thirty-six (36) upregulated pathways, and 368 downregulated pathways with nominalized p-value < 0.05 (Supplementary Table 2).

3.4.3 Gene Expression of HRTPT Cells Exposed to iAs.

Global gene expression was performed between P0 versus P3, P8 and P10 (each group separately) cells and identified 247, 363, and 304 differentially expressed genes, respectively (Supplementary Table 3). For the 3 sets of differentially expressed genes, 106/247 genes were down-regulated and 141/247 up-regulated; 118/363 were down-regulated and 245/363 up-regulated; and 111/304 genes were down-regulated and 193/304 were up-regulated in expression (Supplementary Table 3). An intersection analysis of the 3 gene sets for commonality, identified 167 common genes with 91 genes being up- and 76 genes down-regulated (Supplementary Table 3, Figure 26A).

When analyze all the exposed samples together i.e., iAs+ with respect to the P0, the first 2 components of PCA shows 57.8% and 22.2% of the variance, respectively (Figure 26B) among the phenotypes. The variation between P8 and P10 is very narrow as compare to P3 samples. A total of 280 probes (234 gene symbols) were found differentially expressed between these conditions (Supplementary Table 4) and a subset of the top 25 genes was determined from these differentially expressed genes (Figure 26C). A volcano plot shows the most significant upregulated and downregulated differentially expressed genes (Figure 26D) between iAs+ and P0.

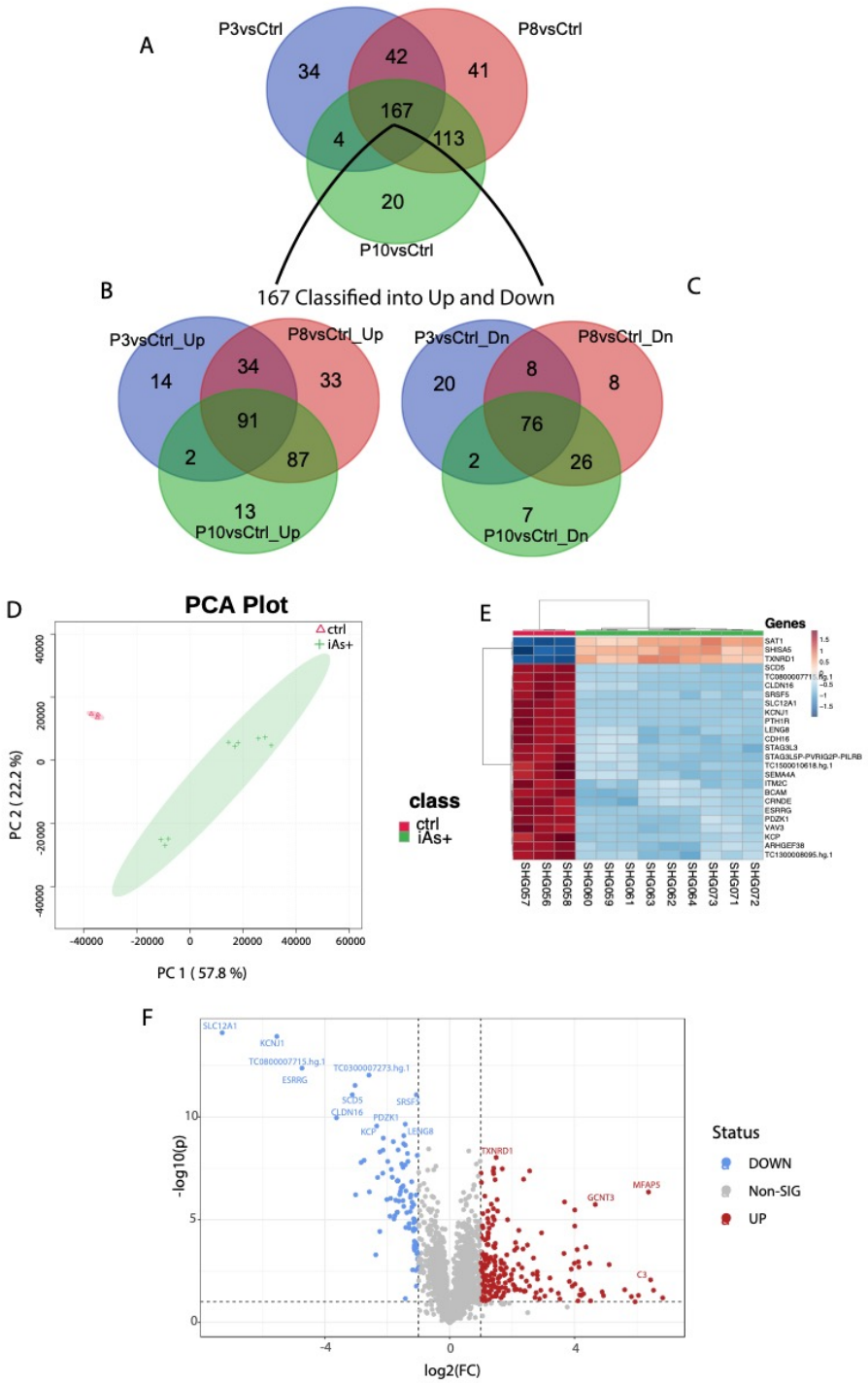


Figure 26: Siginficant genes: iAs+ VS Ctrl

Box 21: A) Common genes among the differentially expressed gene set for P3 (246 gene set), P8 (363 gene set), and P10 (304 gene set) when compared to the control. B) Common genes between the 167 gene set and the up-regulated genes from P3, P8, and P10 genes compared to the control. C) Common genes between the 167 gene set and the down-regulated genes from P3, P8, and P10. D) Principal component analysis of iAs+ cells with the control. E) Hierarchical clustering of the top 25 differentially expressed genes for iAs+ cells with the control. F) Significantly upregulated and downregulated differentially expressed genes based on the iAs+ cells with the control.

3.4.4 Pathway Analysis of HRTPT Cells Exposed to iAs.

GSEA was performed on the 167 gene-set selected as an interaction of significant genes identified from P0 versus each of the iAs exposed cells passage (i.e., P3, P8, P10). We found 37 upregulated, and 129 downregulated pathways that had a nominalized p-value < 0.05 (Supplemental Table 5). Again, 167 common gene set was also analyzed using Reactome and we found the down-regulate pathways were associated with signaling pathways, especially those associated with FGF (Supplemental Table 6). Other associated pathways such as PI3K, the RAF/MAP kinase cascade, and ERBB were also identified using Reactome. The analysis of the 76 down-regulated gene set using Reactome also identified IGF signaling as a pathway. The prominent pathways association with 91 up-regulated gene sets was interleukin signaling (IL4, IL10, IL13, IL18) and chemokine receptors (Supplemental Table 6). An analysis of the 167 gene set by the Panther Classification System also identified signaling pathways as a prominent component (Supplemental Table 6).

In addition, we performed pathway analysis on total of 280 probes (234 gene-symbols, Supplemental Table 4) were found differentially expressed between P0 and iAs+ conditions ($p<0.05$ and FC <0.5 0r >2) (Supplementary Table 4). Of the 234 genes, we found 151 were upregulated pathways and 34 downregulated pathways with nominalized p-value < 0.05 (Supplemental Table 7). One of the upregulated pathways was the Hallmark Epithelial Mesenchymal Transition, a gene set with genes defining the epithelial-mesenchymal transition [36]. IPA analysis on above 234 genes identified other significant pathways associated with exposure to iAs (Supplemental Table 8). Total 533 pathways were demonstrated in this list out of which around 200 were under the $p<0.05$. EIF2, Ferroptosis and mTOR signaling were the top in the list.

3.4.5. Progenitor Cell Properties of HRTPT Cells After Recovery from Exposure to iAs

The HRTPT cells recovered from iAs exposure were shown to retain the ability to differentiate, form nephrospheres, and express PROM1 and CD24 in over 94% of the cells (Figure 27A-F). One noteworthy alteration in tubular differentiation was that the recovered cells demonstrated no significant change in the expression of aquaporin from control cells, but did exhibit a large increase

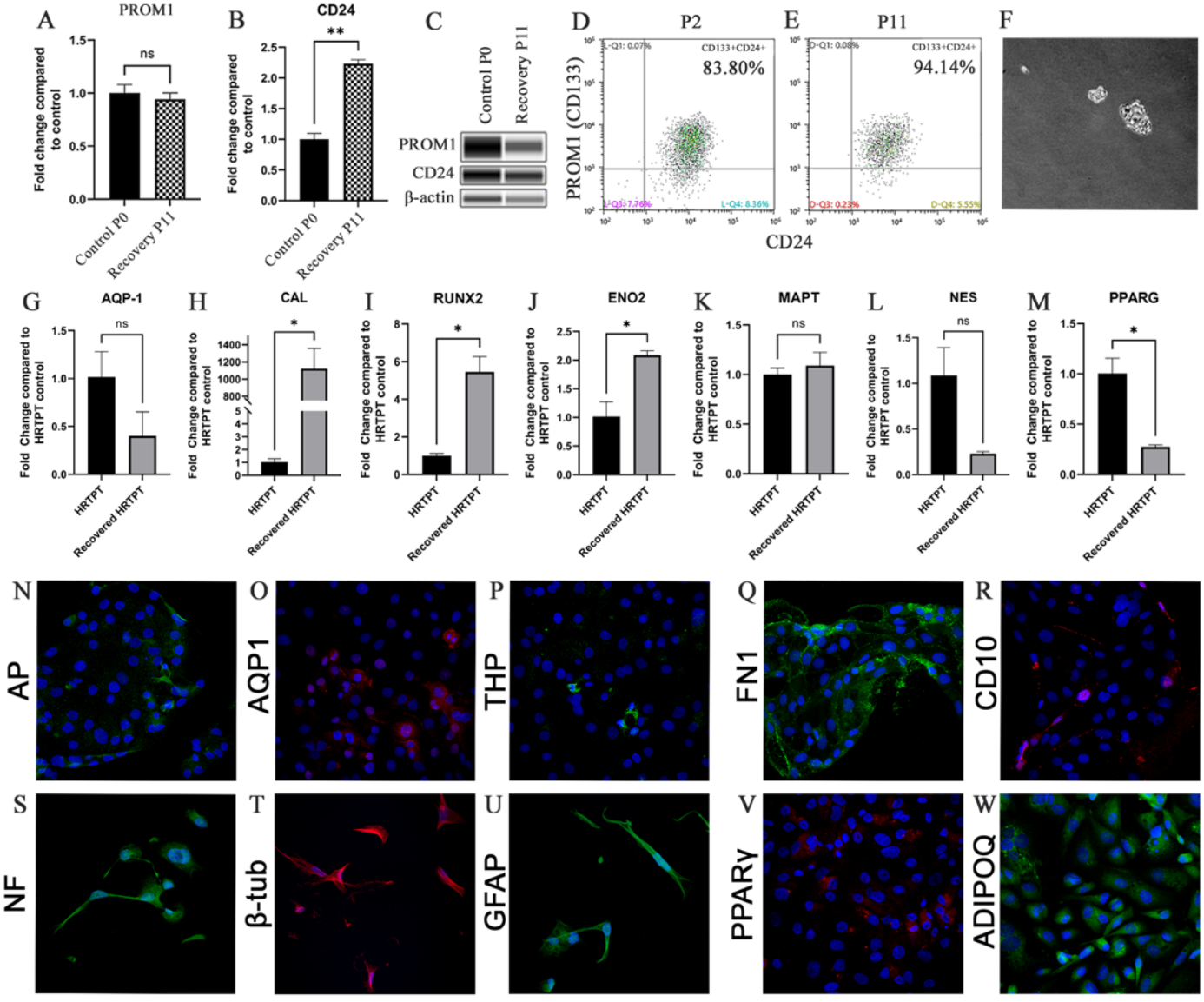


Figure 27: Gene expression for iAs-

*Box 22: A) PROM1 B) CD24 expression for P0C and recovery passages P11AR. C) Western blot of PROM1, CD24, and b-actin for P0 and recovery passages P11. D & E) Flow cytometry expression of PROM1 and CD24. F) Nephrospheres in HRTPT cells. G) AQP-1 H) CAL I) RUNX2 J) ENO2 K) MAPT L) NES M) PPARG gene expression. N) AP O) AQP1 P) THP Q) FN1 R) CD10 S) NF T) b-tub U) GFAP V) PPARg W) ADIPOQ confocal microscopy, scale bar = 21.16 µm, magnification x400. ns indicates no significance, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$.*

in the expression of calbindin (Figure 27G, H). The osteogenic gene RUNX2, neurogenic gene ENO2 showed significant increase (Figure 27I, J); while neurogenic genes MAPT and NES showed no significant change in expression (Figure 27K, L) and adipogenic gene, PPARG showed a decrease in expression when compared to the control HRTPT cells (Figure 27M). The confocal images show expression of AP, AQP1 and THP as tubulogenic marker (Figure 27N-P); FN1 and CD10 as osteogenic markers (Figure 27Q, R); NF, β-tub and GFAP as neurogenic marker (Figure 27S-U); and PPARγ and ADIPOQ as adipogenic markers (Figure 27V, W) expression in recovered cells.

