University of Cincinnati

Date: 5/26/2021

I, Sarah E Davidson, hereby submit this original work as part of the requirements for the degree of Doctor of Philosophy in Biostatistics (Environmental Health).

It is entitled:

Alternative Approach to Dose-Response Modeling of Toxicogenomic Data with an Application in Risk Assessment of Engineered Nanomaterials

Student's name: **Sarah E Davidson**

This work and its defense approved by:

Committee chair: Mario Medvedovic, Ph.D.

Committee member: Michael Borchers, Ph.D.

Committee member: Eileen Kuempel, Ph.D.

Committee member: Siva Sivaganesan, Ph.D.

Committee member: Matthew Wheeler, Ph.D.

39316

Alternative Approach to Dose-Response Modeling of Toxicogenomic Data with an Application in Risk Assessment of Engineered Nanomaterials

A dissertation submitted to the
Graduate School of the University of Cincinnati
in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

in the Department of Environmental &
Public Health Sciences
Division of Biostatistics and Bioinformatics
of the College of Medicine
by

Sarah E. Davidson
B.S. Xavier University

July 2021

Committee Chair: M. Medvedovic, Ph.D.

Abstract

Increased use of genomic data in dose-response (DR) modeling for quantitative risk assessment necessitates the development of new methods which better account for the biological underpinnings leading to adverse health effects or disease. Current genomic dose-response (GDR) modeling methods use parametric models traditionally used for evaluating *in vivo* endpoints. However, assumptions in the current models may be inappropriate (e.g. monotonic response) at the level of gene expression. Additionally, these GDR methods do not take into account other biological phenomenons such as a shared transcription factors, upstream signaling, and feed-back mechanisms which may lead to coordination expression of multiple genes. Coordinated changes in gene expression may result in correlated DR patterns which can be leveraged to better understand the development of adverse health effects and better estimate a dose related to minimal biological response, or benchmark dose (BMD), which can be used as an interim point-of-departure (POD) for risk assessment in the absence of *in vivo* data. The aim of this dissertation is to develop an alternative GDR method which couples shape-constrained spline models and Bayesian clustering models to obtain biologically relevant gene sets sharing similar DR patterns. Here, it is proposed this approach will help to better evaluate the biological mechanisms after an exposure leading to adverse health effects and obtain more cohesive BMDs which can be used as PODs in efficient interim risk assessments. Finally, we demonstrate the utility of the developed method in an evaluation of rodent lung tissue samples after exposure to a set of well-studied engineered nanomaterial exposure and compare our results with those from *in vivo* toxicology endpoints measuring pulmonary inflammation and fibrosis typically used in risk assessment of these exposures.

Preface

This dissertation marks the end of one stage of my life the beginning of another. I am grateful to God for the opportunity to have this be part of my journey in life. It would not have been possible without the blessing of my life and all the graces He has granted me, which I am thankful for with all of my heart. My hope is that the knowledge I gained will prepare me for the work to come. That the memories will be ones to last a lifetime. Finally, that the peaks and the valleys I have journeyed have helped me to grow in virtue for the lessons in life yet to come. With that being said, I would like to thank some of the amazing people God put in my life to aid me in this journey.

I would like to thank all of the teachers and professors that helped to lay a foundation for success and perseverance in my studies. Your encouragement along various parts of my academic journey stuck with me through the end.

I would also like to thank all of the individuals at the National Institute for Occupational Safety and Health (NIOSH), particularly in the Division of Science Integration for helping to inspire and encourage me along the way. The joy and enthusiasm for helping young professionals, like myself, was a constant source of support. I am thankful for the professional development I gained as part of my fellowship opportunity that has prepared me to think critically and become part of a working team. Most of all I am thankful for the relationships I made with the many formal and informal mentors at NIOSH. Your support is something I hope one day to pay forward.

I would like to thank my committee members for their guidance throughout this dissertation. Also, thank you to the the professors here at the University of Cincinnati that helped me to gain a better understanding of statistical methods and biological mechanisms needed for completing this work. Thank you to the department staff who helped to enable my studies, research, and progression through this program.

I would like to thank my family and friends who supported me mentally, emotionally, intellectually, and spiritually along the way. None of this would have been possible without the never-ending support of my parents. Your lessons in perseverance and pursuing my vocation with one-hundred and ten percent helped me to continue going when I felt like quitting. To my extended family and friends, your love and prayers have always given me a reason to strive to challenge myself and remember when it is necessary to fully immerse myself in the present moment.

Finally, to the love of my life Logan. Thank you for supporting me through the ups and downs of my dissertation. In the worst times, you kept me thinking on the bright side and supported me no matter where I ended. Thank you for the freedom to be me, the love to keep me fueled, and for being my teammate to the finish line. JMJ+

0.1 List of Abbreviations

DR - dose-response
GDR - genomic dose-response
BMD - benchmark dose estimate
BMR - benchmark response
BLX - Bayesian local extrema spline
BPF - best parametric fit
AIC - Aikake information criterion
ALOHA - Aggregated Local extrema splines for High-throughput dose-response Analyses
CSIMM - context-specific infinite mixture model
CPM - counts per million
MCMC - Markov Chain Monte Carlo

BAP - Benzo-[α]-pyrene
CPZ - Chlorpromazine
KCl - potassium chloride
AD - all doses
CDR - cytotoxic doses removed

ENM - engineered nanomaterial(s)
CNT - carbon nanotubes
MWCNT - multi-walled carbon nanotubes
CB - carbon black
TiO_2 - titanium dioxide

0.2 List of Definitions

CSIMM* - csimm clustering on the observed normalized expression data
ALOHA - csimm clustering on the BLX estimated dose-response data
BMDExpress* - csimm clustering on the parametric estimated dose-response data

0.3 List of R Packages

temposeqData - TempOSeq Gene Expression Platform Normalization Package
lxsplineBMD - batch modeling and benchmark dose estimation for BLX spline models, add-on package for *lxspines* R package developed by Wheeler et al. (2017)
clusterBLXspline - data preparation package for CSIMM clustering

Contents

Chapter 1

Introduction

Human health risk assessment evaluates the risk associated with environmental exposures (i.e. chemicals, particulates, or other physical exposures) and biological responses, which may indicate the onset of adverse health effects or disease progression at certain exposures. Quantitative evaluations of these relationships are conducted to evaluate potency of the exposure and set standards for protecting individuals in the population from developing adverse health effects (i.e. disease or death). One cornerstone to these assessments is dose-response (DR) modeling analysis.

Dose-response (DR) modeling evaluates the relationship between the exposure level (i.e. dose or concentration) and biological response(s) most relevant for predicting adverse health effects producing a curve estimating the true underlying relationship. The resulting curve fit can be used to estimate a dose (or benchmark dose – BMD) yielding a specified level of response (or benchmark response – BMR) related to minimal biological effect after exposure compared with the background response rate of a unexposed population (1). Data for these analyses are typically collected from *in vivo* toxicological or clinical endpoints from studies in animals or humans. *In vivo* toxicology endpoints may include tumor incidence, development of fibrotic tissue, liver enzymes, or another response relevant for obtaining information about the adverse outcome of concern.

Though *in vivo* toxicology endpoints are indicators of disease onset or adverse biological outcomes, there are several shortcomings to using this kind of data for human health risk assessment. First, collecting these samples for a single exposure compound can be costly and require extended periods of time. Time and money constraints ultimately limit risk assessors to the number of evaluations that can be conducted every year. This can be a cause of concern since the number of existing chemicals and particulates with protective standards is far-outweighed by the number which have little to no toxicological data. Secondly, clinical indicators only provide information about the disease endpoint. This does not allow for investigation of biological events upstream the chain of events to help prevent the onset of the disease development process or further progression of these events at earlier time-points.

Advancement of technologies, laboratory assays, and computational methods for assessing molecular and cellular biology in the past two decades has been adopted by toxicologists and risk assessors to move from whole animal testing to new methods with the potential to obtain better predictability (2). The development of high-throughput technologies measuring molecular endpoints such as gene and protein expression are particularly important to improving upon the limitations of traditional risk assessment practices. Initiatives such as those developed by of Sciences Engineering and Medicine (2), U.S. EPA (3), NTP et al. (4), Krewski et al. (5) encouraged the use of alternative testing methods – such as high-throughput genomic, proteomic, and metabolomic assays – to improve human health risk assessments. Utilization of these methods is aimed at making risk assessments more efficient and allow researchers to investigate upstream molecular events to better predict adverse health outcomes and conduct evaluations on numerous chemicals simultaneously currently without protective

standards or *in vivo* data (4, 6). The efficiency of these methods and their ability to better understand biological mechanisms underpinning the development of adverse outcomes (e.g. cancer, tissue damage, etc.) may help to better estimate protective levels of exposure for individuals potentially exposed (7, 8).

Genomic dose-response (GDR) modeling is one method merging the strengths of advanced high-throughput technologies – in particular gene expression or transcriptomic data – with those of DR modeling for quantitative risk assessment. These analyses are similar to traditional DR modeling, except they use transcriptional responses from *in vivo* animal tissue samples or *in vitro* human cell lines to evaluate the relationship between exposure levels and biological responses. Efficiency and wide availability of data from transcriptomic experiments provides the capacity to rapidly assess and characterize risk associated with perturbagens. It has been proposed that genomic BMD estimates may be used as a surrogate point-of-departure (POD) in cases where there is not sufficient *in vivo* data to obtain a traditional *in vivo* BMD as a POD for risk management (7, 9, 10). Qualitative and quantitative evaluations of perturbed bio-molecular/cellular mechanisms may indicate the level of severity associated with an exposure providing a preliminary understanding of its hazard potential and may help determine priority for further investigation. However, current approaches to GDR modeling rely on modeling approaches used for apical endpoints, which may not be appropriate for gene expression data since the underlying modeling assumptions – e.g. monotonicity – do not always hold, and couple this with the addition of bioinformatics methods to account for relationships between genes.

The goal of this work is to develop a new GDR modeling approach to alleviate some of the limitations in current approaches. In the development of the new method the general hypothesis is that utilizing co-regulated gene expression information across dose-response groups to estimate benchmark dose estimates (BMDs) will provide more biologically relevant estimates for human health risk assessment of preventable perturbagen exposures. That is, transcriptional BMD estimates from clusters sharing similar dose-response patterns, which are enriched for metabolic processes linked with adverse health events, will be predictive of points-of-departure (PODs) estimated from traditional *in vivo* toxicology studies used to protect individuals. Chapter 2 will focus on the Specific Aim 1 of this dissertation which is to develop a two-stage genomic dose-response modeling method that (i) fits genes with shape constrained spline models and (ii) probabilistically clusters genes that have a similar dose-response pattern. We describe our new GDR modeling approach, called Aggregated Local Extrema Splines for High-throughput Dose-response Analysis (ALOHA) (11), and compare our results from analyzing a real-world trascriptomics data set with those from the current state-of-the-art GDR modeling approach BMDExpress (12, 13, 14). In Chapter 2, the biological relevance of BMD estimates is assessed using three major metrics. These metrics include the level of functional coherence among genes in identified gene clusters (i.e. how common clustered genes are also members of the same enriched biological pathways), how tightly the gene-level BMD estimates are grouped within a gene cluster, and enriched

gene sets identified within a cluster are also identified when applying standard enrichment methods. Chapter 3 will focus on the application of the ALOHA method from Specific Aim 1 to publicly-available datasets on mouse lung tissue samples after exposure to engineered nanomaterials (ENMs) and evaluates the doses associated with adverse pulmonary effects in rodents (e.g. chronic inflammation and fibrosis). In this chapter, we investigate two sub-aims to evaluate the feasibility of this approach for risk assessment and obtain biologically relevant information. The first sub-aim is to identify the similarities and differences in the perturbed molecular pathways and potency of various materials, which is done by comparing enriched gene sets and transcriptional points-of-departure (POD) within and between materials. The second sub-aim is to evaluate the predictive ability of our transcriptional PODs for estimating *in vivo* PODs by evaluating the overlap between the transcriptional and traditional *in vivo* PODs. In Chapter 3, biological relevance of BMD estimates is assessed qualitatively by evaluating the enriched gene sets, particularly those that cluster together, and their participation in the etiology or development of chronic pulmonary inflammation or pulmonary fibrosis. These clusters and their core genes are also qualitatively evaluated for their consistency with other published studies on the pathways to the pulmonary adverse outcomes of interest here. Finally, evaluating the overlap between the ALOHA transcriptional PODs with those estimated from *in vivo* pulmonary endpoints allows us to determine their the ability to predict potency. Chapter 4 provides a tour of the software packages developed for implementing these methods with some simple examples in vignettes.

References

[1] U.S. EPA. Benchmark dose technical guidance document. Risk Assessment Forum EPA/100/R-12/001, U.S. Environmental Protection Agency, 2012.

[2] National Academies of Sciences Engineering and Medicine. *Using 21st Century Science to Improve Risk-Related Evaluations*. The National Academies Press, Washington, DC, 2017. ISBN 978-0-309-45348-6. doi: 10.17226/24635. URL `https://www.nap.edu/catalog/24635/using-21st-century-science-to-improve-risk-related-evaluations`.

[3] U.S. EPA. A framework for a computational toxicology research program. Technical Report EPA/600/R-03/065, U.S. Environmental Protection Agency, 2003.

[4] National Toxicology Program NTP et al. A national toxicology program for the 21st century: Roadmap to achieve the ntp vision. *National Institute of Environmental Health Sciences*, 2004.

[5] Daniel Krewski, Melvin E Andersen, Ellen Mantus, and Lauren Zeise. Toxicity testing in the 21st century: implications for human health risk assessment. *Risk Analysis: An International Journal*, 29(4):474–479, 2009.

[6] National Toxicology Program et al. Ntp research report on national toxicology program approach to genomic dose-response modeling: Research report 5. NTP Research Report Series 2473-4756, National Toxicology Program, 2018. URL http://dx.doi.org/10.22427/NTP-RR-5.

[7] Ivy Moffat, Nikolai L Chepelev, Sarah Labib, Julie Bourdon-Lacombe, Byron Kuo, Julie K Buick, France Lemieux, Andrew Williams, Sabina Halappanavar, Amal I Malik, et al. Comparison of toxicogenomics and traditional approaches to inform mode of action and points of departure in human health risk assessment of benzo [a] pyrene in drinking water. *Critical reviews in toxicology*, 45(1):1–43, 2015.

[8] Jeffry L Dean, Q Jay Zhao, Jason C Lambert, Belinda S Hawkins, Russell S Thomas, and Scott C Wesselkamper. Editor's highlight: Application of gene set enrichment analysis for identification of chemically induced, biologically relevant transcriptomic networks and potential utilization in human health risk assessment. *Toxicological Sciences*, 157(1):85–99, 2017.

[9] Russell S. Thomas, Bruce C. Allen, Andy Nong, Longlong Yang, Edilberto Bermudez, III Clewell, Harvey J., and Melvin E. Andersen. A Method to Integrate Benchmark Dose Estimates with Genomic Data to Assess the Functional Effects of Chemical Exposure. *Toxicological Sciences*, 98(1): 240–248, 04 2007. ISSN 1096-6080. doi: 10.1093/toxsci/kfm092. URL https://doi.org/10.1093/toxsci/kfm092.

[10] Russell S. Thomas, III Clewell, Harvey J., Bruce C. Allen, Scott C. Wesselkamper, Nina Ching Y. Wang, Jason C. Lambert, Janet K. Hess-Wilson, Q. Jay Zhao, and Melvin E. Andersen. Application of Transcriptional Benchmark Dose Values in Quantitative Cancer and Noncancer Risk Assessment. *Toxicological Sciences*, 120(1):194–205, 11 2010. ISSN 1096-6080. doi: 10.1093/toxsci/kfq355. URL https://doi.org/10.1093/toxsci/kfq355.

[11] Sarah E. Davidson, Matthew W. Wheeler, Scott S. Auerbach, Siva Sivaganesan, and Mario Medvedovic. Aloha: Aggregated local extrema splines for high-throughput dose-response analysis. *bioRxiv*, 2021. doi: 10.1101/2021.03.29.437588. URL https://www.biorxiv.org/content/early/2021/03/31/2021.03.29.437588.

[12] Longlong Yang, Bruce C Allen, and Russell S Thomas. Bmdexpress: a software tool for the benchmark dose analyses of genomic data. *BMC genomics*, 8(1):387, 2007.

[13] Jason R Phillips, Daniel L Svoboda, Arpit Tandon, Shyam Patel, Alex Sedykh, Deepak Mav, Byron Kuo, Carole L Yauk, Longlong Yang, Russell S Thomas, et al. Bmdexpress 2: enhanced transcriptomic dose-response analysis workflow. *Bioinformatics*, 35(10):1780–1782, 2019.

[14] Scott S. Auerbach and Richard S. Paules. Genomic dose response: Successes, challenges, and next steps. *Current Opinion in Toxicology*, 11-12: 84 – 92, 2018. ISSN 2468-2020. doi: https://doi.org/10.1016/j.cotox.2019.04.002. URL `http://www.sciencedirect.com/science/article/pii/S2468202018301025`. Genomic Toxicology.

Chapter 2

A New Approach to Genomic Dose-Response Modeling

The following chapter contains the manuscript and supplemental materials describing and comparing the developed genomic dose-response method ALOHA with current approaches as part of this dissertation. This manuscript was submitted to the pre-print publication bioRxiv, see Davidson et al. (2021) at https://www.biorxiv.org/content/early/2021/03/31/2021.03.29.437588.

ALOHA: Aggregated Local Extrema Splines for High-throughput Dose-response Analysis

Sarah E. Davidson[1]*, Matthew W. Wheeler[2], Scott S. Auerbach[3], Siva Sivaganesan[4], Mario Medvedovic[1]

[1] University of Cincinnati Department of Environmental Health Sciences Division of Biostatistics and Bioinformatics, Cincinnati, OH
[2] National Institute of Environmental Health Sciences (NIEHS) Division of Intramural Research (DIR) Biostatistics & Computational Biology Branch (BCBB), Durham, NC
[3] National Institute of Environmental Health Sciences (NIEHS) National Toxicology Program Division (DNTP) Biomolecular Screening Branch (BSB), Durham, NC
[4] University of Cincinnati Department of Mathematical Sciences Division of Statistics and Data Science, Cincinnati, OH

Abstract

Computational methods for genomic dose-response integrate dose-response modeling with bioinformatics tools to evaluate changes in molecular and cellular functions related to pathogenic processes. These methods use parametric models to describe each gene's dose-response, but such models may not adequately capture expression changes. Additionally, current approaches do not consider gene co-expression networks. When assessing co-expression networks, one typically does not consider the dose-response relationship, resulting in 'co-regulated' gene sets containing genes having different dose-response patterns. To avoid these limitations, we develop an analysis pipeline called Aggregated Local Extrema Splines for High-throughput Analysis (ALOHA), which computes individual genomic dose-response functions using a flexible class Bayesian shape constrained splines and clusters gene co-regulation based upon these fits. Using splines, we reduce information loss due to parametric lack-of-fit issues, and because we cluster on dose-response relationships, we better identify co-regulation clusters for genes that have co-expressed dose-response patterns from chemical exposure. The clustered pathways can then be used to estimate a dose associated with a pre-specified biological response, i.e., the benchmark dose (BMD), and approximate a point of departure dose corresponding to minimal adverse response in the whole tissue/organism. We compare our approach to current parametric methods and our biologically enriched gene sets to cluster on normalized expression data. Using this methodology, we can more effectively extract the underlying structure leading to more cohesive estimates of gene set potency.

Keywords: Biological Pathways, Genomic Benchmark Dose, High throughput data, Bayesian Clustering

1 Introduction

Genomic dose-response (GDR) modeling evaluates the relationship between chemical exposure and transcriptional responses to identify a level of minimal biological effect, called a point-of-departure

(POD). Traditionally, PODs are estimated using *in vivo* toxicology responses. It has been proposed when minimal *in vivo* data exists, GDR data can be used in an interim risk assessment to estimate a genomic POD [1, 2, 3, 4]. Acceptance of this approach relies on mitigating concerns about the qualitative differences between genomic and traditional PODs. A suitable estimation of transcriptional PODs accounts for differences in biological relevance between perturbed pathways and incorporates mechanistic information [2, 5].

To date, several GDR modeling approaches have been developed with a variety of methodological features, see Supplemental Table 1. Except for BMDExpress [6, 7, 8], which has documented guidelines in NTP et al. [9], none of the other approaches have established an approach for determining a genomic POD linked to gene- and pathway-level biological alterations. Consequently, BMDExpress [6, 7, 8] is still the most widely utilized approach for GDR modeling and will be the focus of our comparison. The platform uses a multi-step framework of parametric modeling and bioinformatics tools to characterize DR patterns and evaluate alterations in cellular function [6, 7, 8, 9]. It addresses common challenges by using genomic data in risk assessment, but some of the assumptions made may not be appropriate given the observed data. In cases where cytotoxicity occurs in high-doses, a non-monotonic (i.e., curve with umbrella orderings) is possible [10]. For such shapes, parametric models are typically inadequate, because most of the models assume monotone changes in expression [9, 11, 12, 13], which may lead to removing genes with biological relevance due to poor model fit. Another limitation of BMDExpress is the enrichment analysis used to detect altered cellular functions and obtain genomic PODs. The enrichment analysis matches genes remaining after model fitting to a list of curated gene sets, such as the Hallmark gene sets [14, 15, 16], determined a priori. The remaining genes' biological relevance is evaluated by their association to these sets. However, relying on curated gene sets disregards the underlying correlation structure of gene responses. These correlations may clarify understanding of mechanisms, such as co-regulation, related to developing a toxicological event.

Allocco et al. [17] demonstrated grouped genes with highly correlated expressions are likely to indicate co-regulation by a common transcription factor. Clustering-based upon similar dose-response patterns may identify co-regulation. Studies have shown Bayesian clustering methods employing Dirichlet Process mixture (DPM) models are preferred for identifying clusters of co-regulated genes, see Dahl [18], Fraley and Raftery [19], Medvedovic and Sivaganesan [20], Medvedovic et al. [21], Yeung et al. [22] among others. These methods do not always consider relationships among samples that may influence co-regulation patterns. DPM clustering methods that incorporate the relationship among samples using latent variables are shown to improve clustering, see Bar-Joseph et al. [23], Kirk et al. [24], Liu et al. [25], Lock and Dunson [26], Luan and Li [27, 28], McDowell et al. [29], Park and Kyung [30], Savage et al. [31] and references therein. However, there is not currently an approach that accounts for the underlying correlation structure of co-expressed genes with similar DR patterns.

Despite the advantage of a latent DPM model that simultaneously includes DR modeling and gene clustering, it is well known these approaches are computationally intensive. Our method, Aggregated Local Extrema Splines for High-throughput dose-response Analysis (ALOHA), first fits DR data, then uses a simpler DPM to cluster genes sharing similar DR patterns. From the DPM analysis, the method uses individual clusters to define PODs for co-regulated gene networks. The first stage of our analysis pipeline fits a flexible DR model using Bayesian local extremum splines [32], which is a shape-constrained non-parametric spline approach providing flexibility in fitting monotone, umbrella, and 'J' shaped responses. The second stage of our analysis inputs these fits into the context-specific infinite mixture model (CSIMM) developed by Liu et al. [25], and clusters

genes with similar expression patterns. By utilizing gene fits, we incorporate GDR data's dose-dependent nature for clustering while removing the scatter typically observed in such experiments. Biologically enriched gene sets are evaluated to obtain a POD possibly related to adverse biological responses.

ALOHA consists of two parts (i) shape-constrained non-parametric modeling on individual genes discussed in §2.1 and (ii) clustering genes with similar DR patterns covered in §2.2. In §2.3, we outline the application of our method to genomic dose-response data from HepaRG cells exposed to various compounds [33] and compare the results with the current GDR modeling and clustering approaches, respectively.

2 Methods

2.1 Dose-Response Modeling

We observe n genes across D dose groups, with each dose group having K replicates such that $K \geq 1$. For simplicity of exposition, K is constant across each dose group but could change at each dose level. Thus, we have $p = D \times K$ total observations for each gene. Let Y be the $n \times p$ matrix of these observations, such that the i^{th} gene is represented by the row-vector $y_i = (y_{(i,1)}, \ldots, y_{(i,p)})$. The expression of gene i over the experimental doses can be represented as:

$$y_i = f_i(x) + \epsilon_i \tag{1}$$

where $x = (x_1, \ldots x_p)$ is a vector of doses, f_i is the DR function, and $\epsilon_i = (\epsilon_{(i,1)}, \ldots, \epsilon_{(i,p)})$ is a vector of measurement errors. Here, we assume $\epsilon_i \sim N(0, \sigma^2 I)$, where I is a $p \times p$ identity matrix and σ^2 is some unknown variance. We use Bayesian shape-constrained non-parametric models to estimate the dose-response curve ($f_i(x)$) for all n genes.

2.1.1 Shape-constrained Non-parametric Modeling

For shape-constrained non-parametric modeling, we use a variant of Bayesian local extrema (BLX) splines [32]. This approach places restrictions on the shape of the model space allowing functions with no more than H extrema (i.e., change-points or local minima/maxima). We define the BLX spline to model functions having at most two extrema. That is, they may increase and then decrease (i.e., have an umbrella ordering), decrease and then increase (i.e., have a 'J' shaped ordering), or increase then decrease and increase again (i.e., have a tilde – '∼' – shaped ordering). In practice, we do not consider the '∼' shape plausible, but it is required to consider 'J' and umbrella orderings simultaneously. When a '∼' shape is observed, it typically represents no dose-response relationship, and as a result, the POD for the given gene is greater than the maximum tested doses.

For gene i, we estimate the dose-response, f_i, with a BLX spline defined over the fixed knot set $T = \{0, 0.5, 1\}$, where knots are chosen to represent the background, midpoint and maximum dose. The approach is different from Wheeler et al. [32] who consider a prior over a set of knots dense in $[0, 1]$. For our application, there are not enough dose groups to use this prior, and the fixed set T is adequate. With $H = 2$,the BLX spline model for f_i is:

$$f_i(x) = \beta_0 + \Sigma_{k=1}^{4} \beta_k B_k^*(x), \tag{2}$$

where β_0 is the intercept, $B^*(x)$ is a BLX spline basis function of degree 2, and β_k for $k = \{1, \ldots, 4\}$ are spline coefficients. For more information on construction, we refer the reader to Wheeler et al. [32].

We assume dose-response relationships are initially increasing by setting the fixed integer of the BLX spline basis to one ($M = 1$). The function models monotonic increasing, monotonic decreasing, J-shaped, umbrella shaped, or '∼' shaped curves (see Supplementary §3.1 and Figures 2 - 7 therein). All model parameters are estimated as in [32].

Though the dose-response of genes may differ, unlike parametric modeling, the BLX spline specification is the same for each gene. This model similarity facilitates clustering on the estimated parameters (i.e., one can cluster the dose-response curve on the BLX fitted parameters). The BLX approach is different from parametric models, where the parameters all have different interpretations making clustering between models difficult.

2.1.2 Parametric Modeling

For the parametric modeling approach, we estimate f_i using all available models, excluding the Polynomial 3 and 4, from BMDExpress (version 2.3) [6, 7, 8]. We refer the reader to Auerbach [6], Phillips et al. [7], Yang et al. [8] for the model equations. Higher-order polynomials in most cases will fit dose-response data similarly to a quadratic model with fewer degrees of freedom and therefore were excluded from our set. This model suite captures a majority of biological response shapes thought to arise in practice. We use the model with the best fit, determined using the Aikake Information Criterion (AIC) Akaike et al. [34], Akaike [35], Sakamoto et al. [36], for inference and comparison to the BLX fit.

Other than using a quadratic dose-response curve, the parametric model suite cannot represent 'J' or umbrella-shaped orderings, which cytotoxic doses can induce. A quadratic model – the only parametric model that can evaluate a 'J'-shaped response – assumes the response is approximately parabolic forcing symmetry. Symmetry around the extrema occurs infrequently. Dose-response relationships that enforce symmetry may be inadequately represented in the low-dose region.

2.1.3 BLX Spline and Parametric Model Comparison

As a comparison between model fitting approaches, we compare the mean square predicted error (MSPE) between both estimates using a leave-one-out cross-validation (LOOCV) approach removing one non-zero dose group at a time, i.e.,

$$MSPE_{(i,m)} = \frac{\Sigma_{d=1}^{D}\Sigma_{j=1}^{n_d}(y_{ik} - f_{(-d)im}(x_j))^2}{\Sigma_{d=1}^{D} n_d} \tag{3}$$

where i denotes the gene, m denotes the best parametric or BLX spline fit, d denotes the dose group (excluding the control group), n_d is the number of samples in d, j denotes the sample, x_j denotes the dose administered for sample j, and $f_{(-d)im}$ is the model fit without dose group d. The MSPE measures how well each models predicts unobserved data, and it provides a comparison between each approach.

2.2 Gene Clustering

We identify co-expression clusters sharing similar DR patterns using the context-specific infinite mixture models (CSIMM) clustering algorithm developed by Liu et al. [25], which allows us to pre-specify sample contexts and evaluate global and local patterns. We modify this approach to model-fit data.

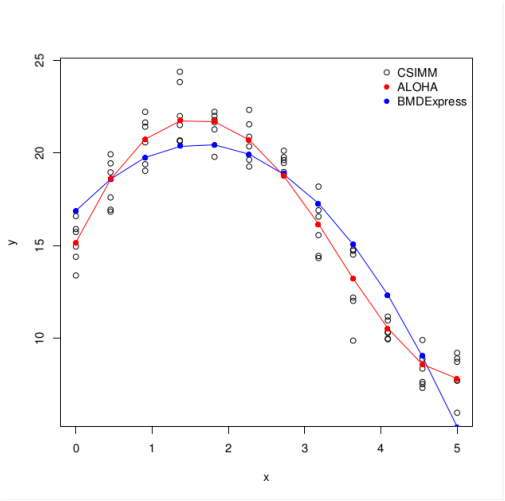


Figure 1: Simulated dose-response data and model fits demonstrating data used for gene clustering approaches. Black points illustrate the observed data values used as input in CSIMM. Red points illustrate the BLX estimated data values used as input in ALOHA. Blue points illustrate the estimated values from a quadratic model fit. $f(x) = 7sin(0.3\pi x) + 15 + \epsilon_i$, where $\epsilon_i \sim N(0, 1)$.

After dose-response modeling, the ALOHA clustering algorithm uses the BLX estimated dose-response data (e.g. red points in Figure 1) as input for the CSIMM MCMC [25]. We compare the ALOHA clustering algorithm with two other alternatives, including CSIMM and clustering on BMDExpress model-fit data. CSIMM uses the observed expression data (e.g. black points in Figure 1) and BMDExpress uses the estimated dose-response data from the best parametric models (e.g. blue points in Figure 1). All three algorithms used are given in Table 1. We compare these clustering results to evaluate their ability to capture biologically relevant co-expression clusters with coherent DR patterns and potency estimates (i.e., benchmark dose) for enriched gene sets. For further details on the CSIMM MCMC and the posterior-pairwise probability (PPP) matrix we refer the reader to §3.3 in the Supplement and Liu et al. [25].

Strategy (Abbreviation)	Data	CSIMM Contexts
Original DR Data (CSIMM)	Normalized expression data with all replicate samples	Dose groups $d \in \{1, \ldots, D\}$ $M = D$
BLX Estimated DR Data (ALOHA)	Average estimated expression for each dose group using individual BLX fits	Dose groups $d \in \{1, \ldots, D\}$ $M = D$
BPF Estimated DR Data (BMDExpress)	Average estimated expression for each dose group using individual best parametric fits (BPF)	Dose groups $d \in \{1, \ldots, D\}$ $M = D$

Table 1: Three approaches for clustering genes to identify co-expression clusters with similar DR patterns.

In addition to the clustering comparisons, we investigate the Hallmark gene sets identified in the pathway enrichment analysis used in BMDExpress. Enrichment analyses are qualitatively different, and we do not compare the enrichment scores from the two approaches. Our focus was on comparing the identified gene sets, the gene set level potency estimates (i.e., BMD and its interval) obtained from the clustering approach compared to BMDExpress, and the core genes contributing to the gene set level BMD.

2.3 Application

2.3.1 Data

To investigate ALOHA, we use datasets on three compounds tested by Ramaiahgari et al. [33]. Of these compounds, Benzo[α]pyrene (BAP) and Chlorpromazine (CPZ) were the positive control chemicals, and potassium chloride (KCl) was the negative control. We evaluate the positive control compounds with and without cytotoxic doses, i.e., "all doses" (AD) and "cytotoxic doses removed" (CDR), respectively. Gene expression was measured using the S1500+ TempO-Seq platform [33, 37, 38]. All datasets' raw RNA read counts are normalized to a log counts per million scale ($log_2(CPM + 1)$). Doses were transformed to a $log_{10}(dose + 1)$ scale for approximately equally spaced dose groups. The probes that did not meet a minimum read count (MRC), i.e., median $log_2(CPM + 1)$ response equal to 0 and variance less than 0.1, were removed from further analysis.

2.3.2 Application of BMD modeling

We fit each probe passing the MRC criteria individually as described above. For the BLX fits, a total of $50,000$ MCMC iterations were taken with the first $10,000$ disregarded as burn-in, which is are the values used in [32]. Trace plots were analyzed and showed quick convergence (usually within 200 iterations) and reasonable mixing. For parametric models, all modeling is performed with BMDExpress (version 2.3) [6, 7, 8].

Estimates of the DR are used to compute a benchmark dose (BMD) for each probe. BMD estimates for both up and down-regulation are computed by solving for BMD in:

$$BMR = \pm \frac{f(BMD) - f(0)}{\sqrt{\frac{1}{\tau}}}. \tag{4}$$

Here, τ^{-1} is estimated variance, BMR is the benchmark response, $f(0)$ is the response at the control dose [39]. In this definition, the BMR indicates a standard deviation change in gene expression from the control. For our application, we consider set $BMR = 1$, which is consistent with previous analyses with these datasets Ramaiahgari et al. [33] and NTP et al. [9].

For monotone curves, we compute the BMD using (4). Most parametric models in BMDExpress assume expression responses are monotone. In the non-monotone cases, the direction of adversity is determined using a linear trend test [6].

For BLX estimates, it is possible an up- and down-regulated BMDs exist. In these cases, we compute a BMD for both directions of adversity, where the curve's shape determines an adverse response. For DR curves having two extrema, we consider the expression changes in the first "bump" to be an artifact of the fitting procedure. Consequently, we choose the largest BMD found after the maximum as the estimate. For DR curves that are 'J' or umbrella-shaped, there may be two solutions for the BMD. In these cases, the smallest of the estimates is chosen to be the BMD

estimate. When no solution to (4) exists in the dosing interval, we set the BMD to be one plus the maximum experimental dose (i.e. $x_D + 1$) to place the BMD above the maximum experimental dose. If up- and down-regulated BMDs exist, we use the smallest of these BMDs as the estimate.

Confidence limits in the BMD are estimated relative to the estimation method. For the BLX spline, calculated using MCMC, we take the 2.5^{th} and 97.5^{th} quantiles of the BMD distribution to estimate the BMDL and BMDU, respectively. For parametric models, estimated using maximum likelihood, the limits are estimated using the profile likelihood. Point and interval estimates of each BMD are transformed from the log-scale to the original dose scale before calculating the ratio of the upper and lower confidence limits (BMDU/BMDL). The BMDU/BMDL ratio is a measure of reliability in the BMD estimate of each probe. Smaller BMDU/BMDL estimates indicate increased certainty in BMD estimates and more reliable fits. We define probes with reliable expression activity as those having a (i) BMD less than the maximum dose and (ii) 95% confidence limit BMDU/BMDL less than 40, which are consistent with the reliability criteria proposed by NTP et al. [9]. When fitting parametric models, the best model fits must also have a goodness-of-fit p-value > 0.1, which indicates no significant difference between the observed expression values and those estimated by the model. Probes meeting this criterion have reliable expression activity and remain in the analysis. This is consistent with NTP et al. [9].

2.3.3 Gene Clustering

After model fitting, model data from reliable probes are used to obtain the inputs, described in §2.2, to the CSIMM clustering algorithm.

In the TempOSeq platform, multiple probes target one gene designated by a common EntrezID, but their targets may be different gene isoforms displaying different co-expression patterns. Consequently, we do not average these probes' responses, and we remove probes without a designated EntrezID or multiple EntrezIDs.

For CSIMM, the MCMC sampler is run for 50,000 iterations, with the first 10,000 samples disregarded as burn-in. The clustering strategies are assessed by their ability to separate DR trends into co-expression clusters, where separation is assessed using a heat map. Biological enrichment of clustering results is analyzed using Cluster Enrichment Analysis (CLEAN) [40]. This approach compares the level of functional coherence in clusters based on previously curated biological pathway lists and potency estimates for enriched gene sets.

To compute CLEAN scores, we use the human Hallmark Gene List and the Computational Cancer Gene Neighborhood Gene List (C4:CGN or CGN) from the Molecular Signatures Database (MSigDB) to evaluate the enrichment of gene clusters [14, 15, 16, 41]. Gene sets containing at least three genes are used for the enrichment analyses consistent with NTP et al. [9]. A gene set is enriched for a cluster if Fischer's exact test Benjamini-Hochberg False Discovery Rate adjusted p-value (FFDR) is less than 0.01. We consider gene clusters with the smallest FFDR p-value for at least one enriched gene set to be the most relevant. For each enriched gene set to be considered for biological interpretation, based on the criteria from NTP et al. [9], the most relevant cluster representing it must contain at the minimum 3 of its genes and populate the set by at least 5%. Though we include the NTP et al. [9] criteria for defining active pathways since we include the FFDR p-value threshold criteria for the clustering approaches, our criteria are fundamentally different from those used in Phillips et al. [7], Yang et al. [8], NTP et al. [9].

Potency estimates for enriched gene sets are obtained using the average BMD estimates from the core genes representing each enriched gene set. We define the core genes to be those in the

most relevant cluster for an enriched gene set belonging to that set. For parametric fits, potency estimates are the median BMD estimate of the enriched gene sets. For the clustering results on the normalized expression data and the BLX spline fits we use the posterior mean BMD estimate for each gene. We obtain the lower and upper bounds on the gene set BMD – BMDL and BMDU, respectively – using the 0.025 and the 0.975 quantiles of the best BMD estimates for core genes to obtain a 95% credible interval.

2.3.4 BMDExpress Functional Classification Analysis

After parametric model fitting and post-model filtering of probes, we performed functional classification analyses in BMDExpress using the human Hallmark Gene List from the Molecular Signatures Database (MSigDB) to compare gene set level BMD estimates with those from each of the clustering approaches [6, 7, 8, 14, 15, 16, 41]. It should be noted, prior to the functional classification analysis, BMDExpress consolidates multiple probes mapping to the same EntrezID. The BMD of these probes are averaged together to obtain one BMD value representing their common EntrezID [6, 7, 8].

From the pathway enrichment results from BMDExpress we obtain a gene set level potency estimate for each of the gene sets. For our analysis we use the median BMDL, median BMD, and median BMDL as estimates of the potency estimates for each of these gene sets. We compare these estimates with those obtained from the most biologically relevant cluster for each of the enriched Hallmark gene sets found with various clustering approaches.