3.4.6 Gene Expression analysis of HRTPT Cells after Recovery from iAs Exposure

The HRTPT cells were assessed for their gene expression at the P2 and P11 passage following removal of iAs. A comparison between the control HRTPT (P0) cells and the P2 cells demonstrated that 166 genes were differentially expressed between the two groups, with 30 upregulated and 136 down regulated genes (Supplemental Table 9). A similar comparison between P0 and P11 cells demonstrated that 71 genes were differentially expressed with 39 up regulated and 32 down regulated (Supplemental Table 9). The common genes between the two gene sets were determined and 36 genes were common (Supplemental Table 9, Figure 28A), with 22 up and 11 down regulated genes (Figure 28B, C). Three genes were found to differ in directionality between the P2 and P11, IGFBP3, NMNAT2, and CYFIP2 (Figure 28D).

PCA found significant separation between the two recovered samples with PC1-74.1%, as compared to control with PC2-18.9% (Figure 28E). Differential gene expression identified 77 probes significantly different between the recovered cells versus control (Table-10). The heat map shows the top 25 differentially expressed genes along with a volcano plot of the results (Figure 28F, G). The gene expression analysis of control versus iAs recovered samples (iAs-) identified 426 probes that were differently expressed with $p<0.05$ (Supplementary Table 11).

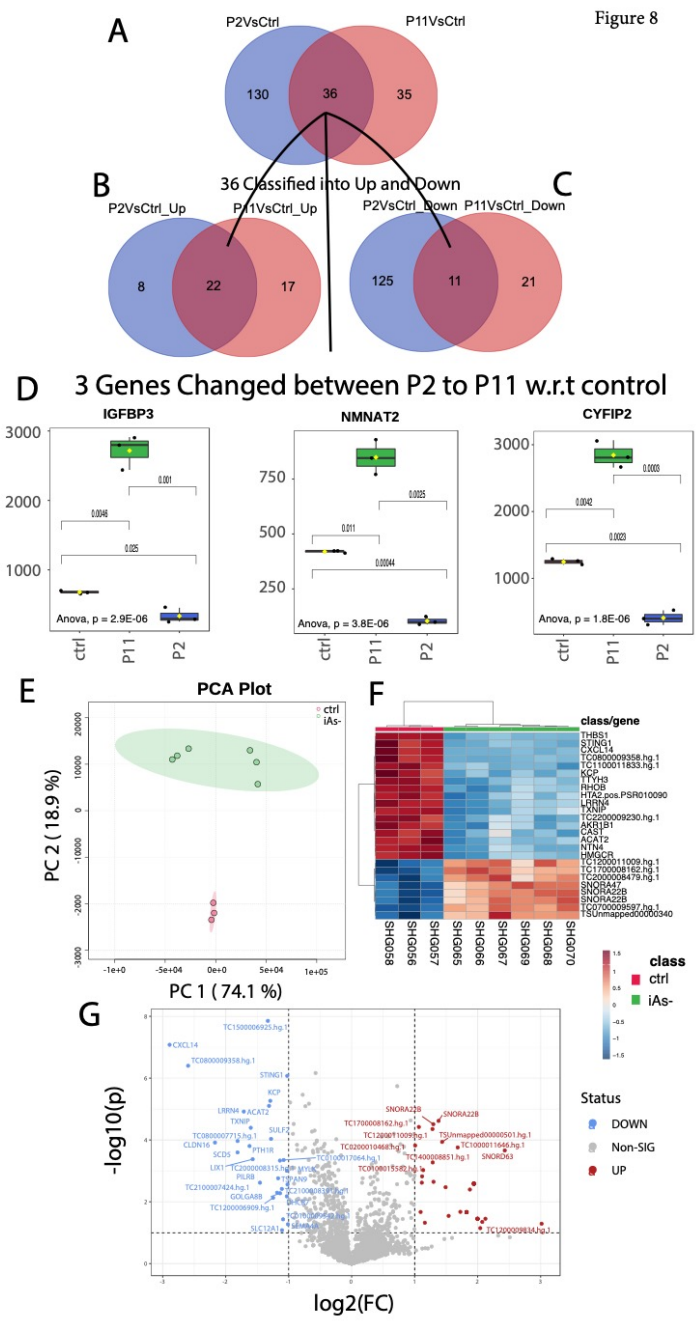


Figure 28: Siginficant genes: iAs- VS Ctrl

Box 23: A) Common genes among the differentially expressed gene set for P11 and P2 when compared to the control. B) Common genes between the 36 gene set and the upregulated genes from P11 and P2. C) Common genes between the 36 gene set and the downregulated genes from P11 and P2. D) Log2 Foldchanges of genes with differing directions between P2 and the control, and P11 and the control. E) Principal component analysis of iAs- cells with the control. F) Hierarchical clustering of the top 25 differentially expressed genes for iAs exposed cells in recovery with the control. G) Significantly upregulated and downregulated differentially expressed genes based on the iAs exposed cells in recovery with the control.

3.4.7 Pathway Analysis of HRTPT Cells Following Recovery from iAs Exposure

36 common genes were examined using Reactome and Panther databases. The Reactome database identified elastic fibres and pathways involved in cell cycle and p53 interactions (Supplementary Table 12). The Panther database identified mostly signaling and regulatory processes. GSEA on 49 genes identified as differentially expressed (Supplementary Table 10) between the control and iAs-. Only two down-regulated pathways were identified at nominalized p-value < 0.05 (Supplementary Table 13). IPA on differentially expressed genes between iAs- vs control (Supplementary Table 10), confirm p53 signaling as a canonical pathway (Figure 29, Supplementary Table 14).

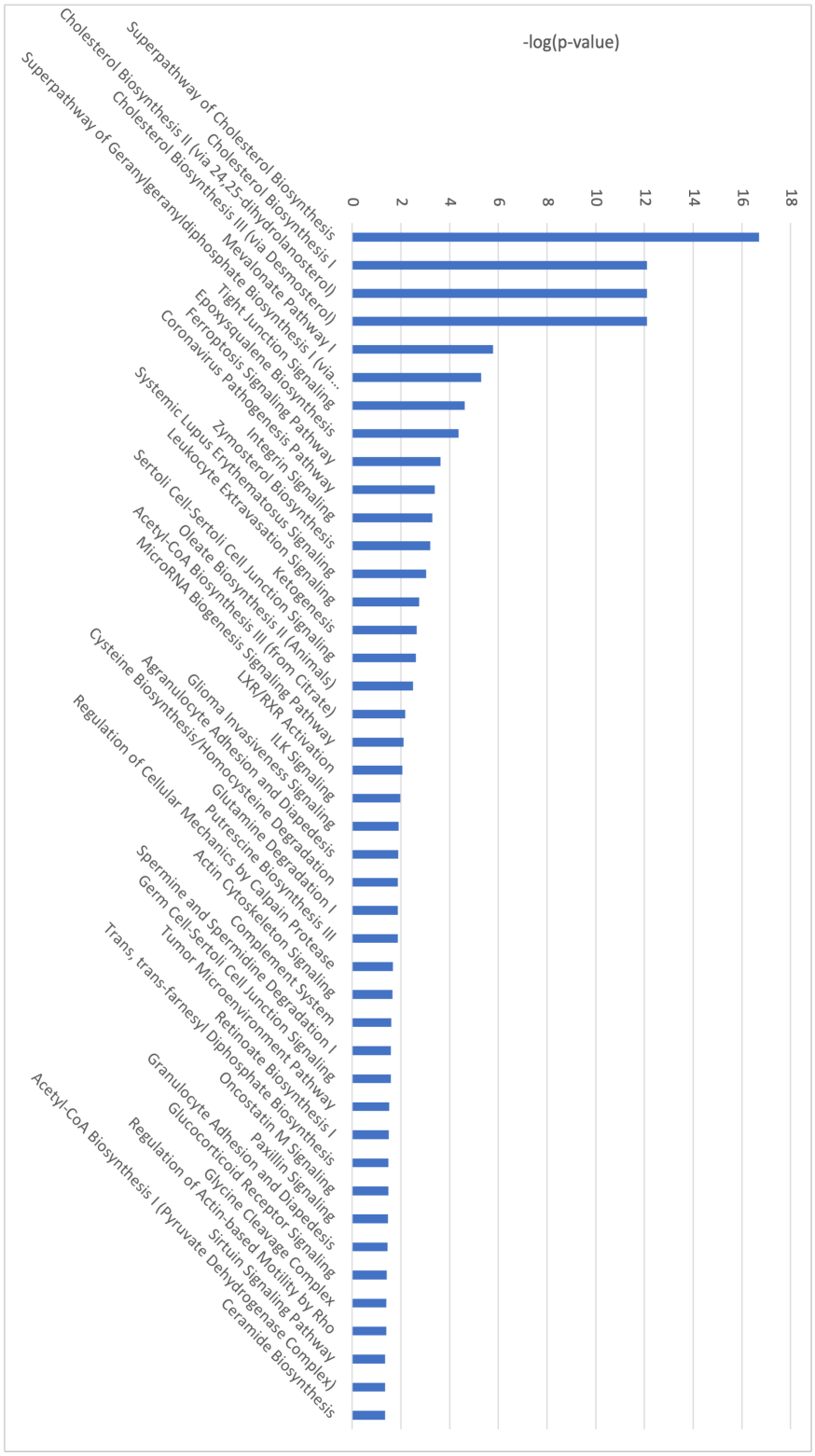


Figure 29 : Ingenuity Pathway Analysis identified canonical pathways for Ctrl vs iAs- conditions.

3.4.8 Comparison of iAs Exposed HRTPT Cells and HRTPT Cells Following Recovery from iAs Exposure

An intersection of differentially expressed genes between iAs exposed HRTPT cells and HRTPT cells following recovery from iAs exposure found 9-genes of interest (Figure 30A). These genes included CLDN16, CTSE, PTH1R, CYFIP2, SCD5, LIX1, MFAP5, KCP, and SH2D1B. PC1 and PC2 were showing 51.8% and 27% variance of the data, respectively (Figure 30B). Total of 305 differentially expressed probes (280 gene-symbols) were identified between P0, P3, P8, P10 and P0, P2, P11 with 41 down regulated and 264 up regulated genes (Figure 30D, Supplementary Table 15). GSEA on the 280 genes (Supplementary Table 15) found 42 downregulated pathways, and 157 upregulated pathways with nominal p-value < 0.05 (Supplementary Table 16). IPA identified FGFR as a significant upstream regulator (Figure 31, Supplementary Table 17).

3.5 DISCUSSION

The HRTPT cell line provides an opportunity to determine how a human renal progenitor cell responds to a nephrotoxic agent and its subsequent removal. The hypothesis being that a nephrotoxin might alter the regenerative capacity of the RPCs to repair tubular damage. The results demonstrated that the HRTPT cells exposed to 4.5 μM iAs displayed those characteristics of a cell undergoing EMT as noted by a change to a mesenchymal morphology and an increase in expression of mesenchymal markers such as ACTA2 and TAGLN. It was also shown that the alteration in morphology and increased expression of smooth muscle actin alpha 2 and transgelin also occurred at lower levels of iAs exposure (1.0 and 2.0μM), albeit at much longer times of exposure, providing evidence that results found with 4.5 μM iAs would translate to lower levels of exposure. Global gene expression was used to further analyze the EMT response when the HRTPT cells were exposed to iAs. GSEA of the common genes expressed from the comparison of P3, P8, and P10 compared to control identified the Hallmark Epithelial Mesenchymal Transition from the MSigDB as an upregulated pathway.[36] Global differently expressed gene set was examined to determine if the iAs treated cells were transitioned to myoepithelial (keratin expressing) or myofibroblast-like (vim expressing) cells. The common gene set did not show the differential expression of any keratin genes or the vimentin gene. To further explore this finding,

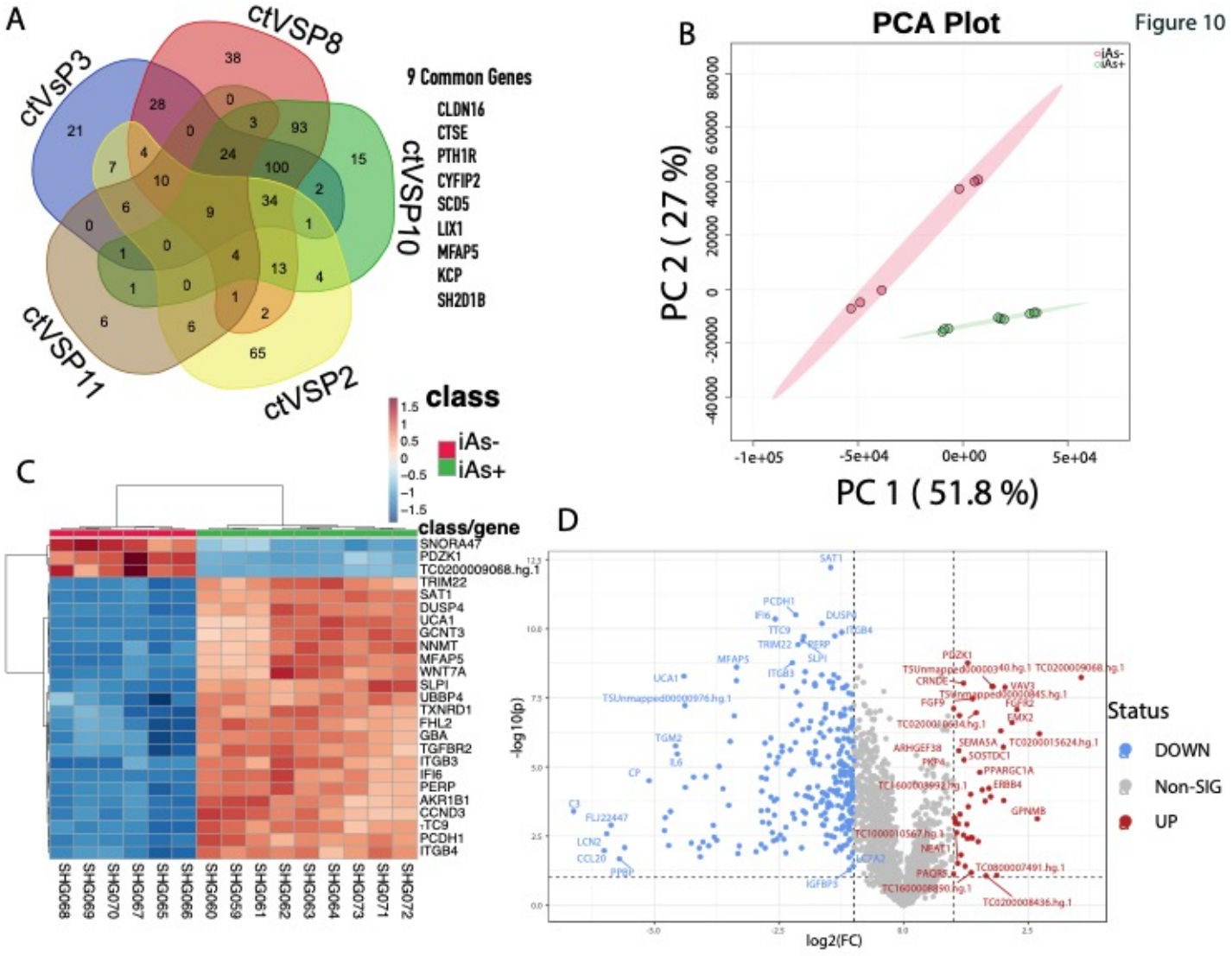


Figure 30: Siginficant genes: iAs+ VS iAs-

Box 24: A) Common genes between P3, P8, P10, P2, and P11 when compared to the control. B) Principal component analysis of the two different conditions, iAs+ and iAs-. C) Hierarchical clustering of the top 25 differentially expressed genes between the two conditions. D) Significant upregulated and downregulated differentially expressed genes based on the iAs+ and iAs- conditions.

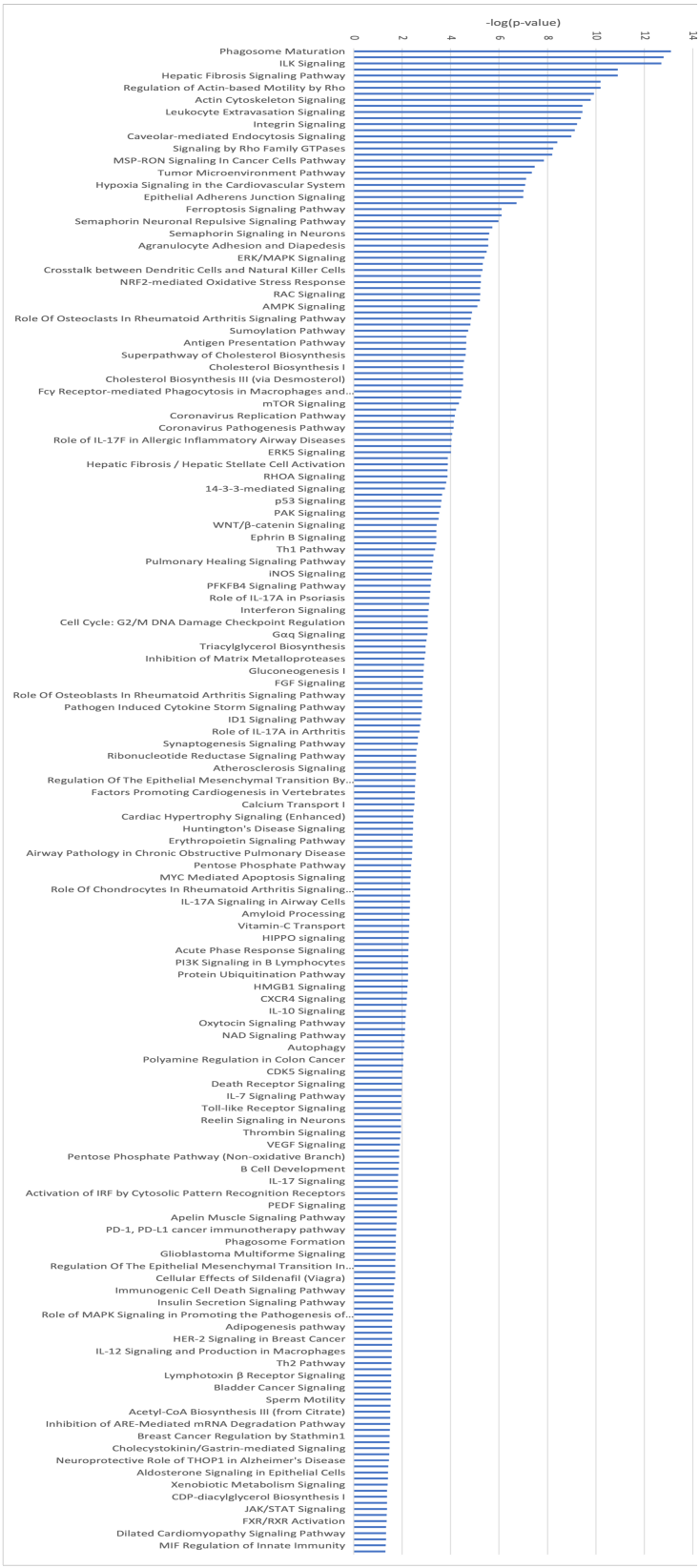


Figure 31 : Ingenuity Pathway Analysis identified canonical pathways for iAs+ vs iAs- conditions.

the P3, P8, P10 were examined separately for keratin and vimentin expression. This confirmed that vimentin was not identified as differentially expressed for any of the 3 passages. In contrast, KRT18 was increased in expression when P3 and P8, but not P10, was compared to the control. This provides evidence that the transition favors the myoepithelial cell. KRT18 has been noted to increase during tubular injury and approximately 20-fold in the early stage following human renal transplantation. [37, 38] To the authors' knowledge this is the first observation that a human RPC can undergo EMT.

Pathway analysis for the 167 gene set identified signaling pathways associated with FGFR2 and chemokine receptors and chemokines as those having strong significance. An analysis of the down-regulated 76 gene set identified strongly with signaling related to the FGFR2 pathway, while the 91 up-regulated gene set was more strongly related to chemokine receptors, chemokines, and related pathways. An interesting feature of the down-regulated 76 gene set is that it was identical for all three time points of iAs exposure. The 76 gene set included FGF 9 and FGF 13 in addition to the FGFR2 receptor. The FGFR2 receptor has been shown to protect against tubular cell death and acute kidney injury involving ERK1/2 signaling in models of renal ischemia and reperfusion.[39, 40] The expression of FGF9 has been shown to maintain the stemness of renal progenitor/stem cells during renal development.[41] FGF9 also has an essential role in the development of mesenchymal components in cells and tissues.[9, 42, 43] The FGF13 is elevated in ischemia/reperfusion in concert with the FGFR2 receptor.[39] The FGF18 gene, the only up-regulated FGF gene in the 167 gene set, has seen only limited study in the kidney, but has been shown to have increased expression in cisplatin-induced murine AKI36. In breast cancer, the FGF18 has been shown to be involved in both cell migration and the EMT [44]. Despite these findings, the individual components of the FGF pathway have seen limited study in renal disease as it relates to agent-induced changes in EMT and MET recovery from those changes. Arguing against any cause-and-effect relationship of the FGF pathway and iAs induced EMT is the observation that iAs increased the activation of ERK1/2 in the HRTPT cells that had been exposed to iAs and undergone EMT. This type of response suggests that the many other ligands that can influence the ERK pathways might be active in the iAs induced EMT. The important observation is that ERK was activated during the EMT process.

The increase in chemokine receptors and their ligands might also play an important role during

iAs induced EMT. The increase in expression of IL4, 10, 13, and 14 and the pathway identification of chemokine receptors bind chemokines would appear to have consequences for renal diseases in the human setting due to their role in immune responses and inflammation. Most studies on EMT involve its involvement in cancer progression. However, a role in renal disease was established a decade ago, indicating that renal epithelial cells could switch to a mesenchymal phenotype. [45-47] The involvement of EMT in inflammation,[48] fibrosis, [49-51] and wound healing [52] suggest a link between chemokines and EMT. The pathway analysis was consistent for a role of FGF and chemokines in the EMT of the HRTPT cells. The only pathway presents in Panther, but not Reactome or David, was the WNT7a pathway. Wnt7a was increased in expression and provides some evidence for up regulation of the non-canonical Wnt-signaling pathway. [53, 54] Both the canonical and non-canonical Wnt pathways have been linked with diabetic nephropathy.[55] Overall, this aspect of the study provides the first demonstration that a renal progenitor cell can undergo EMT when exposed to an environmental toxin. The time course of exposure provides a 167 gene set associated with the iAs induced development of EMT and corresponding 91 and 76 gene sets representing genes up- and down-regulated within the 167 gene set. These 3 sets of genes will be valuable in determining the expression, druggable targets, and prediction value in a wide variety of human renal diseases and other diseases datasets associated with iAs exposure.