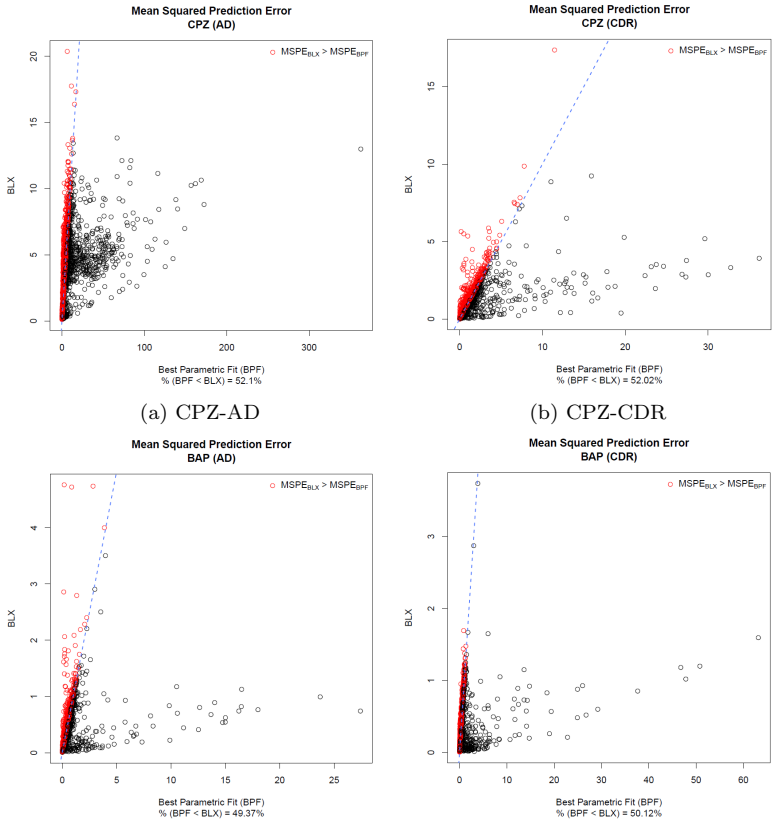
3 Results

3.1 Model Fitting

3.1.1 Model Predictions

We evaluate each modeling approach's ability to obtain the best estimate of the dose-response curve using a leave-one-out cross-validation (LOOCV) prediction. Figure 2 shows the mean squared prediction error (MSPE) for the parametric and BLX fits for each probe attaining the minimum response criteria in §2.3.1 and having less than 75% of their observed responses equal to zero. A smaller MSPE indicates a better prediction. In this figure, the blue dashed line shows situations where the MSPE is the same for a given probe. Black dots, or points to the right of the line, show results where ALOHA outperforms the parametric approach. Red dots, or points to the left of the line, show results where the best parametric model is superior. Here, most of the red points are close to the blue line, which indicates the method fit and predictions are essentially the same for that probe; however, the black points show a greater right skew. ALOHA does a better job describing the dose-responses for these probes, which occurs for approximately 50% of the probes. Further, we observe in general the maximum of the BLX spline MSPEs is an order of magnitude smaller than the maximum of the MSPEs for the best parametric fit, which gives credence to the argument that ALOHA is better representing the true underlying dose-response curve.

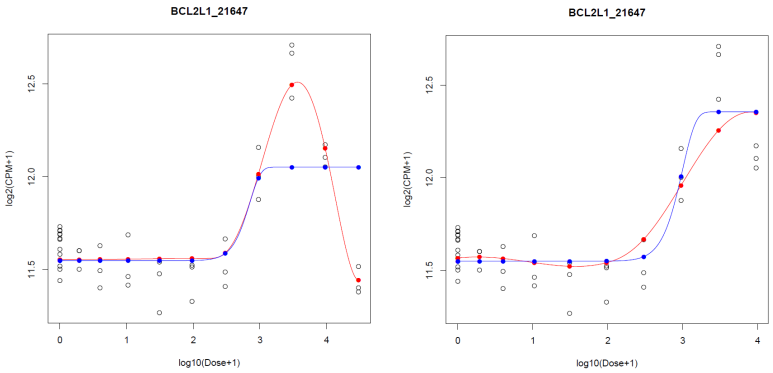
Figure 3 shows why ALOHA is doing better modeling these data and compares the curve fit for a probe from Benzo-[α]-pyrene. The left figure has all doses and while the right one has the highest cytotoxic dose removed. In the left plot, we observe a clear non-monotonic response which the BLX spline fit captures; however, the Exponential 5 model misses this behavior. The right plot shows the fits where the highest dose group is removed since it is cytotoxic. Though this dose's



(a) CPZ-AD (b) CPZ-CDR

(c) BAP-AD (d) BAP-CDR

Figure 2: Scatterplots with the mean squared prediction error (MSPE) for the best parametric model fit (BPF) (x-axis) and the BLX spline (y-axis) for each MRC filtered probe modeled with both methods and having less than 75% percent of observations with a response equal to zero. The blue dashed line indicates where the BLX and BPF MSPE are equivalent. Red points indicate probes with a smaller BPF MSPE than the BLX MSPE. Black points indicate probes with a BLX MSPE smaller than or equal to the BPF MSPE.



(a) All Doses (b) Cytotoxic Doses Removed

Figure 3: Probe passing the reliability criteria with BLX spline model fitting but fails with BMDExpress (version 2.3) parametric modeling for Benzo-[α]-pyrene all doses (left plot) and cytotoxic doses removed (right plot). In both plots the curves display the BLX spline (red) and best parametric model fit from BMDExpress (blue).

removal is to mitigate issues of non-monotone relationships, there is still a roughly non-monotonic shape. In both of these cases, the model constraints of the best parametric fit, i.e. the Exponential 5, induces a plateau in the high doses, which is unable to capture the peak response and the change in response resulting in a poor fit to the observed data (i.e. fit p-value $= 1e-4$ for both Benzo-[α]-pyrene with and without cytotoxic doses). On the other hand, our approach's flexibility captures the non-monotonic response when all doses are included, and it splits the difference between the response in the two highest dose groups when cytotoxic doses are excluded. There is no longer sufficient evidence to determine if there is a true non-monotonic response.

3.1.2 Probe Retention

Out of 2977 probes in the S1500+ TempO-Seq platform, there were 2925, 2703, 2684, 2928, and 2922 probes used in the analysis for Potassium chloride, Benzo-[α]-pyrene with and without cytotoxic doses, and Chlorpromazine with and without cytotoxic doses, respectively, based on MRC filtering described in §2.3.1. Table 2 shows the number of modeled probes passing the criteria described in §2.3.2 with the two modeling approaches. Other than Benzo-[α]-pyrene with all doses and Chlorpromazine without cytotoxic doses, more probes passed with parametric modeling compared to BLX spline modeling. Many of the same probes were retained between the two modeling approaches.

3.2 Gene Clustering Comparison

3.2.1 Separation of Dose-Response Curve Shapes

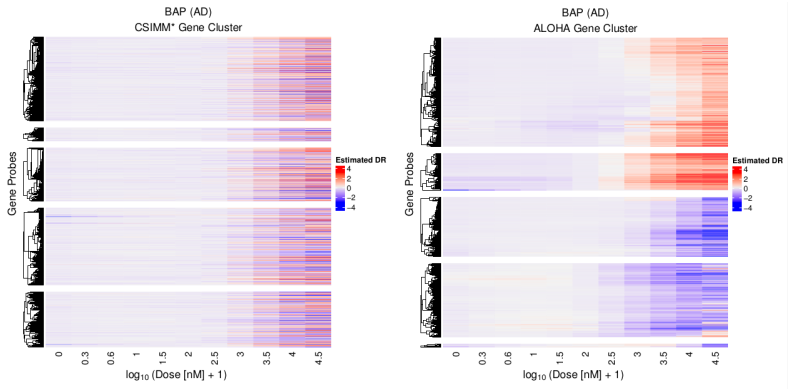
For both Benzo-[α]-pyrene and Chlorpromazine, normalized expression values for up and down-regulated genes were better separated with the CSIMM (see Supplementary Figures 19-22). In this case, the BLX spline parameter estimates showed poor separation, which implies dissimilarity in dose-response curves for genes in the same cluster. For the ALOHA approach, the BLX parameter

Compound	BMDExpress	ALOHA	Shared
Potassium chloride	56	0	0
Benzo[α]pyrene (AD)	877	954	464
Benzo[α]pyrene (CDR)	973	719	413
Chlorpromazine (AD)	1168	1065	570
Chlorpromazine (CDR)	849	862	362

Table 2: Number of modeled probes passing model-fit criteria in BMDExpress (parametric models) and ALOHA (BLX spline model), respectively. Columns 2 & 3 provide the number of probes passing the criteria in §2.3.2 for each dataset. Column 4 provides the number of probes in common between the modeling approaches. [AD - all doses; CDR - cytotoxic doses removed].

estimates were separated, which indicates greater similarity in the shape of dose-response curves for genes belonging to the same.

Separation of BLX estimated dose-response data were also evaluated to confirm the coherence of dose-response trends of genes in the same cluster. We chose to cut the posterior pairwise dendrogram such that five representative clusters resulted. The chosen cut-off was arbitrary, but was done for exposition of dose-response trend coherence within clusters. For both Benzo-[α]-pyrene and Chlorpromazine, the ALOHA clusters had a high level of coherence in the dose-response patterns compared to CSIMM (see Figure 4 & Supplementary Figures 23-26). We also calculated several internal validation scores for the five resulting clusters using the Euclidean distance as the measure of dissimilarity [42, 43, 44, 45] (see Table 3 & Supplementary Tables 2-5). All validity scores were consistent with greater coherence in dose-response trends of ALOHA gene clusters compared with CSIMM. These results indicate ALOHA is better suited for identifying gene groups with co-regulated dose-response patterns.



(a) CSIMM (b) ALOHA

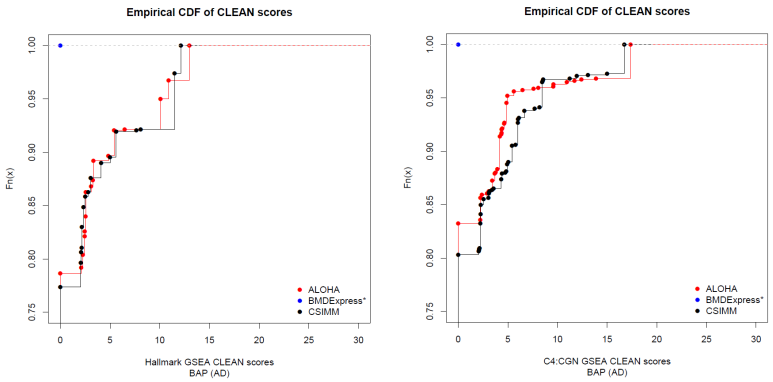
Figure 4: **BAP AD** - Heatmaps displaying clustering results applied to BLX estimated dose-response data showing visual separation of DR trends. (Left - *CSIMM*) CSIMM clustering on observed normalized expression data with all replicates in each dose group. (Right - *ALOHA*) CSIMM clustering on BLX estimated dose-response data.

Benzo-[α]-pyrene (AD)	ALOHA	CSIMM
Avg. Distance Between Clusters	3.84	3.23
Avg. Distance Within Clusters	1.68	3.21
Dunn Index	4.4e -3	9.3e -4
Avg. Silhouette Width	0.20	-0.06
Entropy	1.37	1.49

Table 3: Scores measuring the internal coherence of dose-response trends in the five clusters resulting from ALOHA and CSIMM in the Benzo-[α]-pyrene dataset with all doses, see Figure 4. Larger values for average distance between clusters, Dunn index, and the average silhouette width indicate better separation of genes into similar dose-response groups. Smaller values for the average distance within clusters and entropy indicate more coherence between genes clustered together.

3.2.2 Functional Coherence in Clusters

CLEAN scores [40] were used to compare the quality of functional coherence in the resulting co-expression clusters. Figure 5 compares the empirical cumulative distribution function (ECDF) of CLEAN scores for all clustering approaches on Benzo-[α]-pyrene with all doses. For BMDExpress clustering, there were fewer genes with a non-zero CLEAN score. Further, the CLEAN scores were generally smaller, e.g., the Benzo-[α]-pyrene with all doses results showed no biological significance with BMDExpress clustering. Generally, BMDExpress was the worst of the three approaches at identifying biologically relevant clusters of the three approaches (see Supplementary Figures 27-30). There was little separation between the ECDF curves for CSIMM (black) and ALOHA (red), except in the case of Chlorpromazine with all doses. In that case, there was no clear delineation between any of the curves, which indicates that both approaches obtained similar levels of functional coherence in co-expression clusters (see Supplementary Figures 27-30). Enrichment scores from the most relevant cluster of the enriched gene set showed biological relevance of the clustering approaches was similar (see Supplementary Figures 31-38). Although CSIMM and ALOHA had similar functional coherence, as seen below, CSIMM did not produce clear co-regulated dose-response functions, i.e., the dose responses for genes in a given cluster were not similar (see Figure 4).



(a) Hallmark (b) Cancer Gene Neighborhoods

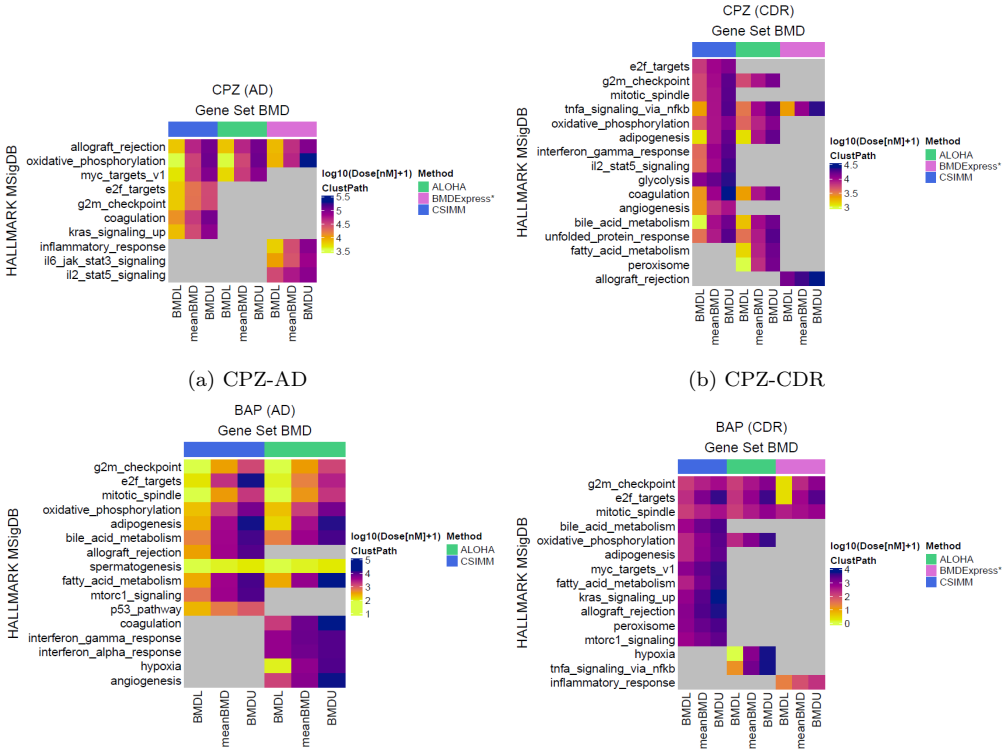
Figure 5: CLEAN score empirical cumulative distribution function (ECDF) plots for Benzo-[α]-pyrene including all doses (BAP-AD), with MSigDB's Hallmark (right) and Cancer Gene Neighborhood (CGN) (left) gene lists on the normalized expression (CSIMM – black), BLX estimated expression (ALOHA – red), and best parametric fit estimated expression (BMDExpress – blue) approaches.

3.2.3 Benchmark Dose (BMD) Coherence

Figure 6 shows the BMD estimate, and the associated 95% credible interval estimate for each enriched Hallmark gene set from the clustering approaches. For enriched gene sets shared between CSIMM and ALOHA, the gene set level BMD estimates were similar. There were a few exceptions to this behavior. For example, in gene set "e2f_targets", the ALOHA cluster BMD interval was smaller, indicating improved BMD coherence. Generally, gene set BMD intervals for BMDExpress clustering were wider than those from ALOHA.

Similar trends were observed for the enriched Cancer Gene Neighborhood gene sets shared between the clustering approaches, except for the Chlorpromazine with all doses which had tighter BMD intervals for many of the gene sets enriched with BMDExpress clustering (see Supplementary Figures 39 - 41). Though there were similar BMD estimates between CSIMM and ALOHA, the gene clusters from CSIMM were not representative of a particular dose-response pattern. As a result, BMD estimates were inconsistent about the direction of regulation, and they provided contradictory information about the adverse direction.

Gene set BMD estimates from the clustering approaches were also compared with those estimated after functional classification in BMDExpress described in §2.3.4. Overall, these comparisons showed more genes contributed to the BMD estimate from the current BMDExpress approach (see Figure 7). With ALOHA and the other clustering approaches, individual gene BMDs were tighter grouped than those from the current BMDExpress approach. Confidence intervals for gene set BMDs with the clustering approaches also overlapped or contained those intervals from the current BMDExpress approach, which indicated clustering resulted in gene groups that produced BMD estimates consistent with BMDExpress.



(a) CPZ-AD (b) CPZ-CDR

(c) BAP-AD (d) BAP-CDR

Figure 6: Heatmaps comparing the posterior mean BMD and 95% credible intervals (i.e. BMDL and BMDU) estimates (columns) for enriched Hallmark gene sets [14, 15, 16] (rows) across the three clustering methods. The BMDs are estimated using core genes in the most relevant cluster, as defined in §2.3.3, mapping to the respective gene sets. (a) Chlorpromazine with all experimental dose groups included, (b) Chlorpromazine with cytotoxic dose groups removed, (c) Benzo-[α]-pyrene with all experimental dose groups, (d) Benzo-[α]-pyrene with cytotoxic dose groups removed. Gray blocks within the heatmap indicate which gene sets are not enriched for a clustering approach. Bars on above the columns indicate the various clustering approaches, i.e. CSIMM (blue), ALOHA (green), and BMDExpress clustering (magenta, described in §2.2.

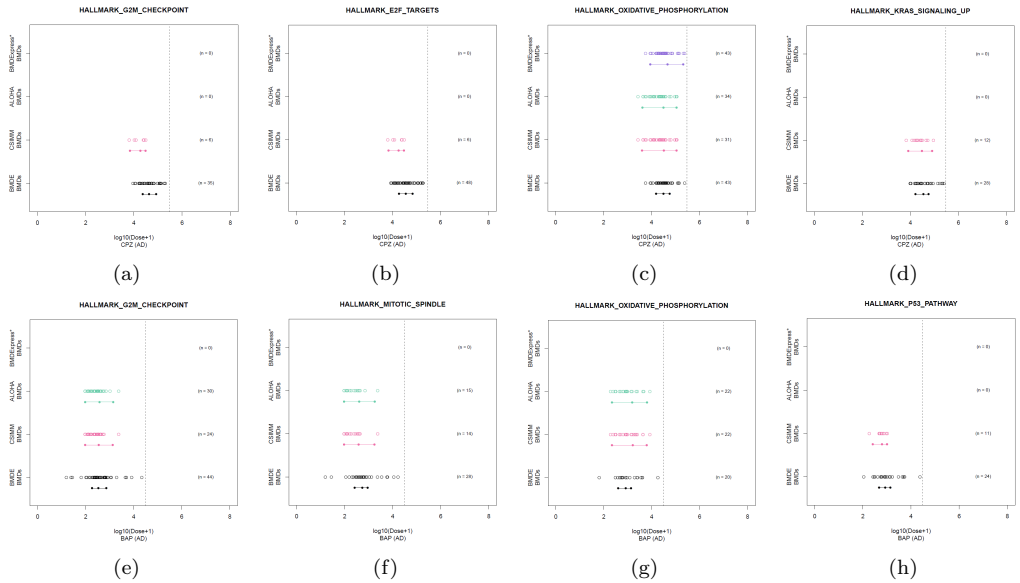


Figure 7: Gene set level BMD intervals from BMDExpress (BMDE), using functional classification without clustering, and the clustering approaches on Chlorpromazine and Benzo-[α]-pyrene with all doses (top and bottom, respectively). Points show the best BMD estimate of individual core probes and the number of core probes displayed on the right (see Supplementary Tables 6-13 for details about these probes). (Note: For BMDExpress, probes mapping to the same EntrezID had their BMD estimates averaged and share a single BMD value in the results, but we consider these probes individually in the count to be consistent with core probe counts with clustering.). The vertical dashed line indicate the maximum experimental dose and the line segments below the individual BMDs display the interval (i.e. BMDL to BMDU) and central tendency for the gene set level BMD as described in §2.3.3 & §2.3.4.

4 Discussion

The use of genomic data in risk assessment relies on developing new methods that account for biological phenomena affecting the underlying data structure. Here, we developed a genomic dose-response (GDR) modeling approach, called Aggregated Local Extrema Splines for High-throughput Dose-Response Analysis (ALOHA), that utilizes model-fit data from Bayesian Local Extremum (BLX) splines and context-specific infinite mixture models (CSIMM) to obtain gene clusters with similar dose-response curves.

For ALOHA, the increased number of probes with a minimum read count (MRC) filtering did not impact the number of probes deemed to have dose-response (DR) information. For example, few probes were removed in the Potassium chloride dataset before fitting, but after DR modeling, there were no remaining probes (see Table 2). This result corresponds to Potassium chloride being a negative control. For probes passing only with parametric modeling, extreme response observations impacted the fit, resulting in an unrealistic fit to the data but an acceptable benchmark dose (BMD) estimate. Here, these observations did not impact the BLX fit, which resulted in greater uncertainty in the BMD estimate or a flat dose-response curve. For probes passing only with BLX, the model

flexibility captured non-monotone responses in high doses where cytotoxicity may occur (see Figure 3). In these cases, parametric models were unable to fit these observations due to model constraints, resulting in poor fit estimates.

Comparing our method's ability to describe the genomic dose-response data with BMDExpress (version 2.3), using a leave-one-out cross-validation (LOOCV) method, demonstrated little difference between the two approaches in the case of monotone dose-response curves. However, in cases where cytotoxicity induced a non-monotonic response, ALOHA produced smaller mean squared predicted error (MSPE) estimates, thus outperforming the predictive ability of BMDExpress (see Figure 2). For many of these genes, the parametric models were unable to appropriately capture the observed response at high doses due to the assumption of monotonicity and model constraints. This assumption led to poor model fits and removal from the analysis. When cytotoxicity is present, a common approach to prevent discarding genes from the analysis is to remove the doses producing the cytotoxic behavior. Though this may resolve some of the genes with non-monotone behavior, it does not guarantee this across all genes. Figure 3 illustrates this case, where removing the high doses in the Benzo-[α]-pyrene data set did not remove the downturn for "BCL2L1_21647". Since the identification of cytotoxic dose groups is an arbitrary decision made by the modeler, any removal approach may result in responsive genes being arbitrarily removed from the analysis. ALOHA does not suffer from this difficulty and retains non-monotone dose-responses for further analyses.

Except for Benzo-[α]-pyrene and Chlorpromazine, with and without cytotoxic doses respectively, more probes passed the post-model fit criteria with BMDExpress compared to ALOHA. Investigation of individual probe fits showed BMDExpres inappropriately identified a portion of these as active (i.e., distinct and reliable DR trend). Here, probes fell into one of two categories. First, observed responses were noisy with a few extreme observations (see Supplementary Figure 10). Second, the response in high doses was split between active and dead cells (i.e. zero expression) – (see Supplementary Figure 11). In the first case, ALOHA estimated a flat curve. In the second case, ALOHA split the difference between active and dead samples' responses in each dose group. The cytotoxic dose range uncertainty did not provide enough evidence to classify the gene as active; either the curve was flat, or there was uncertainty in the BMD estimate near the maximum experimental dose. In both of these cases, constraints in the parametric models were influenced by the most extreme response values from sensitive samples, which produced a curve resulting in a BMD estimate (see examples in Supplementary Figures 10 & 11). Whereas ALOHA was robust to sensitive samples, and the overall fit was unaffected without more evidence of an extreme change in response.

ALOHA's ability to filter active genes is shown with potassium chloride, a negative control compound. ALOHA removed all probes from the analysis, unlike BMDExpress, which retained several genes as active. Some probes in the positive control compound datasets, i.e. Benzo-[α]-pyrene and Chlorpromazine, passed the model-fit criteria with BMDExpress but not with ALOHA. Here, the posterior distribution of BMD estimates from ALOHA was either too uncertain or bimodal with one-mode larger than the maximum dose. In these cases, the 95% credible interval or the posterior mean BMD estimate was unable to pass the recommended criteria presented in the NTP guidance document [9]. Even without differential gene expression (DEG) filtering and increased data, ALOHA retained active probes (i.e. probes with distinct and reliable DR trends) while removing inactive ones.

Assessing ALOHA's clustering behavior, we first compared the separation of dose-response patterns in the normalized expression responses and BLX spline parameter estimates – a proxy measure of model fit shapes – between ALOHA and CSIMM [25] (see Supplementary Figures 19 -

22). ALOHA showed no clear separation in the regulation trends of normalized expression data. However, there was a clear improvement in the separation of the BLX parameter estimates, which implies DR patterns of individuals' genes are more coherent in ALOHA clusters than CSIMM clusters. To verify these results we compared the separation of the BLX estimated dose-response data, which confirmed the ALOHA clusters contained genes with more coherent DR patterns (see Figure 4 or Supplementary Figures 23 & 26).

Comparison of the functional coherence of the resulting gene clusters demonstrated ALOHA performed similarly to CSIMM. CSIMM did obtain more enriched gene sets than ALOHA. However, the clusters enriched for these gene sets were not necessarily reflective of genes with common DR patterns. Members of a gene set may have different DR curve shapes. Consequently, ALOHA breaks these sets into different clusters, whereas CSIMM clusters these genes regardless if they have conflicting DR patterns. ALOHA also showed clear improvements in the biological relevance of clusters obtained compared to BMDExpress clustering. These results were likely the result of qualitative differences in probes passed to the clustering process. Alternatively, the limited flexibility of parametric models and inactive probes' retention may have made it difficult for CSIMM to discern clear DR co-expression clusters using BMDExpress model fit data. In these cases, additional information is needed to distinguish DR patterns better.

Gene set BMD estimates from ALOHA clustering were tighter in most cases than those from BMDExpress clustering. The only exception was for the Cancer Gene Neighborhood (CGN) gene sets enriched in Chlorpromazine with all doses. This improvement was due to clustering probe fits with more similar DR curves producing consistent BMD estimates. BMD intervals from ALOHA also largely overlapped with intervals estimated using the functional classification results from BMDExpress, which places all probes passing the model-fit criteria into their respective gene sets. Though BMDExpress generally produced tighter BMD intervals than ALOHA, the qualitative difference between the two approaches when estimating the confidence intervals was likely the source of this disparity in precision. This result is supported by the observed differences in interval estimates between BMDExpress clustering and the current BMDExpress approach. Because both use functional classification, they should be relatively similar (see "OXIDATIVE_PHOSPHORYLATION" in see Figure 7). Individual gene BMDs that contributed to the overall gene set BMD were also more diffuse than ALOHA when using the current BMDExpress approach. This result shows that even with the addition of clustering, ALOHA was able to obtain BMD estimates consistent with the current approach, BMDExpress, while identifying gene groups with more coherent BMD estimates.

Genomic dose-response (GDR) analyses currently separate modeling dose-dependent gene expression changes from characterizing coordination among responses. ALOHA integrates these analyses to identify co-expressed genes and characterize altered cellular processes. Our study evaluated three chemical compounds, two eliciting strong transcriptional responses and one with limited to no biological response. Here, ALOHA demonstrated better characterization of GDR relationships, particularly for genes with a non-monotonic response. It did this by retaining more biologically active genes while removing inactive or noisy probes by utilizing shape-constrained splines. The coupling of flexible modeling with co-expression clustering resulted in biologically relevant clusters and potency estimates consistent with the current approaches in BMDExpress.

ALOHA stream-lines current methods and better accounts for the underlying data structure. Identifying gene clusters with coordinated DR expressions captures cellular processes related to adverse health effects from which gene set level BMDs can be estimated. These estimates may be used to determine a genomic POD comparable to those derived from *in vivo* toxicological endpoints. In summary, ALOHA provides an alternative step-wise method integrating improved DR models

as well as consideration of coordinated expression changes to obtain BMDs that may be used to estimate a biologically relevant POD.

While we have demonstrated ALOHA is competitive with the current BMDExpress, there are several areas for future research. The first is to evaluate the sensitivity and specificity of ALOHA's GDR modeling strategy by including simulated dose-response data. No simulation method to generate this kind of data exists, and developing such a simulation methodology to evaluate ALOHA is an important area of future research. The second would be to optimize the clustering algorithm to obtain disjoint gene clusters that provide a clear delineation for evaluating the accuracy of similar DR patterns in gene clusters and their related biological mechanisms. The last would be to improve the estimation of uncertainty in BMDs from individual and gene set clusters. Such a refinement would allow ALOHA to better quantify the uncertainty between transcriptional and *in vivo* toxicological points-of-departure.

5 Acknowledgments

We would like to acknowledge Ms. Davidson's mentor Dr. Eileen D. Kuempel for her guidance and input on investigating the use of toxicogenomics data in occupational health risk assessment, Mr. Nathan M. Drew for his comments on technical issues during initial methods development, Mr. Wen Niu for uploading cleaned expression sets into the iLincs database, the CDC OAMD SCBS team for assistance with improving methods for high-performance computing during pipeline development, and thank Drs. Pierre Bushel and Fred Parham for reviewing the manuscript and their comments.

This work has been supported by the contract from The National Institute for Occupational Safety and Health (75D30118P01991), the grants from National Institutes of Health (U54HL127624 and P30ES006096), in part by intermural funds at National Institute of Environmental Health Sciences Division of Intramural Research (NIEHS/DIR), and in part by an appointment to the Research Participation Program at the Centers for Disease Control and Prevention administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the Centers for Disease Control and Prevention.

6 Contributions

Sarah E. Davidson performed methods development, software development, model implementation, evaluation of results, and visualizations.

Matt Wheeler provided input on implementation and interpretation of BLX spline modeling, as well as interpretation of results.

Scott Auerbach provided experimental data, BMDExpress analysis results, and toxicological input on modeling results.

Siva Sivaganesan suggested and provided input on the model fitting leave-one-out cross-validation analysis.

Mario Medvedovic provided input on implementation, interpretation, and visualization of clustering results.

7 Abbreviations

ALOHA, Aggregated LOcal extrema splines for High-throughput Analysis; **AD**, All Doses; **AIC**, Aikake Information Criterion; **BAP**, Benzo-[α]-pyrene; **BLX** Bayesian Local Extrema; **BMD** Benchmark Dose; **BMDL**, lower BMD uncertainty bound; **BMDU**, upper BMD uncertainty bound; **BMR** Benchmark Response; **BPF**, Best Parametric Fit; **CDR**, Cytotoxic Doses Removed; **CGN**, Cancer Gene Neighborhoods; **CLEAN**, Cluster Enrichment Analysis; **CPM** Count per Million; **CPZ**, Chlorpromazine; **CSIMM**, context-specific mixture models; **DR**, dose-response; **DPM**, Dirchlet Proces Mixture Model; **ECDF**, Empirical Cumulative Distribution Function; **FFDR**, Fischer's exact test Benjamini-Hochberg False Discovery Rate adjusted p-value; **GDR**, genomic dose-response; **KCl**, Potassium chloride; **LOOCV**, leave-one-out cross-validation; **MCMC**, Markov Chain Monte Carlo; **MSigDB**, Molecular Signature Database; **MSPE** Mean squared predicted error; **MRC**, Minimum Read Count; **POD** point-of-departure;

References

[1] Reza Farmahin, Andrew Williams, Byron Kuo, Nikolai L Chepelev, Russell S Thomas, Tara S Barton-Maclaren, Ivan H Curran, Andy Nong, Michael G Wade, and Carole L Yauk. Recommended approaches in the application of toxicogenomics to derive points of departure for chemical risk assessment. *Archives of toxicology*, 91(5):2045–2065, 2017.

[2] Ivy Moffat, Nikolai L Chepelev, Sarah Labib, Julie Bourdon-Lacombe, Byron Kuo, Julie K Buick, France Lemieux, Andrew Williams, Sabina Halappanavar, Amal I Malik, et al. Comparison of toxicogenomics and traditional approaches to inform mode of action and points of departure in human health risk assessment of benzo [a] pyrene in drinking water. *Critical reviews in toxicology*, 45(1):1–43, 2015.

[3] Russell S Thomas, Martin A Philbert, Scott S Auerbach, Barbara A Wetmore, Michael J Devito, Ila Cote, J Craig Rowlands, Maurice P Whelan, Sean M Hays, Melvin E Andersen, et al. Incorporating new technologies into toxicity testing and risk assessment: moving from 21st century vision to a data-driven framework. *toxicological sciences*, 136(1):4–18, 2013.

[4] A Francina Webster, Nikolai Chepelev, Rémi Gagné, Byron Kuo, Leslie Recio, Andrew Williams, and Carole L Yauk. Impact of genomics platform and statistical filtering on transcriptional benchmark doses (bmd) and multiple approaches for selection of chemical point of departure (pod). *PLoS One*, 10(8):e0136764, 2015.

[5] Jeffry L Dean, Q Jay Zhao, Jason C Lambert, Belinda S Hawkins, Russell S Thomas, and Scott C Wesselkamper. Editor's highlight: Application of gene set enrichment analysis for identification of chemically induced, biologically relevant transcriptomic networks and potential utilization in human health risk assessment. *Toxicological Sciences*, 157(1):85–99, 2017.

[6] Scott S. Auerbach. BMDExpress 2.3, 2017. URL `https://github.com/auerbachs/BMDExpress-2/wiki`.

[7] Jason R Phillips, Daniel L Svoboda, Arpit Tandon, Shyam Patel, Alex Sedykh, Deepak Mav, Byron Kuo, Carole L Yauk, Longlong Yang, Russell S Thomas, et al. Bmdexpress 2: enhanced transcriptomic dose-response analysis workflow. *Bioinformatics*, 35(10):1780–1782, 2019.

[8] Longlong Yang, Bruce C Allen, and Russell S Thomas. Bmdexpress: a software tool for the benchmark dose analyses of genomic data. *BMC genomics*, 8(1):387, 2007.

[9] National Toxicology Program NTP et al. Ntp research report on national toxicology program approach to genomic dose-response modeling: Research report 5. NTP Research Report Series 2473-4756, National Toxicology Program, 2018. URL http://dx.doi.org/10.22427/NTP-RR-5.

[10] Jui-Hua Hsieh, Alexander Sedykh, Ruili Huang, Menghang Xia, and Raymond R Tice. A data analysis pipeline accounting for artifacts in tox21 quantitative high-throughput screening assays. *Journal of biomolecular screening*, 20(7):887–897, 2015.

[11] U.S. EPA. Benchmark dose technical guidance document. Risk Assessment Forum EPA/100/R-12/001, U.S. Environmental Protection Agency, 2012.

[12] DA Williams. A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics*, pages 103–117, 1971.

[13] DA Williams. The comparison of several dose levels with a zero dose control. *Biometrics*, pages 519–531, 1972.

[14] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[15] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6):417–425, 2015.

[16] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[17] Dominic J Allocco, Isaac S Kohane, and Atul J Butte. Quantifying the relationship between co-expression, co-regulation and gene function. *BMC bioinformatics*, 5(1):18, 2004.

[18] David B Dahl. Model-based clustering for expression data via a dirichlet process mixture model. *Bayesian inference for gene expression and proteomics*, 4:201–218, 2006.

[19] Chris Fraley and Adrian E. Raftery. Mclust: Software for model-based cluster analysis. *Journal of Classification*, 16(2):297–306, 1999. URL https://EconPapers.repec.org/RePEc:spr:jclass:v:16:y:1999:i:2:p:297-306.

[20] Mario Medvedovic and Siva Sivaganesan. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18(9):1194–1206, 2002.

[21] Mario Medvedovic, Ka Yee Yeung, and Roger Eugene Bumgarner. Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, 20(8):1222–1232, 2004.

[22] Ka Yee Yeung, Chris Fraley, Alejandro Murua, Adrian E. Raftery, and Walter L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987, 2001.

[23] Ziv Bar-Joseph, Georg Gerber, David K Gifford, Tommi S Jaakkola, and Itamar Simon. A new approach to analyzing gene expression time series data. In *Proceedings of the sixth annual international conference on Computational biology*, pages 39–48, 2002.

[24] Paul Kirk, Jim E Griffin, Richard S Savage, Zoubin Ghahramani, and David L Wild. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24):3290–3297, 2012.

[25] Xiangdong Liu, Siva Sivaganesan, Ka Yee Yeung, Junhai Guo, Roger Eugene Bumgarner, and Mario Medvedovic. Context-specific infinite mixtures for clustering gene expression profiles across diverse microarray dataset. *Bioinformatics*, 22(14):1737–1744, 2006.

[26] Eric F Lock and David B Dunson. Bayesian consensus clustering. *Bioinformatics*, 29(20):2610–2616, 2013.

[27] Yihui Luan and Hongzhe Li. Clustering of time-course gene expression data using a mixed-effects model with b-splines. *Bioinformatics*, 19(4):474–482, 2003.

[28] Yihui Luan and Hongzhe Li. Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. *Bioinformatics*, 20(3):332–339, 2004.

[29] Ian C McDowell, Dinesh Manandhar, Christopher M Vockley, Amy K Schmid, Timothy E Reddy, and Barbara E Engelhardt. Clustering gene expression time series data using an infinite gaussian process mixture model. *PLoS computational biology*, 14(1):e1005896, 2018.

[30] Ju-Hyun Park and Minjung Kyung. Bayesian curve fitting and clustering with dirichlet process mixture models for microarray data. *Journal of the Korean Statistical Society*, 48(2):207–220, 2019.

[31] Richard S Savage, Zoubin Ghahramani, Jim E Griffin, Bernard J De la Cruz, and David L Wild. Discovering transcriptional modules by bayesian data integration. *Bioinformatics*, 26(12):i158–i167, 2010.

[32] MW Wheeler, DB Dunson, and AH Herring. Bayesian local extremum splines. *Biometrika*, 104(4): 939–952, 2017.

[33] Sreenivasa C Ramaiahgari, Scott S Auerbach, Trey O Saddler, Julie R Rice, Paul E Dunlap, Nisha S Sipes, Michael J DeVito, Ruchir R Shah, Pierre R Bushel, Bruce A Merrick, et al. The power of resolution: contextualized understanding of biological responses to liver injury chemicals using high-throughput transcriptomics and benchmark concentration modeling. *Toxicological Sciences*, 169(2): 553–566, 2019.

[34] Hirotsugu Akaike, Boris Nikolaevich Petrov, and F Csaki. Second international symposium on information theory, 1973.

[35] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[36] Yosiyuki Sakamoto, Makio Ishiguro, and Genshiro Kitagawa. Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, 81, 1986.

[37] Deepak Mav, Ruchir R Shah, Brian E Howard, Scott S Auerbach, Pierre R Bushel, Jennifer B Collins, David L Gerhold, Richard S Judson, Agnes L Karmaus, Elizabeth A Maull, et al. A hybrid gene selection approach to create the s1500+ targeted gene sets for use in high-throughput transcriptomics. *PloS one*, 13(2), 2018.

[38] Joanne M Yeakley, Peter J Shepard, Diana E Goyena, Harper C VanSteenhouse, Joel D McComb, and Bruce E Seligmann. A trichostatin a expression signature identified by tempo-seq targeted whole transcriptome profiling. *PLoS One*, 12(5), 2017.

[39] Kenny S Crump. Calculation of benchmark doses from continuous data. *Risk Analysis*, 15(1):79–89, 1995.

[40] Johannes M Freudenberg, Vineet K Joshi, Zhen Hu, and Mario Medvedovic. Clean: Clustering enrichment analysis. *BMC bioinformatics*, 10(1):234, 2009.