The second aspect of this study was to determine, once iAs exposure was stopped, if the iAs treated HRTPT cells would retain their mesenchymal properties. The results showed that by the second passage following iAs removal the cells had regained an epithelial morphology indistinguishable from the control HRTPT cells. This represents the initial observation that RPCs that have undergone EMT due to toxin exposure, can undergo MET back to an epithelial morphology after toxin removal. This ability is consistent with the observation that renal epithelial cells arise during embryogenesis by mesenchymal-to-epithelial transition (MET). [56, 57] It was confirmed that the cells undergoing MET retained the co-expression of PROM1 and CD24 and the ability to form nephron spheres and to undergo osteogenic, neurogenic, lipogenic, and tubulogenic differentiation. A difference in tubulogenic differentiation was found for the recovered cells in that they expressed high levels of calbindin and low levels of aquaporin whereas the control unexposed

cells had the opposite expression levels. To further explore this finding, global gene expression was performed at 2 and 11 passages following iAs removal. Following the removal of iAs, the cells at both P2 and P11 showed a marked divergence from the iAs exposed HRTPT cells at P8 and a return to an expression profile more in line with the control HRTPT cells. This was especially noticeable at passage 11. The common genes between the control HRTPT cells compared to both the P2 and P11 cells was 36. Of these 36 genes, 3 had a reverse in expression between the control and recovered cells (CYFIP2, IGFBP3, and NMNAT2). To determine if iAs exposure might have a lasting, or potentially permanent, effect on gene expression, a common gene set was identified for iAs exposed cells at P3, P8, P10 with those unexposed through P11. One could speculate that epigenetic modification due to iAs exposure might produce long lasting alterations in the genome after iAs removal. The 33 gene set did identify interactions with p53 and the cell cycle. The possible interactions with p53 and the cell cycle would be consistent with the long-term carcinogenic effects of iAs.

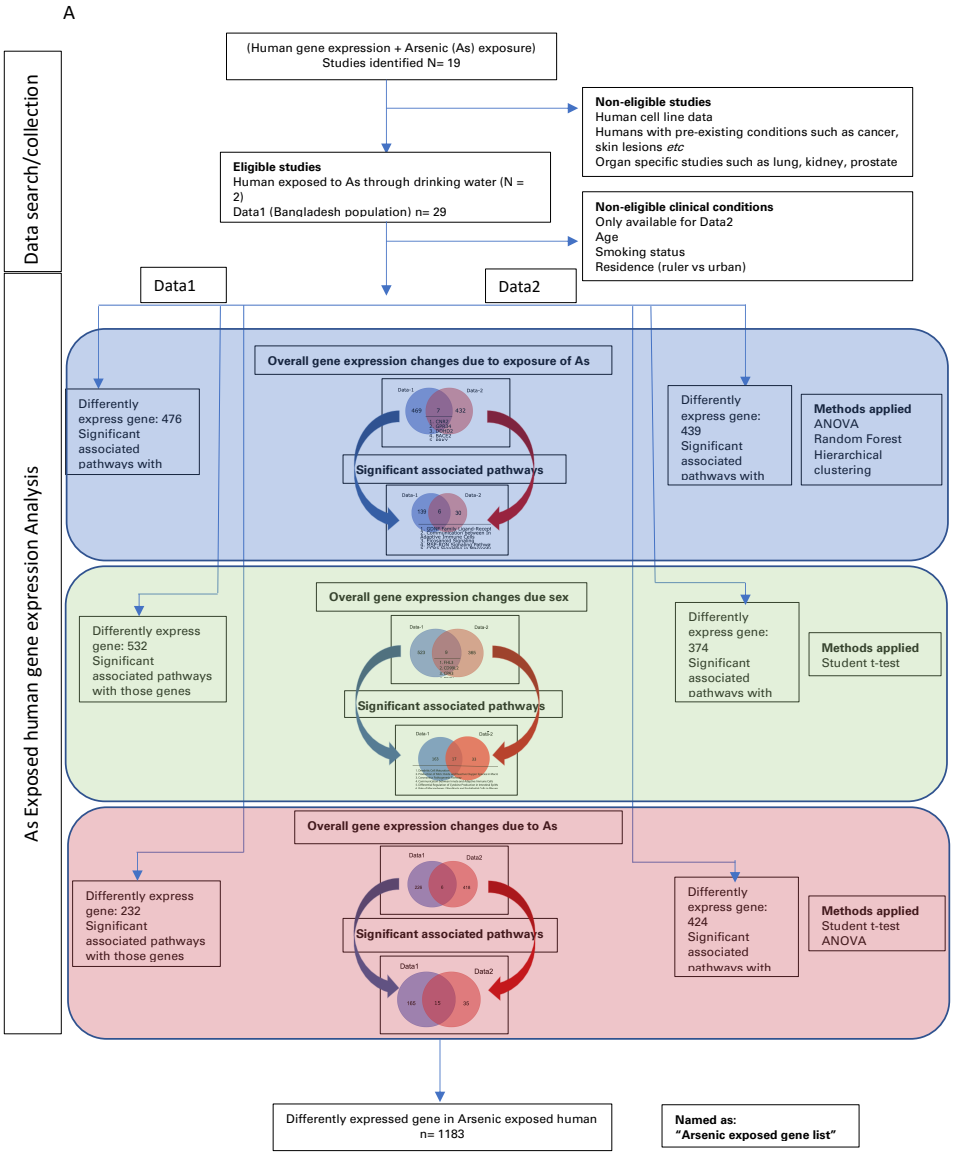
The obvious limitation of the study is that it is performed using cells in culture. The results will require validation in the human kidney.

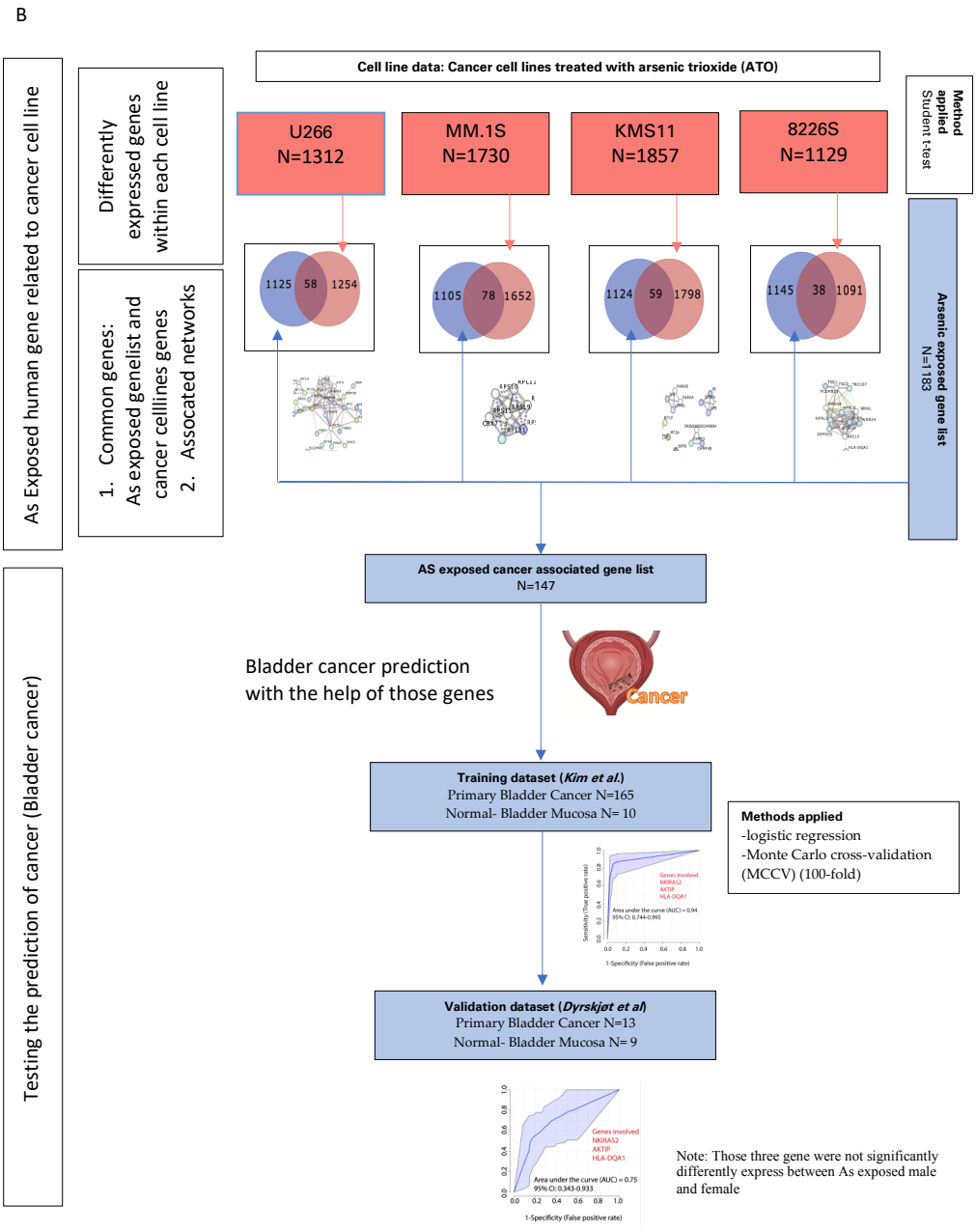
3.6 CONCLUSION

This study shows that human renal progenitor cells, in vitro, undergo EMT when exposed to a nephrotoxin and undergo MET upon toxin removal. In addition, this study identified several significant genes and pathways of interest associated with inorganic arsenic exposure/removal and their linkage with renal disease. These genes provide robust sets of biological functions that can be further validated to predict their association in different diseases. In this study, a variety of machine learning and statistical analysis approaches have been taken to establish in-vitro to in-silico concordance, including an unsupervised analysis of genes across different phenotypic conditions, which can be used as an analytical guideline for other researchers.