[41] Arthur Liberzon. A description of the molecular signatures database (msigdb) web site. In *Stem Cell Transcriptional Networks*, pages 153–160. Springer, 2014.

[42] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1973. doi: 10.1080/01969727308546046. URL `https://doi.org/10.1080/01969727308546046`.

[43] Joseph C Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1): 95–104, 1974.

[44] Marina Meilă and William Pentney. Clustering by weighted cuts in directed graphs. In *Proceedings of the 2007 SIAM international conference on data mining*, pages 135–144. SIAM, 2007.

[45] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Supplemental Material:
ALOHA - Aggregated Local Extrema Splines for High-throughput Dose-response Analysis

Sarah E. Davidson[1], Matthew W. Wheeler[2], Scott S. Auerbach[3]
Siva Sivaganesan[4], Mario Medvedovic[1]

1 Available Genomic Dose-Response Methods

Modeling Features	ALOHA	BMDExpress [1, 2, 3]	tcplFit [4, 5]	Reynolds et al. (2020) [6]	GRAVEE [7]
Differential Expression Pre-filtering		X	X		
Non-monotonic Fits	X			X	
Gene Level Fits	X	X		X	
Gene Level BMDs	X	X		X	
Gene Set Level Trends	X		X		X
Gene Set Level BMDs	X	X	X		X
Parametric Models		X	X		
Non-parametric Models	X			X	X
Bayesian Modeling	X			X	
Frequentist Modeling		X	X		X
A Priori Gene Groups		X	X		X
Self-Organizing Gene Groups	X				

Table 1: Comparison of features in various genomic dose-response (GDR) methods. Different GDR modeling methods are listed in the columns and rows list key features in the modeling methods. "X" indicates a given GDR method has the listed feature. (BMD = benchmark dose estimate).

2 Comparison Outline Supplement

The comparison between the the two genomic dose-response (GDR) modeling approaches – BMDExpress and ALOHA – described the paper is summarized by the following figure.

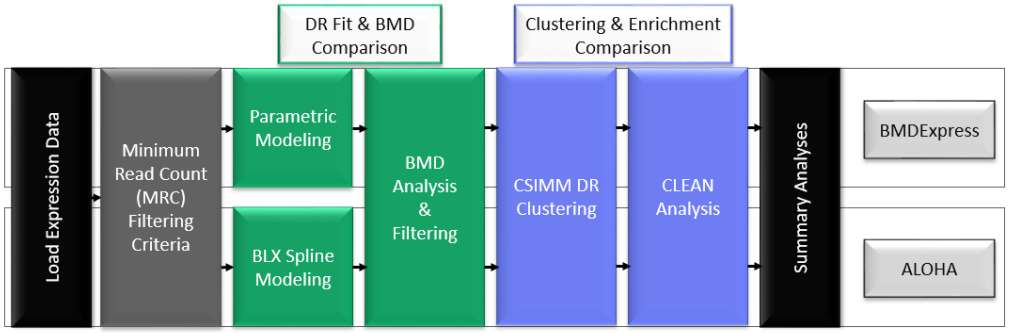


Figure 1: Workflow chart outlining our key comparative analyses for evaluating our GDR modeling approach ALOHA in comparison with BMDExpress – adapted from the "Basic Workflow" figure in Auerbach [1].

3 Methods Supplement Material

3.1 BLX Spline Curve Shapes (Extrema = 2 & Increasing = TRUE)

Example BLX spline curve shapes and posterior samples for changepoints. Location of posterior samples of the two change-points in the parameter space with respect to each other provides information about the resulting curve shape from the BLX spline fit. The following equations were used to simulate the five possible curve shapes fit by the BLX spline specifications mentioned in the main paper in §2.1.1:

$$g_1(x) = 2.75x + 6 \tag{1}$$

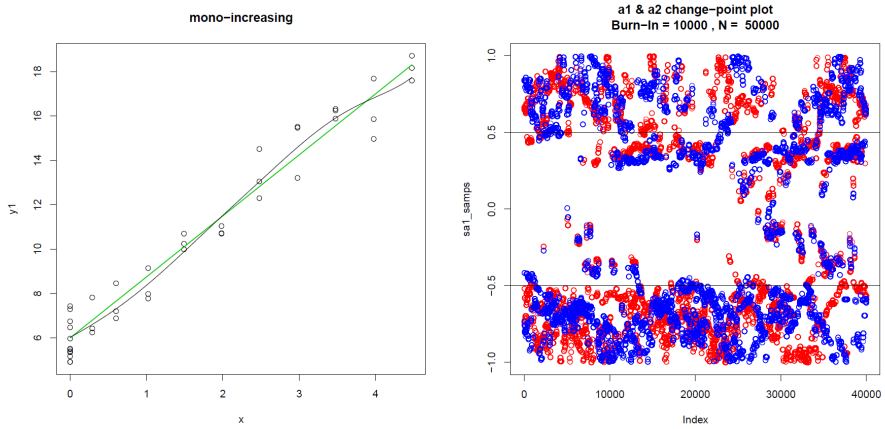
$$g_2(x) = -2.5x + 15 \tag{2}$$

$$g_3(x) = 7sin(0.3\pi x) + 15 \tag{3}$$

$$g_4(x) = 2.3(x - 1.245)^2 \tag{4}$$

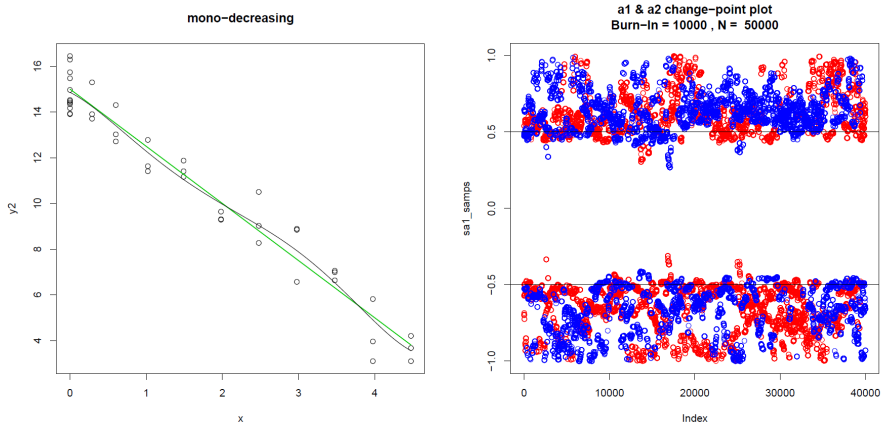
$$g_5(x) = 5.75cos(0.06\pi x - 5) + (x + 5)^1.4 \tag{5}$$

$$g_6(x) = 7.75 + 0.05x + 0.002x^2 \tag{6}$$



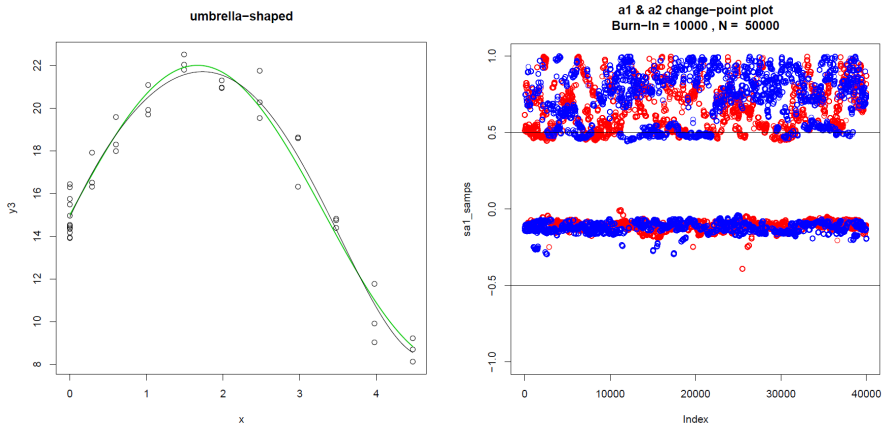
(a) (b)

Figure 2: BLX spline fit (black line) to monotonic increasing trend $g_1(x)$ (green line). (Left) The observed data points and BLX spline estimation. (Right) The MCMC chain of the posterior samples for both change points (i.e. α_1 in red and α_2 in blue).



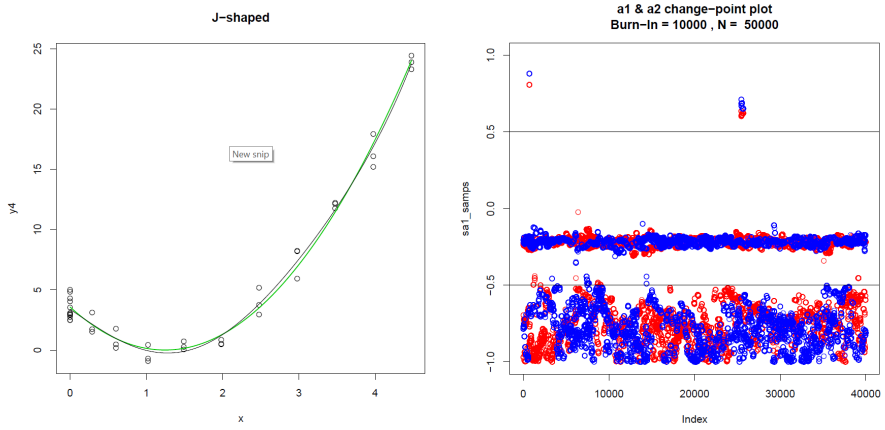
(a) (b)

Figure 3: BLX spline fit (black line) to monotonic decreasing trend $g_2(x)$ (green line). (Left) The observed data points and BLX spline estimation. (Right) The MCMC chain of the posterior samples for both change points (i.e. α_1 in red and α_2 in blue).



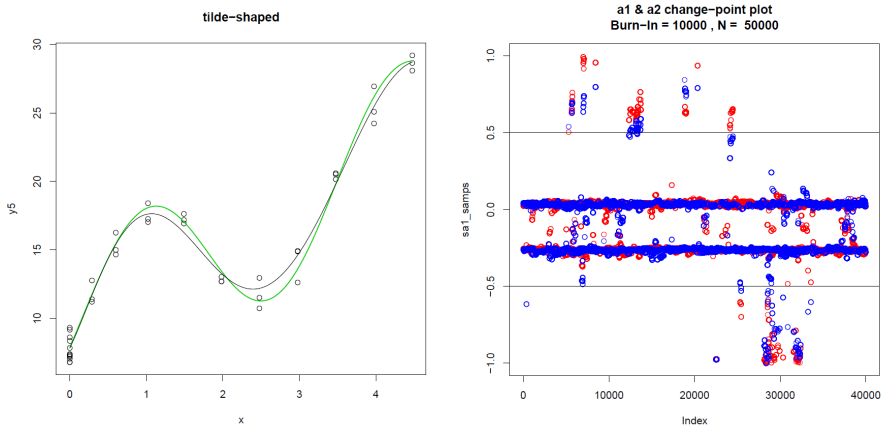
(a) (b)

Figure 4: BLX spline fit (black line) to umbrella-shaped trend $g_3(x)$ (green line). (Left) The observed data points and BLX spline estimation. (Right) The MCMC chain of the posterior samples for both change points (i.e. α_1 in red and α_2 in blue).



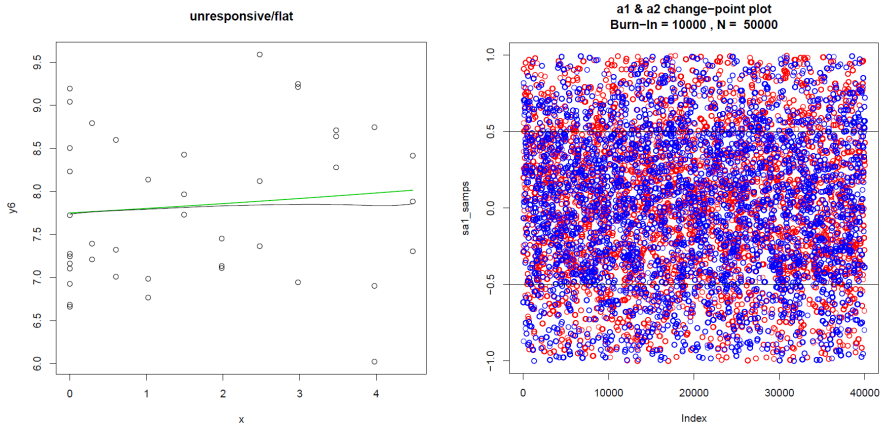
(a) (b)

Figure 5: BLX spline fit (black line) to J-shaped trend $g_4(x)$ (green line). (Left) The observed data points and BLX spline estimation. (Right) The MCMC chain of the posterior samples for both change points (i.e. α_1 in red and α_2 in blue).



(a) (b)

Figure 6: BLX spline fit (black line) to tilde-shaped trend $g_5(x)$ (green line). (Left) The observed data points and BLX spline estimation. (Right) The MCMC chain of the posterior samples for both change points (i.e. α_1 in red and α_2 in blue).



(a) (b)

Figure 7: BLX spline fit (black line) to unresponsive/flat trend $g_6(x)$ (green line). (Left) The observed data points and BLX spline estimation. (Right) The MCMC chain of the posterior samples for both change points (i.e. α_1 in red and α_2 in blue).

3.2 BMD Estimation BLX Spline Modeling

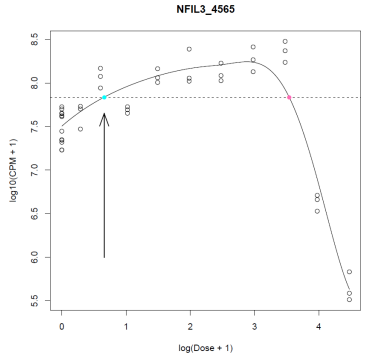


Figure 8: Example probe from BAP (AD) with an umbrella-shaped DR curve illustrating BMD estimation for BLX spline estimating umbrella or J-shaped DR curves (i.e. BMR obtains 2 possible BMD estimates). The dotted horizontal line is the BMR value and the aqua blue and pink dots are the potential BMD estimates. The arrow pointing to the aqua blue dot indicates the dose we chose to be the BMD estimate for this curve.

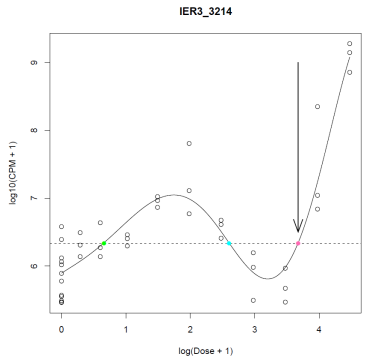


Figure 9: Example probe from BAP (AD) with an umbrella-shaped DR curve illustrating BMD estimation for BLX spline estimating tilde-shaped DR curves (i.e. BMR obtains 3 possible BMD estimates). The dotted horizontal line is the BMR value and the green, aqua blue, and pink dots are the potential BMD estimates. The arrow pointing to the pink dot indicates the dose we chose to be the BMD estimate for this curve.

3.3 CSIMM Gene Clustering

Previously, we observed n genes; assume $n^* \leq n$ of these genes can be clustered. That is, after the fitting process n^* genes have acceptable fits. Consider each sample $j \in \{1, \ldots, p\}$ falls into one of M experimental contexts. Here, we specify the contexts as their respective dose groups (i.e.

$M = D$). As before, let Y be the $n^* \times p$ matrix of observations we are clustering. Each gene i is evaluated for membership to global and local clusters via a MCMC algorithm.

For each iteration, CSIMM first evaluates the probability y_i is a member of the global cluster $q \in \{1, \ldots, Q\}$ described as the multivariate-normal (MVN):

$$y_i \sim N(\mu_q, \Sigma_q) \tag{7}$$

where μ_q is the mean vector and Σ_q is the variance-covariance matrix for global cluster q. Once all genes are assigned to a global cluster, each global cluster q is evaluated for membership to a local cluster in some experimental context $f \in \{1, \ldots, M\}$. Without loss of generality, for context f CSIMM evaluates the probability that $\mu_{(q,f)}$ – a sub-vector of global cluster q – belongs to a local cluster w. The local cluster w is described by the MVN:

$$\mu_{(q,f)} \sim N(\mu_w^*, \Sigma_w^*) \tag{8}$$

with the mean vector μ_w^* and the context-specific variance-covariance matrix Σ_w^*.

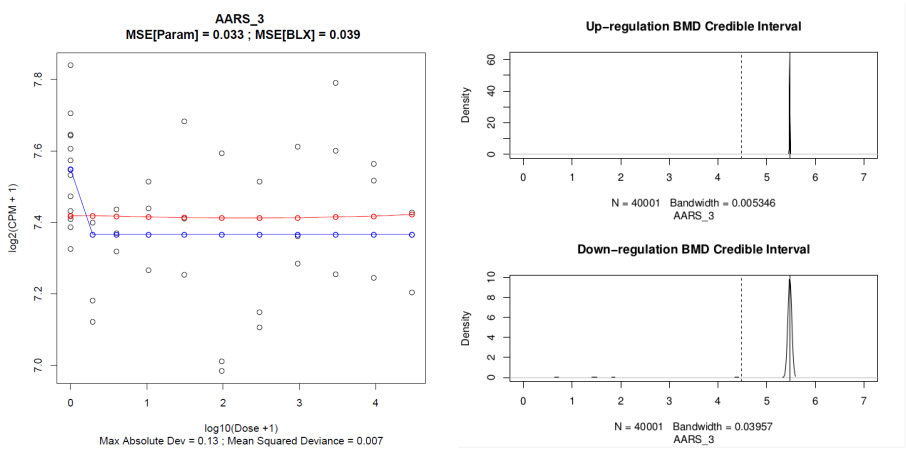
After genes are assigned to global and local clusters, parameters for all global MVNs, all local MVNs, and the hyperparameters are updated based on membership in each cluster.

At the end of the MCMC chain the posterior pairwise probabilities (PPP) is calculated for genes belonging to the same global and local clusters. For example, consider gene a and b are assigned the same group labels for 85% of the MCMC iterations after burn-in, then the PPP for these genes is 0.85. Calculating the PPP of genes avoids finding the "correct" number of resulting gene sets and circumvents the label-switching problem, common with DPMs (see Celeux et al. [8] and Stephens [9], among others). PPP's are then used to construct a hierarchical tree using 1-PPP as the distance measure. Using 1-PPP as the distance measure this places gene sets with a high PPP closer to the leaves of the tree and are obtained with smaller cut-off values (e.g. if PPP = 0.95, then h = 0.05).

4 Results Supplement Material

4.1 Fit Comparison for Cases where Parametric Fit Passes but BLX Spline Fails to Pass

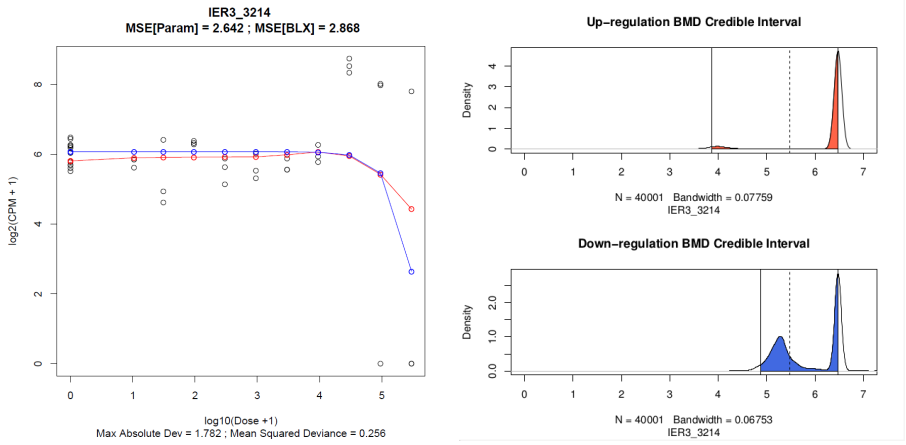
Case 1: Noisy expression Response Data.



(a) (b)

Figure 10: (Left) Plot containing the best parametric (blue) and BLX spline (red) fits on the observed expression response (points) for the "AARS_3" probe from the BAP-AD dataset. (Right) The BLX spline BMD posterior distribution plots, for both the up (top) and down (bottom) regulation directions. The black dotted line indicates the maximum experimental dose on the $log_{10}(Dose + 1)$ scale. Both regulation directions have a BMD posterior distribution highly concentrated around 1 plus the maximum experimental dose, as indicated by the sharp peaks right of the black dotted line. The credible interval, for both regulations directions, is indicated by a solid black like located at 1 plus the maximum experimental dose ($BMD_{95\%CI} = (5.48, 5.48)$). The down-regulation BMD estimate plot (bottom-right) provides the best illustration.

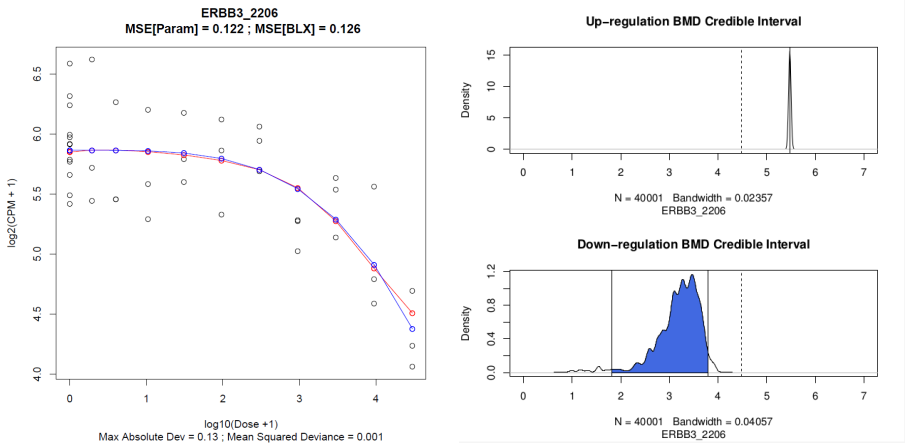
Case 2: Active and Inactive Samples in the Cytotoxic Dose Groups.



(a) (b)

Figure 11: (Left) Plot containing the best parametric (blue) and BLX spline (red) fits on the observed expression response (points) for the "IER3_3214" probe from the CPZ-AD dataset. (Right) The BLX spline BMD posterior distribution plots, for both the up (top) and down (bottom) regulation directions. The black dotted line indicates the maximum experimental dose on the $log_{10}(Dose + 1)$ scale. Credible intervals are indicated by the red (up-regulation – $BMD_{95\%CI} = (3.87, 6.48)$) and blue (down-regulation – $BMD_{95\%CI} = (4.88, 6.48)$) shaded regions.

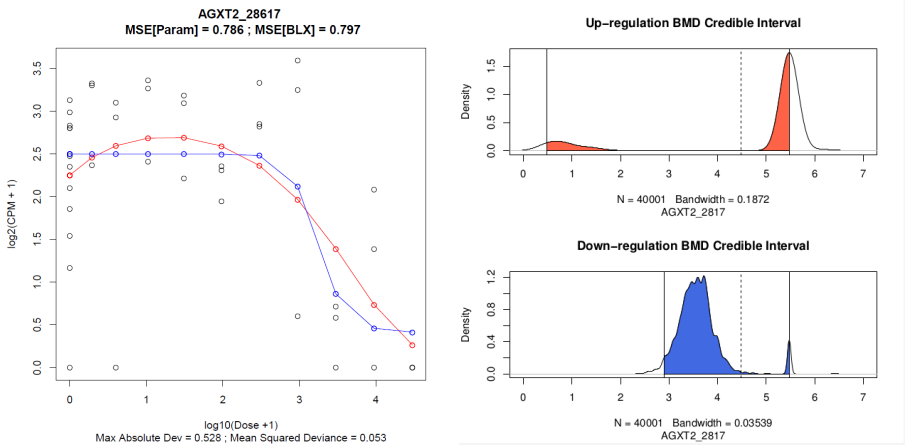
Case 3: Unreliable (or Wide) BLX spline BMD credible intervals Otherwise Seemingly Active.



(a) (b)

Figure 12: (Left) Plot containing the best parametric (blue) and BLX spline (red) fits on the observed expression response (points) for the "ERBB3_2206" probe from the BAP-AD dataset. (Right) The BLX spline BMD posterior distribution plots, for both the up (top) and down (bottom) regulation directions. The black dotted line indicates the maximum experimental dose on the $log_{10}(Dose + 1)$ scale. In the top-right plot, the credible interval for the BMD estimate in the up-regulation direction is indicated by a solid black line at 1 plus the maximum dose ($BMD_{95\%CI} = (5.48, 5.48)$). In the bottom-right plot, the blue shaded region indicates the credible interval for the BMD estimate in the down-regulation direction ($BMD_{95\%CI} = (1.81, 3.79)$).

Case 4: Bi-modal BLX spline BMD credible intervals with at Least One Mode Greater than the Maximum Dose.



(a) (b)

Figure 13: (Left) Plot containing the best parametric (blue) and BLX spline (red) fits on the observed expression response (points) for the "AGXT2_28617" probe from the BAP-AD dataset. (Right) The BLX spline BMD posterior distribution plots, for both the up (top) and down (bottom) regulation directions. The black dotted line indicates the maximum experimental dose on the $log_{10}(Dose + 1)$ scale. Credible intervals are indicated by the red (up-regulation – $BMD_{95\%CI} = (0.48, 5.48)$) and blue (down-regulation – $BMD_{95\%CI} = (2.89, 5.48)$) shaded regions.

4.2 BMD Estimate Comparison

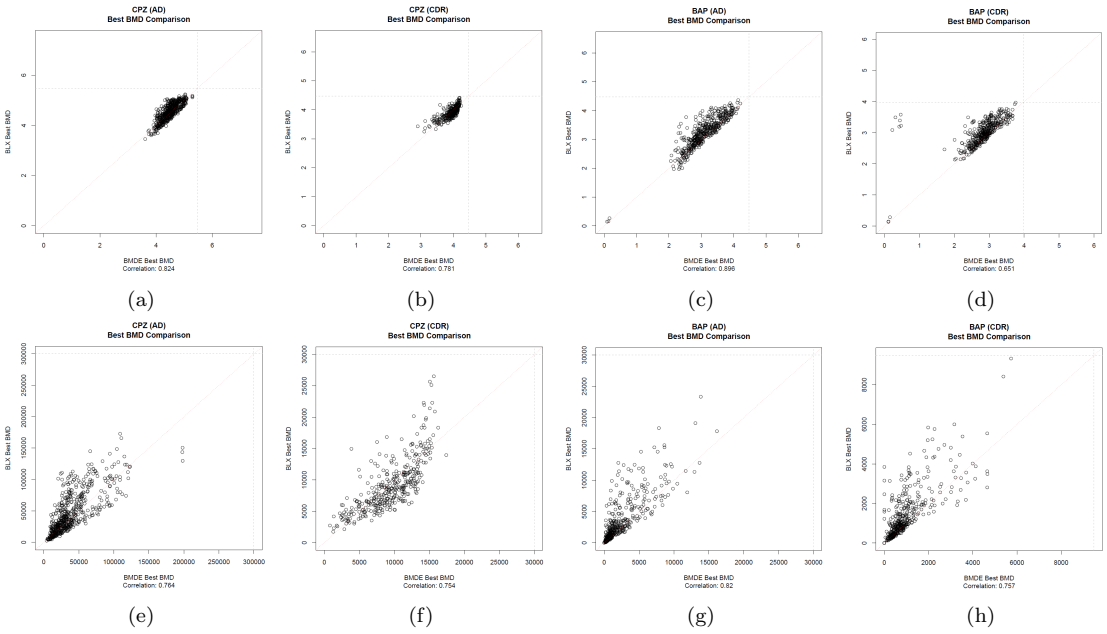
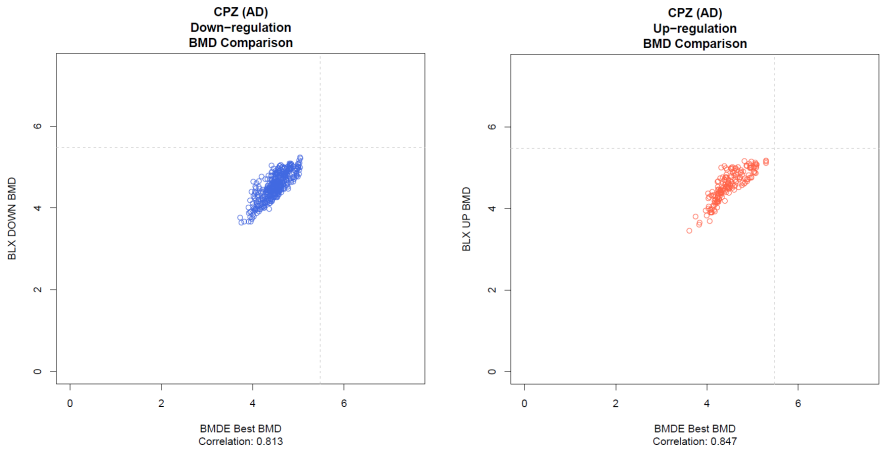


Figure 14: Scatterplot of the best BMD estimates from the best parametric fit (x-axis) and the BLX spline fit (y-axis) for all datasets, only for those probes with the best BMD estimate less than the maximum experimental dose and BMDU/BMDL less than 40 in both BMDExpress (version 2.3) and the BLX spline modeling. (Top row) Logarithmic scale BMD values. (Bottom row) Original dose scale BMD values in nM.

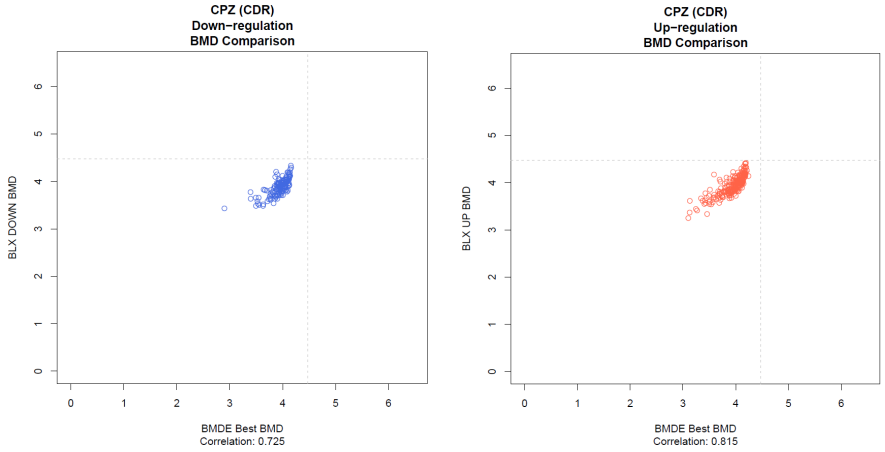
Chlorpromazine All Doses



(a) Down-regulation (b) Up-regulation

Figure 15: Scatterplot of the BMD estimates from the best parametric fit (x-axis) and the BLX spline fit (y-axis), based on the regulation direction from BMDExpress (version 2.3), for CPZ all doses, only for those probes passing "reliability" criteria in both BMDExpress and the BLX spline modeling. (Left) Down-regulated BMD estimate comparison, only probes where BMDExpress adverse direction = -1 (down-regulation). (Right) Up-regulation BMD estimate comparison, only probes where BMDExpress adverse direction = 1 (up-regulation).

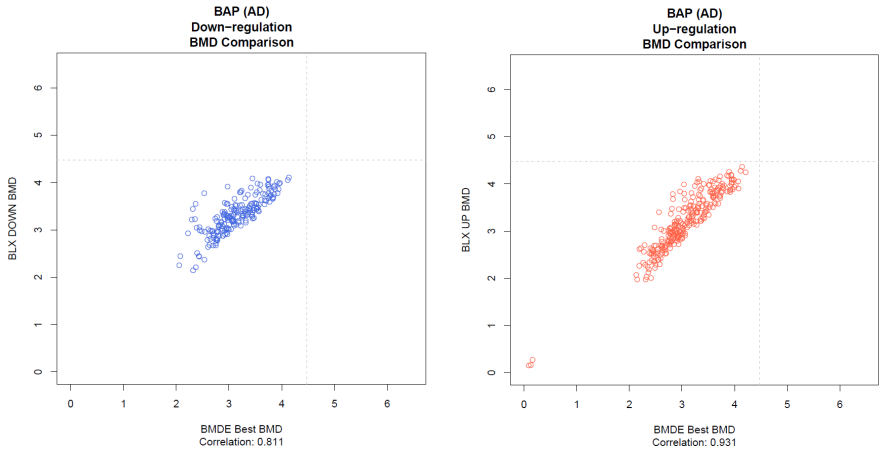
Chlorpromazine Cytotoxic Doses Removed



(a) Down-regulation (b) Up-regulation

Figure 16: Scatterplot of the BMD estimates from the best parametric fit (x-axis) and the BLX spline fit (y-axis), based on the regulation direction from BMDExpress (version 2.3), for CPZ cytotoxic doses removed, only for those probes passing "reliability" criteria in both BMDExpress and the BLX spline modeling. (Left) Down-regulated BMD estimate comparison, only probes where BMDExpress adverse direction = -1 (down-regulation). (Right) Up-regulation BMD estimate comparison, only probes where BMDExpress adverse direction = 1 (up-regulation).

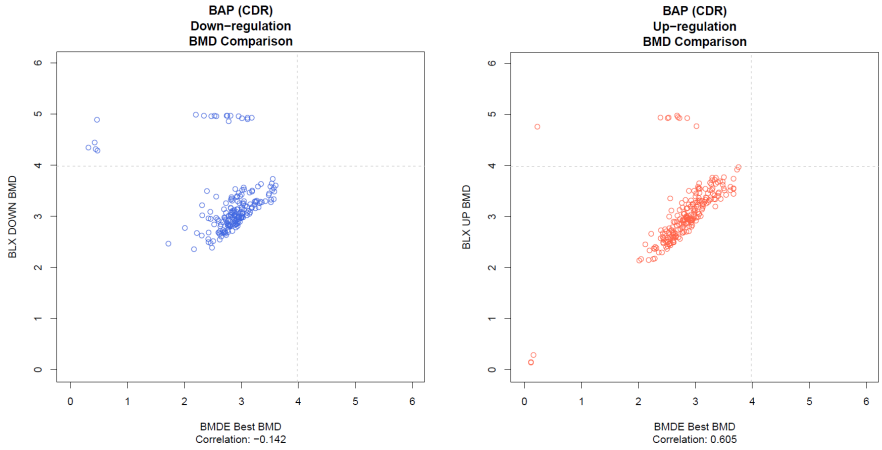
Benzo-[α]-pyrene All Doses



(a) Down-regulation (b) Up-regulation

Figure 17: Scatterplot of the BMD estimates from the best parametric fit (x-axis) and the BLX spline fit (y-axis), based on the regulation direction from BMDExpress (version 2.3), for BAP all doses, only for those probes passing "reliability" criteria in both BMDExpress and the BLX spline modeling. (Left) Down-regulated BMD estimate comparison, only probes where BMDExpress adverse direction = -1 (down-regulation). (Right) Up-regulation BMD estimate comparison, only probes where BMDExpress adverse direction = 1 (up-regulation).

Benzo-[α]-pyrene Cytotoxic Doses Removed



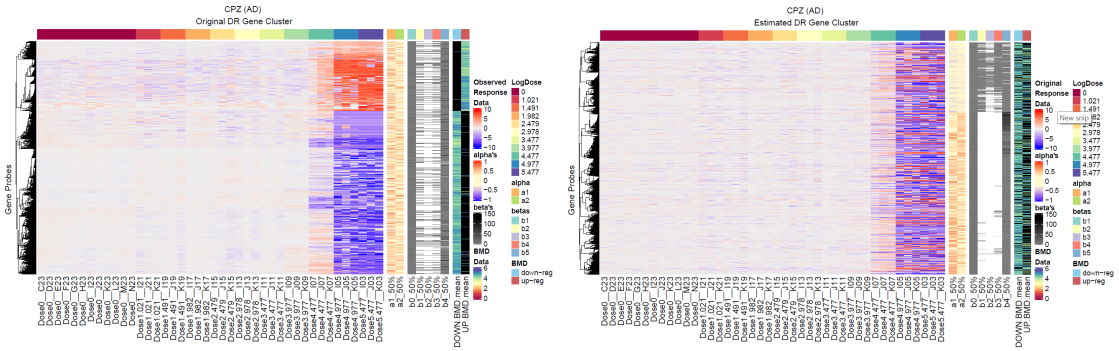
(a) Down-regulation (b) Up-regulation

Figure 18: Scatterplot of the BMD estimates from the best parametric fit (x-axis) and the BLX spline fit (y-axis), based on the regulation direction from BMDExpress (version 2.3), for BAP cytotoxic doses removed, only for those probes passing "reliability" criteria in both BMDExpress and the BLX spline modeling. (Left) Down-regulated BMD estimate comparison, only probes where BMDExpress adverse direction = -1 (down-regulation). (Right) Up-regulation BMD estimate comparison, only probes where BMDExpress adverse direction = 1 (up-regulation).

4.3 Dose-Response Clustering Comparison

Plot description: Heatmaps of the CSIMM clusters (left column) and ALOHA clusters (right column) for all datasets assessed. (From left to right in each heatmap plot) The main plot displays the normalized expression values with all replicates. Next are the heatmaps displaying BLX spline parameter estimates, i.e. change-points and spline coefficients respectively. Finally, down and up regulated mean BMD estimates for each probe are displayed in the right-most heatmap (BMDs above the maximum dose are NA's and are displayed in black).

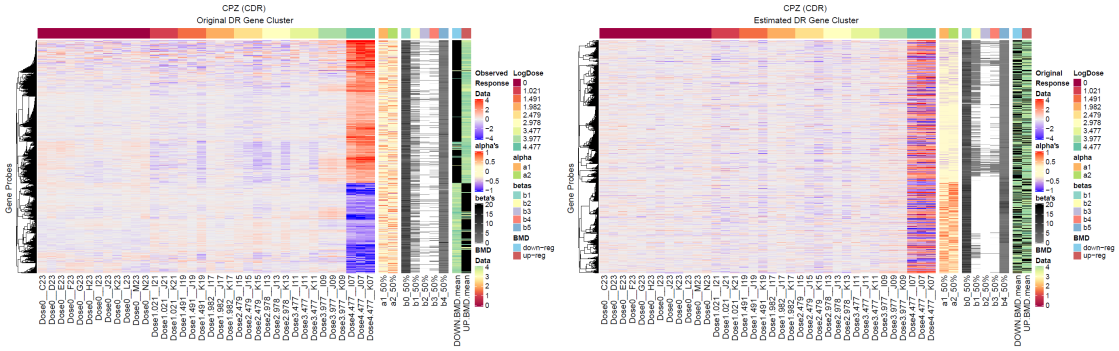
Chlorpromazine All Doses



(a) CSIMM (b) ALOHA

Figure 19: **CPZ AD** - Heatmaps displaying clustering results applied to normalized expression data showing visual separation of DR trends. (Left - *CSIMM*) CSIMM clustering on normalized expression data with all replicates in each dose group. (Right - *ALOHA*) CSIMM clustering on estimated response for each experimental dose group.