SUPPLEMENTARY DATA

CHAPTER 2





Supplementary 1. Figure S1: study flow chart part. (A) As exposed significant gene selection. (B) Find association with multiple melanoma and predictive modeling of bladder cancer.

https://downloads.hindawi.com/journals/omcl/2022/3459855.f1.pdf

Supplementary 2. Table S1: disease as well as molecular and cellular functions associated with statistically significant differentially expressed genes ($p < 0:05$) selected in different phenotypic conditions.

https://downloads.hindawi.com/journals/omcl/2022/3459855.f2.pdf

Supplementary 3. Table S2: significant pathways in Data1 and Data2 and the overlap between two (highlighted in green).

https://downloads.hindawi.com/journals/omcl/2022/3459855.f3.pdf

Supplementary 4. Table S3: significant pathways in Data1 and Data2 (male and female sex) and the overlap between two (highlighted in pink).

https://downloads.hindawi.com/journals/omcl/2022/3459855.f4.pdf

Supplementary 5. Table S4: significant pathways in Data1 and Data2 (low, medium, and high concentrations) and the overlap between various combinations (highlighted in green).

https://downloads.hindawi.com/journals/omcl/2022/3459855.f5.pdf

Supplementary 6. Table S5: significant pathways in Data1 (low, high concentration) and Data2 (low, medium, and high concentrations) and the overlap between two (highlighted in pink).

https://downloads.hindawi.com/journals/omcl/2022/3459855.f6.pdf

Supplementary 7. Table S6: significant genes in Data1 (low and high concentrations) and Data2 (low, medium, and high concentrations) and the overlap between two (highlighted in green).

https://downloads.hindawi.com/journals/omcl/2022/3459855.f7.pdf

Supplementary 8. Table S7: significant pathways in Data1 (low and high concentrations) and Data2 (low, medium, and high concentrations) and the overlap between two (highlighted in pink).

https://downloads.hindawi.com/journals/omcl/2022/3459855.f8.pdf

Supplementary 9. Table S8: significant genes in Data1 (low vs. high) and Data2 (low, medium, and high concentrations) and the overlap between various combinations.

https://downloads.hindawi.com/journals/omcl/2022/3459855.f9.pdf

Supplementary 10. Table S9: significant pathways in Data1 and Data2 (low, medium, and high concentrations) and the overlap between various combinations.

https://downloads.hindawi.com/journals/omcl/2022/3459855.f10.pdf

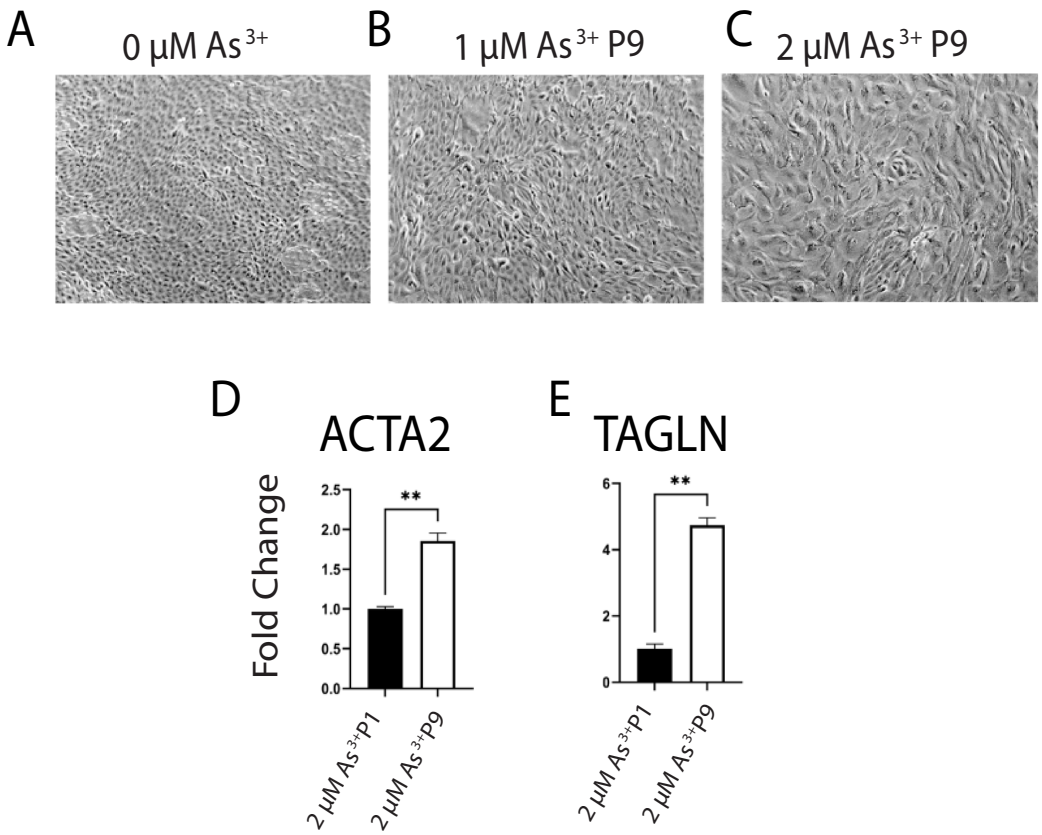
Supplementary 11. Table S10: list of 147 unique genes.

https://downloads.hindawi.com/journals/omcl/2022/3459855.f11.pdf

Supplementary 12. Table S11: survival information of 18 genes differentially expressed between sex and common between ATO and arsenic exposed human.

https://downloads.hindawi.com/journals/omcl/2022/3459855.f12.pdf

CHAPTER 3



Supplementary 1. Figure S1: Light microscopy of lower iAs concentrations A) 0mM As3+ B) 1mM As3+ P9 C) 2mM As3+ P9 and fold change compared to the control of genes D) ACTA2 and E) TAGLN for P1 and P9. Scale bar = 50 µm and Magnification x10

Supplementary 2. Table S1: The list of most significant genes (p<0.05) using ANOVA across all possible conditions (i.e., Control, P3, P8, P10, p2, P11)

https://www.mdpi.com/article/10.3390/ijms24065092/s1

Supplementary 3. Table 2A: Upregulated Gene Set Enrichment Analysis for 2478 probes identified by ANOVA with nominal p-value < 0.05.

https://www.mdpi.com/article/10.3390/ijms24065092/s1

Supplementary 4. Table 2B: Downregulated Gene Set Enrichment Analysis for 2478 probes identified by ANOVA with nominal p-value < 0.05.

https://www.mdpi.com/article/10.3390/ijms24065092/s1

Supplementary 5. Table 3: An examination of the differential expression ($p<0.05$ and $|FC| >2$) between the control HRTPT cells and those exposed to 4.5μM iAs for 3, 8, and 10 passages identified 247, 363, and 304 genes, respectively.

https://www.mdpi.com/article/10.3390/ijms24065092/s1

Supplementary 6. Table 4: Differential gene expression analysis between all HRTPT cells exposed to iAs passage (i.e., combined P3, P8, P10) and control.

https://www.mdpi.com/article/10.3390/ijms24065092/s1

Supplementary 7. Table 5A: Upregulated Gene Set Enrichment Analysis for 167 gene set with nominal p-value < 0.05.

https://www.mdpi.com/article/10.3390/ijms24065092/s1

Supplementary 8. Table 5B: Downregulated Gene Set Enrichment Analysis for 167 gene set with nominal p-value
< 0.05.

https://www.mdpi.com/article/10.3390/ijms24065092/s1

*Supplementary 9. Table 6: Reactome pathway analysis of 167 Common HRTPT cells exposed to iAs passage P3, P8, P10 Genes * indicates $p < 0.05$, ** indicates $p < 0.01$.*

https://www.mdpi.com/article/10.3390/ijms24065092/s1

Supplementary 10. Table 7A: Upregulated Gene Set Enrichment Analysis for significant genes between the control vs iAs+ with nominal p-value < 0.05.

https://www.mdpi.com/article/10.3390/ijms24065092/s1

Supplementary 11. Table 7B: Downregulated Gene Set Enrichment Analysis for significant genes between the control vs iAs+ with nominal p-value < 0.05.

https://www.mdpi.com/article/10.3390/ijms24065092/s1

Supplementary 12. Table 8: Ingenuity Pathway Analysis performed on 234 genes from Supplementary Table 4.

https://www.mdpi.com/article/10.3390/ijms24065092/s1

Supplementary 13. Table 9: Common differentially expressed genes between P2, P11, and the control.

https://www.mdpi.com/article/10.3390/ijms24065092/s1

Supplementary 14. Table 10: Differentially expressed genes between iAs- and the control.

https://www.mdpi.com/article/10.3390/ijms24065092/s1

Supplementary 15. Table 11: Differentially expressed genes between HRTPT cells recovery to iAs passage (P2, P11) vs control.

https://www.mdpi.com/article/10.3390/ijms24065092/s1

Supplementary 16. Table 12: Reactome and Panther pathway analysis of 36 Common gene HRTPT cells recovery to iAs passage P2, P11.

https://www.mdpi.com/article/10.3390/ijms24065092/s1

Supplementary 17. Table 13: Downregulated Gene Set Enrichment Analysis for significant genes between the control and iAs with nominal p-value < 0.05.

https://www.mdpi.com/article/10.3390/ijms24065092/s1

Supplementary 18. Table 14: Ingenuity Pathway Analysis of Supplementary Table 10.

https://www.mdpi.com/article/10.3390/ijms24065092/s1

Supplementary 19. Table 15: Differentially expressed genes between iAs exposed HRTPT cells (P3, P8, P10) and HRTPT cells following recovery from iAs exposure (P2, P11).

https://www.mdpi.com/article/10.3390/ijms24065092/s1

Supplementary 20. Table 16A: Upregulated Gene Set Enrichment Analysis for significant genes between iAs+ and iAs- with nominal p-value < 0.05.

https://www.mdpi.com/article/10.3390/ijms24065092/s1

Supplementary 21. Table 16B: Downregulated Gene Set Enrichment Analysis for significant genes between iAs+ and iAs- with nominal p-value < 0.05.

https://www.mdpi.com/article/10.3390/ijms24065092/s1

Supplementary 22. Table 17: Ingenuity Pathway Analysis of Supplementary Table 15.

https://www.mdpi.com/article/10.3390/ijms24065092/s1

REFERENCES

CHAPTER 1

1. Tomczak, K., P. Czerwinska, and M. Wiznerowicz. "The Cancer Genome Atlas (Tcga): An Immeasurable Source of Knowledge." *Contemp Oncol (Pozn)* 19, no. 1A (2015): A68-77.
2. Harrow, J., A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo, and T. J. Hubbard. "Gencode: The Reference Human Genome Annotation for the Encode Project." *Genome Res* 22, no. 9 (2012): 1760-74.
3. Genomes Project, Consortium, G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, and G. A. McVean. "An Integrated Map of Genetic Variation from 1,092 Human Genomes." *Nature* 491, no. 7422 (2012): 56-65.
4. Musa, I. H., L. O. Afolabi, I. Zamit, T. H. Musa, H. H. Musa, A. Tassang, T. Y. Akintunde, and W. Li. "Artificial Intelligence and Machine Learning in Cancer Research: A Systematic and Thematic Analysis of the Top 100 Cited Articles Indexed in Scopus Database." *Cancer Control* 29 (2022): 10732748221095946.
5. Bumgarner, R. "Overview of DNA Microarrays: Types, Applications, and Their Future." *Curr Protoc Mol Biol* Chapter 22 (2013): Unit 22 1.
6. Edgar, R., M. Domrachev, and A. E. Lash. "Gene Expression Omnibus: Ncbi Gene Expression and Hybridization Array Data Repository." *Nucleic Acids Res* 30, no. 1 (2002): 207-10.
7. Schena, M., D. Shalon, R. W. Davis, and P. O. Brown. "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray." *Science* 270, no. 5235 (1995): 467-70.
8. Grant, G. R., E. Manduchi, and C. J. Stoeckert, Jr. "Analysis and Management of Microarray Gene Expression Data." *Curr Protoc Mol Biol* Chapter 19 (2007): Unit 19 6.
9. Gautier, L., L. Cope, B. M. Bolstad, and R. A. Irizarry. "Affy--Analysis of Affymetrix Genechip Data at the Probe Level." *Bioinformatics* 20, no. 3 (2004): 307-15.
10. McCall, M. N., B. M. Bolstad, and R. A. Irizarry. "Frozen Robust Multiarray Analysis (Frma)." *Biostatistics* 11, no. 2 (2010): 242-53.
11. Trevino, V., F. Falciani, and H. A. Barrera-Saldana. "DNA Microarrays: A Powerful Genomic Tool for Biomedical and Clinical Research." *Mol Med* 13, no. 9-10 (2007): 527-41.
12. MA., D'Agostino RB and Stevens. "Goodness of Fit Techniques." *Statistics, textbooks and monographs* 68, no. 519.5'6 (1986).
13. *International Encyclopedia of Statistical Science*. Springer, 2011.
14. Kruskal, William H., and W. Allen Wallis. "Use of Ranks in One-Criterion Variance Analysis." *Journal of the American statistical Association* 47, no. 260 (1952): 583-621.

15. Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proc Natl Acad Sci U S A* 102, no. 43 (2005): 15545-50.
16. Huang da, W., B. T. Sherman, and R. A. Lempicki. "Systematic and Integrative Analysis of Large Gene Lists Using David Bioinformatics Resources." *Nat Protoc* 4, no. 1 (2009): 44-57.

CHAPTER 2

1. Humans, IARC Working Group on the Evaluation of Carcinogenic Risks to. "Arsenic, Metals, Fibres, and Dusts. A Review of Human Carcinogens." *The International Agency for Research on Cancer* (2012).
2. Services., U.S. Department of Health and Human. "Report on Carcinogens, Twelfth Edition." *National Toxicology Program* (2011).
3. Lindau, L. "Emissions of Arsenic in Sweden and Their Reduction." *Environ Health Perspect* 19 (1977): 25-9.
4. Faudel, Arun D. Shendrikar and Gerald B. "Distribution of Trace Metals During Oil Shale Retorting." *nvironmental Science & Technology* no. 3 (1978): 332-34.
5. Hansen, L. D., D. Silberman, G. L. Fisher, and D. J. Eatough. "Chemical Speciation of Elements in Stack-Collected, Respirable-Size, Coal Fly Ash." *Environ Sci Technol* 18, no. 3 (1984): 181-6.
6. Murphy, E. A., and M. Aucott. "An Assessment of the Amounts of Arsenical Pesticides Used Historically in a Geographical Area." *Science of The Total Environment* 218, no. 2 (1998): 89-101.
7. Smith, A. H., M. Goycolea, R. Haque, and M. L. Biggs. "Marked Increase in Bladder and Lung Cancer Mortality in a Region of Northern Chile Due to Arsenic in Drinking Water." *Am J Epidemiol* 147, no. 7 (1998): 660-9.
8. Rahman, M. M., R. Naidu, and P. Bhattacharya. "Arsenic Contamination in Groundwater in the Southeast Asia Region." *Environ Geochem Health* 31 Suppl 1 (2009): 9-21.
9. Singh, P., Zhang, W., Robins, R. and Muir, D. " Arsenic in the Asia-Pacific Region: Managing Arsenic for Our Future." *1st International Workshop on Arsenic in the Asia-Pacific region: Managing arsenic for our future* (2001).
10. Vahter, M. "Effects of Arsenic on Maternal and Fetal Health." *Annu Rev Nutr* 29 (2009): 381-99.
11. Zhang, L., Y. Gao, S. Wu, S. Zhang, K. R. Smith, X. Yao, and H. Gao. "Global Impact of Atmospheric Arsenic on Health Risk: 2005 to 2015." *Proc Natl Acad Sci U S A* 117, no. 25 (2020): 13975-82.
12. Jones, M. M. "Antagonists for Toxic Heavy Metals." *Proc West Pharmacol Soc* 27 (1984): 163-7.
13. Haller, J. S. "Therapeutic Mule: The Use of Arsenic in the Nineteenth Century Materia Medica." *Pharm Hist* 17, no. 3 (1975): 87-100.
14. He, X., K. Yang, P. Chen, B. Liu, Y. Zhang, F. Wang, Z. Guo, X. Liu, J. Lou, and H. Chen. "Arsenic Trioxide-Based Therapy in Relapsed/Refractory Multiple Myeloma Patients: A Meta-Analysis and Systematic Review." *Onco Targets Ther* 7 (2014): 1593-9.
15. Kim, J. H., J. H. Kim, Y. S. Yu, D. H. Kim, C. J. Kim, and K. W. Kim. "Antitumor Activity of Arsenic Trioxide on Retinoblastoma: Cell Differentiation and Apoptosis Depending on Arsenic Trioxide Concentration." *Invest Ophthalmol Vis Sci* 50, no. 4 (2009): 1819-23.
16. Hughes, M. F. "Biomarkers of Exposure: A Case Study with Inorganic Arsenic." *Environ Health Perspect* 114, no. 11 (2006): 1790-6.
17. Bommarito, P. A., R. Beck, C. Douillet, L. M. Del Razo, G. G. Garcia-Vargas, O. L. Valenzuela, L. C. Sanchez-Pena, M. Styblo, and R. C. Fry. "Evaluation of Plasma

Arsenicals as Potential Biomarkers of Exposure to Inorganic Arsenic." *J Expo Sci Environ Epidemiol* 29, no. 5 (2019): 718-29.
18. Maki-Paakkanen, J., P. Kurttio, A. Paldy, and J. Pekkanen. "Association between the Clastogenic Effect in Peripheral Lymphocytes and Human Exposure to Arsenic through Drinking Water." *Environ Mol Mutagen* 32, no. 4 (1998): 301-13.
19. Sardana, M. K., G. S. Drummond, S. Sassa, and A. Kappas. "The Potent Heme Oxygenase Inducing Action of Arsenic and Parasiticidal Arsenicals." *Pharmacology* 23, no. 5 (1981): 247-53.
20. Cantor, K. P., and J. H. Lubin. "Arsenic, Internal Cancers, and Issues in Inference from Studies of Low-Level Exposures in Human Populations." *Toxicol Appl Pharmacol* 222, no. 3 (2007): 252-7.
21. Chiou, H. Y., Y. M. Hsueh, K. F. Liaw, S. F. Horng, M. H. Chiang, Y. S. Pu, J. S. Lin, C. H. Huang, and C. J. Chen. "Incidence of Internal Cancers and Ingested Inorganic Arsenic: A Seven-Year Follow-up Study in Taiwan." *Cancer Res* 55, no. 6 (1995): 1296-300.
22. Luster, M. I., and P. P. Simeonova. "Arsenic and Urinary Bladder Cell Proliferation." *Toxicol Appl Pharmacol* 198, no. 3 (2004): 419-23.
23. Steinmaus, C., L. Moore, C. Hopenhayn-Rich, M. L. Biggs, and A. H. Smith. "Arsenic in Drinking Water and Bladder Cancer." *Cancer Invest* 18, no. 2 (2000): 174-82.
24. Tsuda, T., A. Babazono, E. Yamamoto, N. Kurumatani, Y. Mino, T. Ogawa, Y. Kishi, and H. Aoyama. "Ingested Arsenic and Internal Cancer: A Historical Cohort Study Followed for 33 Years." *Am J Epidemiol* 141, no. 3 (1995): 198-209.
25. Lenis, A. T., P. M. Lec, K. Chamie, and M. D. Mshs. "Bladder Cancer: A Review." *JAMA* 324, no. 19 (2020): 1980-91.
26. Hu, Y., J. Li, B. Lou, R. Wu, G. Wang, C. Lu, H. Wang, J. Pi, and Y. Xu. "The Role of Reactive Oxygen Species in Arsenic Toxicity." *Biomolecules* 10, no. 2 (2020).
27. Jomova, K., Z. Jenisova, M. Feszterova, S. Baros, J. Liska, D. Hudecova, C. J. Rhodes, and M. Valko. "Arsenic: Toxicity, Oxidative Stress and Human Disease." *J Appl Toxicol* 31, no. 2 (2011): 95-107.
28. Sawicka, E., A. Lisowska, P. Kowal, and A. Dlugosz. "[the Role of Oxidative Stress in Bladder Cancer]." *Postepy Hig Med Dosw (Online)* 69 (2015): 744-52.
29. Wigner, P., B. Szymanska, M. Bijak, E. Sawicka, P. Kowal, Z. Marchewka, and J. Saluk-Bijak. "Oxidative Stress Parameters as Biomarkers of Bladder Cancer Development and Progression." *Sci Rep* 11, no. 1 (2021): 15134.
30. Boonla, Chanchai. "Oxidative Stress, Epigenetics, and Bladder Cancer." *Cancer Biomark* Second Edition (2021): 67-75.
31. He, J., G. Zhu, G. Wang, and F. Zhang. "Oxidative Stress and Neuroinflammation Potentiate Each Other to Promote Progression of Dopamine Neurodegeneration." *Oxid Med Cell Longev* 2020 (2020): 6137521.
32. Im, M., and L. Dagnino. "Protective Role of Integrin-Linked Kinase against Oxidative Stress and in Maintenance of Genomic Integrity." *Oncotarget* 9, no. 17 (2018): 13637-51.
33. Munoz, A., Y. Chervona, M. Hall, T. Kluz, M. V. Gamble, and M. Costa. "Sex-Specific Patterns and Deregulation of Endocrine Pathways in the Gene Expression Profiles of Bangladeshi Adults Exposed to Arsenic Contaminated Drinking Water." *Toxicol Appl Pharmacol* 284, no. 3 (2015): 330-8.
34. Rehman, M. Y. A., M. van Herwijnen, J. Krauskopf, A. Farooqi, J. C. S. Kleinjans, R. N. Malik, and J. J. Briede. "Transcriptome Responses in Blood Reveal Distinct Biological

Pathways Associated with Arsenic Exposure through Drinking Water in Rural Settings of Punjab, Pakistan." *Environ Int* 135 (2020): 105403.
35. Matulis, S. M., A. A. Morales, L. Yehiayan, C. Croutch, D. Gutman, Y. Cai, K. P. Lee, and L. H. Boise. "Darinaparsin Induces a Unique Cellular Response and Is Active in an Arsenic Trioxide-Resistant Myeloma Cell Line." *Mol Cancer Ther* 8, no. 5 (2009): 1197-206.
36. Lee, J. S., S. H. Leem, S. Y. Lee, S. C. Kim, E. S. Park, S. B. Kim, S. K. Kim, Y. J. Kim, W. J. Kim, and I. S. Chu. "Expression Signature of E2f1 and Its Associated Genes Predict Superficial to Invasive Progression of Bladder Tumors." *J Clin Oncol* 28, no. 16 (2010): 2660-7.
37. Kim, W. J., E. J. Kim, S. K. Kim, Y. J. Kim, Y. S. Ha, P. Jeong, M. J. Kim, S. J. Yun, K. M. Lee, S. K. Moon, S. C. Lee, E. J. Cha, and S. C. Bae. "Predictive Value of Progression-Related Gene Classifier in Primary Non-Muscle Invasive Bladder Cancer." *Mol Cancer* 9 (2010): 3.
38. Dyrskjot, L., M. Kruhoffer, T. Thykjaer, N. Marcussen, J. L. Jensen, K. Moller, and T. F. Orntoft. "Gene Expression in the Urinary Bladder: A Common Carcinoma in Situ Gene Expression Signature Exists Disregarding Histopathological Classification." *Cancer Res* 64, no. 11 (2004): 4040-8.
39. Lin, W. J., H. M. Hsueh, and J. J. Chen. "Power and Sample Size Estimation in Microarray Studies." *BMC Bioinformatics* 11 (2010): 48.
40. Gromski, P. S., H. Muhamadali, D. I. Ellis, Y. Xu, E. Correa, M. L. Turner, and R. Goodacre. "A Tutorial Review: Metabolomics and Partial Least Squares-Discriminant Analysis--a Marriage of Convenience or a Shotgun Wedding." *Anal Chim Acta* 879 (2015): 10-23.
41. Cutler, A., and J. R. Stevens. "Random Forests for Microarrays." *Methods Enzymol* 411 (2006): 422-32.
42. Fieller, E. C., and E. S. Pearson. "Tests for Rank Correlation Coefficients: Ii." *Biometrika* 48, no. 1/2 (1961): 29-40.
43. Susan J. Devlin, R. GNANADESIKAN, J. R. KETTENRING. "Robust Estimation and Outlier Detection with Correlation Coefficients." *Biometrika* 62, no. 3 (1975): 531-45.
44. Wilkinson, Leland, and Michael Friendly. "History Corner the History of the Cluster Heat Map." *American Statistician* (2009).
45. Nielsen, Frank. "Introduction to Hpc with Mpi for Data Science." *Springer* (2016).
46. Rokach L., Maimon O. "Lustering Methods." *Data Mining and Knowledge Discovery Handbook, Springer* (2005).
47. Moore, David S, and Stephane Kirkland. "The Basic Practice of Statistics." *WH Freeman New York* 2 (2007).
48. Tabachnick, Barbara G, and Linda S Fidell. "Experimental Designs Using Anova." *Thomson/Brooks/Cole Belmont, CA* (2007).
49. Abdi, Hervé, and Lynne J Williams. "Tukey's Honestly Significant Difference (Hsd) Test." *Encyclopedia of Research Design* 3, no. 1 (2010): 1-5.
50. Kramer, A., J. Green, J. Pollard, Jr., and S. Tugendreich. "Causal Analysis Approaches in Ingenuity Pathway Analysis." *Bioinformatics* 30, no. 4 (2014): 523-30.
51. Kanehisa, M., S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. "Kegg for Integration and Interpretation of Large-Scale Molecular Data Sets." *Nucleic Acids Res* 40, no. Database issue (2012): D109-14.