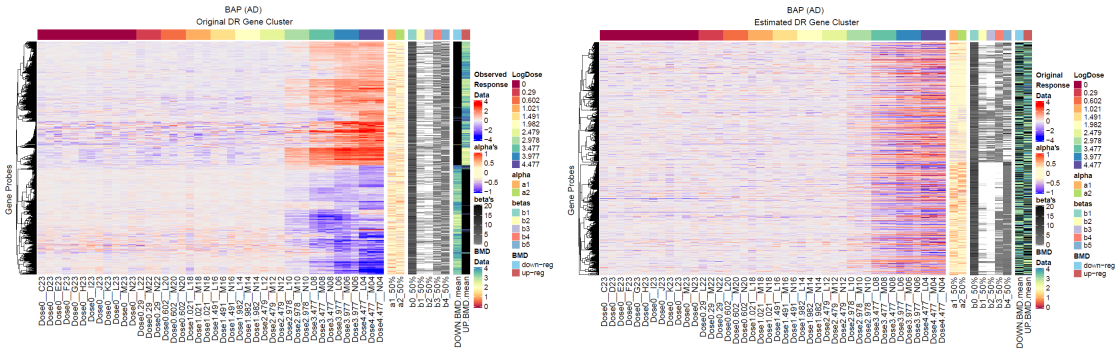
Chlorpromazine Cytotoxic Doses Removed



(a) CSIMM (b) ALOHA

Figure 20: **CPZ CDR** - Heatmaps displaying clustering results applied to normalized expression data showing visual separation of DR trends. (Left - *CSIMM*) CSIMM clustering on normalized expression data with all replicates in each dose group. (Right - *ALOHA*) CSIMM clustering on estimated response for each experimental dose group.

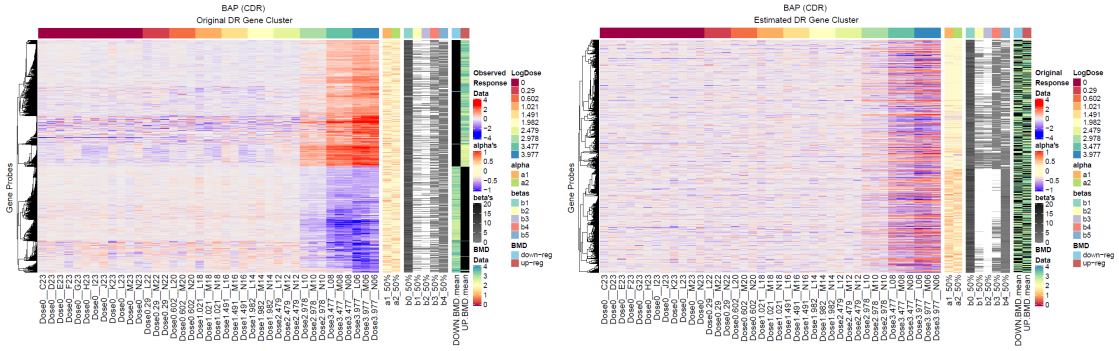
Benzo-[α]-pyrene All Doses



(a) CSIMM (b) ALOHA

Figure 21: **BAP AD** - Heatmaps displaying clustering results applied to normalized expression data showing visual separation of DR trends. (Left - *CSIMM*) CSIMM clustering on normalized expression data with all replicates in each dose group. (Right - *ALOHA*) CSIMM clustering on estimated response for each experimental dose group.

Benzo-[α]-pyrene Cytotoxic Doses Removed



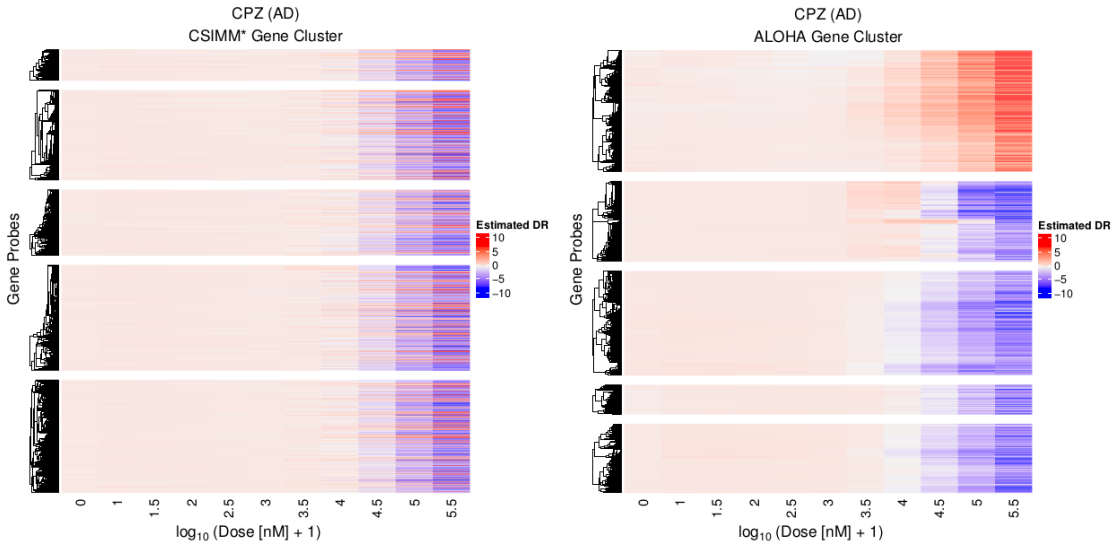
(a) CSIMM (b) ALOHA

Figure 22: **BAP CDR** - Heatmaps displaying clustering results applied to normalized expression data showing visual separation of DR trends. (Left - *CSIMM*) CSIMM clustering on normalized expression data with all replicates in each dose group. (Right - *ALOHA*) CSIMM clustering on estimated response for each experimental dose group.

Plot description: Heatmaps of the CSIMM clusters (left column) and ALOHA clusters (right column) for all datasets assessed. (From left to right in each heatmap plot) The main plot displays the BLX estimated dose-response data. The posterior pairwise probability dendrogram is cut at the height resulting in five clusters. The chosen cut-off was arbitrary, since there many possible cut-offs exist when evaluating clustering approaches that result in a dendrogram, but this was done for exposition of dose-response coherence within resulting clusters from the CSIMM and ALOHA

algorithms.

Chlorpromazine All Doses



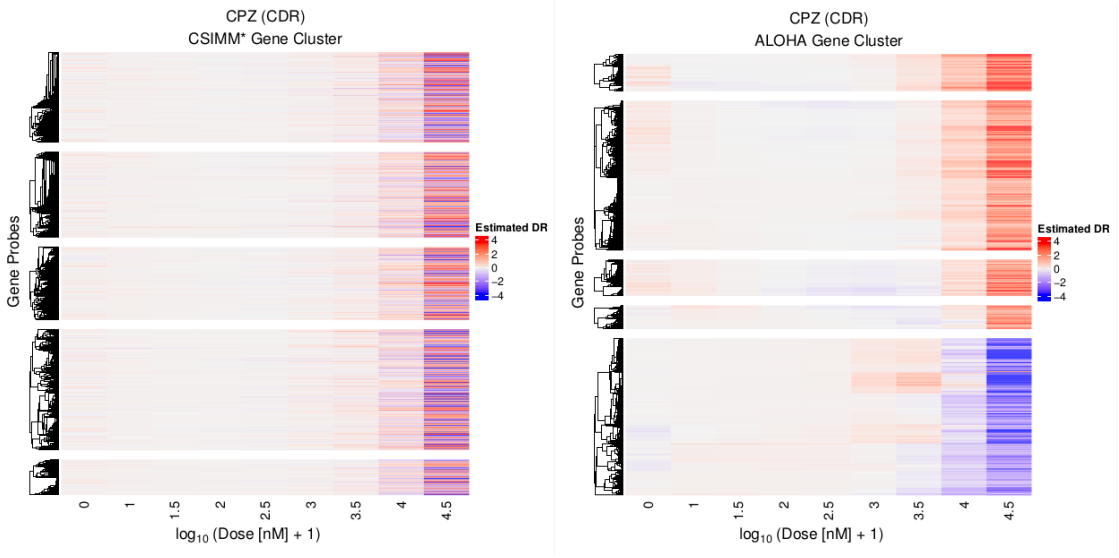
(a) CSIMM (b) ALOHA

Figure 23: **CPZ AD** - Heatmaps displaying clustering results applied to BLX estimated dose-response data showing visual separation of DR trends. (Left - *CSIMM*) CSIMM clustering on observed normalized expression data with all replicates in each dose group. (Right - *ALOHA*) CSIMM clustering on BLX estimated dose-response data.

Chlorpromazine (AD)	ALOHA	CSIMM
Avg. Distance Between Clusters	8.66	7.35
Avg. Distance Within Clusters	2.97	7.33
Dunn Index	6.0e-3	9.5e-4
Avg. Silhouette Width	0.19	-0.07
Entropy	1.52	1.53

Table 2: Scores measuring the internal coherence of dose-response trends in the five clusters resulting from ALOHA and CSIMM in the Chlorpromazine dataset with all doses, see Figure 23. Larger values for average distance between clusters, Dunn index, and the average silhouette width indicate better separation of genes into similar dose-response groups. Smaller values for the average distance within clusters and entropy indicate more coherence between genes clustered together.

Chlorpromazine Cytotoxic Doses Removed



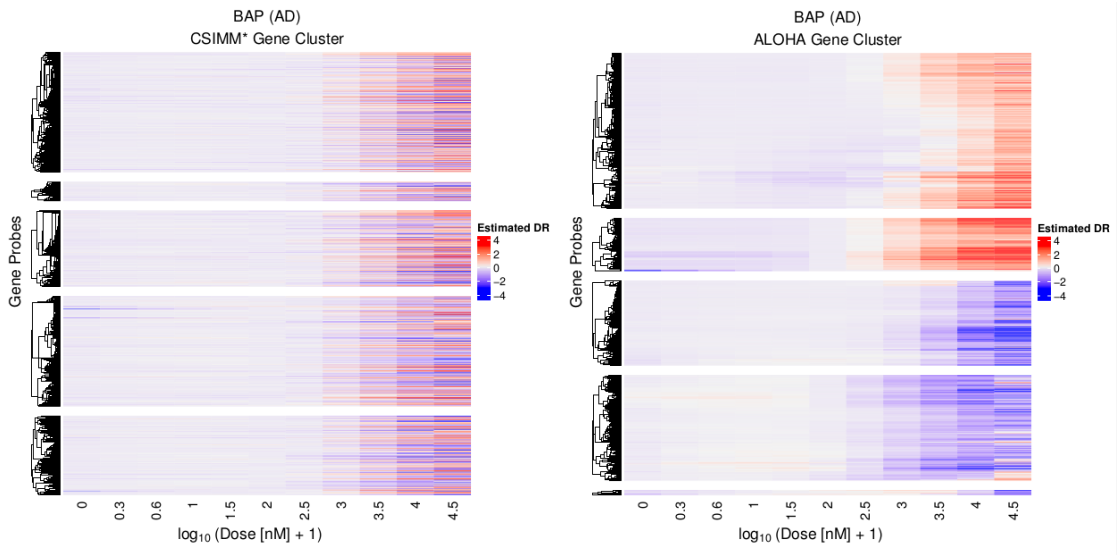
(a) CSIMM (b) ALOHA

Figure 24: **CPZ CDR** - Heatmaps displaying clustering results applied to BLX estimated dose-response data showing visual separation of DR trends. (Left - *CSIMM*) CSIMM clustering on observed normalized expression data with all replicates in each dose group. (Right - *ALOHA*) CSIMM clustering on BLX estimated dose-response data.

Chlorpromazine (CDR)	ALOHA	CSIMM
Avg. Distance Between Clusters	3.85	3.11
Avg. Distance Within Clusters	1.42	3.10
Dunn Index	3.03-3	1.1e-3
Avg. Silhouette Width	0.23	-0.10
Entropy	1.34	1.55

Table 3: Scores measuring the internal coherence of dose-response trends in the five clusters resulting from ALOHA and CSIMM in the Chlorpromazine dataset without cytotoxic doses, see Supplementary Figure 24. Larger values for average distance between clusters, Dunn index, and the average silhouette width indicate better separation of genes into similar dose-response groups. Smaller values for the average distance within clusters and entropy indicate more coherence between genes clustered together.

Benzo-[α]-pyrene All Doses



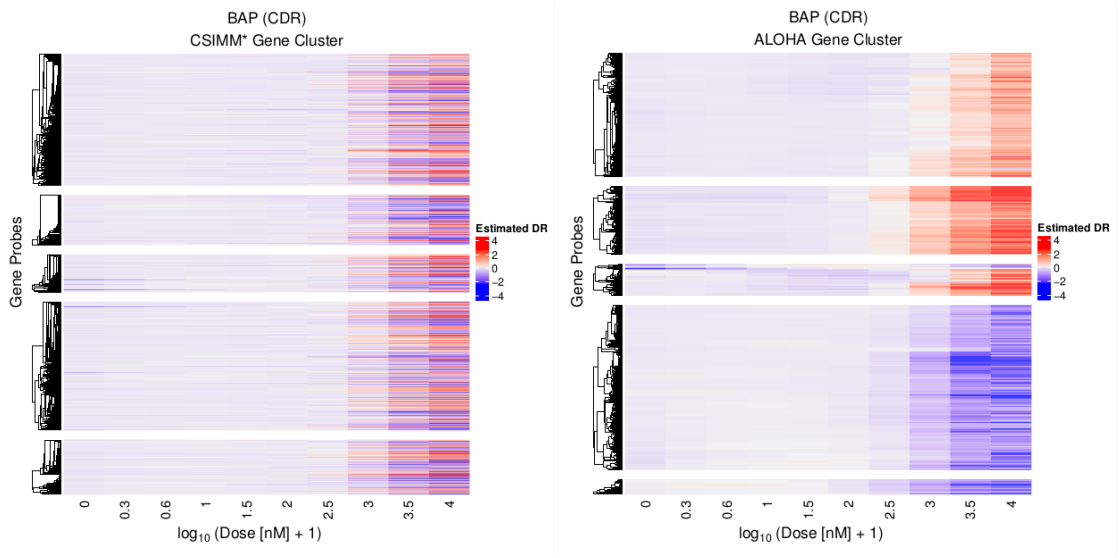
(a) CSIMM (b) ALOHA

Figure 25: **BAP AD** - Heatmaps displaying clustering results applied to BLX estimated dose-response data showing visual separation of DR trends. (Left - *CSIMM*) CSIMM clustering on observed normalized expression data with all replicates in each dose group. (Right - *ALOHA*) CSIMM clustering on BLX estimated dose-response data.

Benzo-[α]-pyrene (AD)	ALOHA	CSIMM
Avg. Distance Between Clusters	3.84	3.23
Avg. Distance Within Clusters	1.68	3.21
Dunn Index	4.4e-3	9.3e-4
Avg. Silhouette Width	0.20	-0.06
Entropy	1.37	1.49

Table 4: Scores measuring the internal coherence of dose-response trends in the five clusters resulting from ALOHA and CSIMM in the Benzo-[α]-pyrene dataset with all doses, see Figure 25. Larger values for average distance between clusters, Dunn index, and the average silhouette width indicate better separation of genes into similar dose-response groups. Smaller values for the average distance within clusters and entropy indicate more coherence between genes clustered together.

Benzo-[α]-pyrene Cytotoxic Doses Removed



(a) CSIMM (b) ALOHA

Figure 26: **BAP CDR** - Heatmaps displaying clustering results applied to BLX estimated dose-response data showing visual separation of DR trends. (Left - *CSIMM*) CSIMM clustering on observed normalized expression data with all replicates in each dose group. (Right - *ALOHA*) CSIMM clustering on BLX estimated dose-response data.

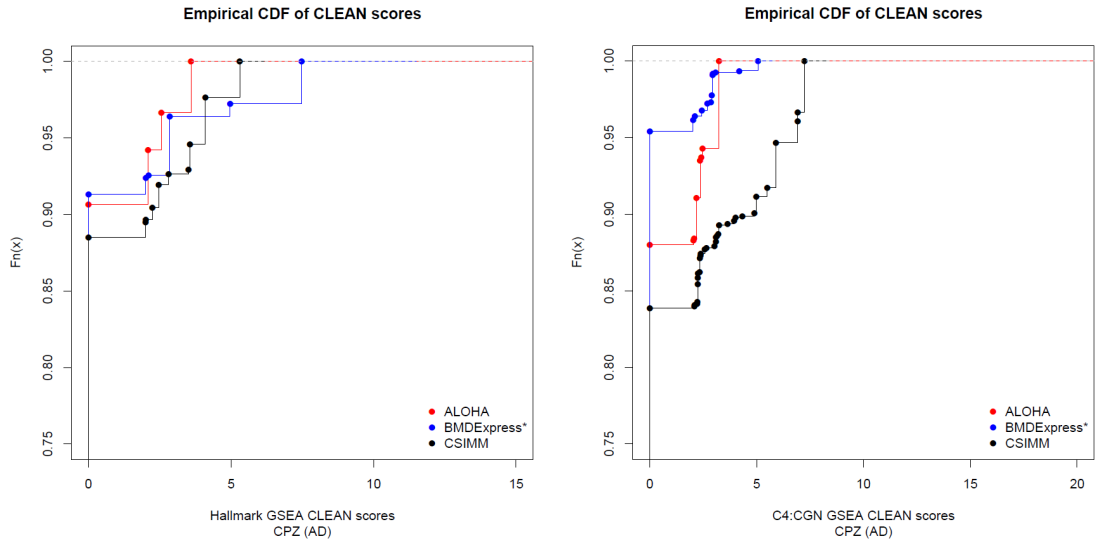
Benzo-[α]-pyrene (CDR)	ALOHA	CSIMM
Avg. Distance Between Clusters	3.21	2.59
Avg. Distance Within Clusters	1.11	2.57
Dunn Index	5.6e-3	1.5e-3
Avg. Silhouette Width	0.23	-0.04
Entropy	1.34	1.48

Table 5: Scores measuring the internal coherence of dose-response trends in the five clusters resulting from ALOHA and CSIMM in the Benzo-[α]-pyrene dataset without cytotoxic doses, see Supplementary Figure 26. Larger values for average distance between clusters, Dunn index, and the average silhouette width indicate better separation of genes into similar dose-response groups. Smaller values for the average distance within clusters and entropy indicate more coherence between genes clustered together.

4.4 Gene Level Functional Coherence

Figures 27-30 compare the empirical cumulative distribution function (ECDF) of CLEAN scores for each strategy on all the datasets. Functional coherence is indicated by ECDF plots with high CLEAN scores (i.e. slowest to reach 1 and furthest to the right on the ECDF plot). In general,

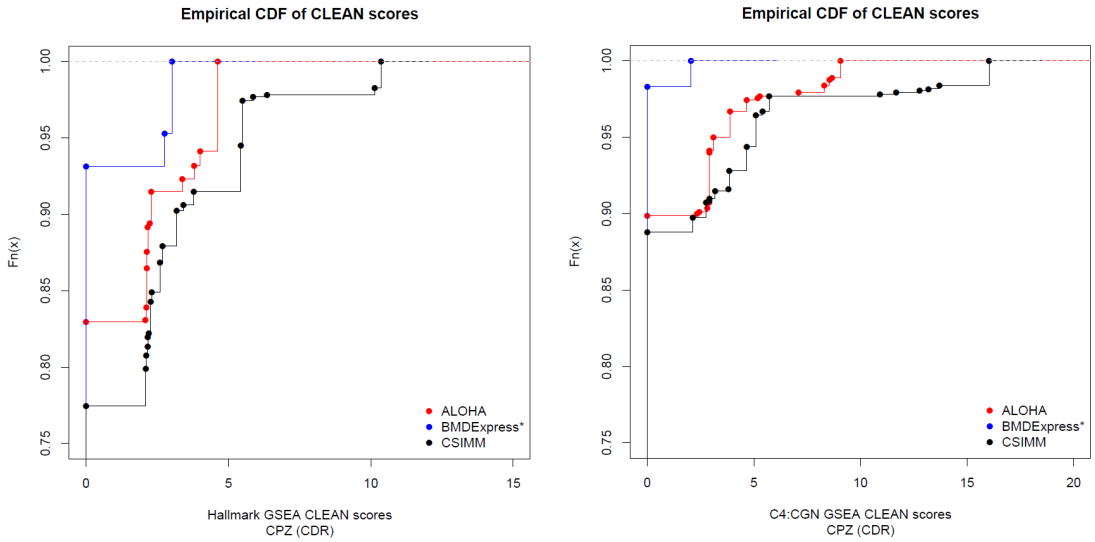
the BMDExpress approach (blue) showed the lowest level of functional coherence. For the other two approaches, other than Chlorpromazine with all doses, there is little separation between the ECDF curves. For Chlorpromazine without cytotoxic doses, there was greater separation at the highest scores; however, these constitute less than 5% of the genes, and subsequently, it may not indicate a difference between the approaches. For Chlorpromazine with all doses, the ECDF curves for the Hallmark gene set overlap indicating the methods are equivocal. Using the Cancer Gene Neighborhood gene list, there were clear improvements using the CSIMM approach. Overall, similar levels of functional coherence were seen using CSIMM and ALOHA. However, since the genes are grouped differently, as described in §3.2.1 incorporation of the model fit data does not reduce our ability to obtain biologically relevant clusters. The flexibility of the BLX spline obtains more accurate fits and when coupled with clustering helps identify biologically relevant clusters.



(a) Hallmark

(b) Cancer Gene Neighborhoods

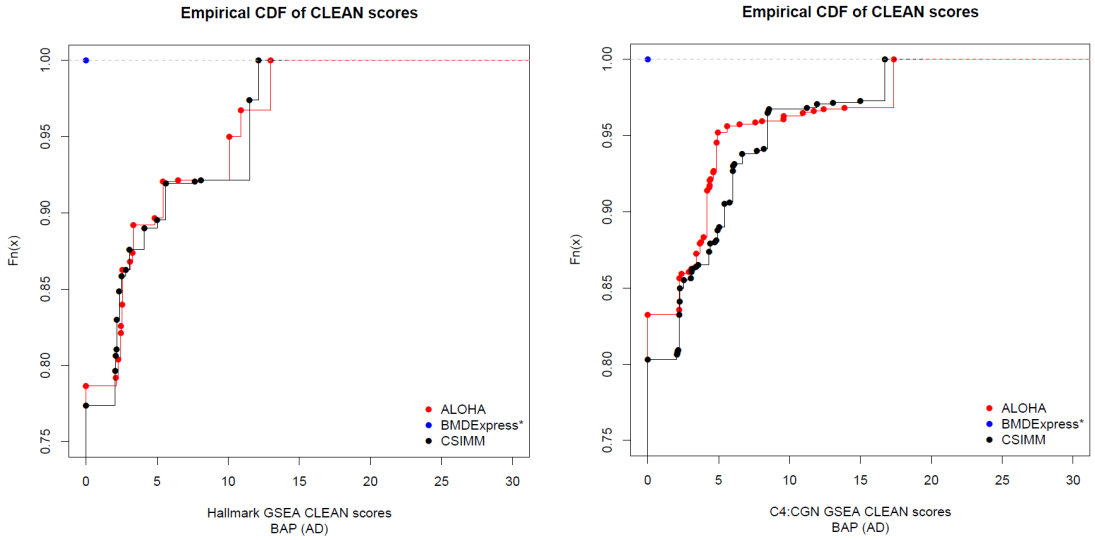
Figure 27: CLEAN score empirical cumulative distribution function (ECDF) plots for Chlorpromazine including all doses (CPZ-AD), with MSigDB's Hallmark (right) and Cancer Gene Neighborhood (CGN) (left) gene lists on the normalized expression (CSIMM – black), BLX estimated expression (ALOHA – red), and best parametric fit estimated expression (BMDExpress – blue) approaches.



(a) Hallmark

(b) Cancer Gene Neighborhoods

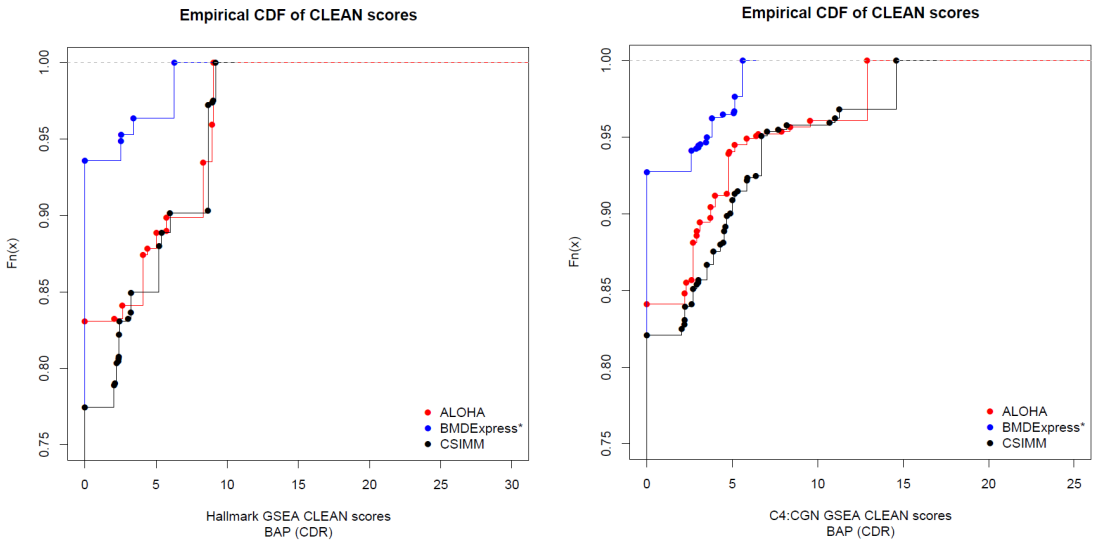
Figure 28: CLEAN score empirical cumulative distribution function (ECDF) plots for Chlorpromazine without cytotoxic doses (CPZ-CDR), with MSigDB's Hallmark (right) and Cancer Gene Neighborhood (CGN) (left) gene lists on the normalized expression (CSIMM – black), BLX estimated expression (ALOHA – red), and best parametric fit estimated expression (BMDExpress – blue) approaches.



(a) Hallmark

(b) Cancer Gene Neighborhoods

Figure 29: CLEAN score empirical cumulative distribution function (ECDF) plots for Benzo-[α]-pyrene including all doses (BAP-AD), with MSigDB's Hallmark (right) and Cancer Gene Neighborhood (CGN) (left) gene lists on the normalized expression (CSIMM – black), BLX estimated expression (ALOHA – red), and best parametric fit estimated expression (BMDExpress – blue) approaches.



(a) Hallmark

(b) Cancer Gene Neighborhoods

Figure 30: CLEAN score empirical cumulative distribution function (ECDF) plots for Benzo-[α]-pyrene without cytotoxic doses (BAP-CDR), with MSigDB's Hallmark (right) and Cancer Gene Neighborhood (CGN) (left) gene lists on the normalized expression (CSIMM – black), BLX estimated expression (ALOHA – red), and best parametric fit estimated expression (BMDExpress – blue) approaches.

4.5 Gene Set Level Functional Coherence & Enriched Sets

4.5.1 MSigDB Hallmark Gene Set Significant Categories

Plot description: Heatmaps comparing the enrichment scores of the clusters with smallest FDR adjusted Fisher's p-value for the MSigDB Hallmark gene sets found to be significant (FDR adjusted Fisher's p-value < 0.01) with the three different data inputs for CSIMM clustering, namely the normalized expression data with all replicates (CSIMM), parametric model estimated response data (BMDExpress), and the BLX spline fit estimated response data (ALOHA). Clusters with the smallest enrichment score not having a minimum of 3 genes mapping to the enriched gene set or populating the gene set by a minimum of 5% were biologically irrelevant due to lack of appropriate information and were removed therefore not displayed in the heatmap. (Left) Heatmap shows $-log_{10}(FisherFDR)$ of each enriched gene set from each approach. (Right) Heatmap shows $log(OddsRatio)$ of each enriched gene set from each approach. Rows are unique Hallmark gene sets and columns indicate clustering approaches.

Chlorpromazine All Doses

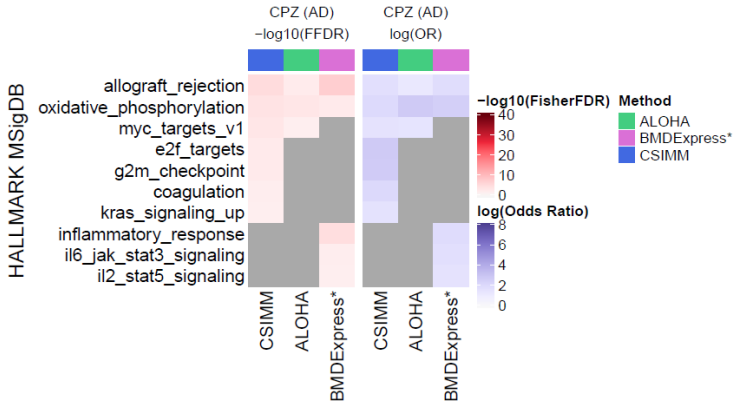


Figure 31: **CPZ AD** Heatmaps of the significant Hallmark gene sets comparing the most statistically significant clusters, displaying negative log-scale Fisher FDR and log odds ratio, respectively.

Chlorpromazine Cytotoxic Doses Removed

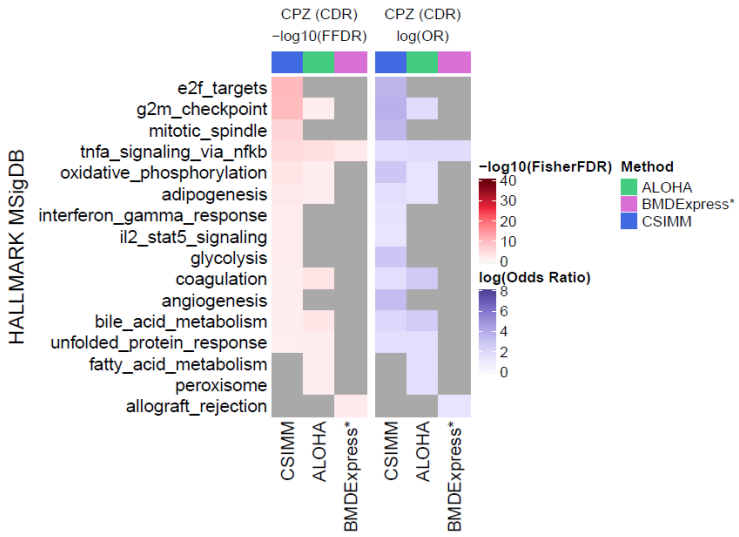


Figure 32: **CPZ CDR** Heatmaps of the significant Hallmark gene sets comparing the most statistically significant clusters, displaying negative log-scale Fisher FDR and log odds ratio, respectively.

Benzo-[α]-pyrene All Doses

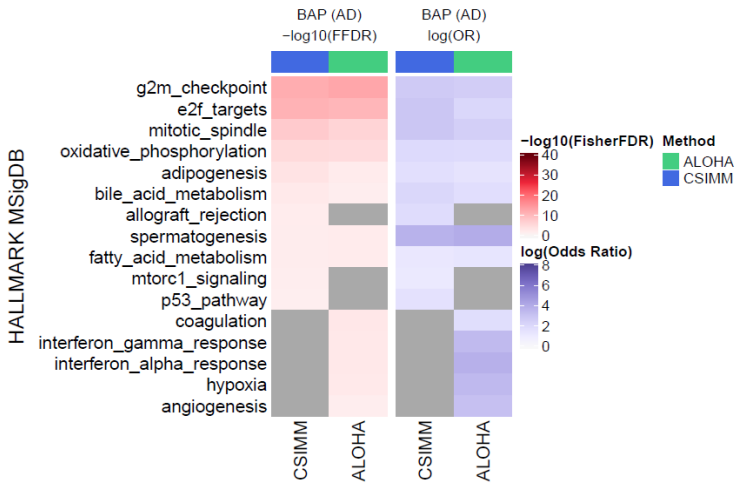


Figure 33: **BAP AD** Heatmaps of the significant Hallmark gene sets comparing the most statistically significant clusters, displaying negative log-scale Fisher FDR and log odds ratio, respectively.

Benzo-[α]-pyrene Cytotoxic Doses Removed

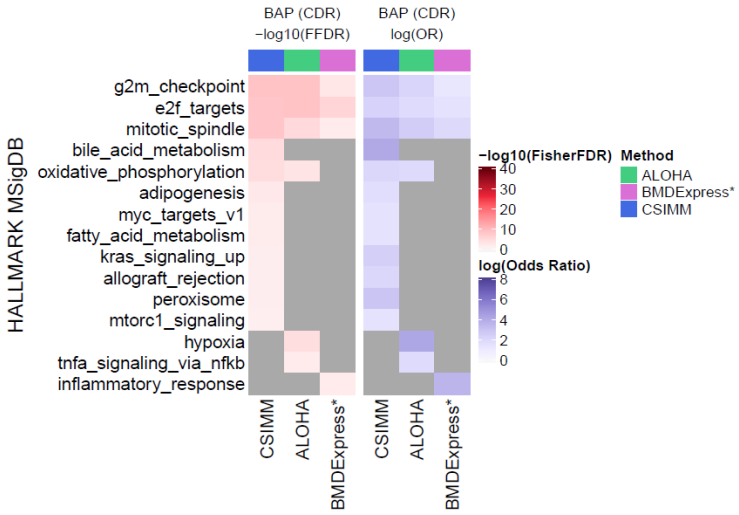


Figure 34: **BAP CDR** Heatmaps of the significant Hallmark gene sets comparing the most statistically significant clusters, displaying negative log-scale Fisher FDR and log odds ratio, respectively.

4.6 MSigDB C4:CGN Gene Set Significant Categories

Plot description: Heatmaps comparing the enrichment scores of the clusters with smallest FDR adjusted Fisher's p-value for the MSigDB Computational Cancer Gene Neighborhood gene sets (C4:CGN) found to be significant (FDR adjusted Fisher's p-value < 0.01) with the three data inputs, namely the normalized expression data with all replicates (CSIMM), the parametric model estimated response data (BMDExpress*), and the BLX spline fit estimated response data (ALOHA). Clusters with the smallest enrichment score not having a minimum of 3 genes mapping to the enriched gene set or populating the gene set by a minimum of 5% were biologically irrelevant due to lack of appropriate information and removed and are not displayed in the heatmap. (Left) Heatmap shows $-log_{10}(FisherFDR)$ of each enriched gene set from each approach. (Right) Heatmap shows $log(OddsRatio)$ of each enriched gene set from each approach. Rows are unique C4:CGN gene sets and columns indicate clustering approaches.

Chlorpromazine All Doses

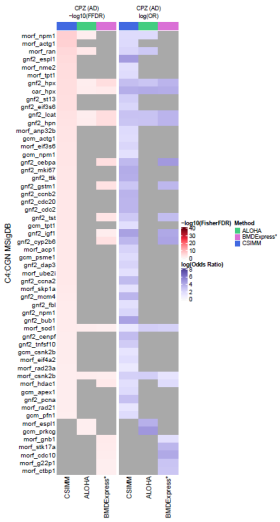


Figure 35: **CPZ AD** Heatmaps of the significant Cancer Gene Neighborhood gene sets comparing the most statistically significant clusters, displaying negative log-scale Fisher FDR and log odds ratio, respectively. Gray indicates which gene sets were not enriched for a given clustering approach.

Chlorpromazine Cytotoxic Doses Removed

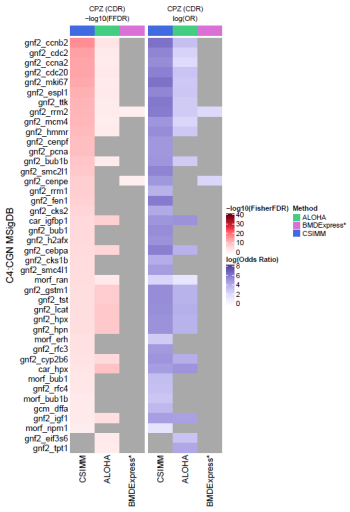


Figure 36: **CPZ CDR** Heatmaps of the significant Cancer Gene Neighborhood gene sets comparing the most statistically significant clusters, displaying negative log-scale Fisher FDR and log odds ratio, respectively. Gray indicates which gene sets were not enriched for a given clustering approach.

Benzo-[α]-pyrene All Doses

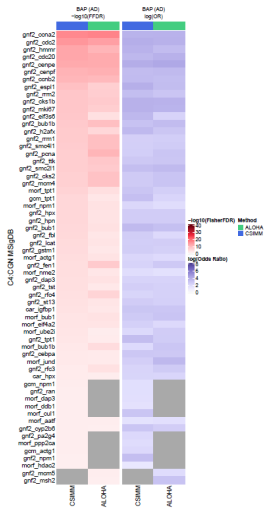


Figure 37: **BAP AD** Heatmaps of the significant Cancer Gene Neighborhood gene sets comparing the most statistically significant clusters, displaying negative log-scale Fisher FDR and log odds ratio, respectively. Gray indicates which gene sets were not enriched for a given clustering approach.

Benzo-[α]-pyrene Cytotoxic Doses Removed

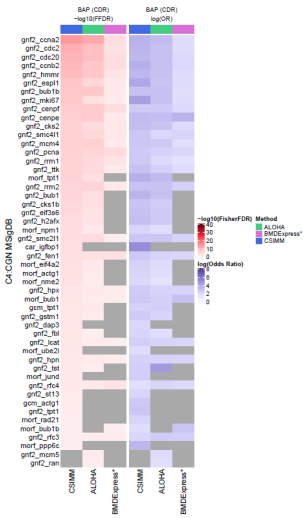


Figure 38: **BAP AD** Heatmaps of the significant Cancer Gene Neighborhood gene sets comparing the most statistically significant clusters, displaying negative log-scale Fisher FDR and log odds ratio, respectively. Gray indicates which gene sets were not enriched for a given clustering approach.