52. Finn, R. D., A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. Sonnhammer, J. Tate, and M. Punta. "Pfam: The Protein Families Database." *Nucleic Acids Res* 42, no. Database issue (2014): D222-30.
53. UniProt, Consortium. "Uniprot: A Worldwide Hub of Protein Knowledge." *Nucleic Acids Res* 47, no. D1 (2019): D506-D15.
54. Joshi-Tope, G., M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein. "Reactome: A Knowledgebase of Biological Pathways." *Nucleic Acids Res* 33, no. Database issue (2005): D428-32.
55. Szklarczyk, D., A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, and C. V. Mering. "String V11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets." *Nucleic Acids Res* 47, no. D1 (2019): D607-D13.
56. Philo, J. S. "A Critical Review of Methods for Size Characterization of Non-Particulate Protein Aggregates." *Curr Pharm Biotechnol* 10, no. 4 (2009): 359-72.
57. Tolles, J., and W. J. Meurer. "Logistic Regression: Relating Patient Characteristics to Outcomes." *JAMA* 316, no. 5 (2016): 533-4.
58. Uhlen, M., C. Zhang, S. Lee, E. Sjostedt, L. Fagerberg, G. Bidkhori, R. Benfeitas, M. Arif, Z. Liu, F. Edfors, K. Sanli, K. von Feilitzen, P. Oksvold, E. Lundberg, S. Hober, P. Nilsson, J. Mattsson, J. M. Schwenk, H. Brunnstrom, B. Glimelius, T. Sjoblom, P. H. Edqvist, D. Djureinovic, P. Micke, C. Lindskog, A. Mardinoglu, and F. Ponten. "A Pathology Atlas of the Human Cancer Transcriptome." *Science* 357, no. 6352 (2017).
59. Huang, C. Y., Y. C. Lin, H. S. Shiue, W. J. Chen, C. T. Su, Y. S. Pu, P. L. Ao, and Y. M. Hsueh. "Comparison of Arsenic Methylation Capacity and Polymorphisms of Arsenic Methylation Genes between Bladder Cancer and Upper Tract Urothelial Carcinoma." *Toxicol Lett* 295 (2018): 64-73.
60. Zhou, C. Y., L. Y. Gong, R. Liao, N. N. Weng, Y. Y. Feng, Y. P. Dong, H. Zhu, Y. Q. Zhao, Y. Y. Zhang, Q. Zhu, and S. X. Han. "Evaluation of the Target Genes of Arsenic Trioxide in Pancreatic Cancer by Bioinformatics Analysis." *Oncol Lett* 18, no. 5 (2019): 5163-72.
61. Zhang, L., Y. Huang, J. Ling, Y. Xiang, and W. Zhuo. "Screening of Key Genes and Prediction of Therapeutic Agents in Arsenic-Induced Lung Carcinoma." *Cancer Biomark* 25, no. 4 (2019): 351-60.
62. Zhang, L., Y. Zhou, J. Zhang, A. Chang, and X. Zhuo. "Screening of Hub Genes and Prediction of Putative Drugs in Arsenic-Related Bladder Carcinoma: An in Silico Study." *J Trace Elem Med Biol* 62 (2020): 126609.
63. Bettiga, A., M. Aureli, G. Colciago, V. Murdica, M. Moschini, R. Luciano, D. Canals, Y. Hannun, P. Hedlund, G. Lavorgna, R. Colombo, R. Bassi, M. Samarani, F. Montorsi, A. Salonia, and F. Benigni. "Bladder Cancer Cell Growth and Motility Implicate Cannabinoid 2 Receptor-Mediated Modifications of Sphingolipids Metabolism." *Sci Rep* 7 (2017): 42157.
64. Jin, Z. T., K. Li, M. Li, Z. G. Ren, F. S. Wang, J. Y. Zhu, X. S. Leng, and W. D. Yu. "G-Protein Coupled Receptor 34 Knockdown Impairs the Proliferation and Migration of Hgc-27 Gastric Cancer Cells in Vitro." *Chin Med J (Engl)* 128, no. 4 (2015): 545-9.

65. Husi, H., R. J. Skipworth, A. Cronshaw, K. C. Fearon, and J. A. Ross. "Proteomic Identification of Potential Cancer Markers in Human Urine Using Subtractive Analysis." *Int J Oncol* 48, no. 5 (2016): 1921-32.
66. Uhlen, M., P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester, S. Hober, H. Wernerus, L. Bjorling, and F. Ponten. "Towards a Knowledge-Based Human Protein Atlas." *Nat Biotechnol* 28, no. 12 (2010): 1248-50.
67. Ren, R., K. Tyryshkin, C. H. Graham, M. Koti, and D. R. Siemens. "Comprehensive Immune Transcriptomic Analysis in Bladder Cancer Reveals Subtype Specific Immune Gene Expression Patterns of Prognostic Relevance." *Oncotarget* 8, no. 41 (2017): 70982-1001.
68. Weakley, S. M., H. Wang, Q. Yao, and C. Chen. "Expression and Function of a Large Non-Coding Rna Gene Xist in Human Cancer." *World J Surg* 35, no. 8 (2011): 1751-6.
69. Zhu, J., F. Kong, L. Xing, Z. Jin, and Z. Li. "Prognostic and Clinicopathological Value of Long Noncoding Rna Xist in Cancer." *Clin Chim Acta* 479 (2018): 43-47.
70. Wei, W., Y. Liu, Y. Lu, B. Yang, and L. Tang. "Lncrna Xist Promotes Pancreatic Cancer Proliferation through Mir-133a/Egfr." *J Cell Biochem* 118, no. 10 (2017): 3349-58.
71. Gutschner, T., M. Hammerle, and S. Diederichs. "Malat1 -- a Paradigm for Long Noncoding Rna Function in Cancer." *J Mol Med (Berl)* 91, no. 7 (2013): 791-801.
72. Zhou, K., J. Yang, X. Li, and W. Chen. "Long Non-Coding Rna Xist Promotes Cell Proliferation and Migration through Targeting Mir-133a in Bladder Cancer." *Exp Ther Med* 18, no. 5 (2019): 3475-83.
73. Li, C., Y. Cui, L. F. Liu, W. B. Ren, Q. Q. Li, X. Zhou, Y. L. Li, Y. Li, X. Y. Bai, and X. B. Zu. "High Expression of Long Noncoding Rna Malat1 Indicates a Poor Prognosis and Promotes Clinical Progression and Metastasis in Bladder Cancer." *Clin Genitourin Cancer* 15, no. 5 (2017): 570-76.
74. Cai, X., J. Wang, X. Huang, W. Fu, W. Xia, M. Zou, Y. Wang, J. Wang, and D. Xu. "Identification and Characterization of Mt-1x as a Novel Fhl3-Binding Partner." *PLoS One* 9, no. 4 (2014): e93723.
75. Manara, M. C., M. Pasello, and K. Scotlandi. "Cd99: A Cell Surface Protein with an Oncojanus Role in Tumors." *Genes (Basel)* 9, no. 3 (2018).
76. Li, M. Z., D. H. Lai, H. B. Zhao, Z. Chen, Q. X. Huang, and J. Situ. "Socs3 Overexpression Enhances Adm Resistance in Bladder Cancer T24 Cells." *Eur Rev Med Pharmacol Sci* 21, no. 13 (2017): 3005-11.
77. Chen, C., W. He, J. Huang, B. Wang, H. Li, Q. Cai, F. Su, J. Bi, H. Liu, B. Zhang, N. Jiang, G. Zhong, Y. Zhao, W. Dong, and T. Lin. "Lnmat1 Promotes Lymphatic Metastasis of Bladder Cancer Via Ccl2 Dependent Macrophage Recruitment." *Nat Commun* 9, no. 1 (2018): 3826.
78. Neelam Mukherjee, Niannian Ji, Zhen-Ju Shu, Tyler J Curiel and Robert S Svatek. "Ccl2/Ccr2 Signaling Protects against Bladder Cancer Growth in a T Cell Dependent Manner." *Journal of Immunol* 204 (2020).

CHAPTER 3

[1] B. Bussolati *et al.*, "Isolation of Renal Progenitor Cells from Adult Human Kidney," *The American journal of pathology,* vol. 166, no. 2, pp. 545-555, 2005, doi: 10.1016/S0002-9440(10)62276-6.

[2] C. Sagrinati *et al.*, "Isolation and characterization of multipotent progenitor cells from the bowman's capsule of adult human kidneys," *Journal of the American Society of Nephrology,* vol. 17, no. 9, pp. 2443-2456, 2006, doi: 10.1681/ASN.2006010089.

[3] B. Bussolati *et al.*, "Hypoxia modulates the undifferentiated phenotype of human renal inner medullary CD133+ progenitors through Oct4/miR-145 balance," *American journal of physiology. Heart and circulatory physiology,* vol. 71, no. 1, 2012.

[4] B. Smeets *et al.*, "Proximal tubular cells contain a phenotypically distinct, scattered cell population involved in tubular regeneration: Phenotypically distinct proximal tubular cells," *The Journal of pathology,* vol. 229, no. 5, pp. 645-659, 2013, doi: 10.1002/path.4125.

[5] P. Romagnani and G. Remuzzi, "CD133+ renal stem cells always co-express CD24 in adult human kidney tissue," *Stem cell research,* vol. 12, no. 3, pp. 828-829, 2014, doi: 10.1016/j.scr.2013.12.011.

[6] E. Ronconi *et al.*, "Regeneration of Glomerular Podocytes by Human Renal Progenitors," *Journal of the American Society of Nephrology,* vol. 20, no. 2, pp. 322-332, 2009, doi: 10.1681/ASN.2008070709.

[7] P. Romagnani, L. Lasagni, and G. Remuzzi, "Renal progenitors: an evolutionary conserved strategy for kidney regeneration," (in eng), *Nat Rev Nephrol,* vol. 9, no. 3, pp. 137-46, Mar 2013, doi: 10.1038/nrneph.2012.290.

[8] K. Berger and M. J. Moeller, "Podocytopenia, parietal epithelial cells and glomerulosclerosis," *Nephrology, dialysis, transplantation,* vol. 29, no. 5, pp. 948-950, 2014, doi: 10.1093/ndt/gft511.

[9] D. Lindgren *et al.*, "Isolation and Characterization of Progenitor-Like Cells from Human Renal Proximal Tubules," *The American journal of pathology,* vol. 178, no. 2, pp. 828-837, 2011, doi: 10.1016/j.ajpath.2010.10.026.

[10] S. Shrestha *et al.*, "Human renal tubular cells contain CD24/CD133 progenitor cell populations: Implications for tubular regeneration after toxicant induced damage using cadmium as a model," *Toxicology and applied pharmacology,* vol. 331, pp. 116-129, 2017, doi: 10.1016/j.taap.2017.05.038.

[11] S. Shrestha, S. H. Garrett, D. A. Sens, X. D. Zhou, R. Guyer, and S. Somji, "Characterization and determination of cadmium resistance of CD133+/CD24+ and CD133−/CD24+ cells isolated from the immortalized human proximal tubule cell line, RPTEC/TERT1," *Toxicology and applied pharmacology,* vol. 375, pp. 5-16, 2019, doi: 10.1016/j.taap.2019.05.007.

[12] S. Shrestha *et al.*, "Role of HRTPT in kidney proximal epithelial cell regeneration: Integrative differential expression and pathway analyses using microarray and scRNA-seq," *Journal of cellular and molecular medicine,* vol. 25, no. 22, pp. 10466-10479, 2021, doi: 10.1111/jcmm.16976.

[13] M. Wieser *et al.*, "hTERT alone immortalizes epithelial cells of renal proximal tubules without changing their functional characteristics," (in eng), *Am J Physiol Renal Physiol,* vol. 295, no. 5, pp. F1365-75, Nov 2008, doi: 10.1152/ajprenal.90405.2008.
[14] M. F. Hughes, "Arsenic toxicity and potential mechanisms of action," *Toxicology letters,* vol. 133, no. 1, pp. 1-16, 2002, doi: 10.1016/S0378-4274(02)00084-X.
[15] D. K. Nordstrom, "Public health. Worldwide occurrences of arsenic in ground water," *Science (American Association for the Advancement of Science),* vol. 296, no. 5576, p. 2143, 2002.
[16] A. H. Smith and C. M. Steinmaus, "Arsenic in drinking water," *BMJ,* vol. 342, no. may05 2, pp. d2248-d2248, 2011, doi: 10.1136/bmj.d2248.
[17] S. M. Cohen, L. L. Arnold, M. Eldan, A. S. Lewis, and B. D. Beck, "Methylated Arsenicals: The Implications of Metabolism and Carcinogenicity Studies in Rodents to Human Risk Assessment," *Critical reviews in toxicology,* vol. 36, no. 2, pp. 99-133, 2006, doi: 10.1080/10408440500534230.
[18] M. F. Hughes, B. D. Beck, Y. Chen, A. S. Lewis, and D. J. Thomas, "Arsenic Exposure and Toxicology: A Historical Perspective," *Toxicological sciences,* vol. 123, no. 2, pp. 305-332, 2011, doi: 10.1093/toxsci/kfr184.
[19] L.-I. Hsu *et al.*, "Arsenic Exposure From Drinking Water and the Incidence of CKD in Low to Moderate Exposed Areas of Taiwan: A 14-Year Prospective Study," *American journal of kidney diseases,* vol. 70, no. 6, pp. 787-797, 2017, doi: 10.1053/j.ajkd.2017.06.012.
[20] C. G. Sotomayor *et al.*, "Circulating Arsenic is Associated with Long-Term Risk of Graft Failure in Kidney Transplant Recipients: A Prospective Cohort Study," *Journal of clinical medicine,* vol. 9, no. 2, p. 417, 2020, doi: 10.3390/jcm9020417.
[21] M. L. Robles-Osorio, E. Sabath-Silva, and E. Sabath, "Arsenic-mediated nephrotoxicity," *Renal failure,* vol. 37, no. 4, pp. 542-547, 2015, doi: 10.3109/0886022X.2015.1013419.
[22] B. A. Peters *et al.*, "Creatinine, arsenic metabolism, and renal function in an arsenic-exposed population in Bangladesh," (in eng), *PLoS One,* vol. 9, no. 12, p. e113760, 2014, doi: 10.1371/journal.pone.0113760.
[23] Z. Qi, Q. Wang, H. Wang, and M. Tan, "Metallothionein Attenuated Arsenic-Induced Cytotoxicity: The Underlying Mechanism Reflected by Metabolomics and Lipidomics," *Journal of agricultural and food chemistry,* vol. 69, no. 18, pp. 5372-5380, 2021, doi: 10.1021/acs.jafc.1c00724.
[24] T. T. Ngu and M. J. Stillman, "Arsenic Binding to Human Metallothionein," *Journal of the American Chemical Society,* vol. 128, no. 38, pp. 12473-12483, 2006, doi: 10.1021/ja062914c.
[25] M. T. Rahman and M. De Ley, "Arsenic Induction of Metallothionein and Metallothionein Induction Against Arsenic Cytotoxicity," (in eng), *Rev Environ Contam Toxicol,* vol. 240, pp. 151-168, 2017, doi: 10.1007/398_2016_2.
[26] D. S. Moore and S. Kirkland, *The basic practice of statistics*. New York: WH Freeman, 2007.
[27] J. L. Myers, W. A., and R. F. Lorch, "Research design and statistical analysis," 3rd ed. New York: Routledge, 2010, p. 809.
[28] W. Kirch, "Pearson's Correlation Coefficient in Encyclopedia of Public Health," ed: Springer Netherlands: Dordrecht, 2008, pp. 1090-1091.

[29] A. Krämer, J. Green, J. J. Pollard, and S. Tugendreich, "Causal analysis approaches in Ingenuity Pathway Analysis," *Bioinformatics,* vol. 30, no. 4, pp. 523-530, 2014, doi: 10.1093/bioinformatics/btt703.
[30] A. Subramanian *et al.*, "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles," *Proceedings of the National Academy of Sciences - PNAS,* vol. 102, no. 43, pp. 15545-15550, 2005, doi: 10.1073/pnas.0506580102.
[31] M. Gillespie *et al.*, "The reactome pathway knowledgebase 2022," *Nucleic acids research,* vol. 50, no. D1, pp. D687-D692, 2022, doi: 10.1093/nar/gkab1028.
[32] P. D. Thomas, D. Ebert, A. Muruganujan, T. Mushayahama, L. P. Albou, and H. Mi, "PANTHER: Making genome-scale phylogenetics accessible to all," *Protein science,* vol. 31, no. 1, pp. 8-22, 2022, doi: 10.1002/pro.4218.
[33] B. T. Sherman *et al.*, "DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update)," *Nucleic acids research,* 2022, doi: 10.1093/nar/gkac194.
[34] D. W. Huang, R. A. Lempicki, and B. T. Sherman, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature protocols,* vol. 4, no. 1, pp. 44-57, 2008, doi: 10.1038/nprot.2008.211.
[35] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdottir, P. Tamayo, and J. P. Mesirov, "Molecular signatures database (MSigDB) 3.0," *Bioinformatics,* vol. 27, no. 12, pp. 1739-1740, 2011, doi: 10.1093/bioinformatics/btr260.
[36] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo, "The Molecular Signatures Database (MSigDB) hallmark gene set collection," (in eng), *Cell Syst,* vol. 1, no. 6, pp. 417-425, Dec 23 2015, doi: 10.1016/j.cels.2015.12.004.
[37] M. G. Snoeijs *et al.*, "Tubular Epithelial Injury and Inflammation After Ischemia and Reperfusion in Human Kidney Transplantation," *Annals of surgery,* vol. 253, no. 3, pp. 598-604, 2011, doi: 10.1097/SLA.0b013e31820d9ae9.
[38] S. Djudjaj *et al.*, "Keratins are novel markers of renal epithelial cell injury," *Kidney international,* vol. 89, no. 4, pp. 792-808, 2016, doi: 10.1016/j.kint.2015.10.015.
[39] Z. Xu, X. Zhu, M. Wang, Y. Lu, and C. Dai, "FGF/FGFR2 Protects against Tubular Cell Death and Acute Kidney Injury Involving Erk1/2 Signaling Activation," *Kidney diseases,* vol. 6, no. 3, pp. 181-194, 2020, doi: 10.1159/000505661.
[40] X. Tan *et al.*, "Fibroblast Growth Factor 2 Attenuates Renal Ischemia-Reperfusion Injury via Inhibition of Endoplasmic Reticulum Stress," *Frontiers in cell and developmental biology,* vol. 8, pp. 147-147, 2020, doi: 10.3389/fcell.2020.00147.
[41] H. Barak *et al.*, "FGF9 and FGF20 Maintain the Stemness of Nephron Progenitors in Mice and Man," *Developmental cell,* vol. 22, no. 6, pp. 1191-1207, 2012, doi: 10.1016/j.devcel.2012.04.018.
[42] J. S. Colvin, A. C. White, S. J. Pratt, and D. M. Ornitz, "Lung hypoplasia and neonatal death in Fgf9-null mice identify this gene as an essential regulator of lung mesenchyme," *Development (Cambridge),* vol. 128, no. 11, pp. 2095-2106, 2001, doi: 10.1242/dev.128.11.2095.
[43] I. H. Hung, K. Yu, K. J. Lavine, and D. M. Ornitz, "FGF9 regulates early hypertrophic chondrocyte differentiation and skeletal vascularization in the developing stylopod," *Developmental biology,* vol. 307, no. 2, pp. 300-313, 2007, doi: 10.1016/j.ydbio.2007.04.048.

[44] N. Song *et al.*, "FGF18 Enhances Migration and the Epithelial-Mesenchymal Transition in Breast Cancer by Regulating Akt/GSK3β/B-Catenin Signaling," *Cellular physiology and biochemistry,* vol. 49, no. 3, pp. 1019-1073, 2018, doi: 10.1159/000493286.
[45] J. Gros and C. J. Tabin, "Vertebrate Limb Bud Formation Is Initiated by Localized Epithelial-to-Mesenchymal Transition," *Science (American Association for the Advancement of Science),* vol. 343, no. 6176, pp. 1253-1256, 2014, doi: 10.1126/science.1248228.
[46] K. Stark, S. Vainio, G. Vassileva, and A. P. McMahon, "Epithelial transformation of metanephric mesenchyme in the developing kidney regulated by Wnt-4," *Nature (London),* vol. 372, no. 6507, pp. 679-683, 1994, doi: 10.1038/372679a0.
[47] J. P. Thiery, H. Acloque, R. Y. J. Huang, and M. A. Nieto, "Epithelial-Mesenchymal Transitions in Development and Disease," *Cell,* vol. 139, no. 5, pp. 871-890, 2009, doi: 10.1016/j.cell.2009.11.007.
[48] E. G. Neilson, "Mechanisms of disease: Fibroblasts--a new look at an old problem," (in eng), *Nat Clin Pract Nephrol,* vol. 2, no. 2, pp. 101-8, Feb 2006, doi: 10.1038/ncpneph0093.
[49] F. Strutz and E. G. Neilson, "New insights into mechanisms of fibrosis in immune renal injury," *Springer seminars in immunopathology,* vol. 24, no. 4, pp. 459-476, 2003, doi: 10.1007/s00281-003-0123-5.
[50] S. Lovisa *et al.*, "Epithelial-to-mesenchymal transition induces cell cycle arrest and parenchymal damage in renal fibrosis," *Nature medicine,* vol. 21, no. 9, pp. 998-1009, 2015, doi: 10.1038/nm.3902.
[51] M. T. Grande *et al.*, "Snail1-induced partial epithelial-to-mesenchymal transition drives renal fibrosis in mice and can be targeted to reverse established disease," *Nature medicine,* vol. 21, no. 9, pp. 989-997, 2015, doi: 10.1038/nm.3901.
[52] P. Banerjee, S. Venkatachalam, M. K. Mamidi, R. Bhonde, K. Shankar, and R. Pal, "Vitiligo patient-derived keratinocytes exhibit characteristics of normal wound healing via epithelial to mesenchymal transition," *Experimental dermatology,* vol. 24, no. 5, pp. 391-393, 2015, doi: 10.1111/exd.12671.
[53] A. Gajos-Michniewicz and M. Czyz, "WNT Signaling in Melanoma," *International journal of molecular sciences,* vol. 21, no. 14, p. 4852, 2020, doi: 10.3390/ijms21144852.
[54] I. Ackers and R. Malgor, "Interrelationship of canonical and non-canonical Wnt signalling pathways in chronic metabolic diseases," *Diabetes & vascular disease research,* vol. 15, no. 1, pp. 3-13, 2018, doi: 10.1177/1479164117738442.
[55] H. Wang *et al.*, "The Wnt Signaling Pathway in Diabetic Nephropathy," (in eng), *Front Cell Dev Biol,* vol. 9, p. 701547, 2021, doi: 10.3389/fcell.2021.701547.
[56] P. Galichon, S. Finianos, and A. Hertig, "EMT–MET in renal disease: Should we curb our enthusiasm?," *Cancer letters,* vol. 341, no. 1, pp. 24-29, 2013, doi: 10.1016/j.canlet.2013.04.018.
[57] E. D. Hay and A. Zuk, "Transformations between epithelium and mesenchyme: normal, pathological, and experimentally induced," (in eng), *Am J Kidney Dis,* vol. 26, no. 4, pp. 678-90, Oct 1995, doi: 10.1016/0272-6386(95)90610-x.