4.7 Gene Set Level Benchmark Dose (BMD) Coherence

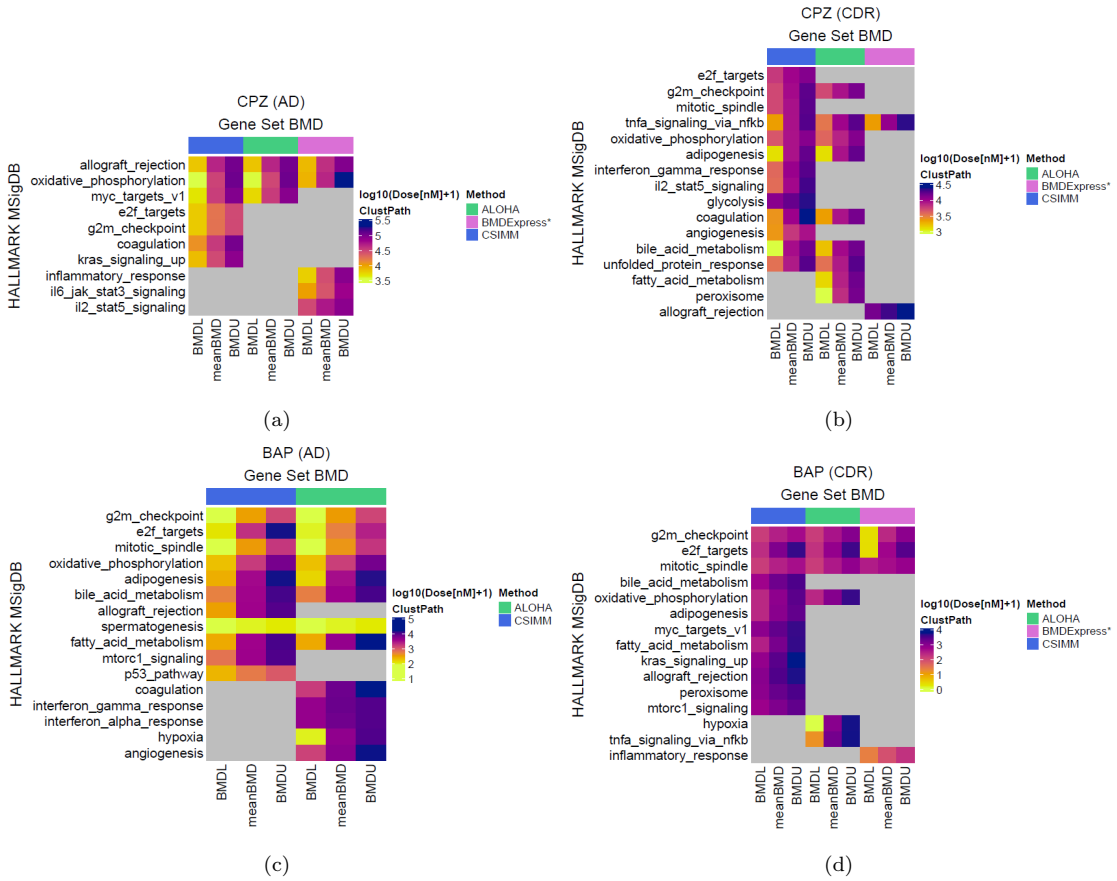
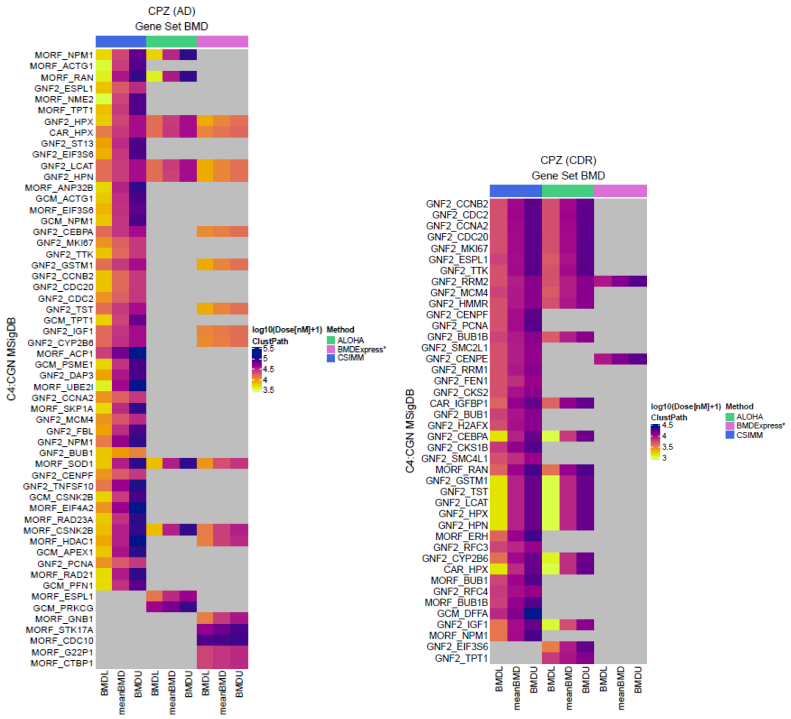
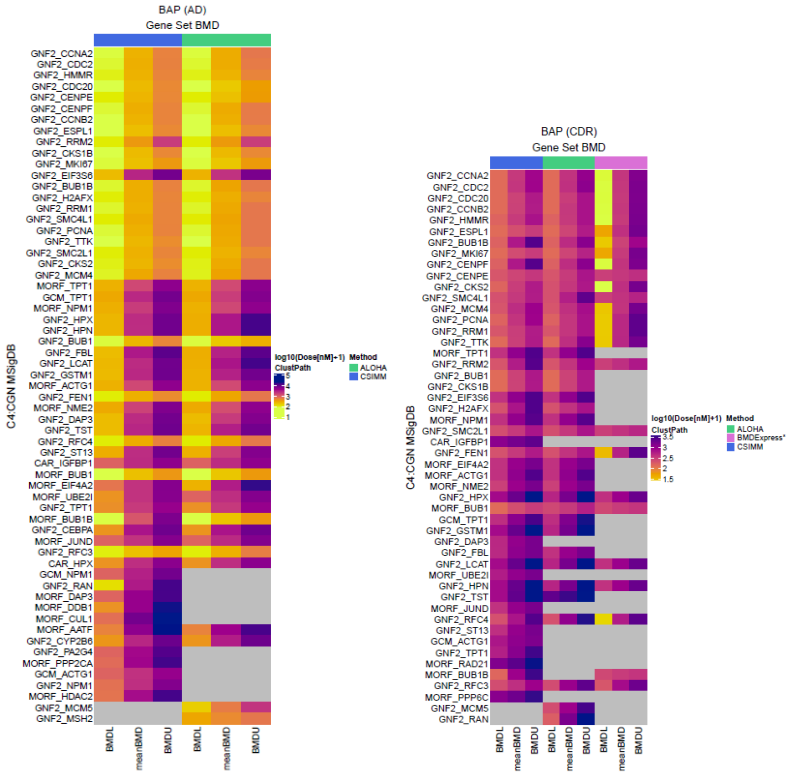


Figure 39: Heatmaps comparing the posterior mean BMD and 95% credible intervals (i.e. BMDL and BMDU) estimates (columns) for enriched Hallmark gene sets [10, 11, 12] (rows) across the three clustering methods. The BMDs are estimated using core genes in the most relevant cluster, as defined in §2.3.3, mapping to the respective gene sets. (a) Chlorpromazine with all experimental dose groups included. (b) Chlorpromazine with cytotoxic dose groups removed. (c) Benzo-[α]-pyrene with all experimental dose groups. (d) Benzo-[α]-pyrene with cytotoxic dose groups removed. Gray blocks within the heatmap indicate which gene sets are not enriched for a clustering approach. Bars on above the columns indicate the various clustering approaches, i.e. CSIMM (blue), ALOHA (green), and BMDExpress clustering (magenta), described in §2.2.



(a) (b)

Figure 40: Heatmaps comparing the posterior mean BMD and 95% credible interval (i.e. BMDL and BMDU) estimates (columns) for enriched Cancer Gene Neighborhoods (CGN) gene sets [10, 11, 12] (rows) across the three clustering methods. The BMDs are estimated using core genes in the most relevant cluster, as defined in §2.3.3, mapping to the respective gene sets. (a) Chlorpromazine with all experimental dose groups included. (b) Chlorpromazine with cytotoxic dose groups removed. Gray blocks within the heatmap indicate which gene sets are not enriched for a clustering approach. Bars on above the columns indicate the various clustering approaches, i.e. CSIMM (blue), ALOHA (green), and BMDExpress clustering (magenta), described in §2.2.



(a) (b)

Figure 41: Heatmaps comparing the posterior mean BMD and 95% credible intervals (i.e. BMDL and BMDU) estimates (columns) for enriched Cancer Gene Neighborhoods (CGN) gene sets [10, 11, 12] (rows) across the three clustering methods. The BMDs are estimated using core genes in the most relevant cluster, as defined in §2.3.3, mapping to the respective gene sets. (a) Benzo-[α]-pyrene with all experimental dose groups included. (b) Benzo-[α]-pyrene with cytotoxic dose groups removed. Gray blocks within the heatmap indicate which gene sets are not enriched for a clustering approach. Bars on above the columns indicate the various clustering approaches, i.e. CSIMM (blue), ALOHA (green), and BMDExpress clustering (magenta), described in §2.2.

4.8 Core Gene Comparison Tables

The following tables contain information about the core probes used to estimate the gene-set level BMDs in Figure 7.

Chlorpromazine All Doses

BMDE genes	CSIMM genes	ALOHA genes	BMDExpress genes
ATF5_501,	CDC20_1157,		
AURKA_586,	HMMR_3040,		
CCND1_1062,	KIF2C_3603,		
CCNF_1076,	MAD2L1_3930,		
CDC7_19277,	TOP2A_7277,		
CENPF_1252,	TPX2_7308		
CKS1B_1362,			
CKS2_19000,			
CTCF_19656,			
DBF4_20417,			
E2F2_2009,			
G3BP1_2542,			
H2AFX_26072,			
H2AFZ_2864,			
HSPA8_3142,			
MAD2L1_3930,			
MAPK14_4000,			
MARCKS_4021,			
MT2A_4334,			
MTF2_4342,			
ORC6_4832,			
PBK_19921,			
PLK1_5203,			
PRMT5_28286,			
PTTG1_5623,			
RACGAP1_19398,			
RAD21_5698,			
SMAD3_27880,			
SRSF10_24587,			
STIL_15150,			
STMN1_6857,			
TFDP1_7045,			
TGFB1_27949,			
TOP2A_7277,			
TPX2_7308			

Table 6: Core probes from the most significant cluster(s) enriched for the Hallmark gene set *"G2M Checkpoint"* in each clustering approach and BMDExpress use to estimate the gene set level BMD estimate and interval for the **CPZ-AD** dataset.

BMDE genes	CSIMM genes	ALOHA genes	BMDExpress genes
ANP32E_26838, ATAD2_22038, AURKA_586, BRCA1_757, BRMS1L_22068, BUB1B_800, CCP110_26981, CDCA3_19307, CDCA8_1185, CDKN2A_28291, CKS1B_1362, CKS2_19000, CTCF_19656, CTPS1_16302, DCK_1776, DEK_1841, DEPDC1_16570, DEPDC1_27104, DLGAP5_1882, H2AFX_26072, H2AFZ_2864, KIF18B_3598, LYAR_12352, MAD2L1_3930, MSH2_19521, NME1_27589, NOP56_4635, ORC6_4832, PAICS_4908, PCNA_4987, PLK1_5203, PSIP1_27736, PSIP1_5500, PTTG1_5623, RACGAP1_19398, RAD21_5698, RAD50_12781, RAD51C_5708, RAN_5725, RFC2_5805, RPA3_5906, SLBP_20689, SPAG5_17378, SPC25_6704, STMN1_6857, TIPIN_11359, TOP2A_7277, UBE2T_7492	CDC20_1157, HMMR_3040, KIF2C_3603, MAD2L1_3930, SPC25_6704, TOP2A_7277		

Table 7: Core probes from the most significant cluster(s) enriched for the Hallmark gene set *"E2F Targets"* in each clustering approach and BMDExpress use to estimate the gene set level BMD estimate and interval for the **CPZ-AD** dataset.

BMDE genes	CSIMM genes	ALOHA genes	BMDExpress genes
ACADVL_20681,	AIFM1_183,	ACAA1_48,	ACADVL_20681,
ACAT1_22692,	ATP5B_26124,	AIFM1_183,	ACAT1_22692,
AIFM1_183,	ATP5C1_17193,	ATP5B_26124,	AIFM1_183,
ATP1B1_527,	ATP6AP1_14351,	ATP5C1_17193,	ATP1B1_527,
ATP5B_26124,	ATP6V0B_562,	ATP6AP1_14351,	ATP5B_26124,
ATP5C1_17193,	ATP6V0E1_23077,	ATP6V0B_562,	ATP5C1_17193,
ATP6AP1_14351,	ATP6V1D_574,	ATP6V0E1_23077,	ATP6AP1_14351,
ATP6V0E1_23077,	COX8A_19058,	ATP6V1D_574,	ATP6V0E1_23077,
ATP6V1D_574,	CYCS_1692,	COX8A_19058,	ATP6V1D_574,
BAX_28222,	DECR1_1831,	CYC1_27087,	BAX_28222,
CASP7_985,	ECH1_2022,	CYCS_1692,	CASP7_985,
CYCS_1692,	ECI1_22689,	DECR1_1831,	CYCS_1692,
DLAT_1875,	ETFB_2245,	ECH1_2022,	DLAT_1875,
ECH1_2022,	GPX4_17664,	ECI1_22689,	ECH1_2022,
ECI1_22689,	LDHA_3766,	ETFB_2245,	ECI1_22689,
ETFA_2244,	MDH1_4087,	GLUD1_2678,	ETFA_2244,
ETFB_2245,	MRPL15_4264,	GPX4_17664,	ETFB_2245,
GLUD1_2678,	NDUFA1_4514,	LDHA_3766,	GLUD1_2678,
GPI_2753,	NDUFB3_18124,	MDH1_4087,	GPI_2753,
GPX4_17664,	NDUFV2_4530,	MRPL15_4264,	GPX4_17664,
HSD17B10_3112,	PDHX_5033,	NDUFA1_4514,	HSD17B10_3112,
IDH3A_15224,	PRDX3_27724,	NDUFB3_18124,	IDH3A_15224,
MDH1_4087,	SDHB_6173,	NDUFV2_4530,	MDH1_4087,
MDH2_20964,	SLC25A4_6406,	PDHX_5033,	MDH2_20964,
MRPL15_4264,	SLC25A5_20900,	PRDX3_27724,	MRPL15_4264,
NDUFA1_4514,	SLC25A6_14758,	SDHB_6173,	NDUFA1_4514,
NDUFB3_18124,	SUPV3L1_6897,	SLC25A4_6406,	NDUFB3_18124,
NDUFV2_4530,	UQCRC1_13094,	SLC25A5_20900,	NDUFV2_4530,
PDHB_13121,	UQCRFS1_7566,	SLC25A6_14758,	PDHB_13121,
PDHX_5033,	VDAC1_7642,	SUPV3L1_6897,	PDHX_5033,
POR_23019,	VDAC2_28051	UQCRC1_13094,	POR_23019,
PRDX3_27724,		UQCRFS1_7566,	PRDX3_27724,
SDHB_6173,		VDAC1_7642,	SDHB_6173,
SLC25A4_6406,		VDAC2_28051	SLC25A4_6406,
SLC25A5_20900,			SLC25A5_20900,
SLC25A6_14758,			SLC25A6_14758,
SUCLA2_6870,			SUCLA2_6870,
UQCRC1_13094,			UQCRC1_13094,
UQCRQ_7569,			UQCRQ_7569,
VDAC1_7642,			VDAC1_7642,
VDAC2_11431,			VDAC2_11431,
VDAC2_28051,			VDAC2_28051,
VDAC2_28052			VDAC2_28052

Table 8: Core probes from the most significant cluster(s) enriched for the Hallmark gene set *"Oxidative Phosphorylation"* in each clustering approach and BMDExpress use to estimate the gene set level BMD estimate and interval.

BMDE genes	CSIMM genes	ALOHA genes	BMDExpress genes
ACE_26794, ADAM8_12335, CA2_908, CFH_27007, CTSS_25157, DOCK2_20159, DUSP6_1993, ELTD1_21057, ELTD1_28345, EMP1_18941, ENG_13441, ETS1_27185, ETV1_2251, ETV1_27186, FCER1G_11058, G0S2_13343, GLRX_2674, IGFBP3_3270, IKZF1_3293, IL1B_3325, IL2RG_15624, INHBA_21251, IRF8_20223, KIF5C_3612, LAPTM5_23860, SERPINA3_17487, SNAP25_6607, TLR8_16795	ACE_26793, DUSP6_1993, ELTD1_21057, ELTD1_28345, ENG_13441, ETV1_2251, IL2RG_15624, INHBA_21251, KIF5C_3612, LY96_18964, MAP7_3988, TLR8_16795		

Table 9: Core probes from the most significant cluster(s) enriched for the Hallmark gene set *"KRAS Signaling Up"* in each clustering approach and BMDExpress use to estimate the gene set level BMD estimate and interval.

Benzo-[α]-pyrene All Doses

BMDE genes	CSIMM genes	ALOHA genes	BMDExpress genes
AURKA_586,	AURKB_587,	AURKB_587,	
AURKB_587,	CCNA2_1050,	CCNA2_1050,	
BIRC5_709,	CCNF_1076,	CCNF_1076,	
CCNF_1076,	CDC20_1157,	CDC20_1157,	
CDC25B_1161,	CDC45_26987,	CDC45_26987,	
CENPA_1247,	CDC6_1177,	CDC6_1177,	
CENPF_1252,	CDK1_1196,	CDK1_1196,	
CHEK1_1288,	CENPF_1252,	CENPF_1252,	
CKS1B_1362,	CKS1B_1362,	CKS1B_1362,	
CKS2_19000,	EZH2_2282,	CKS2_19000,	
EGF_2055,	FBXO5_10555,	EZH2_2282,	
EXO1_17006,	HMMR_3040,	FBXO5_10555,	
EZH2_2282,	KIF11_3588,	HMMR_3040,	
FBXO5_10555,	KIF23_14714,	KIF11_3588,	
G3BP1_2542,	KIF2C_3603,	KIF23_14714,	
H2AFX_26072,	MKI67_28355,	KIF2C_3603,	
HIF1A_2945,	NEK2_20563,	MCM3_4072,	
HMMR_3040,	NUSAP1_4789,	MKI67_28355,	
KIF15_3595,	PLK1_5203,	MT2A_4334,	
KIF23_14714,	PRC1_14588,	NEK2_20563,	
KIF2C_3603,	SMC1A_6580,	NUSAP1_4789,	
KIF4A_28251,	SMC4_6586,	ORC6_4832,	
MAD2L1_3930,	TACC3_15729,	PLK1_5203,	
MCM3_4072,	TPX2_7308	PRC1_14588,	
MKI67_28355,		RACGAP1_19398,	
MT2A_4334,		RAD54L_14089,	
MTF2_4342,		SMC1A_6580,	
NOTCH2_18421,		SMC4_6586,	
ORC6_4832,		TACC3_15729,	
PLK1_5203,		TPX2_7308	
PRC1_14588,			
RACGAP1_19398,			
RAD21_5698,			
RAD54L_14089,			
RPA2_5905,			
SLC7A5_13711,			
SMAD3_27880,			
SMARCC1_6575,			
SMC1A_6580,			
SMC2_12897,			
STIL_15150,			
TACC3_15729,			
TTK_7394,			
UBE2C_7468			

Table 10: Core probes from the most significant cluster(s) enriched for the Hallmark gene set *"G2M Checkpoint"* in each clustering approach and BMDExpress use to estimate the gene set level BMD estimate and interval for the **BAP-AD** dataset.

BMDE genes	CSIMM genes	ALOHA genes	BMDExpress genes
ANLN_12002, ARHGEF2_423, AURKA_586, BCL2L11_26886, BIRC5_709, CDC42_1168, CENPF_1252, CEP57_1257, DLGAP5_1882, FBXO5_10555, HDAC6_2901, KIF15_3595, KIF23_14714, KIF2C_3603, KIF4A_28251, NCK1_4491, NOTCH2_18421, OPHN1_28310, PCM1_4980, PLK1_5203, PRC1_14588, PXN_5642, RACGAP1_19398, SMC1A_6580, SMC3_6583, SOS1_10682, SPTAN1_6727, TTK_7394	ANLN_12002, CDK1_1196, CENPF_1252, FBXO5_10555, KIF11_3588, KIF23_14714, KIF2C_3603, NEK2_20563, NUSAP1_4789, PLK1_5203, PRC1_14588, SMC1A_6580, SMC4_6586, TPX2_7308	ANLN_12002, CDK1_1196, CENPF_1252, FBXO5_10555, KIF11_3588, KIF23_14714, KIF2C_3603, NEK2_20563, NUSAP1_4789, PLK1_5203, PRC1_14588, RACGAP1_19398, SMC1A_6580, SMC4_6586, TPX2_7308	

Table 11: Core probes from the most significant cluster(s) enriched for the Hallmark gene set *"Mitotic Spindle"* in each clustering approach and BMDExpress use to estimate the gene set level BMD estimate and interval for the **BAP-AD** dataset.

BMDE genes	CSIMM genes	ALOHA genes	BMDExpress genes
ACAA1_48,	ATP1B1_527,	ACAA1_48,	
ACADVL_20681,	ATP5C1_17193,	ATP1B1_527,	
ALAS1_28273,	COX6B1_1519,	COX6B1_1519,	
ATP5J_13202,	CYC1_27087,	CYC1_27087,	
ATP6AP1_14351,	CYCS_1692,	CYCS_1692,	
ATP6V1D_574,	DLAT_1875,	DECR1_1831,	
DLAT_1875,	GPI_2753,	DLAT_1875,	
ECI1_22689,	GPX4_17664,	GPI_2753,	
ETFA_2244,	LDHA_3766,	GPX4_17664,	
MDH1_4087,	MAOB_13383,	LDHA_3766,	
NDUFA1_4514,	MDH2_20964,	MDH2_20964,	
NDUFB3_18124,	MRPL15_4264,	NDUFA1_4514,	
NDUFV2_4530,	NDUFA1_4514,	NDUFA4_14125,	
POR_23019,	NDUFA4_14125,	NDUFB3_18124,	
PRDX3_27724,	NDUFB3_18124,	NDUFV2_4530,	
SDHB_6173,	NDUFV2_4530,	NNT_4618,	
SUPV3L1_6897,	NNT_4618,	SDHB_6173,	
TIMM9_7106,	SDHB_6173,	SLC25A4_6406,	
VDAC1_7642,	SLC25A4_6406,	SLC25A5_20900,	
VDAC2_11431	SLC25A6_14758,	SLC25A6_14758,	
	SUCLA2_6870,	SUCLA2_6870,	
	UQCRQ_7569	UQCRQ_7569	

Table 12: Core probes from the most significant cluster(s) enriched for the Hallmark gene set *"Oxidative Phosphorylation"* in each clustering approach and BMDExpress use to estimate the gene set level BMD estimate and interval for the **BAP-AD** dataset.

BMDE genes	CSIMM genes	ALOHA genes	BMDExpress genes
APP_367,	AEN_18198,		
ATF3_499,	CASP1_26966,		
BLCAP_713,	CCP110_26981,		
BTG2_13191,	DDB2_1798,		
CASP1_26966,	FAS_2345,		
CCP110_26981,	GADD45A_2569,		
DDIT3_16736,	PPM1D_25317,		
FAS_2345,	PRKAB1_14557,		
FOXO3_2481,	PROCR_24511,		
GPX2_11201,	SDC1_6160,		
IKBKAP_3288,	XPC_28078		
IL1A_22714,			
MDM2_23384,			
MDM2_27496,			
PHLDA3_18856,			
RAD51C_5708,			
RALGDS_5721,			
RNF19B_24383,			
SAT1_17598,			
SFN_6269,			
SLC7A11_14100,			
TM7SF3_22780,			
TP53_7287,			
TRIAP1_21159			

Table 13: Core probes from the most significant cluster(s) enriched for the Hallmark gene set *"P53 Pathway"* in each clustering approach and BMDExpress use to estimate the gene set level BMD estimate and interval for the **BAP-AD** dataset.

References

[1] Scott S. Auerbach. BMDExpress 2.3, 2017. URL `https://github.com/auerbachs/BMDExpress-2/wiki`.

[2] Jason R Phillips, Daniel L Svoboda, Arpit Tandon, Shyam Patel, Alex Sedykh, Deepak Mav, Byron Kuo, Carole L Yauk, Longlong Yang, Russell S Thomas, et al. Bmdexpress 2: enhanced transcriptomic dose-response analysis workflow. *Bioinformatics*, 35(10):1780–1782, 2019.

[3] Longlong Yang, Bruce C Allen, and Russell S Thomas. Bmdexpress: a software tool for the benchmark dose analyses of genomic data. *BMC genomics*, 8(1):387, 2007.

[4] Dayne L Filer, Parth Kothiya, R Woodrow Setzer, Richard S Judson, and Matthew T Martin. tcpl: the toxcast pipeline for high-throughput screening data. *Bioinformatics*, 33(4):618–620, 2017.

[5] Joshua A. Harrill. Going broad: A high throughput transcriptomics (httr) bioactivity screen of the toxcast chemical library in mcf7 cells using targeted rna-seq. URL `https://cfpub.epa.gov/si/si_public_file_download.cfm?p_download_id=536111`.

[6] Joe Reynolds, Sophie Malcomber, and Andrew White. A bayesian approach for inferring global points of departure from transcriptomics data. *Computational Toxicology*, 16:100138, 2020.

[7] Lyle D Burgoon, Ingrid L Druwe, Kyle Painter, and Erin E Yost. Using in vitro high-throughput screening data for predicting benzo [k] fluoranthene human health hazards. *Risk Analysis*, 37(2): 280–290, 2017.

[8] Gilles Celeux, Merrilee Hurn, and Christian P Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970, 2000.

[9] Matthew Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

[10] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[11] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6):417–425, 2015.

[12] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

Chapter 3

Dose-Response Co-expression Clustering of Gene Responses in Mouse Lung after Engineered Nanomaterial Exposure

The following chapter contains the draft manuscript and supplemental materials describing the application of ALOAH to publicly-available gene expression data on engineered nanomaterial (ENM) exposures.

Dose-Response Co-expression Clustering of Gene Responses in Mouse Lung after Engineered Nanomaterial Exposure

Sarah E. Davidson[1], Michael Borchers[2], Eileen D. Keumpel[3]

[1] University of Cincinnati Department of Environmental Health Sciences Division of Biostatistics and Bioinformatics, Cincinnati, OH
[2] University of Cincinnati Department of Internal Medicine College of Medicine Internal Medicine Pulmonary Division, Cincinnati, OH
[3] National Institute for Occupational Safety and Health (NIOSH) Emerging Technologies Branch, Cincinnati, OH

Abstract

Over the past couple of decades, engineered nanomaterials (ENMs) have become more widely utilized in industrial applications and are under continual development. Similarities in the physical-chemical properties of nanomaterials with other micro-scale particulates, e.g. asbestos, known to cause adverse pulmonary health effects in humans. This makes them a priority for risk assessment and necessitates the use of more efficient methods for assessing the risk of developing adverse health effects. Here, we evaluate publicly available genomic dose-response data from nanomaterial exposure studies utilizing a novel genomic dose-response modeling approach, which utilizes flexible model fitting with gene clustering to estimate a point-of-departure (POD) from gene groups with similar dose-response patterns. We compared the genomic benchmark dose estimates to those obtained from *in vivo* endpoints indicative of pulmonary inflammation and fibrosis. Our results showed we were able to identify gene clusters that were biologically relevant for understanding the development of pulmonary inflammation and fibrosis after exposure to nanomaterials. Genomic benchmark dose estimates from the most sensitive gene clusters aligned well with PODs obtained from the *in vivo* pulmonary endpoints. In this study, we demonstrated dose-response co-expression clusters can be used for evaluating biological mechanisms linked with adverse health while estimating a POD, which can be leveraged to prioritize nanomaterial exposures for further risk assessment.

Keywords: Engineered nanomaterials, Genomic dose-response, Gene clustering, Pulmonary Inflammation & Fibrosis

1 Introduction

Engineered nanomaterials (ENMs) are particles with at least one dimension measuring less than 100 nm. Their size induces unique physical and chemical properties are useful for developing innovative technologies in fields of industry ranging from aerospace engineering to biomedical devices (1). Due to their small size, when aerosolized and inhaled these particles can deposit in alveolar regions in the lung where adverse cellular level interactions can occur as a result of the unique properties, see Oberdörster et al. (2), Schulte et al. (3) and references therein. Shared structural similarities between nanomaterials and known micro-scale toxicants, e.g. asbestos, also indicate

potential hazard to the pulmonary health of exposed individuals, see NIOSH (4, 5, 6) and references therein. Increased use and development of nanomaterials necessitates the development of efficient risk assessment evaluations since traditional approaches are time and cost intensive, see Fadeel et al. (7), Labib et al. (8), Williams and Halappanavar (9, 10) and references there-in.

Developing risk assessment methods using high-throughput data, such as genomics, transcriptomics, and proteomics, was sparked by the Tox21 initiative, see (11, 12, 13, 14, 15, 16) and references therein. Since its inception the initiative has encouraged many researchers to produce new high-throughput testing and assessment methods for evaluating existing perturbagens currently without any *in vivo* toxicology data. Comparative studies with gene expression and traditional *in vivo* toxicity biomarkers have shown high-throughput approaches can be used to estimate transcriptional points-of-departure (POD) comparable to those from traditional biomarkers (17, 18, 19, 20, 21, 22). High-throughput analyses of molecular and cellular endpoints also allow researchers to explore the mechanisms of action (MOA) (18) and etiology of disease development (23). Mechanistic information can be used to then develop more generalized adverse outcome pathways (AOPs) independent of a particular perturbagen (24, 25). This information can be used to help group nanomaterials inducing adverse pulmonary outcomes via a common AOP and determine similarities in their physicochemical properties which produce these effects.

Several high-throughput analysis approaches have been used to evaluate bio-molecular responses after exposure to nanomaterials. One approach is to identify altered genes and pathways in disease models, for example pulmonary fibrosis or lung cancer, to compare with samples from cell lines or animals exposed to nanomaterials. Williams and Halappanavar (9, 10) use bi-clustering to derive gene sets shared among related lung disease models and DAVID (26, 27) to identify the biological functionality of these sets, i.e. functional classification. These gene sets are then used to perform gene set enrichment analyses (GSEA) (28, 29) on the samples obtained after a nanomaterial exposure (9, 10). The results could then be used to identify commonly perturbed mechanisms, which may indicate adverse pulmonary effects down-stream nanomaterial inhalation exposure. The meta-analysis conducted by Nikota et al. (30) used hierarchical clustering of nanomaterials exposure and lung disease models to identify perturbed pathways and key genes within clusters to better understand the induced adverse pulmonary effects. Though these approaches provide mechanistic information about nanomaterial exposures and potential adverse pulmonary effects the analyses do not allow for quantitative evaluation of the relationship between dose and transcriptional responses. Quantitative analyses are required to estimate a dose producing a minimal biological response (i.e. benchmark dose – BMD) for a human health risk assessment.

Studies that have reported quantitative dose-response modeling of genomic data have used BMDExpress (31, 32, 33) to estimate BMDs by modeling individual genes/probes and evaluating perturbed molecular mechanisms using functional classification tools and/or adverse outcome pathways (AOPs), see Labib et al. (8), Halappanavar et al. (34, 35), Rahman et al. (36) and references therein. The functional classification approach used in BMDExpress (31, 32, 33) place genes remaining after dose-response modeling into predetermined gene sets which they are a member. However, this approach is limited since it does not account for relationships between dose-response curves of co-regulated genes. AOPs may also be used in conjunction with functional classification results to incorporate known events and relationships along the pathway of progression towards a specified adverse pulmonary outcome. For example, Labib et al. (8) compares the gene sets with the most sensitive BMD estimates from BMDExpress with their proposed AOP pulmonary fibrosis after a nanomaterial exposure. However, currently developed AOPs for may not be applicable for new materials or target outcomes being evaluated, and development of a new AOP can be time

intensive. Thus, this will create a bottle-neck in the evaluation process of new materials that are developed and utilized in industry potentially impacting the health of exposed individuals.

In this study we evaluate co-expressed genomic dose-response patterns in the mouse lung tissue after exposure to three commonly produced and relatively well-studied nanomaterials using the Aggregated Local Extrema Splines for High-throughput dose-response analysis (ALOHA) method (37). This group of nanomaterials includes multiwalled carbon nanotubes (MWCNT), titanium dioxide (TiO_2), and carbon black (CB). Enriched gene sets and their corresponding BMD estimates are compared among the three materials to evaluate similarities and differences in altered pathways. Finally, we compare gene set BMDs with BMDs from *in vivo* pulmonary endpoints indicative of chronic inflammation and/or fibrosis development to evaluate the biological relevance of the transcriptional BMD estimates from dose-response co-expression clusters. Identification of perturbed molecular mechanisms with transcriptional BMD estimates predictive of traditional toxicological endpoints for chronic pulmonary inflammation and fibrosis can help to better understand the mechanisms of developing these adverse health effects. We consider transcriptional BMDs to be predictive if the estimates are similar to the POD estimates from the traditional toxicological endpoints or their uncertainty intervals overlap. This can then be used to establish a framework for future evaluations of nanomaterials with high-throughput dose-response modeling methods.

2 Methods

2.1 Gene Expression Data

We use three publicly available microarray datasets from the Gene Expression Omnibus (GEO) Database (38, 39) for our genomic dose response modeling analyses (see Table 1). These studies were selected from results of a systematic data search in GEO based on their inclusion of the commonly produced nanomaterials (40) –namely MWCNT, TiO_2, and CB – with adequate data for dose-response modeling (minimum of three dose groups and untreated controls), same or similar routes of exposure (e.g. pharyngeal aspiration or intratracheal instillation), and same rodent species (mouse C57BL/6 or C57BL/6J). Two studies used female mice and the other male mice (Table 1). We refer the reader to §1 of the Supplementary Material for further details on the search dates, terms, and results.

All processing and analyses are conducted with R statistical programming software (41, 42). Data pre-processing is performed using the limma Bioconductor package (43). Raw expressions are background corrected using the normal-exponential ("normexp") convolution method (44), normalized with the loess method (45) and log2 transformed. Then, probes are mapped to their corresponding Entrez Gene ID (EntrezID). If a probe maps to multiple or no EntrezID it is removed from the analysis since the gene mapping is ambiguous, unknown, or a control probe. We average the expressions of probes mapping to the same EntrezID, since microarray probes typically capture different sections of a gene transcript. Pre-processed data are formatted into dose-response datasets for individual post-exposure time-points using a custom R script, see (46). Each dose-response dataset is then analyzed with ALOHA (37).

GEO Series (Platform) (Publication*)	Material	Doses (μg)	Route of Exposure	Species Strain (Sex)	Tissue
GSE29042 (GPL4134) (47)	Multi-walled Carbon Nanotube (Mitsui-7)	0,10,20, 40,80	Pharyngeal Aspiration	Mice C57BL/6J (Male)	Whole Lung
GSE35193 (GPL4134) (48)	Carbon Black Nanoparticles (Printex 90)	0,18, 54,162	Intratracheal Instillation	Mice C57BL/6 (Female)	Whole Lung
GSE41041 (GPL7202) (49)	Titanium Dioxide (UV Titan)	0,18, 54,162	Intratracheal Instillation	Mice C57BL/6 (Female)	Whole Lung

Table 1: Publicly available microarray gene expression datasets on ENM exposure that were included in our analyses. * The primary related source for each GEO dataset.

2.2 ALOHA genomic dose-response analysis

Genomic dose-response (GDR) modeling with ALOHA (37) is performed with the 1- and 28-days post-exposure (PE) time points. ALOHA first fits individual genes with a BLX spline taking 50,000 iterations and discarding the first 10,000 as burn-in (37, 50). We use the BLX fits to estimate benchmark doses (BMDs) for individual genes with a benchmark response (BMR) of 1 standard deviation change from control (BMR = 1) which is consistent with current approaches (21, 37, 51). Genes with a posterior mean BMD estimates greater than the maximum dose and 95% credible bound ratio (BMDU/BMDL) greater than 40 are removed from further analyses (37, 51).

After model fitting, ALOHA clusters genes using BLX estimated dose-response data and the Bayesian context-specific infinite mixture model (CSIMM) developed by (52). The clustering process is run for 50,000 iterations with the first 10,000 iterations discarded as burn-in (37). Enriched pathways in resulting clusters are determined with Cluster Enrichment Analysis (CLEAN) (53). We use three *Mus musculus* (mouse) gene lists for evaluating biological relevance, including the Hallmark (H) and Computational Cancer Gene Neighborhoods (C4:CGN or CGN) lists from the Molecular Signatures Database (MSigDB) (28, 54, 55, 56) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathways list (57, 58, 59). All gene sets we use in the CLEAN analysis must contain at least 3 genes from the input data, and gene sets not meeting this criterion are removed (37, 51).

We classify enriched gene sets to be those with a Benjamini-Hochberg False Discovery Rate Adjusted Fisher's exact test p-value (FFDR) less than 0.01 for at least one cluster (53, 60, 61, 62, 63). The cluster with the smallest FFDR for an enriched gene set is the most relevant since it is the cluster of genes that most represents the set if the total number of genes were chosen at random. We remove gene sets with a representative cluster that does not contain at least 3 of its genes or populate the gene set by at least 5% since there is not enough information for biological interpretation in accordance with current guidance in NTP et al. (51). Gene set BMDs are estimated from individual BMD estimates from core genes; we refer the reader to (37) for further details.

Study	Material	Doses (μg)	Route of Exposure	Species Strain (Sex)	Endpoint	PE Time-point
Porter et al. (64)	MWCNT (Mitsui-7)	0, 10, 20, 40, 80	Pharyngeal aspiration	Mice C57BL/6J (Male)	% PMN	1 & 28 days
Mercer et al. (65)	MWCNT (Mitsui-7)	0, 10, 20, 40, 80	Pharyngeal aspiration	Mice C57BL/6J (Male)	Alveolar thickness	56 days
Bourdon et al. (66)	CB (Printex 90)	0, 18, 54, 162	Intratracheal instillation	Mice C57BL/6 (Female)	% Neutrophils$^{\triangle}$	1 & 28 days
Elder et al. (67)	CB (Printex 90)	0, 15.98, 245.3, 1388*	Inhalation	Mice B6C3F1	% PMN	1 day
Saber et al. (68)	TiO_2 (UV Titan)	0, 18, 54, 162	Intracheal instillation	Mice C57BL/6 (Female)	% Neutrophils$^{\triangle}$	1 & 28 days
Bermudez et al. (69)	Ultrafine TiO_2	0, 10.03, 38.82, 293.7*	Inhalation	Mice B3C3F1/ CrlBR (Female)	% Neutrophils	0 weeks

Table 2: Published nanomaterial exposure studies and *in vivo* pulmonary endpoint data from whole mouse lung included in our analyses. Each *in vivo* dataset is matched by material to the genomic datasets listed in Table 1.
* Comparable dose from inhalation to pharyngeal aspiration or intratracheal instillation is the estimated lung burden in wet lung at the end of exposure (see Supplement §2 Supplement for details).
$^{\triangle}$ Mean and standard deviation for percent neutrophils are estimated from summarized count data (see Supplement §3 for details).

2.3 *In vivo* Pulmonary Endpoint Data

In vivo pulmonary endpoint data and/or BMDs are obtained from the literature. Some BMD estimates we include are previously derived values from existing risk assessment documents (5). Otherwise, we estimate BMDs using *in vivo* pulmonary data reported in publications either related to the studies in Table 1 or those with similar experimental designs. Endpoints of interest include neutrophil (or polymorphonuclear leukocyte – PMN) cell composition in bronchioalveolar lavage fluid (BALF) or alveolar epithelial tissue thickness, which are related to pulmonary inflammation and fibrosis, respectively. Table 2 provides details about the *in vivo* endpoints used in our analyses and the publications.

2.4 *In vivo* Pulmonary Endpoint Benchmark Dose Modeling

We fit the inflammatory endpoints with all continuous models in Benchmark Dose Modeling Software (BMDS – Version 2.7.0.4) (70), considering both constant and non-constant variance. For each

inflammatory endpoint, we estimate two BMDs one specifying the benchmark response (BMR) as 1 standard deviation change from the control response (71) and another specifying the BMR as an added 4% response from the control response (4). For the fibrosis endpoint, i.e. alveolar thickness data from Mercer et al. (65), we are using the BMD estimate reported in National Institute for Occupational Safety and Health Current Intelligence Bulletin on carbon nanotubes (CNT) and carbon nanofibers (5), a quantitative risk assessment was performed using data from published toxicology studies in rats or mice, including MWCNT (Mitsui-7) included in this analysis (65), see NIOSH (5) for details.

3 Results

3.1 Between Material Transcriptional Response

We compared transcriptional responses between materials using the enriched gene sets at 1-day and 28-days post-exposure. Figure 1 shows a heatmap of the enriched Hallmark gene sets and their corresponding gene set level BMD estimates for each material at 1- and 28-days post-exposure.

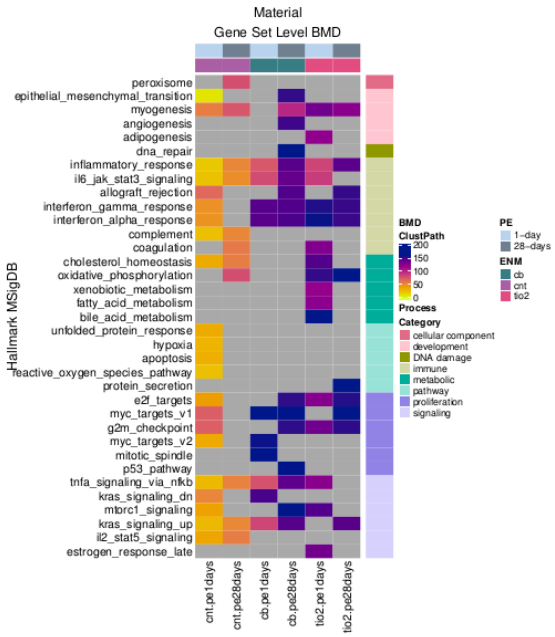


Figure 1: Heatmap displaying benchmark dose (BMD) estimates of enriched gene sets from their most relevant cluster, see §2.2, for lung tissue of mice exposed to multiwalled-carbon nanotubes (MWCNT - GSE29042 (47)), carbon black (CB - GSE35193 (66)), and titanium dioxide ($TiO2$ - GSE41041 (49)). Rows display the enriched gene sets across all materials and time points, which are grouped by designated "Process Categories" for Hallmark gene sets in Subramanian et al. (28), Liberzon et al. (54), Liberzon (55), Liberzon et al. (56). Columns display the material and the post-exposure time-point datasets, which are indicated by colored bars at top of the heatmap. Gray boxes in the heatmap indicate where a gene set is not enriched for a given material and time-point.

The most represented biological processes among all three materials and time points were immune and signaling processes. In addition to these processes we also observed several cell cycle and tissue repair mechanisms shared across the materials, namely myogenesis, E2F targets, and MYC targets version 1. Of the three materials, carbon nanotubes at 1-day post-exposure had the highest number of enriched gene sets in the shared biological processes. There were some Hallmark gene sets enriched for only one of the three materials. For example the angiogenesis, DNA-repair, and P53 pathway Hallmark gene sets were exclusively enriched for carbon black at 28-day post-exposure. Except for carbon black, there was a decrease in the number of enriched Hallmark gene sets from 1- to 28-days post-exposure. This suggests there is either resolution in the molecular response to CNT and TiO_2 after 28-days, or the responses at 1-day post-exposure activated a protective response to evade lung injury. We observed similar trends with the enriched KEGG and CGN gene sets (see Figures in §6 in the Supplemental Material). Though we observe a reduction in the number of enriched gene sets and genes 28-days after exposure to CNT, the remaining genes and gene sets are related pathways that play a role in the development of pulmonary fibrosis.

CNT had the smallest BMD estimates among all enriched Hallmark gene sets at both time points, indicating it was the most potent of the three materials. Gene set BMD estimates for CB and TiO_2 were near their maximum experimental dose ($162\mu g$) except for the immune and signaling processes at 1-day post-exposure ($BMD = \{68.87\mu g, 72.91\mu g, 83.57\mu g, 99.42\mu g\}$) and myogenesis at 28-days post-exposure ($BMD = 95.39\mu g$). Overall, for gene sets enriched at both time points within a material BMD estimates increased from 1- to 28-days post-exposure. These results were also observed with the enriched KEGG Pathway and CGN gene sets (see Figures in §6 in the Supplemental Material).

3.2 Within Material Transcriptional & *In vivo* Pulmonary Responses

To evaluate the relevance of gene set level BMD estimates concerning pulmonary inflammation and fibrosis we used the most sensitive gene set (i.e. smallest gene set level BMD estimate) from each of the three gene lists. Figures 2-4 show the BMD estimates for the *in vivo* pulmonary endpoints and the most sensitive gene sets at both post-exposure times. In most cases, the gene set level BMD and BMDL estimates from the most sensitive pathways were higher than the BMD estimates for *in vivo* pulmonary inflammation endpoints. For MWCNT, which was the only material to elicit a fibrotic response in mouse lungs *in vivo*, the most sensitive gene set level BMDLs at 28-days post-exposure were in the BMD (BMDL) interval obtained from the 56-day fibrosis endpoint NIOSH (5), Mercer et al. (65), see Figure 2. The BMD and/or BMDL estimates from the most sensitive gene sets 1-day after exposure to MWCNT were also either contained in the BMD (BMDL) interval for the *in vivo* fibrosis endpoint or they were less than the interval, see Figure 2.

Though we were able to estimate BMDs (BMDLs) from *in vivo* pulmonary inflammation endpoints many of these were unreliable. Uncertainty in the estimates was the result of one of three cases. First, the fit statistics for the best model fit indicated a good model fit, but the data were highly variable making the model fit and subsequent BMD (BMDL) estimates unreliable, see §4.1 in the Supplemental Material. Second, the best model fit was saturated (i.e. Goodness-of-fit=NA) and resulted in BMD and BMDL estimates near zero, see §4.4 in the Supplemental Material. In these cases, the best model fit does not provide any additional information, and the estimated BMD (BMDL) near zero is unsuitable for risk assessment. Third, none of the parametric models provide a reasonable fit of the data with fit statistics indicating a good model fit, see §4.5 in the Supplemental Material. In this case, parametric model fitting does not provide any useful information.

One explanation for these results is the observed percent PMN or neutrophil response went from a near-zero response to a near maximal response in subsequent dose groups. This induces a hockey-stick shaped dose-response curve, which misses the change in response, and results in unreliable parametric model fits and BMD (BMDL) estimates.

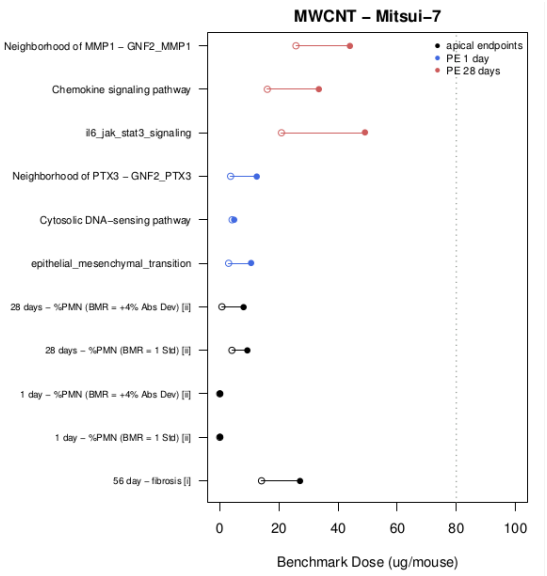


Figure 2: *In vivo* pulmonary and gene set BMD comparison plot for multi-walled carbon nanotube (MWCNT) exposure in mice. Black segments indicate the *in vivo* pulmonary BMD (BMDL) intervals. Blue and red indicate the most sensitive BMD (BMDL) intervals for 1- and 28-days post-exposure, respectively. The vertical dashed line indicates the maximum experimental dose from the genomic dataset (GSE29042). Represented data are from [i]Mercer et al. (65) & NIOSH (5), and [ii] Porter et al. (64), see §4 and Table 19 in the Supplemental Material for details.

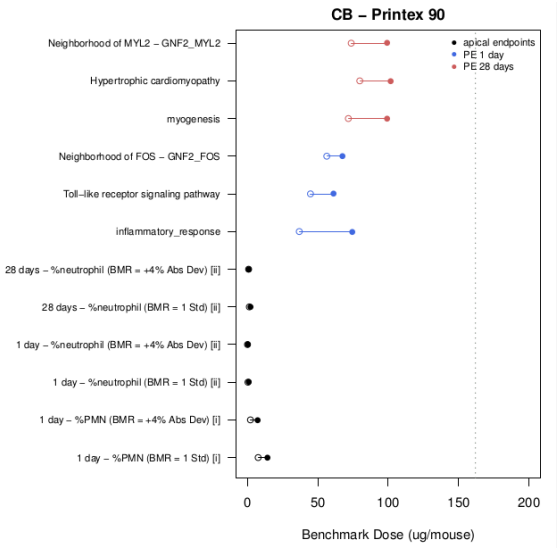


Figure 3: *In vivo* pulmonary and gene set BMD comparison plot for carbon black (CB) exposure in mice. Black segments indicate the *in vivo* pulmonary BMD (BMDL) intervals. Blue and red indicate the most sensitive BMD (BMDL) intervals for 1- and 28-days post-exposure, respectively. The vertical dashed line indicates the maximum experimental dose from the genomic dataset (GSE35193). Represented data are from [i]Elder et al. (67) and [ii] Bourdon et al. (66), see §4 and Table 19 in the Supplemental Material for details.

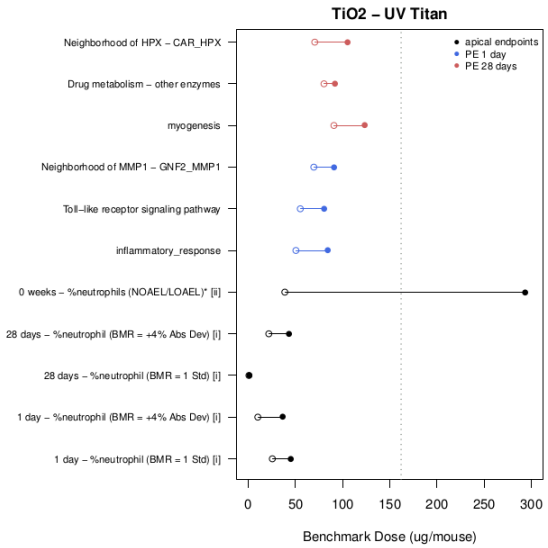
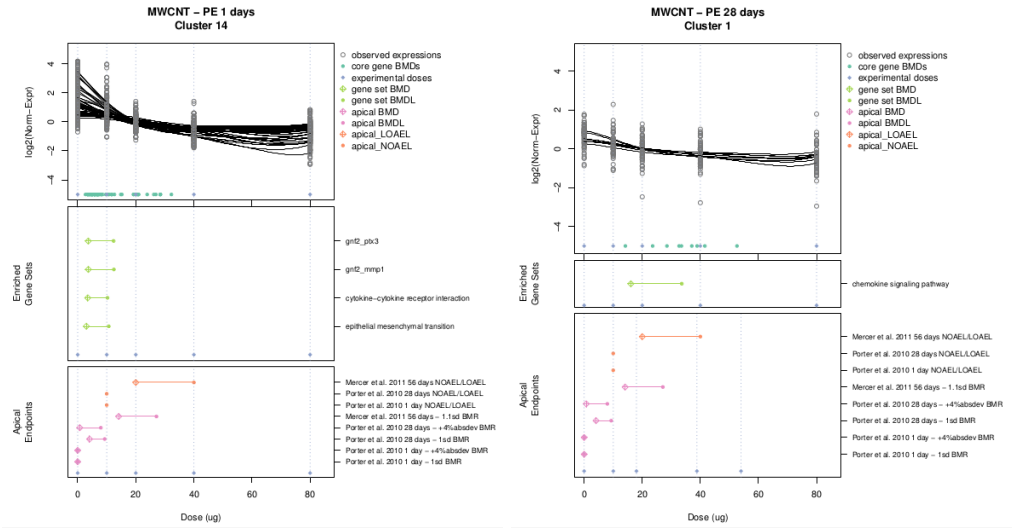


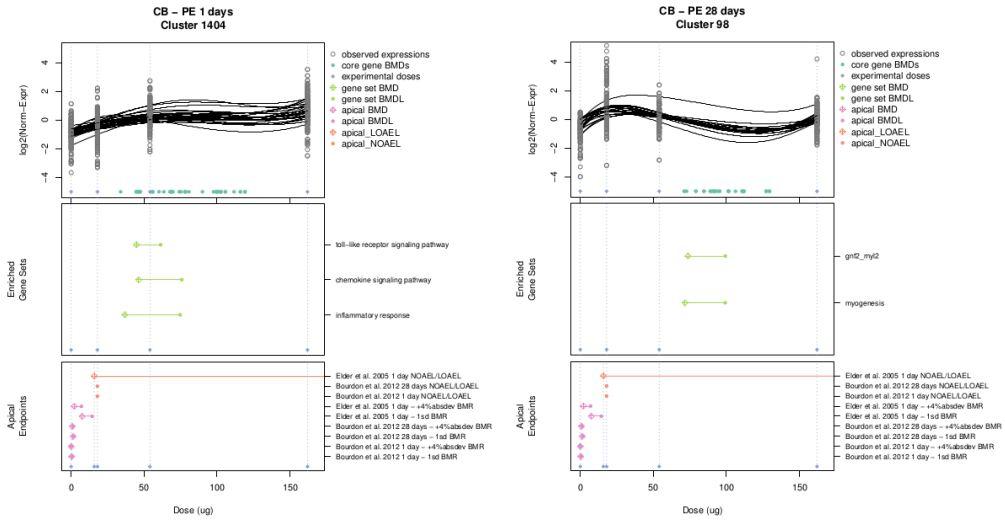
Figure 4: *In vivo* pulmonary and gene set BMD comparison plot for titanium dioxide (TiO_2) exposure in mice. Black segments indicate the *in vivo* pulmonary BMD (BMDL) intervals. Blue and red indicate the most sensitive BMD (BMDL) intervals for 1- and 28-days post-exposure, respectively. The vertical dashed line indicates the maximum experimental dose from the genomic dataset (GSE41041). Represented data are from [i]Saber et al. (68) and [ii] Bermudez et al. (69), see §4 and Table 21 in the Supplemental Material for details.

Since the BMD (BMDL) estimates from the *in vivo* pulmonary inflammation endpoints were unreliable, we also compared the NOAEL and LOAEL for all endpoints, as an alternative POD, with the gene set level BMD estimates (see Supplementary Materials §4 & §5 for details). Here, we only compared the gene set level BMD estimates from the cluster representing the most sensitive and relevant gene set for pulmonary inflammation and fibrosis. Figures 5-7 compare the clusters at the core gene level showing their dose-response and individual BMD estimates, gene set summary level showing the BMD (BMDL) estimates for all enriched gene sets represented by the cluster, and the *in vivo* endpoint level showing the BMD (BMDL) as well as the LOAEL (NOAEL). Except for carbon black, which had small BMD (BMDL) and LOAEL (NOAEL) estimates for most *in vivo* endpoints, the gene set level BMD (BMDL) estimates in the most sensitive and relevant cluster coincided with the *in vivo* LOAEL (NOAEL).



(a) 1-day post-exposure (b) 28-days post-exposure

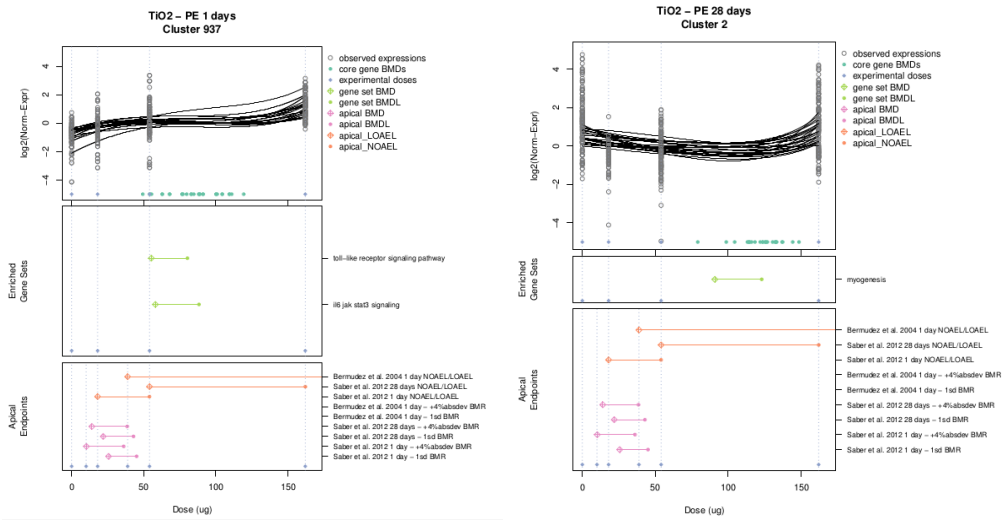
Figure 5: **CNT** The most sensitive gene clusters for 1- and 28-days after exposure to CNTs, (a) and (b) respectively. (Top) Dose-response fits for all core genes that are a member of the enriched gene sets and the gene cluster with the most sensitive gene set (i.e. smallest BMD estimate). Gray points indicate observed expression values and the black curves provide the model fit estimated from ALOHA (37). (Middle) BMD (BMDL) interval plots for each of the enriched gene sets for the displayed gene cluster. (Bottom) the points-of-departure estimated from *in vivo* pulmonary endpoints.



(a) 1-day post-exposure

(b) 28-days post-exposure

Figure 6: **CB** The most sensitive gene clusters for 1- and 28-days after exposure to CB, (a) and (b) respectively. (Top) Dose-response fits for all core genes that are a member of the enriched gene sets and the gene cluster with the most sensitive gene set (i.e. smallest BMD estimate). Gray points indicate observed expression values and the black curves provide the model fit estimated from ALOHA (37). (Middle) BMD (BMDL) interval plots for each of the enriched gene sets for the displayed gene cluster. (Bottom) the points-of-departure estimated from *in vivo* pulmonary endpoints. The LOAEL estimate for Elder et al. (67) at 1-day post-exposure is outside the bounds of the genomic dose-response experimental dose range, $LOAEL = 245.3\mu g$. See §2 in the Supplemental Material for details.



(a) 1-day post-exposure

(b) 28-days post-exposure

Figure 7: TiO_2 The most sensitive gene clusters for 1- and 28-days after exposure to CB, (a) and (b) respectively. (Top) Dose-response fits for all core genes that are a member of the enriched gene sets and the gene cluster with the most sensitive gene set (i.e. smallest BMD estimate). Gray points indicate observed expression values and the black curves provide the model fit estimated from ALOHA (37). (Middle) BMD (BMDL) interval plots for each of the enriched gene sets for the displayed gene cluster. (Bottom) the points-of-departure estimated from *in vivo* pulmonary endpoints. The LOAEL estimate for Bermudez et al. (69) at 1-days post-exposure is outside the bounds of the genomic dose-response experimental dose range, $LOAEL = 293.7\mu g$. See §2 in the Supplemental Material for details.

Clusters enriched for the most sensitive gene set not only had gene set level BMD estimates which coincide with *in vivo* pulmonary PODs, but they also showed the dose-response curves of core genes had similar trends and commonalities in the enriched gene sets. For MWCNT, the most relevant cluster at 1-day post-exposure indicated inflammation and response for lung injury. The cluster for 28-days post-exposure also indicated inflammation. The enriched pathways in these clusters, along with the BMD estimates, indicate there is some resolution of pulmonary inflammation after the initial lung tissue repair response from 1- to 28-days post-exposure. However, the enrichment of the "Chemokine Signaling Pathway" at 28-days post-exposure indicates there is still a sustained inflammatory response which indicates there may be further lung tissue damage. For carbon black, two of the most sensitive gene sets at 1-day post-exposure were enriched in the same cluster and indicate an inflammatory response. The ALOHA clustering approach was able to identify a cluster of genes that are of high importance for the initial response after exposure to carbon black. Myogenesis in the most sensitive and relevant cluster at 28-days post-exposure indicates either the activation of the repair mechanisms after lung injury or the beginning stages of pulmonary fibrosis. Titanium dioxide provided similar patterns in the clusters at 1- and 28-days post-exposure.

4 Discussion

Risk evaluation of nanomaterial exposures with gene expression data rely on the use of biostatistical methods that account for the underlying data structure in order to obtain biologically relevant points-of-departure and a better understanding of the biomolecular response. In this study, we assessed gene expression in mouse lung tissue after exposure to three well-studied nanomaterials using the ALOHA genomic dose-response method (37). ALOHA was used to identify similarities and differences between the perturbed pathways of the three materials and to obtain integrated benchmark dose estimates corresponding to key biological events in the development of chronic pulmonary inflammation and fibrosis.

The most sensitive co-expression dose-response clusters ascertained by ALOHA for all three materials were related to immune, signaling, and repair processes. Clusters related to immune and inflammatory responses for all three materials consistently contained core genes that encoded chemokines, cytokines, or related signaling/regulatory genes. These results are similar to the findings in other nano-toxicogenomic studies (8, 9, 10, 30, 48) and observed pulmonary inflammation in traditional toxicological endpoints with CNT, CB, and TiO_2, especially at the early post-exposure time-points (64, 65, 66, 67, 68, 69). Additionally, we observed some clusters were enriched for gene sets myogenesis and epithelial mesenchymal transition shown to be related with repair and/or mechanisms of pulmonary fibrosis (72, 73, 74, 75, 76, 77, 78). The detection of myogenesis is likely linked with the the production or infiltration of myofibroblasts that play a significant role in wound healing process of muscle and other tissues. Though their impact is largely postive, pro-collagen myofibroblasts have been shown to be the primary driver of increased expression at the peak of lung collagen deposition in fibrotic response which may be linked with the result of excess collagen deposition after a fibrotic response, see Zhang et al. (78) and references there in. El Agha et al. (72) after the resolution of fibroproliferation myofibroblasts and their fate cells are not cleared from the site of wound repair. Thus, the remaining myofibroblast cell population if not de-differentiated as described in El Agha et al. (72) may continue to participate in lung remodeling via promotion of collagen deposition. CNT and CB, but not TiO_2, were enriched for similar gene sets identified in biclusters derived by Williams and Halappanavar (9, 10) using lung disease models related to pulmonary fibrosis. Their most relevant clusters for CNT and CB, bicluster 8 and 10, which consisted of some pulmonary fibrosis lung sample models also were enriched for reactive oxidative stress, inflammation, and response to wounding (9, 10). However, it should be noted in *in vivo* studies evidence for pulmonary fibrosis was only reported in mice exposed to CNT, and not those exposed to CB or TiO_2 (64, 67, 69). Other subchronic inhalation studies examined several species – including mice, rats, and hamsters – to be used for mimicking human pulmonary health and they showed after exposure to CB and TiO_2 pulmonary fibrosis developed in rats but not in mice or hamsters (67, 69).

Though the development of fibrosis was only observed in mice exposed to CNT, further investigation into the gene clusters enriched for gene sets related to pulmonary fibrosis in the other nanomaterial exposure may provide additional information to understand the differences in downstream mechanisms and impact on pulmonary health effects. It is unclear whether the perturbation of these pathways in humans after exposure to MWCNT, CB, or TiO_2 would indicate early markers of fibrosis or activation of lung tissue repair mechanisms inhibiting further damage to lung tissue. Additionally, given the reported differences in the adverse outcomes among traditional *in vivo* rodent models, further research utilizing appropriate pulmonary cell lines or the rodent models that best mimic biological outcomes in humans is needed for utilizing this data for risk assessment of

nanomaterials.

For carbon nanotubes, the clusters resulting from ALOHA at both 1- and 28-days post-exposure both contained genes overlapping with genes in the fibrotic model gene sets presented by Brass et al. (79), see Table 19 in the Supplemental Material. Some of these genes are known to play a key role in negatively regulating fibroproliferation. For example, CXCL10 is found in two of the three most sensitive gene clusters at 28-days post-exposure. It has been shown that expression of CXCL10 helps to mitigate fibrosis development by acting as a chemoattractant for myofibroblasts (80), aiding in inhibiting the process of angiogenesis Strieter et al. (81), and helps to mediate the inflammatory response (82). We also observed IL1B and TIMP1 at 1-day post-exposure, CTSK and MMP12 at 28-days post-exposure, and IL6 at both time points were among the core genes with the smallest BMD estimates. These genes are also part of the time-series gene sets evaluated at the carbon nanotube 80 μg dose presented in Dymacek et al. (83) which they found to be significantly related to fibrosis. If we compare our resulting core gene set with the gene networks results in Rahman et al. (36) for Mitsui-7 we find at center of these networks IL6, MMP12, CXCL10, and TIMP1 which are all differentially expressed in their results both in the inflammatory as well as the fibrotic pathway. This indicates the regulation of these genes may play a key role in the development of fibrosis observed after exposure to carbon nanotubes. IL6 is a known upstream regulator for activating inflammatory response in response to purbation. Though MMP12 is identified and related to pulmonary fibrosis, including gene signatures derived from nanotoxicology studies – see (36, 84, 85, 86), there is conflicting evident as to the impact it has as an upstream regulator in pulmonary fibrosis Craig et al. (87), Pardo et al. (88). Pardo et al. (88) also notes MMP activity is inhibited by other proteins including TIMP family proteins. Thus, the presence of TIMP1 in co-expression clusters with MMP12 and other MMPs may indicate its major role as a homeostatic check after lung perturbation.

The most sensitive and biologically relevant clusters at 1- and 28-days post-exposure were largely enriched for multiple related pathways and had similar gene set BMDs. This indicated ALOHA was able to identify gene sets possibly regulated simultaneously as part of more generalized key events in disease progression. For example, in the carbon nanotube clusters the most biologically relevant clusters at 1- and 28-days post-exposure, enriched for more than one gene set, aligned well with the key-events in the AOP proposed by Labib et al. (8) and the Lung Fibrosis (*Homo sapiens*) Wiki Pathway, see (89, 90, 91, 92, 93) and references therein. Cluster 105 at 1-day post-exposure provides a clear demonstration of this results. This cluster was enriched for both the nod-like and toll-like receptor signaling both of which are related to molecular initiating event of a pulmonary fibrotic response Nymark et al. (94). We refer the reader to Supplemental Figures 12-16 for further details. Similar results were observed in the biologically relevant clusters for the other materials (results not shown). The three most relevant clusters at 1-day post-exposure were enriched for il2 stat5 signaling, apoptosis, tnfa signaling via nfkb, toll-like receptor signaling, and chemokine signaling all of which indicate the activation and progression through the initial key events in pulmonary fibrosis AOPs (8, 30). Toll-like signaling is identified by Nymark et al. (94) an references there-in as the molecular initiating event of the fibrotic response which sets off the rest of the signaling cascades. Activation of this molecular event can be followed by a number of different follow-up key events one of these possible key events being the recruitment of cytokines and chemokines via the chemokine signaling pathway or il2 stat5 signaling (8, 30, 95). Recruited macrophages that are part of a larger inflammatory response then set off another wave of events which directly activate fibroblasts, e.g. activation of $IL1\beta$, or indirectly, e.g. activation of $NF-\kappa B$ which induces $TNF-\alpha$ followed by the activation of $TNF-\beta$, see Vietti et al. (95) and references there-in. The identification of all of these

pathways within a small set of gene clusters sharing similar dose-response patterns provides evidence to the high level of co-regulation and concerted effort of multiple events simultaneously during a biological response to a perturbagen. Similarly, at 28-days post-exposure the most relevant clusters were enriched for complement, il6 jak stat3 signaling, inflammatory response, and toll-like signaling pathways which demonstrate a sustained effect in some of the key events in chronic pulmonary inflammation and the pulmonary fibrosis AOPs, see (8, 94) and references there-in. These results from ALOHA suggest that dose-response co-expression clusters enriched for multiple gene sets allow us to investigate and identify co-regulated pathways that may inform the development of new AOPs by providing information highly co-regulated gene sets which may be key events along the pathway of disease progression without any prior knowledge. Compared with approaches that use the gene set with the smallest BMD estimated like Labib et al. (8), then aligning it with the proposed AOP, ALOHA allows us to evaluate multiple related gene sets simultaneously while obtaining an integrated gene set level BMD estimates from shared genes.

In each material, gene set BMDs increased from 1- to 28-days post-exposure indicating some resolution in lung perturbation after exposure to multi-walled carbon nanotubes, carbon black, and titanium dioxide. Gene set BMD (BMDL) estimates in the most sensitive clusters were higher than BMD estimates from *in vivo* pulmonary inflammation endpoints for all three nanomaterials. The one exception where we did not observe this result was for multi-walled carbon nanotubes and the *in vivo* pulmonary fibrosis endpoint (alveolar thickness). In this case, some gene set BMDs coincided with those estimated from the *in vivo* pulmonary fibrosis endpoint. Evaluating the gene sets within enriched clusters and their corresponding BMDs, particularly those indicating pulmonary fibrosis, suggest the potency ranking for the three materials to be $CNT > CB \geq TiO_2$, see §3.2.

These analyses were limited by insufficient data for dose-response modeling and poor parametric fits in the *in vivo* pulmonary inflammation endpoints. This was due to inclusion of few dose groups and poor dose spacing in the rodent studies resulting in hockey-stick-shaped dose-response curves unsuitable for dose-response modeling (96, Chapter 5 Second Edition 2020). These factors inhibited us from obtaining an reliable estimate of the true underlying dose-response relationship, especially in the low dose-region where the lack of response information resulted in unreliable estimates of the true BMD between the control and first dose group (i.e. $10\mu g$ for CNT and $18\mu g$ for CB and TiO_2). Post-model filtering criteria for ALOHA retained genes with minimal change in response across dose groups except for a slight increase in the highest dose group. This produced posterior mean BMD estimates near, but not larger than, the maximum experimental dose which resulted in the retention of genes that were otherwise unresponsive. In these cases, when there are few non-zero dose groups either the current gene-level BMD criteria for ALOHA are not robust enough or there is insufficient evidence to distinguish a clear response pattern. Statistical power for detecting perturbed pathways may vary between materials since the FDR adjustment for the Fisher's exact test is dependent upon the number and size of clusters. As a result, there may be some additional similarities or differences between materials that were not picked up due to more stringent FDR correction in one material versus the others. Finally, mice did not develop pulmonary fibrosis after exposure to Printex 90 carbon black or UV-Titan nano-titanium dioxide. Several studies have demonstrated fibrosis development may differ between the various rodent species for certain nanomaterial types, see Elder et al. (67), Bermudez et al. (69), Hext et al. (97), Heinrich et al. (98), Oberdorster (99) and references therein. Thus, we were only able to include pulmonary inflammation responses for many of the *in vivo* mouse studies, which at 1-day after exposure is not chronic and was resolving by day 28 post-exposure. Publicly available microarray datasets using rats were limited or did not exist for the three materials included in our study, see §1 in the Supplement for details.

In this study we demonstrated the ALOHA genomic dose-response modeling approach (37) was able to identify gene clusters with co-expressed dose-response patterns related to common pathways linked with *in vivo* pulmonary inflammatory and fibrosis endpoints. Additionally, ALOHA estimates genomic benchmark dose estimates of molecular/cellular mechanisms by collectively leveraging the benchmark dose estimates of core genes belonging both to the pathway and the representative gene cluster. Unlike other current approaches which use individual gene benchmark dose estimates and gene sets or AOPs determined *a priori*, we are able to simultaneously account for the underlying correlations in the data to identify perturbed pathways and obtain biologically relevant benchmark dose estimates. These unique features will allow risk assessors to evaluate underlying biological mechanisms when there is little to no *in vivo* toxicological data, generate hypotheses for further mechanistic assay testing, and obtain a point-of-departure for preliminary risk assessment prioritization that incorporate the underpinning biological pathways.

5 Contributions

6 Acknowledgments

We would like to acknowledge the following individuals for their comments and guidance. Dr. Matthew Wheeler for his comments on the limitations in dose-response modeling of the *in vivo* endpoint data used in these analyses. Mr. Randy Smith and Mrs. Theresa Boots with their comments on conversion of summary cell count to summary percent cell type data. Mr. Nathan Drew with his comments on cell count data conversion and statistical limitations during the dose-response modeling of the *in vivo* endpoints. Dr. Dale Porter for his comments on identified gene sets in dose-response co-expression clusters and biological response to nanomaterials with different physicochemical properties. The CDC OAMD SCBS team for their assistance with ensuring efficient evaluation of datasets using the high-performance computing system at the Centers for Disease Control and Prevention. Finally, thank the reviewers for their comments.

We note to the reader Supplementary Figure 12 contains Figure 2 which was originally produced an presented in the manuscript by Labib et al. (8) published by Springer Nature in their journal *Particle and Fibre Toxicology* on March 15, 2016 and is under a Creative Commons license. Additionally, Supplementary Figures 13 & 14 contains the lung fibrosis pathway originally published in WikiPathways (89, 90, 91, 92, 93) and was constructed from a collection of papers, please see other references within WikiPathways Page for Lung fibrosis (Mus musculus). Thank you to the authors of these figures for providing your work under a Creative Commons license and/or a free for use forum enabling comparison of results in this paper and other publications to further the scientific knowledge base.

This work was supported by the contract from The National Institute for Occupational Safety and Health (75D30118P01991), and in part by an appointment to the Research Participation Program at the Centers for Disease Control and Prevention administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the Centers for Disease Control and Prevention.

References

[1] N Pidgeon, J Porritt, J Ryan, A Seaton, S Tendler, M Welland, and R Whatmore. Nanoscience and nanotechnologies: opportunities and uncertainties. *The Royal Society, The Royal Academy of Engineering*, 29(07):2004, 2004.

[2] Günter Oberdörster, Eva Oberdörster, and Jan Oberdörster. Nanotoxicology: an emerging discipline evolving from studies of ultrafine particles. *Environmental health perspectives*, 113 (7):823–839, 2005.

[3] Paul Schulte, Charles Geraci, Ralph Zumwalde, Mark Hoover, and Eileen Kuempel. Occupational risk management of engineered nanoparticles. *Journal of occupational and environmental hygiene*, 5(4):239–249, 2008.

[4] National Institute for Occupational Safety & Health NIOSH. Current intelligence bulletin 63: Occupational exposure to titanium dioxide, 2011.

[5] National Institute for Occupational Safety & Health NIOSH. Current intelligence bulletin 65: Occupational exposure to carbon nanotubes and nanofibers, 2013.

[6] National Institute of Occupational Safety & Health NIOSH. Nioshtic-2 publications search, Dec 2020. URL `https://www2a.cdc.gov/nioshtic-2/advsearch2.asp`.

[7] Bengt Fadeel, Lucian Farcal, Barry Hardy, Socorro Vázquez-Campos, Danail Hristozov, Antonio Marcomini, Iseult Lynch, Eugenia Valsami-Jones, Harri Alenius, and Kai Savolainen. Advanced tools for the safety assessment of nanomaterials. *Nature nanotechnology*, 13(7): 537–543, 2018.

[8] Sarah Labib, Andrew Williams, Carole L Yauk, Jake K Nikota, Håkan Wallin, Ulla Vogel, and Sabina Halappanavar. Nano-risk science: application of toxicogenomics in an adverse outcome pathway framework for risk assessment of multi-walled carbon nanotubes. *Particle and fibre toxicology*, 13(1):15, 2015.

[9] Andrew Williams and Sabina Halappanavar. Application of biclustering of gene expression data and gene set enrichment analysis methods to identify potentially disease causing nanomaterials. *Beilstein journal of nanotechnology*, 6(1):2438–2448, 2015.

[10] Andrew Williams and Sabina Halappanavar. Application of bi-clustering of gene expression data and gene set enrichment analysis methods to identify potentially disease causing nanomaterials. *Data in brief*, 15:933–940, 2017.

[11] Melvin E Andersen and Daniel Krewski. Toxicity testing in the 21st century: bringing the vision to life. *Toxicological sciences*, 107(2):324–330, 2009.

[12] David J Dix, Kathryn Gallagher, William H Benson, Brenda L Groskinsky, J Thomas McClintock, Kerry L Dearfield, and William H Farland. A framework for the use of genomics data at the epa. *Nature biotechnology*, 24(9):1108–1111, 2006.

[13] Daniel Krewski, Melvin E Andersen, Ellen Mantus, and Lauren Zeise. Toxicity testing in the 21st century: implications for human health risk assessment. *Risk Analysis: An International Journal*, 29(4):474–479, 2009.

[14] National Research Council et al. *Toxicity testing in the 21st century: a vision and a strategy*. National Academies Press, 2007.

[15] National Academies of Sciences Engineering and Medicine. *Using 21st Century Science to Improve Risk-Related Evaluations*. The National Academies Press, Washington, DC, 2017. ISBN 978-0-309-45348-6. doi: 10.17226/24635. URL `https://www.nap.edu/catalog/24635/using-21st-century-science-to-improve-risk-related-evaluations`.

[16] Russell S Thomas, Richard S Paules, Anton Simeonov, Suzanne C Fitzpatrick, Kevin M Crofton, Warren M Casey, and Donna L Mendrick. The us federal tox21 program: A strategic and operational plan for continued leadership. *Altex*, 35(2):163, 2018.

[17] Reza Farmahin, Anne Marie Gannon, Rémi Gagné, Andrea Rowan-Carroll, Byron Kuo, Andrew Williams, Ivan Curran, and Carole L Yauk. Hepatic transcriptional dose-response analysis of male and female fischer rats exposed to hexabromocyclododecane. *Food and Chemical Toxicology*, 133:110262, 2019.

[18] Ivy Moffat, Nikolai L Chepelev, Sarah Labib, Julie Bourdon-Lacombe, Byron Kuo, Julie K Buick, France Lemieux, Andrew Williams, Sabina Halappanavar, Amal I Malik, et al. Comparison of toxicogenomics and traditional approaches to inform mode of action and points of departure in human health risk assessment of benzo [a] pyrene in drinking water. *Critical reviews in toxicology*, 45(1):1–43, 2015.

[19] Russell S Thomas, Bruce C Allen, Andy Nong, Longlong Yang, Edilberto Bermudez, Harvey J Clewell III, and Melvin E Andersen. A method to integrate benchmark dose estimates with genomic data to assess the functional effects of chemical exposure. *Toxicological sciences*, 98 (1):240–248, 2007.

[20] Russell S Thomas, Harvey J Clewell III, Bruce C Allen, Scott C Wesselkamper, Nina Ching Y Wang, Jason C Lambert, Janet K Hess-Wilson, Q Jay Zhao, and Melvin E Andersen. Application of transcriptional benchmark dose values in quantitative cancer and noncancer risk assessment. *Toxicological sciences*, 120(1):194–205, 2011.

[21] Russell S Thomas, Harvey J Clewell III, Bruce C Allen, Longlong Yang, Eric Healy, and Melvin E Andersen. Integrating pathway-based transcriptomic data into quantitative chemical risk assessment: a five chemical case study. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, 746(2):135–143, 2012.

[22] Russell S Thomas, Scott C Wesselkamper, Nina Ching Y Wang, Q Jay Zhao, Dan D Petersen, Jason C Lambert, Ila Cote, Longlong Yang, Eric Healy, Michael B Black, et al. Temporal concordance between apical and transcriptional points of departure for chemical risk assessment. *toxicological sciences*, 134(1):180–194, 2013.

[23] Jeffry L Dean, Q Jay Zhao, Jason C Lambert, Belinda S Hawkins, Russell S Thomas, and Scott C Wesselkamper. Editor's highlight: Application of gene set enrichment analysis for identification of chemically induced, biologically relevant transcriptomic networks and potential utilization in human health risk assessment. *Toxicological Sciences*, 157(1):85–99, 2017.

[24] Daniel L Villeneuve, Doug Crump, Natalia Garcia-Reyero, Markus Hecker, Thomas H Hutchinson, Carlie A LaLone, Brigitte Landesmann, Teresa Lettieri, Sharon Munn, Malgorzata Nepelska, et al. Adverse outcome pathway (aop) development i: strategies and principles. *Toxicological Sciences*, 142(2):312–320, 2014.

[25] Daniel L Villeneuve, Doug Crump, Natalia Garcia-Reyero, Markus Hecker, Thomas H Hutchinson, Carlie A LaLone, Brigitte Landesmann, Teresa Lettieri, Sharon Munn, Malgorzata Nepelska, et al. Adverse outcome pathway development ii: best practices. *Toxicological Sciences*, 142(2):321–330, 2014.

[26] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13, 2009.

[27] da W Huang, BT Sherman, and RA Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44, 2009.

[28] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[29] Vamsi K Mootha, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, et al. Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, 34(3):267–273, 2003.

[30] Jake Nikota, Andrew Williams, Carole L Yauk, Håkan Wallin, Ulla Vogel, and Sabina Halappanavar. Meta-analysis of transcriptomic responses as a means to identify pulmonary disease outcomes for engineered nanomaterials. *Particle and fibre toxicology*, 13(1):1–18, 2015.

[31] Scott S. Auerbach. BMDExpress 2.3, 2017. URL `https://github.com/auerbachs/BMDExpress-2/wiki`.

[32] Jason R Phillips, Daniel L Svoboda, Arpit Tandon, Shyam Patel, Alex Sedykh, Deepak Mav, Byron Kuo, Carole L Yauk, Longlong Yang, Russell S Thomas, et al. Bmdexpress 2: enhanced transcriptomic dose-response analysis workflow. *Bioinformatics*, 35(10):1780–1782, 2019.

[33] Longlong Yang, Bruce C Allen, and Russell S Thomas. Bmdexpress: a software tool for the benchmark dose analyses of genomic data. *BMC genomics*, 8(1):387, 2007.

[34] Sabina Halappanavar, James D Ede, Jo Anne Shatkin, and Harald F Krug. A systematic process for identifying key events for advancing the development of nanomaterial relevant adverse outcome pathways. *NanoImpact*, 15:100178, 2019.

[35] Sabina Halappanavar, Luna Rahman, Jake Nikota, Sarah S Poulsen, Yaobo Ding, Petra Jackson, Hakan Wallin, Otmar Schmid, Ulla Vogel, and Andrew Williams. Ranking of nanomaterial potency to induce pathway perturbations associated with lung responses. *NanoImpact*, 14:100158, 2019.

[36] Luna Rahman, Andrew Williams, Krishna Gelda, Jake Nikota, Dongmei Wu, Ulla Vogel, and Sabina Halappanavar. 21st century tools for nanotoxicology: Transcriptomic biomarker panel and precision-cut lung slice organ mimic system for the assessment of nanomaterial-induced lung fibrosis. *Small*, page 2000272, 2020.

[37] Sarah E. Davidson, Matthew W. Wheeler, Scott S. Auerbach, Siva Sivaganesan, and Mario Medvedovic. Aloha: Aggregated local extrema splines for high-throughput dose-response analysis. *bioRxiv*, 2021. doi: 10.1101/2021.03.29.437588. URL `https://www.biorxiv.org/content/early/2021/03/31/2021.03.29.437588`.

[38] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research*, 41 (D1):D991–D995, 2012.

[39] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[40] World Health Organization WHO et al. *WHO guidelines on protecting workers from potential risks of manufactured nanomaterials*. World Health Organization, Geneva, 2017. Licence: CC BY-NC-SA 3.0 IGO.

[41] R Core Team. R: A language and environment for statistical computing, 2018. URL `https://www.R-project.org/`.

[42] RStudio Team and others. Rstudio: integrated development for r. 42:14, 2015. URL `http://www.rstudio.com`.

[43] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

[44] Matthew E Ritchie, Jeremy Silver, Alicia Oshlack, Melissa Holmes, Dileepa Diyagama, Andrew Holloway, and Gordon K Smyth. A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, 23(20):2700–2707, 2007.

[45] Yee Hwa Yang, Sandrine Dudoit, Percy Luu, David M Lin, Vivian Peng, John Ngai, and Terence P Speed. Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic acids research*, 30(4):e15–e15, 2002.

[46] Sarah E. Davidson. Enm aloha analyses. `https://github.com/davidsss-uc/ENM_ALOHA`, 2021.

[47] Nancy L Guo, Ying-Wooi Wan, James Denvir, Dale W Porter, Maricica Pacurari, Michael G Wolfarth, Vincent Castranova, and Yong Qian. Multiwalled carbon nanotube-induced gene signatures in the mouse lung: potential predictive value for human lung cancer risk and prognosis. *Journal of Toxicology and Environmental Health, Part A*, 75(18):1129–1153, 2012.

[48] Julie A Bourdon, Sabina Halappanavar, Anne T Saber, Nicklas R Jacobsen, Andrew Williams, Håkan Wallin, Ulla Vogel, and Carole L Yauk. Hepatic and pulmonary toxicogenomic profiles in mice intratracheally instilled with carbon black nanoparticles reveal pulmonary inflammation, acute phase response, and alterations in lipid homeostasis. *Toxicological sciences*, 127(2):474–484, 2012.

[49] Mainul Husain, Anne T Saber, Charles Guo, Nicklas R Jacobsen, Keld A Jensen, Carole L Yauk, Andrew Williams, Ulla Vogel, Hakan Wallin, and Sabina Halappanavar. Pulmonary instillation of low doses of titanium dioxide nanoparticles in mice leads to particle retention and gene expression changes in the absence of inflammation. *Toxicology and applied pharmacology*, 269(3):250–262, 2013.

[50] MW Wheeler, DB Dunson, and AH Herring. Bayesian local extremum splines. *Biometrika*, 104(4):939–952, 2017.

[51] National Toxicology Program NTP et al. Ntp research report on national toxicology program approach to genomic dose-response modeling: Research report 5. NTP Research Report Series 2473-4756, National Toxicology Program, 2018. URL `http://dx.doi.org/10.22427/NTP-RR-5`.

[52] Xiangdong Liu, Siva Sivaganesan, Ka Yee Yeung, Junhai Guo, Roger Eugene Bumgarner, and Mario Medvedovic. Context-specific infinite mixtures for clustering gene expression profiles across diverse microarray dataset. *Bioinformatics*, 22(14):1737–1744, 2006.

[53] Johannes M Freudenberg, Vineet K Joshi, Zhen Hu, and Mario Medvedovic. Clean: Clustering enrichment analysis. *BMC bioinformatics*, 10(1):234, 2009.

[54] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[55] Arthur Liberzon. A description of the molecular signatures database (msigdb) web site. In *Stem Cell Transcriptional Networks*, pages 153–160. Springer, 2014.

[56] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6):417–425, 2015.

[57] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[58] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1): D353–D361, 2017.

[59] Minoru Kanehisa, Yoko Sato, Miho Furumichi, Kanae Morishima, and Mao Tanabe. New approach for understanding genome variations in kegg. *Nucleic acids research*, 47(D1):D590–D595, 2019.

[60] Alan Agresti et al. A survey of exact inference for contingency tables. *Statistical science*, 7(1): 131–153, 1992.

[61] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[62] Ronald A Fisher. On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

[63] Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics*, pages 66–70. Springer, 1992.

[64] Dale W Porter, Ann F Hubbs, Robert R Mercer, Nianqiang Wu, Michael G Wolfarth, Krishnan Sriram, Stephen Leonard, Lori Battelli, Diane Schwegler-Berry, Sherry Friend, et al. Mouse pulmonary dose-and time course-responses induced by exposure to multi-walled carbon nanotubes. *Toxicology*, 269(2-3):136–147, 2010.

[65] Robert R Mercer, Ann F Hubbs, James F Scabilloni, Liying Wang, Lori A Battelli, Sherri Friend, Vincent Castranova, and Dale W Porter. Pulmonary fibrotic response to aspiration of multi-walled carbon nanotubes. *Particle and fibre toxicology*, 8(1):1–12, 2011.

[66] Julie A Bourdon, Anne T Saber, Nicklas R Jacobsen, Keld A Jensen, Anne M Madsen, Jacob S Lamson, Håkan Wallin, Peter Møller, Steffen Loft, Carole L Yauk, et al. Carbon black nanoparticle instillation induces sustained inflammation and genotoxicity in mouse lung and liver. *Particle and fibre toxicology*, 9(1):1–14, 2012.

[67] Alison Elder, Robert Gelein, Jacob N Finkelstein, Kevin E Driscoll, Jack Harkema, and Gunter Oberdorster. Effects of subchronically inhaled carbon black in three species. i. retention kinetics, lung inflammation, and histopathology. *Toxicological Sciences*, 88(2):614–629, 2005.

[68] Anne Thoustrup Saber, Keld Alstrup Jensen, Nicklas Raun Jacobsen, Renie Birkedal, Lone Mikkelsen, Peter Møller, Steffen Loft, Håkan Wallin, and Ulla Vogel. Inflammatory and genotoxic effects of nanoparticles designed for inclusion in paints and lacquers. *Nanotoxicology*, 6 (5):453–471, 2012.

[69] Edilberto Bermudez, James B Mangum, Brian A Wong, Bahman Asgharian, Paul M Hext, David B Warheit, and Jeffrey I Everitt. Pulmonary responses of mice, rats, and hamsters to subchronic inhalation of ultrafine titanium dioxide particles. *Toxicological sciences*, 77(2): 347–357, 2004.

[70] J Allen Davis, Jeffrey S Gift, and Q Jay Zhao. Introduction to benchmark dose methods and us epa's benchmark dose software (bmds) version 2.1. 1. *Toxicology and applied pharmacology*, 254(2):181–191, 2011.

[71] Kenny S Crump. Calculation of benchmark doses from continuous data. *Risk Analysis*, 15(1): 79–89, 1995.

[72] Elie El Agha, Alena Moiseenko, Vahid Kheirollahi, Stijn De Langhe, Slaven Crnkovic, Grazyna Kwapiszewska, Marten Szibor, Djuro Kosanovic, Felix Schwind, Ralph T Schermuly, et al. Two-way conversion between lipogenic and myogenic fibroblastic phenotypes marks the progression and resolution of lung fibrosis. *Cell stem cell*, 20(2):261–273, 2017.

[73] Andreas Günther, Martina Korfei, Poornima Mahavadi, Daniel von der Beck, Clemens Ruppert, and Philipp Markart. Unravelling the progressive pathophysiology of idiopathic pulmonary fibrosis. *European Respiratory Review*, 21(124):152–160, 2012.

[74] Charlotte Hill, Mark G Jones, Donna E Davies, and Yihua Wang. Epithelial-mesenchymal transition contributes to pulmonary fibrosis via aberrant epithelial/fibroblastic cross-talk. *Journal of lung health and diseases*, 3(2):31, 2019.

[75] Arun Lingampally, Matthew R Jones, Shirisha Bagari, Chengshui Chen, Stefano Rivetti, and Saverio Bellusci. Use of the reversible myogenic to lipogenic transdifferentiation switch for the design of pre-clinical drug screening in idiopathic pulmonary fibrosis. *Frontiers in Bioengineering and Biotechnology*, 8:1083, 2020.

[76] Brigham C Willis, Janice M Liebler, Katherine Luby-Phelps, Andrew G Nicholson, Edward D Crandall, Roland M Du Bois, and Zea Borok. Induction of epithelial-mesenchymal transition in alveolar epithelial cells by transforming growth factor-β1: potential role in idiopathic pulmonary fibrosis. *The American journal of pathology*, 166(5):1321–1332, 2005.

[77] Yuechong Xia, Cheng Lei, Danhui Yang, and Hong Luo. Construction and validation of a bronchoalveolar lavage cell-associated gene signature for prognosis prediction in idiopathic pulmonary fibrosis. *International Immunopharmacology*, 92:107369, 2021. ISSN 1567-5769. doi: https://doi.org/10.1016/j.intimp.2021.107369. URL `https://www.sciencedirect.com/science/article/pii/S1567576921000059`.

[78] Kai Zhang, Mark D Rekhter, David Gordon, and Sem H Phan. Myofibroblasts and their role in lung collagen gene expression during pulmonary fibrosis. a combined immunohistochemical and in situ hybridization study. *The American journal of pathology*, 145(1):114, 1994.

[79] David M Brass, Ivana V Yang, Marcus P Kennedy, Gregory S Whitehead, Holly Rutledge, Lauranell H Burch, and David A Schwartz. Fibroproliferation in lps-induced airway remodeling and bleomycin-induced fibrosis share common patterns of gene expression. *Immunogenetics*, 60(7):353–369, 2008.

[80] Andrew M Tager, Richard L Kradin, Peter LaCamera, Scott D Bercury, Gabriele SV Campanella, Carol P Leary, Vasiliy Polosukhin, Long-Hai Zhao, Hideo Sakamoto, Timothy S Blackwell, et al. Inhibition of pulmonary fibrosis by the chemokine ip-10/cxcl10. *American journal of respiratory cell and molecular biology*, 31(4):395–404, 2004.

[81] Robert M Strieter, Brigitte N Gomperts, Michael P Keane, et al. The role of cxc chemokines in pulmonary fibrosis. *The Journal of clinical investigation*, 117(3):549–556, 2007.

[82] Georgina Xanthou, Cécile Emmanuelle Duchesnes, Timothy John Williams, and James Edward Pease. Ccr3 functional responses are regulated by both cxcr3 and its ligands cxcl9, cxcl10 and cxcl11. *European journal of immunology*, 33(8):2241–2250, 2003.

[83] Julian Dymacek, Brandi N Snyder-Talkington, Dale W Porter, Robert R Mercer, Michael G Wolfarth, Vincent Castranova, Yong Qian, and Nancy L Guo. mrna and mirna regulatory networks reflective of multi-walled carbon nanotube-induced lung inflammatory and fibrotic pathologies in mice. *Toxicological Sciences*, 144(1):51–64, 2015.

[84] Katsuhide Fujita, Makiko Fukuda, Hiroko Fukui, Masanori Horie, Shigehisa Endoh, Kunio Uchida, Mototada Shichiri, Yasuo Morimoto, Akira Ogami, and Hitoshi Iwahashi. Intratracheal instillation of single-wall carbon nanotubes in the rat lung induces time-dependent changes in gene expression. *Nanotoxicology*, 9(3):290–301, 2015. doi: 10.3109/17435390.2014.921737. URL https://doi.org/10.3109/17435390.2014.921737. PMID: 24911292.

[85] Sarah S. Poulsen, Anne T. Saber, Andrew Williams, Ole Andersen, Carsten Købler, Rambabu Atluri, Maria E. Pozzebon, Stefano P. Mucelli, Monica Simion, David Rickerby, Alicja Mortensen, Petra Jackson, Zdenka O. Kyjovska, Kristian Mølhave, Nicklas R. Jacobsen, Keld A. Jensen, Carole L. Yauk, Håkan Wallin, Sabina Halappanavar, and Ulla Vogel. Mwcnts of different physicochemical properties cause similar inflammatory responses, but differences in transcriptional and histological markers of fibrosis in mouse lungs. *Toxicology and Applied Pharmacology*, 284(1):16–32, 2015. ISSN 0041-008X. doi: https://doi.org/10.1016/j.taap.2014.12.011. URL https://www.sciencedirect.com/science/article/pii/S0041008X14004499.

[86] Damiën van Berlo, Verena Wilhelmi, Agnes W Boots, Maja Hullmann, Thomas AJ Kuhlbusch, Aalt Bast, Roel PF Schins, and Catrin Albrecht. Apoptotic, inflammatory, and fibrogenic effects of two different types of multi-walled carbon nanotubes in mouse lung. *Archives of toxicology*, 88(9):1725–1737, 2014.

[87] Vanessa J Craig, Li Zhang, James S Hagood, and Caroline A Owen. Matrix metalloproteinases as therapeutic targets for idiopathic pulmonary fibrosis. *American journal of respiratory cell and molecular biology*, 53(5):585–600, 2015.

[88] Annie Pardo, Sandra Cabrera, Mariel Maldonado, and Moisés Selman. Role of matrix metalloproteinases in the pathogenesis of idiopathic pulmonary fibrosis. *Respiratory research*, 17(1): 1–10, 2016.

[89] Friederike Ehrhart, Penny Nymark, Linda Rieswijk, Kristina Hanspers, Egon Willighagen, and Jonathan Melius. Lung fibrosis (homo sapiens), 2018. URL https://www.wikipathways.org/index.php/Pathway:WP3624.

[90] Marvin Martens, Ammar Ammar, Anders Riutta, Andra Waagmeester, Denise N Slenter, Kristina Hanspers, Ryan A. Miller, Daniela Digles, Elisson N Lopes, Friederike Ehrhart, Lauren J Dupuis, Laurent A Winckers, Susan L Coort, Egon L Willighagen, Chris T Evelo, Alexander R Pico, and Martina Kutmon. WikiPathways: connecting communities. *Nucleic Acids Research*, 49(D1):D613–D621, 11 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa1024. URL https://doi.org/10.1093/nar/gkaa1024.

[91] Thomas Kelder, Martijn P. van Iersel, Kristina Hanspers, Martina Kutmon, Bruce R. Conklin, Chris T. Evelo, and Alexander R. Pico. WikiPathways: building research communities on biological pathways. *Nucleic Acids Research*, 40(D1):D1301–D1307, 11 2011. ISSN 0305-1048. doi: 10.1093/nar/gkr1074. URL https://doi.org/10.1093/nar/gkr1074.

[92] Martina Kutmon, Anders Riutta, Nuno Nunes, Kristina Hanspers, Egon L. Willighagen, Anwesha Bohler, Jonathan Mélius, Andra Waagmeester, Sravanthi R. Sinha, Ryan Miller, Susan L. Coort, Elisa Cirillo, Bart Smeets, Chris T. Evelo, and Alexander R. Pico. WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Research*, 44(D1):D488–D494,

10 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv1024. URL https://doi.org/10.1093/nar/gkv1024.

[93] Denise N Slenter, Martina Kutmon, Kristina Hanspers, Anders Riutta, Jacob Windsor, Nuno Nunes, Jonathan Mélius, Elisa Cirillo, Susan L Coort, Daniela Digles, Friederike Ehrhart, Pieter Giesbertz, Marianthi Kalafati, Marvin Martens, Ryan Miller, Kozo Nishida, Linda Rieswijk, Andra Waagmeester, Lars M T Eijssen, Chris T Evelo, Alexander R Pico, and Egon L Willighagen. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, 46(D1):D661–D667, 11 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx1064. URL https://doi.org/10.1093/nar/gkx1064.

[94] Penny Nymark, Linda Rieswijk, Friederike Ehrhart, Nina Jeliazkova, Georgia Tsiliki, Haralambos Sarimveis, Chris T Evelo, Vesa Hongisto, Pekka Kohonen, Egon Willighagen, and Roland C Grafström. A Data Fusion Pipeline for Generating and Enriching Adverse Outcome Pathway Descriptions. *Toxicological Sciences*, 162(1):264–275, 11 2017. ISSN 1096-6080. doi: 10.1093/toxsci/kfx252. URL https://doi.org/10.1093/toxsci/kfx252.

[95] Giulia Vietti, Dominique Lison, and Sybille van den Brule. Mechanisms of lung fibrosis induced by carbon nanotubes: towards an adverse outcome pathway (aop). *Particle and fibre toxicology*, 13(1):1–23, 2015.

[96] World Health Organization WHO et al. *Principles and methods for the risk assessment of chemicals in food*, volume 240. World Health Organization, Geneva, Switzerland, 2009.

[97] P. M. Hext, D. B. Warheit, J. Mangum, B. Asgharian, B. Wong, E. Bermudez, and J. Everitt. Comparison of the Pulmonary Responses to Inhaled Pigmentary and Ultrafine Titanium Dioxide Particles in the Rat, Mouse and Hamster. *The Annals of Occupational Hygiene*, $46(\text{suppl}_1)$: $191 - -196, 122002$. $ISSN 0003 - 4878$. *doi* : . URL https://doi.org/10.1093/annhyg/46.suppl_1.191.

Uwe Heinrich, R Fuhst, S Rittinghausen, O Creutzenberg, B Bellmann, W Koch, and K Levsen. Chronic inhalation exposure of wistar rats and two different strains of mice to diesel engine exhaust, carbon black, and titanium dioxide. *Inhalation toxicology*, 7(4):533–556, 1995.

G. Oberdorster. Lung particle overload: Implications for occupational exposures to particles. *Regulatory Toxicology and Pharmacology*, 21(1):123–135, 1995. ISSN 0273-2300. https://doi.org/10.1006/rtph.1995.1017. URL https://www.sciencedirect.com/science/article/pii/S0273230085710173.

Supplemental Material: Dose-Response Co-expression Clustering of Gene Responses in Mouse Lung after Engineered Nanomaterial Exposure

Sarah E. Davidson[1], Michael Borchers PhD[2], Eileen D. Kuempel PhD[3]

1 Nanomaterial GDR Data Searches

We conducted two searches for genomic dose-response data evaluating lung tissue response after exposure to engineered nanomaterials or other related particles. These searches were conducted at two different time points. The first search was done in the Gene Expression Omnibus (GEO) Database (1, 2, 3) in 2018 and a second updated data search was performed in 2020 over a set of three public repositories including the GEO Database, National Toxicology Database - Chemical Effects in Biological Systems (NTP-CEBS) (4), and National Institute for Occupational Safety and Health Publications Search Database (NIOSHTIC-2) (5).

1.1 2018 Genomic Dose-Response Data Search

The 2018 GEO data search, conducted on June 7, 2018, was focused on capturing a set of RAW microarray genomic dose-response datasets on nanoscale and microscale particulates which produce adverse pulmonary effects. The list of included materials for our search terms was determined by a well-defined set of criteria developed by Dr. Eileen Kuempel, Mr. Nathan Drew, and in part by Dr. Aleks Stefaniak from the National Institute for Occupational Safety and Health (NIOSH). Our search was systematically structured and conducted manually in the GEO repository search engine. The generalized search query was structured as follows:

> (Material[Title]) AND ("gse"[Filter]) AND (mouse[Organism] OR rat[Organism] OR human[Organism])

"Material" was the only variable term in the query statement. Below is a list of the material terms we used in our search, they are organized by their results status.

Positive for Hits	Negatvie for Hits
Nanoparticle, Nanoparticles, Nanoparticulate, Nanomaterial, Nanomaterials, Nanofiber, Nanofibers, Nanotube, Nanotubes, Particulate, Particulates, Particle, Particles, Nickel subsulfide, Vanadium pentoxide, Talc, MONOCROTALINE, 5-Azacytidine, PHENOBARBITAL, CYCLOPHOSPHAMIDE, ISONIAZID, THIOACETAMIDE, Chromium, Cobalt, Rotenone, Naphthalene, Multiwalled Carbon Nanotube, Carbon, Multiwalled nanotubes, Fullerene	Nanoparticulates, Nanosphere, Nanospheres, Microscale particulate, Microscale particulates, Cobalt sulfate heptahydrate, Nickel sulfate hexahydrate, Ferrocene, Gallium arsenide, Antimony trioxide, Molybdenum trioxide, Nickel (II) oxide, Calcium chromate, Wollastonite calcium silicates, 1,2-Dihydro-2,2,4-trimethylquinoline (monomer), 1,2-Dihydro-2,2,4-trimethylquinoline, Indium phosphide, o-Chlorobenzalmalononitrile (CS), o-Chlorobenzalmalononitrile, O-Chlorobenzalmalononitrile, Emetine hydrochloride, IPD (3,3'-iminobis-1-propanol dimethanesulfonate (ester) hydrochloride), 3,3'-iminobis-1-propanol dimethanesulfonate (ester) hydrochloride, IPD, Procarbazine hydrochloride, Isophosphamide, IFOSFAMIDE, tris(Aziridinyl)-phosphine sulfide (thio-tepa), 2-Chloroacetophenone (CN), 2-Chloroacetophenone, Amsacrine, Epinephrine hydrochloride, Phenoxybenzamine hydrochloride, ortho-Phthalaldehyde, Acronycine, Polyurethane, Abrasive Blasting Agents: Blasting Sand, Blasting Sand, Abrasive blasting agents (coal slag), Abrasive blasting agents, coal slag, Abrasive blasting agents (crushed glass), crushed glass, Abrasive blasting agents (garnet), garnet, Abrasive Blasting Agents: Specular Hematite, Specular Hematite, 1020 Long Multiwalled Carbon Nanotube, Long Multiwalled Carbon Nanotube, Multiwalled nanotube, Singlewalled nanotube, Singlewalled nanotubes, Buckyballs, Buckyball, Fullerenes

Table 1: This table includes all of the key terms used as input for "Material" in the query statement. Material key terms are separated based on their results status from GEO (i.e. positive for hits - had at least one resulting dataset & negative for hits - had zero datasets returned).

1.1.1 Data Criteria

For the studies to be considered relevant for risk assessment of nanomaterial exposures, it needed to contain samples from human cell line cultures or rodents exposed to at least one materials of interest. Out of these studies we select only those using only microarray platforms, excluding those using non-coding RNA microarray, for measuring gene expression. Studies using Affymetrix, Agilent, and Illumina Beadchip microarrays were of particular interest. We chose to use only to include microarray studies since there was a limited number of genomic studies on nanomaterial

exposure that utilized a RNA-sequencing platform. Additionally, studies that did not include the material term in the submitted summary or title were removed since this is pertient experimental design information typically provided.

Once the initial selection was performed we we exported additional study information for manual curation. The related GEO webpages, abstracts, citations, and papers were obtained for further review of the unique study hits relevant to our research aim. We manually assessed these resources to identify studies for removal. This included studies utilizing nanomaterials as a pharmaceutical delivery system, had fewer than three dose-groups, no untreated or vehicle control samples, and those that did not contain samples obtained from cell line or rodents exposed to nanomaterials. Our manual curation provided us a list of studies that may be useful for human health risk assessment of nanomaterial exposures.

1.1.2 Results

Figure 1 provides a diagram of the total number of studies that we obtained from our GEO search terms and those that passed each selection step.

1.2 2018 Study Search & Selection

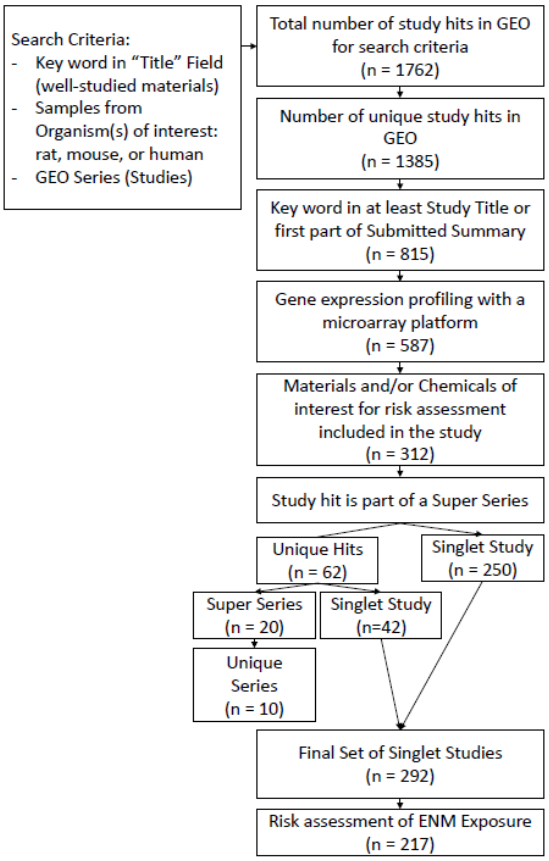


Figure 1: **2018 Study Search & Selection:** Diagram of study hit and selection results for identifying genomic dose-response datasets on rat, mouse, or human exposure to a variety of nanomaterials that may be useful for risk assessment using genomic dose-response data.

Of the 217 unique studies that resulted, the 10 singlet studies in Williams and Halappanavar (6, 7) were contained in our final search results indicating that our study selection criteria was not too strict. We evaluated the overall experimental design used in the 207 new microarray gene expression datasets, see Tables 2 & 3 and Figure 2.

Platform Company	Number of GEO Study Hits	Number of Unique Platforms
Agilent	95	28
Affymetrix	74	26
Illumina Beadchip	32	10
Other	16	6

Table 2: Breakdown selected study hits by platform company. The first column provides the company that produces the microarray platform. The second column provides the number of selected study hits that utilize the platform. Finally, column three provides the number of unique platforms represented among the selected study hits (i.e. unique geo platform ID's - GPL).

Species	Number of GEO Study Hits
Human	103
Mouse	84
Rat	28
Human & Mouse	1
Human & Rat	1

Table 3: Breakdown selected study hits by the represented subject species. The first column provides the species included in a single study. The second column provides the number of selected study hits the species (or species combination) in their samples.

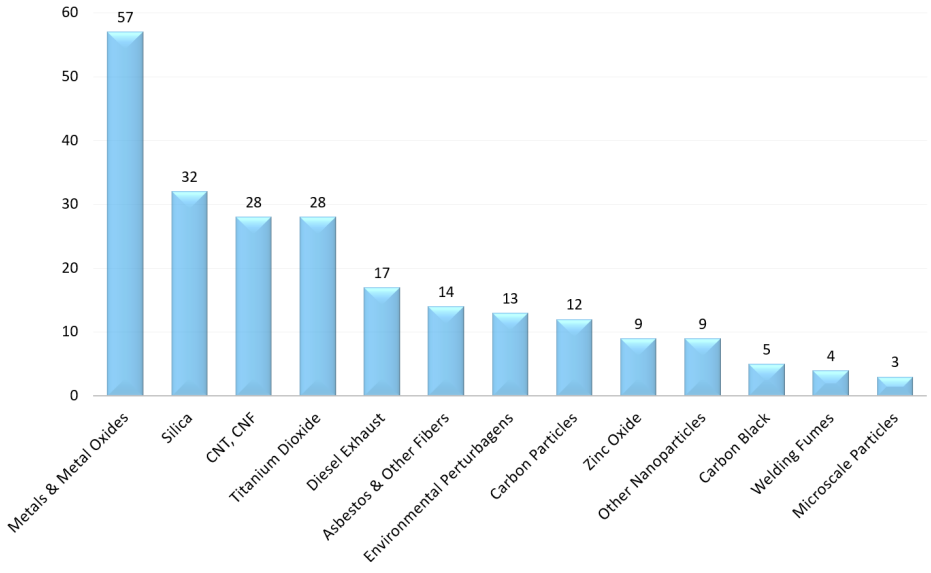


Figure 2: Number of selected study hits per material type included. Some GEO studies include more than one material so the sum of study hits for each material will be greater than the final set of 217 selected studies in Figure 1.

1.3 Final Use of Datasets

These results provide an overview of the materials with experimental designs providing sufficient data for genomic dose-response data. This information was used to inform the choice of materials and datasets to evaluate dose-dependent genomic responses after nanomaterial exposures.

1.4 2020 Genomic Dose-Response Data Search

The 2020 data search, conducted from May 6 to June 1, 2020, was focused on a smaller set of well-studied materials with potential to compare genomic dose-response data results with frank pulmonary effects from related publications. Here we included three databases including the GEO, NTP-CEBS, and NIOSHTIC-2 databases. Each of these search engines had different query structures and thus we will provide the various query statements in their respective orders. It should be noted, the updated GEO searches were able to be conducted via a command line query to the database which retrieved the data, while the other two databases the query had to be performed manually.

We were looking for nanomaterial exposure studies with a sub-chronic to chronic time period (e.g. inhalation or multiple administration of dose), benchmark materials pertinent to nanomaterial exposures (i.e. nano/micro-scale particulates), multiple dose groups, lung tissue/cell-line endpoints, and mice, rats or humans as the subject organism. It is important to note the experimental design whether it is intratracheal instillation (IT), pharyngeal aspiration (PA), or inhalation needed to contain more than one post-exposure time point and/or longer recovery times (e.g. post-exposure time of 1, 3, or more months) to fulfill the data criteria for our analyses. Our focus was specifically on titanium dioxide (TiO2), carbon black (CB), carbon nanotubes (CNT), silica (SiO2), or graphene.

1.4.1 Data Criteria

For studies to be considered for our analyses they had to include evaluation of at least one material of interest in the specified rodent subjects or cell line models. These studies also had to contain sufficient genomic dose-response information. We manually assessed the titles returned from the queries to determine which hits were "true hits" and which were not. True hits were those indicating the study included materials of interest from the set of our well studied materials.

Once the initial identification of true hits was performed we consolidated results found by multiple query statements. The related papers, citations, and/or abstracts were obtained for further review of the unique study hits relevant to our research aim.

1.4.2 GEO Database Search

Queries were conducted via command line in GEO using R (8, 9) and the 'reutils' package (10) we queried GEO using a command line statement. Queries were constructed as:

> esearch ("Material[Title] AND gse[Filter] AND (mouse[Organism] OR rat[Organism] OR human[Organism])", db = "gds")

The same search terms and query statements from the 2018 GEO query search were used resulting in a total of 90 query statements.

1.4.3 GEO 2020 Data Search Results

Query results underwent a series of filtering steps including (i) focus on key terms related to benchmark materials in pilot datasets (e.g. nano, particulates, oxides, tio2, sio2, etc), (ii) removal of duplicated study results, (iii) removal of Super Series, (iv) extraction of relevant studies with key search terms in the summary, and (v) removal of irrelevant studies (i.e. false positives). After the initial filtering (i - iv) the remaining studies were assessed for relevance using a primary set of criteria to identify studies of interest to toxicogenomic analysis for ENM risk assessment.

The updated GEO search for studies with relevant well-studied materials and experimental designs resulted in 1055 total unique hits after the initial filtering steps. Of these hits there were 140 Super Series, which were removed, and 915 singlet studies which were retained for further evaluation. These resulted from the following material terms, "nano*", "carbon", "particl* OR particu*", "fine particl* OR fine particu*", "ultrafine particl* OR ultrafine particu*", "oxide nano*", "oxide OR dioxide OR trioxide", "TiO2", "titanium, silica", "silic* OR SiO2", and "graph*". A regular expression search of the submitted summary returned the following results for each search term:

String Search Term in Submitted Summary	Number of Remaining Studies
Lung	120
Inhalation	23
Carbon black OR CB	18
TiO2 OR titanium	32
SiO2 OR silica OR silic.+dioxide	33
CNT OR carbon nanotube	37
Graph OR Graphene	19

Table 4: Study hit results from the string search for key terms/words in the "Submitted Summary" in the GEO study description provided by the authors of the original genomic data.

The submitted summary, GEO page/sample pages, and materials/methods section in the first associated publication with each of the GEO study results, excluding the lung results, were manually evaluated for relevance. The criteria for a pass included:

1. Material of interest were evaluated.
2. Tissue or cell line used for gene expression were from lung tissue sample or related cell lines.
3. Minimum of 3 non-zero dose groups and an untreated control group.
4. Route of exposure was inhalation, intratracheal instillation (IT), or pharyngeal aspiration (PA).

Lung and inhalation studies were not descriptive enough for identifying studies of interest and were likely to overlap with the studies I the material specific string searches. Thus, these studies were not investigated further.

1.4.4 NTP-CEBS Database Search

Queries were manually performed in the Data & Resources Chemical Effects in Biological Systems NTP Database (NTP-CEBS) (4) `https://manticore.niehs.nih.gov/cebssearch/`. These

searches were initially conducted on May 13, 2020 and updated on June 1, 2020. In the search bar with the databases/studies drop-down option to “ALL” we used the following query terms, with no additional structure to the query statement, they are organized in the columns by their results status (i.e. ”positive” and ”negative for hits).

Positive for Hits	Negatvie for Hits
Nanoscale, carbon black, carbon nanotube, nanotube, titanium, tio2, silica	Nanomaterial, CNT, sio2

Table 5: This table includes all of the key terms used as input for ”Material” in the query statement. Material key terms are separated based on their results status from the NTP-CEBS Database (4) (i.e. positive for hits - had at least one resulting dataset & negative for hits - had zero datasets returned).

1.4.5 NTP-CEBS Database Search Results

The following results table is only concerned with the query statements which resulted in at least one hit in the database.

Search Term	Total of Study Hits	True Hits
Nanoscale	2	2
Carbon black	69	1
Carbon nanotube	49	1
Nanotube	1	1
Titanium	4	1
Silica	4	2

Table 6: Study hit results from our NTP-CEBS search for genomic dose-response datasets by search term. Displayed are the total number of hits and number of true hits after evaluation of the study description for each hit.

Many of the false positive hits were the result of the string search used in the database, e.g. “Carbon black” resulted in hits like “Carbon tetrabromide” and “black cohosh”. Though the NTP-CEBS database does contain dose-response data for genetic markers none of the true hits in our search contained gene expression dose-response information. The dose-response studies we found contain *in vivo* endpoints and/or genetic marker information not transcription (i.e. gene expression) data.

1.4.6 NIOSHTIC-2 Database Search

Queries were conducted manually in the NIOSHTIC-2 database (5) `https://www2a.cdc.gov/nioshtic-2/n2info.asp`. These were conducted on May 13, 2020. In the search bar we used a query statement constructed as:

Key term[All Fields] AND gene expression [All Fields]

The key term was the only variable in the statement and was replaced by one of the following query terms, all of them had at least one study hit.

Key Terms (All Positive for at Least One Study Hit)
Carbon black OR CB, Carbon nanotube OR CNT, Nanomaterial, Nanoscale, Nanotube, Silica OR SiO2, Titanium OR TiO2

Table 7: This table includes all of the key terms used as input for "Material" in the query statement. All material key terms had at least one study hit in the NIOSH-TIC2 Database (5).

1.4.7 NIOSHTIC-2 Database Search Results

The following results in Table 8 are not necessarily mutually exclusive with the other sets of study hits (i.e. the set of studies collectively may contain duplicates).

Search Term	Total of Study Hits	True Hits
Carbon black OR CB	8	0
Carbon nanotube OR CNT	7	5
Nanomaterial	7	1
Nanoscale	1	1
Nanotube	7	5
Silica OR SiO2	17	5
Titanium OR TiO2	1	0

Table 8: Study hit results from our NIOSH-TIC2 search for genomic dose-response datasets by search term. Displayed are the total number of hits and number of true hits after evaluation of the study titles and descriptions for each hit.

Of the true hits from "Carbon nanotube OR CNT" and "Nanotube" all 5 overlap. Studies found in in NIOSH-TIC2 overlapped with the results from our search in GEO. Thus, these were duplicates were further explored in the the selection process of the GEO studies.

1.5 2020 Study Search & Selection

Evaluation of the study hits from all three databases combined we found that the GEO database was still the most comprehensive for genomic dose-response datasets. Study hits found in NTP-CEBS or NIOSH-TIC2 were generally reported in GEO. Thus, we chose to make the final study selection out of the 2020 updated GEO results.

The final selection process consisted of six major steps including the removal of duplicated study hits, removal of Super Series, search study summary for key terms, initial study design evaluation, and final dataset determination.

The purpose of removing duplicated study hits and Super Series was to eliminate studies that may be double counted. Double counting may result when there are multiple material exposures included in the study and when there is an overlap between Super Series datasets and "singlet"

studies (i.e. individual GEO series that make up the Super Series but can stand on their own). Once the duplicates are removed, searching the submitted summary for key terms was done to identify studies that were clearly applicable and should be considered for further evaluation. For example, if a study did not contain the nanomaterial in the submitted summary it was considered inapplicable for our analyses since it is reasonable to assume the summary contains minimal experimental design information about the samples contained in the dataset.

Initial study design evaluation was conducted to do a brief evaluation of the submitted summary and sample descriptions to ensure relevance for our analyses. Following this step, we also conducted an evaluation of the listed publication and GEO study page for clarification of experimental design details. Verification of the experimental design details allowed us to determine if a given study was relevant for out analysis goals. Once the initial set of studies was narrowed down. We had a final evaluation step for identifying if the resulting studies were useful for comparison between datasets, quantitative risk assessment of nanomaterial exposures, and representative of publicly available data that exists. Figure 3 shows the number of remaining study hits after each step.

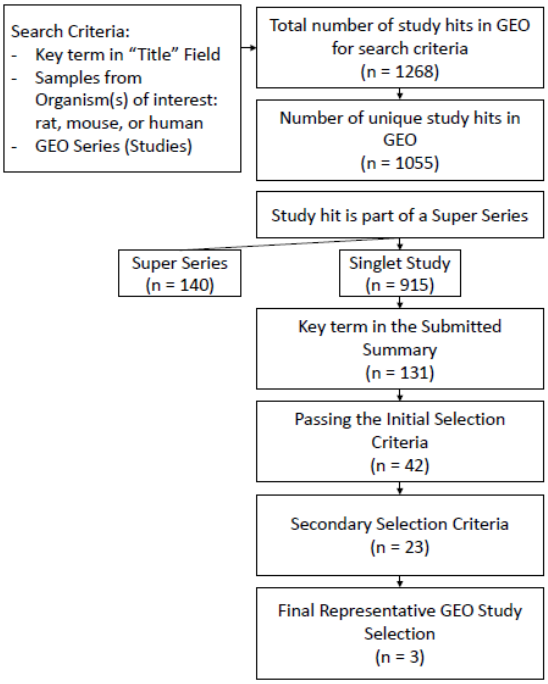


Figure 3: **2020 Study Search & Selection:** Diagram of study hit and selection results for choosing the final genomic dose-response datasets on rat, mouse, or human exposure to well-studied nanomaterials for inclusion to our study analyses.

1.6 Choice of Final Datasets for Analyses

We chose to use three representative datasets out of our updated search results for analyses with ALOHA, one for carbon nanotubes, carbon black, and titanium dioxide. These datasets were

chosen for their similar experimental designs including the microarray platforms used, subject organism (species and strain), tissue samples, method for exposure administration, and common post-exposure time points. These studies had a minimum of 3 non-zero dose groups and an untreated control which could be used for genomic dose-response analyses. Though there were other relevant studies the various materials were not represented in a balanced manner. Thus, we chose one representative for each material that were most closely aligned. The final set of datasets included three GEO series, namely GSE29042 (11), GSE35193 (12), and GSE41041 (13), which evaluated gene expression response in mouse lung tissue after exposure to multi-walled carbon nanotubes, carbon black, and titanium dioxide, respectively.

2 *In vivo* Dose Estimates & Transformation

Bermudez et al. (14) and Elder et al. (15) both utilized inhalation as the route of exposure to nanomaterials. To estimate comparable BMDs we had to convert the inhalation dose values to those comparable to intratracheal instillation or phayngeal aspiration dose values. Comparable dose estimates for the inhalation studies were obtained from the nanomaterial lung burdens in the mouse lung at the end of exposure. These values were reported in figures within each of these studies. Since we were unable to obtain the lung burden information directly from the original authors we used the WebPlotDigitizer online software (16) to estimate the lung burdens reported in the plots, see the results and specifications in Table 9. Since lung burdens for Bermudez et al. (14) were reported as $\frac{mg}{g*lung}$ of dry lungs we estimated the wet lung burdens using the following conversion:

$$d_{LB}\frac{\mu g}{lung} = \frac{x\ mg}{g\ dry-lung} * \frac{1g\ dry-lung}{5g\ wet-lung} * \frac{0.14g}{lung} * \frac{1000\mu g}{1mg} \tag{1}$$

where x is the dry lung burden at the end of exposure. Then we multiply the measured (or estimated) value by a set of biological factors. We assume the average dry-lung weight to wet-lung weight is about $1g$ dry to $5g$ wet, the default lung weight for female mice from 6 to 8 weeks old is approximately $0.14\frac{g}{lung}$. Finally, is the conversion from mg to μg.

	Elder et al. (15)	Bermudez et al. (14)
Nanomaterial	CB	TiO_2
Inhalation Doses	(0, 1, 7, 50) $\frac{mg}{m^3}$	(0, 0.5, 2, 10) $\frac{mg}{m^3}$
Log-Scale X-axis	FALSE	FALSE
Log-Scale Y-axis	FALSE	TRUE
Dry Lung	FALSE	TRUE
Lung Burdens Estimates	(0, 15.8, 245.3, 1388)$\frac{\mu g}{lung}$	(0, 0.358, 0.545, 10.489) $\frac{mg}{g*lung}$
Est. Lung Burden Dose	(0, 15.8, 245.3, 1388) $\frac{\mu g}{lung}$	(0, 10.03, 38.82, 293.7) $\frac{\mu g}{lung}$

Table 9: Dose estimate information reported in the Bermudez et al. (14) & Elder et al. (15) studies, specifications for estimating values using the WebPlotDigitizer (16), and the resulting lung burden estimates.

3 Neutrophil Count Summary Data Transformation

Bourdon et al. (12) and Saber et al. (17) did not provide the summary statistics for the percent

neutrophil in the bronchioalveolar lavage fluid (BALF). We estimated the percent neutrophils using the summary count data tables in BALF for a consistent comparison with the other *in vivo* pulmonary inflammation endpoints and our genomic results. The following equation provides the estimated mean of the percent neutrophils in BALF:

$$\widehat{p_{ct}} = \frac{\widehat{x_{ct}}}{\widehat{x_t}} \tag{2}$$

where $\widehat{p_{ct}}$ is the estimated percent cell type in BALF, $\widehat{x_{ct}}$ is the estimated mean count of the cell type in the BALF, $\widehat{x_t}$ is the estimated mean count of the total number of cells in BALF, and where ct in this case refers to neutophils as the cell type of interest. We then estimated the standard deviation of the percent neutrophil in BALF using the following equation:

$$\widehat{s_{p_{ct}}} = \sqrt{\left(\frac{\widehat{x_{ct}}}{\widehat{x_t}}\right)^2 \left(\frac{(\widehat{s_{ct}})^2}{\widehat{x_{ct}}} - \frac{(\widehat{s_t})^2}{\widehat{x_t}} - 2\frac{\widehat{\rho_{ct,t}}\widehat{s_{ct}}\widehat{s_t}}{\widehat{x_{ct}}\widehat{x_t}}\right)} \tag{3}$$

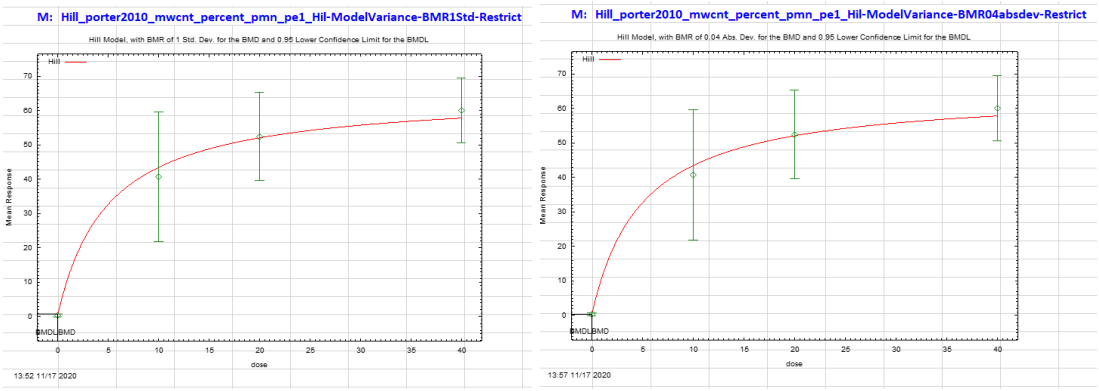
where $\widehat{s_{p_{ct}}}$ is the estimated standard deviation for percent cell type in BALF, $\widehat{s_{ct}}$ is the standard deviation of the observed cell type count, $\widehat{s_t}$ is the standard deviation of the observed total cell count, and $\widehat{\rho_{ct,t}}$ is the estimated correlation between the mean cell type count and mean of total cell count. Then, $\widehat{x_{ct}}$,$\widehat{x_t}$, and ct are the same as specified for Eq. 2.

4 *In vivo* Endpoint Dose-Response Modeling Results

This section contains all of the parametric model fitting and benchmark dose (BMD) estimation for *in vivo* endpoints indicative of pulmonary inflammation.

4.1 Porter et al. (18) Percent PMN Best BMDS Fits

4.1.1 Post-exposure Timepoint: 1-day



(a) BMR = 1 Std. Dev.

(b) BMR = + 4% Abs. Dev.

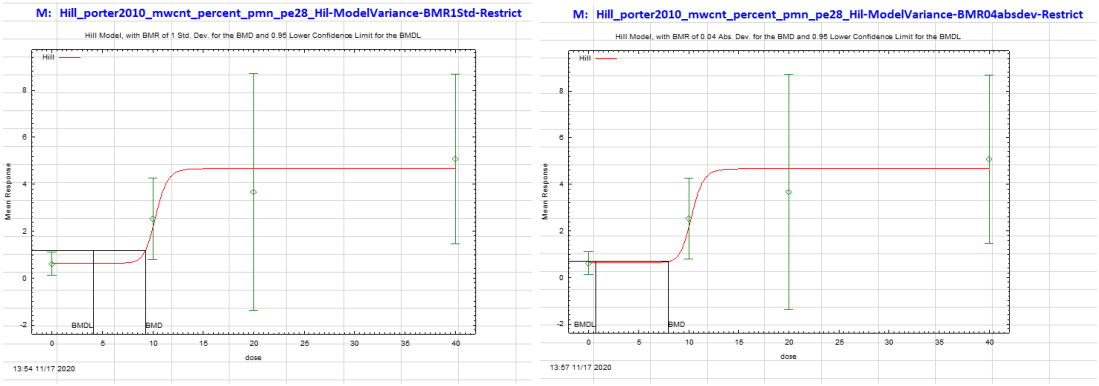
Figure 4: **MWCNT** Best parametric dose-response model fit (Hill with model variance) of observed percent PMN in mouse lungs 1-day after exposure to carbon nanotubes, see Porter et al. (18). (a) Fit with benchmark dose estimate using benchmark response of 1 standard deviation difference from mean response of untreated control group. (b) Fit with benchmark dose estimate using benchmark response of +4% absolute deviance from mean response of untreated control group.

Figure 4	(a)	(b)
Model	Hill	Hill
Variance Model	Model Variance	Model Variance
BMR	1 *Std. Dev.*	+4% *Abs. Dev.*
BMD (BMDL)	0.004 (0.000)	0.003 (0.000)
Lack of Dose Response	< 0.0001	< 0.0001
Constant Variance	< 0.0001	< 0.0001
Good variance model	0.071	0.071
Goodness-of-fit	0.912	0.912
AIC	148.16	148.16

Table 10: **MWCNT** Best parametric dose-response model fit information of observed percent PMN in mouse lungs 1-day after exposure to carbon nanotubes, see Porter et al. (18).

The BMD estimates are extremely low and the first non-zero dose group is not really within the region of any reasonable benchmark response values. Though the fit may be plausible, it is possible it is inappropriately modeling the true dose-response relationship in the low dose region. The best parametric model fit is hockey-shaped indicating we likely do not have enough data to reliably estimate the benchmark dose and estimate the minimal biological effect.

4.1.2 Post-exposure Timepoint: 28-days



(a) BMR = 1 Std. Dev.

(b) BMR = + 4% Abs. Dev.

Figure 5: **MWCNT** Best parametric dose-response model fit (Hill with model variance) of observed percent PMN in mouse lungs 28-days after exposure to carbon nanotubes, see Porter et al. (18). (a) Fit with benchmark dose estimate using benchmark response of 1 standard deviation difference from mean response of untreated control group. (b) Fit with benchmark dose estimate using benchmark response of +4% absolute deviance from mean response of untreated control group.

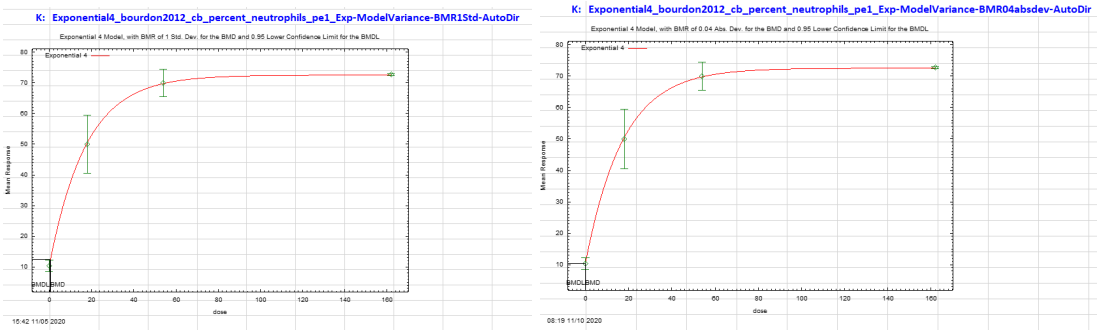
Figure 5	(a)	(b)
Model	Hill	Hill
Variance Model	Model Variance	Model Variance
BMR	1 *Std. Dev.*	+4% *Abs. Dev.*
BMD (BMDL)	9.295 (4.092)	7.990 (0.716)
Lack of Dose Response	< 0.0001	< 0.0001
Constant Variance	< 0.0001	< 0.0001
Good variance model	0.455	0.455
Goodness-of-fit	0.59	0.59
AIC	91.09	91.09

Table 11: **MWCNT** Best parametric dose-response model fit information of observed percent PMN in mouse lungs 28-days after exposure to carbon nanotubes, see Porter et al. (18).

The two high dose groups are extremely variable introducing uncertainty whether the fit should be trusted regardless if the model fit statistics suggest the best model fit is appropriate. If there is a dose-response pattern it is likely between the first three dose groups. The BMD interval estimate is uncertain as it spans almost the whole distance between the untreated control group and the first non-zero dose group. With this kind of fit and the BMD interval estimate the no and lowest observed adverse effect levels – NOAEL & LOAEL, respectively – seem to be as informative as the BMD (BMDL) estimates here.

4.2 Bourdon et al. (12) Percent PMN Best BMDS Fits

4.2.1 Post-exposure Timepoint: 1-day



(a) BMR = 1 Std. Dev.

(b) BMR = + 4% Abs. Dev.

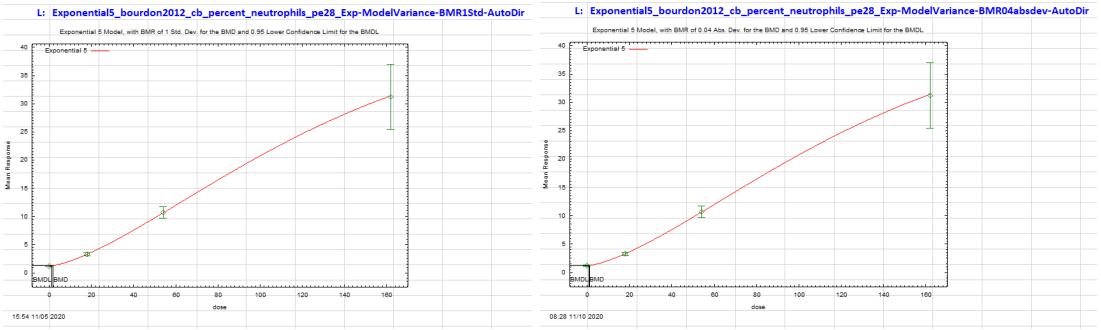
Figure 6: **CB** Best parametric dose-response model fit (Hill with model variance) of observed percent neutrophils in mouse lungs 1-day after exposure to carbon black, see Bourdon et al. (12). (a) Fit with benchmark dose estimate using benchmark response of 1 standard deviation difference from mean response of untreated control group. (b) Fit with benchmark dose estimate using benchmark response of +4% absolute deviance from mean response of untreated control group.

Figure 6	(a)	(b)
Model	Exponential 4	Exponential 4
Variance Model	Model Variance	Model Variance
BMR	1 *Std. Dev.*	+4% *Abs. Dev.*
BMD (BMDL)	0.574 (0.335)	0.011 (0.009)
Lack of Dose Response	< 0.0001	< 0.0001
Constant Variance	< 0.0001	< 0.0001
Good variance model	< 0.0001	< 0.0001
Goodness-of-fit	0.959	0.959
AIC	103.73	103.73

Table 12: **CB** Best parametric dose-response model fit information of observed percent neutrophils in mouse lungs 1-day after exposure to carbon black, see Bourdon et al. (12).

The BMD estimates are extremely low and the first non-zero dose group is not really within the region of reasonable benchmark responses. The fit might be plausible, but there is a possibility it is inappropriately modeling the true dose-response relationship in the low dose region. The best parametric model fit is hockey-shaped indicating we likely do not have enough data to reliably estimate the benchmark dose and estimate the minimal biological effect.

4.2.2 Post-exposure Timepoint: 28-days



(a) BMR = 1 Std. Dev. (b) BMR = + 4% Abs. Dev.

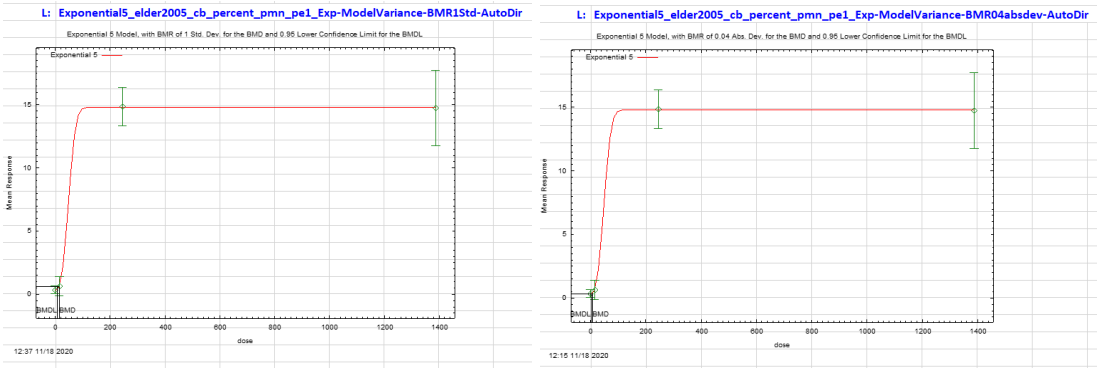
Figure 7: **CB** Best parametric dose-response model fit (Hill with model variance) of observed percent neutrophils in mouse lungs 28-days after exposure to carbon black, see Bourdon et al. (12). (a) Fit with benchmark dose estimate using benchmark response of 1 standard deviation difference from mean response of untreated control group. (b) Fit with benchmark dose estimate using benchmark response of +4% absolute deviance from mean response of untreated control group.

Figure 7	(a)	(b)
Model	Exponential 5	Exponential 5
Variance Model	Model Variance	Model Variance
BMR	1 *Std. Dev.*	+4% *Abs. Dev.*
BMD (BMDL)	1.946 (1.321)	1.172 (0.799)
Lack of Dose Response	< 0.0001	< 0.0001
Constant Variance	< 0.0001	< 0.0001
Good variance model	0.59	0.59
Goodness-of-fit	*NA*	*NA*
AIC	11.06	11.06

Table 13: **CB** Best parametric dose-response model fit information of observed percent neutrophils in mouse lungs 28-days after exposure to carbon black, see Bourdon et al. (12).

The BMD estimates are extremely low. The model fit may be plausible, but it is likely we are missing the true dose response in the low dose region with the inclusion of the highest dose group.

4.3 Elder et al. (15) Percent PMN Best BMDS Fits



(a) BMR = 1 Std. Dev.

(b) BMR = + 4% Abs. Dev.

Figure 8: **CB** Best parametric dose-response model fit (Hill with model variance) of observed percent PMNs in mouse lungs 28-days after exposure to carbon black, see Elder et al. (15). (a) Fit with benchmark dose estimate using benchmark response of 1 standard deviation difference from mean response of untreated control group. (b) Fit with benchmark dose estimate using benchmark response of +4% absolute deviance from mean response of untreated control group.

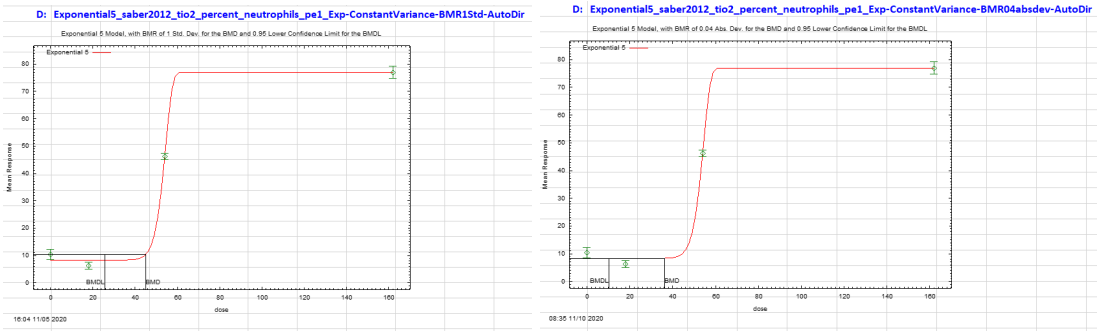
Figure 8	(a)	(b)
Model	Exponential 4	Exponential 4
Variance Model	Model Variance	Model Variance
BMR	1 *Std. Dev.*	+4% *Abs. Dev.*
BMD (BMDL)	14.372 (7.584)	7.014 (2.098)
Lack of Dose Response	< 0.0001	< 0.0001
Constant Variance	0.000	0.000
Good variance model	0.188	0.188
Goodness-of-fit	*NA*	*NA*
AIC	23.25	23.25

Table 14: **CB** Best parametric dose-response model fit information of observed percent neutrophils in mouse lungs 1-day after exposure to carbon black, see Elder et al. (15).

The fit is inappropriate for modeling the true dose-response relationship, especially in the low dose region. The best parametric model fit is hockey-shaped indicating we likely do not have enough data to reliably estimate the benchmark dose and estimate the minimal biological effect.

4.4 Saber et al. (17) Percent PMN Best BMDS Fits

4.4.1 Post-exposure Timepoint: 1-day



(a) BMR = 1 Std. Dev.

(b) BMR = + 4% Abs. Dev.

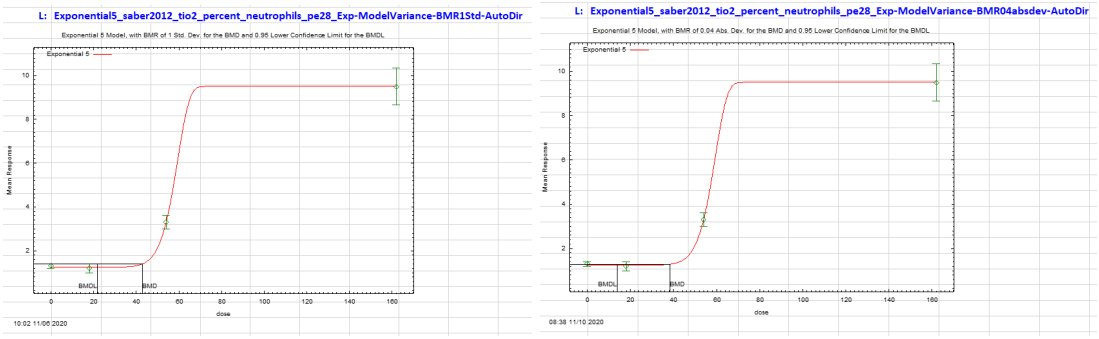
Figure 9: TiO_2 Dose-response fit of percent PMNs in mouse lungs 1-day after exposure to titanium dioxide nanoparticles, see Saber et al. (17).

Figure 9	(a)	(b)
Model	Exponential 5	Exponential 5
Variance Model	Constant Variance	Constant Variance
BMR	1 *Std. Dev.*	+4% *Abs. Dev.*
BMD (BMDL)	45.031 (25.599)	36.144 (10.211)
Lack of Dose Response	< 0.0001	< 0.0001
Constant Variance	0.332	0.332
Good variance model	0.332	0.332
Goodness-of-fit	NA	NA
AIC	68.71	68.71

Table 15: TiO_2 Best parametric dose-response model fit information of observed percent neutrophils in mouse lungs 1-day after exposure to titanium dioxide, see Bourdon et al. (12).

The background level of neutrophils is higher than one would expect with the untreated control group. Inconsistencies in the response from animals in the low dose region. For example, the percent neutrophil is depressed in the first dose group compared to control.

4.4.2 Post-exposure Timepoint: 28-days



(a) BMR = 1 Std. Dev. (b) BMR = + 4% Abs. Dev.

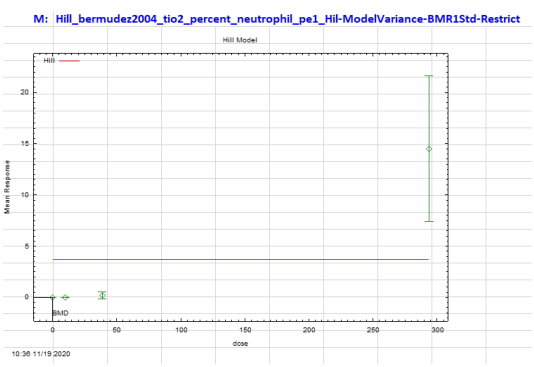
Figure 10: TiO_2 Dose-response fit of percent PMNs in mouse lungs 28-days after exposure to titanium dioxide nanoparticles, see Saber et al. (17) .

Figure 10	(a)	(b)
Model	Exponential 5	Exponential 5
Variance Model	Model Variance	Model Variance
BMR	1 *Std. Dev.*	+4% *Abs. Dev.*
BMD (BMDL)	42.926 (21.857)	38.456 (14.001)
Lack of Dose Response	< 0.0001	< 0.0001
Constant Variance	< 0.0001	< 0.0001
Good variance model	0.195	0.195
Goodness-of-fit	NA	NA
AIC	-28.18	-28.18

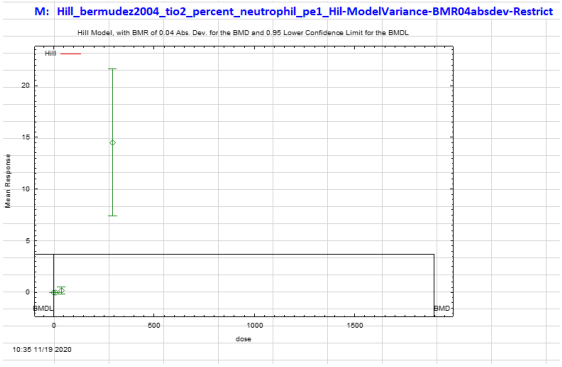
Table 16: TiO_2 Best parametric dose-response model fit information of observed percent neutrophils in mouse lungs 28-days after exposure to titanium dioxide, see Saber et al. (17).

This may be a plausible fit, but looks to inappropriately estimate the true dose-response in the low dose region. The best parametric model fit is hockey-shaped indicating we likely do not have enough data to reliably estimate the benchmark dose and estimate the minimal biological effect.

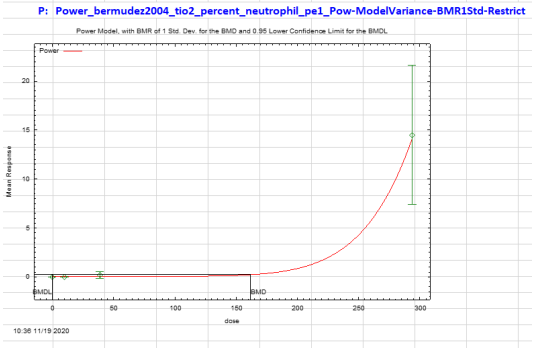
4.5 Bermudez et al. (14) Percent PMN Best BMDS Fits



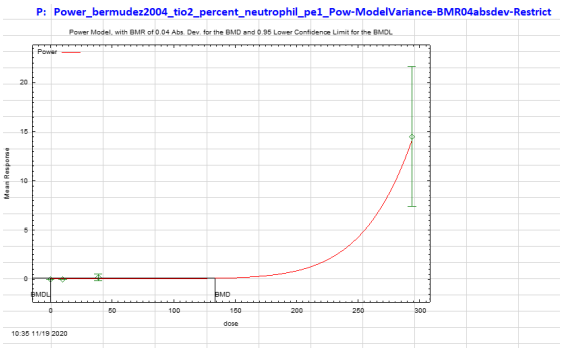
(a) BMR = 1 Std. Dev. (Hill)



(b) BMR = + 4% Abs. Dev. (Hill)



(c) BMR = 1 Std. Dev. (Power)



(d) BMR = + 4% Abs. Dev. (Power)

Figure 11: TiO_2 Best possible parametric dose-response model fits of the observed percent PMNs in mouse lungs 0 weeks after exposure to titanium dioxide, see Bermudez et al. (14). (a) Saturated Hill model fit with only model goodness-of-fit not less than 0.0001 and having a calculated AIC. No benchmark dose estimate using a benchmark response of 1 standard deviation difference from mean response of untreated control group. (b) Saturated Hill model fit with only model goodness-of-fit not less than 0.0001 and having a calculated AIC, and benchmark dose estimate using a benchmark response of +4% absolute deviance from mean response of untreated control group. (c) Best model fit by AIC and calculated benchmark dose estimate using benchmark response 1 standard deviation difference from mean response of untreated control group. (d) Best model fit by AIC and calculated benchmark dose estimate using benchmark response of +4% absolute deviance from mean response of untreated control group.

Figure 11	(a)	(b)	(c)	(d)
Model	Hill	Hill	Power	Power
Variance Model	Model Variance	Model Variance	Model Variance	Model Variance
BMR	1 *Std. Dev.*	+4% *Abs. Dev.*	1 *Std. Dev.*	+4% *Abs. Dev.*
BMD (BMDL)	–	1895.67 (1.133)	161.897 ($2.39e-05$)	133.582 (0.067)
Lack of Dose Response	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Constant Variance	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Good variance model	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Goodness-of-fit	NA	NA	< 0.0001	< 0.0001
AIC	108.42	108.42	-7.12	-7.12

Table 17: TiO_2 Model fit information for the most appropriate parametric dose-response models of observed percent PMNs in mouse lungs 0 weeks after exposure to titanium dioxide, see Bermudez et al. (14).

The observed data is not appropriate for dose-response model fitting and none of the parametric models are appropriate for estimating the dose-response curve. Any BMD estimates from these models would be unreliable. Thus, the NOAEL and LOAEL are the best potency estimates for this data.

5 *In Vivo* Pulmonary Endpoint NOAEL/LOAEL

Study	Material	PE Timepoint	NOAEL (μg)	LOAEL (μg)
Porter et al. (18)	MWCNT	1-day	–	10
Porter et al. (18)	MWCNT	7-days	10	20
Porter et al. (18)	MWCNT	28-days	–	10
Porter et al. (18)	MWCNT	56-days	10	20
Mercer et al. (19)	MWCNT	56-days	20	40
Bourdon et al. (12)	CB	1-day	–	18
Bourdon et al. (12)	CB	3-days	–	18
Bourdon et al. (12)	CB	28-days	–	18
Elder et al. (15)	CB	1-day	15.8*	245.3*
Saber et al. (17)	TiO_2	1-day	18	54
Saber et al. (17)	TiO_2	3-days	18	54
Saber et al. (17)	TiO_2	28-days	54	162
Bermudez et al. (14)	TiO_2	0-weeks	38.82*	293.7*

Table 18: No and lowest adverse effect levels – NOAEL and LOAEL, respectively – as potency estimates for the *in vivo* pulmonary endpoints. *Indicates the experimental dose estimated was converted to μg from the original experimental dose units (e.g. mg/m^3).

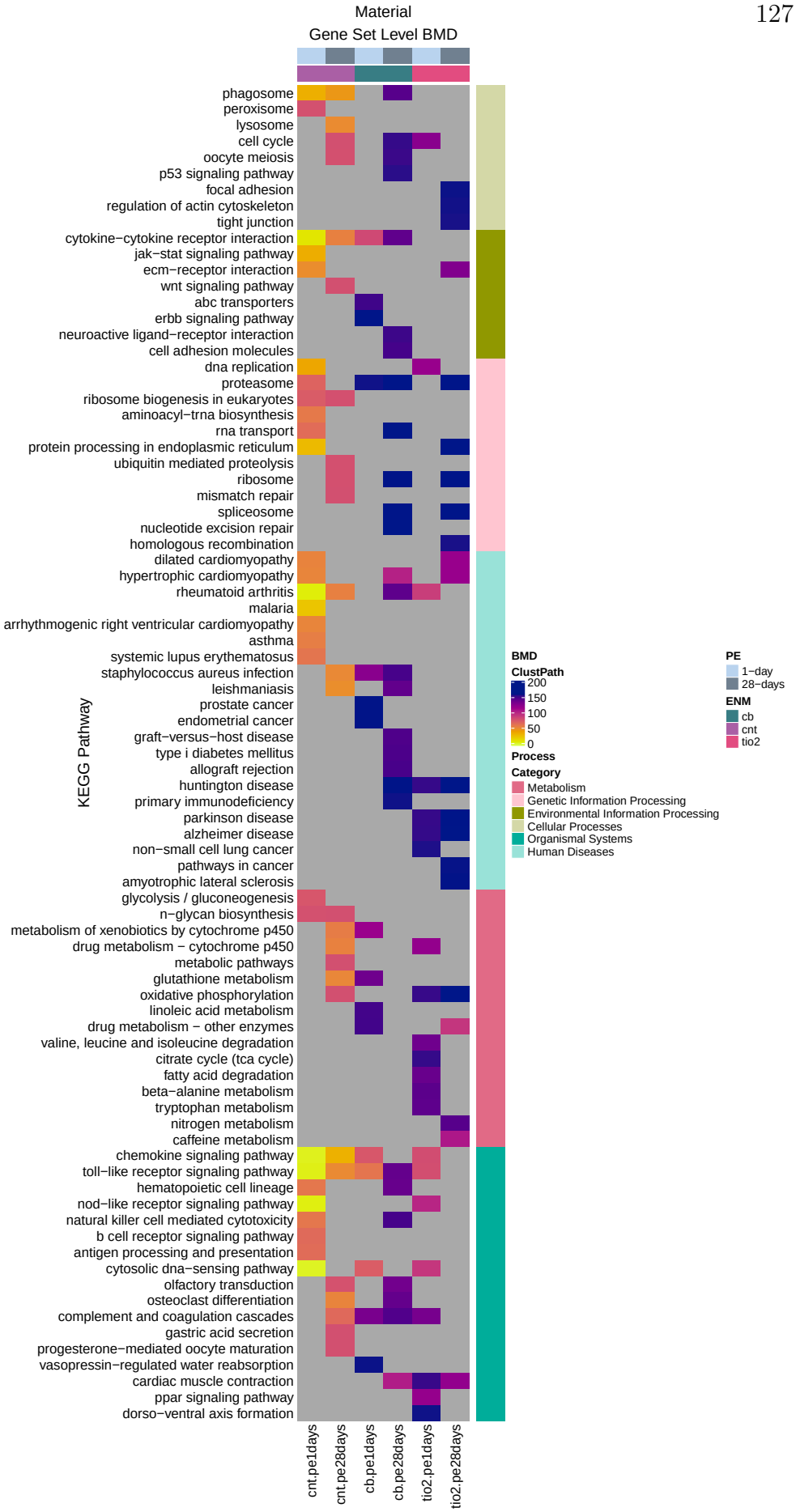
6 Enriched Gene Set Heatmaps

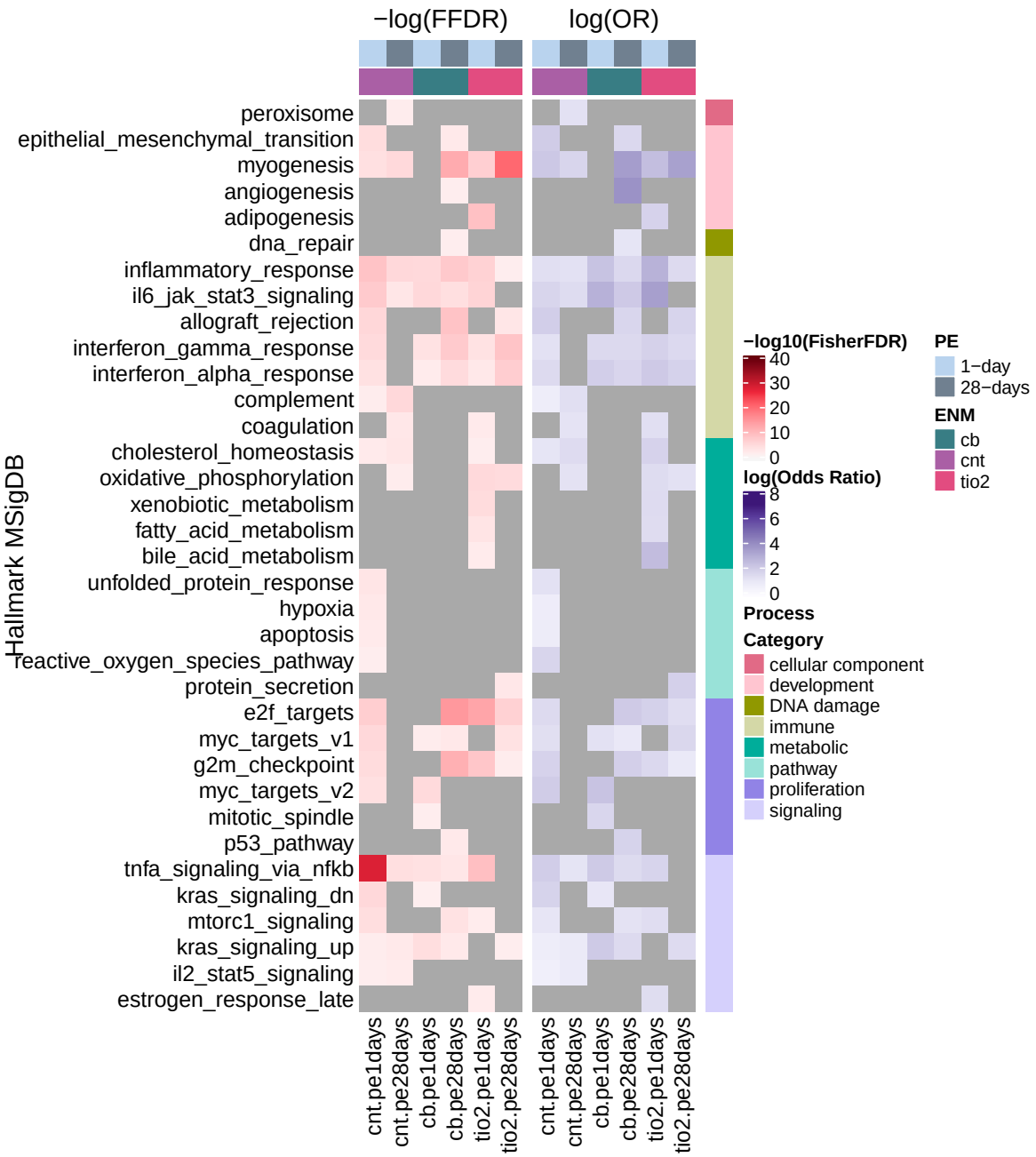
Plot descriptions: The following plots in this section of the supplement contain the enriched gene sets for multi-walled carbon nanotubes (MWCNT - GSE29052 (11)), carbon black (CB - GSE35193 (20)), and titanium dioxide (TiO_2 - GSE41041 (13)), respectively, at both 1 and 28 days post-exposure (PE). For each datasets we display the biological enrichment scores ($-log(FFDR)$ and $log(OddsRatio)$ – red and purple, respectively) and the gene set level BMD estimate from the gene cluster with the smallest $log(FFDR)$. We display a heatmap for each gene list included in the analyses, namely KEGG Pathways (21, 22, 23) and MSigDB's Hallmark and Cancer Gene Neighborhoods (CGN) (24, 25, 26, 27), which are displayed in the respective order.

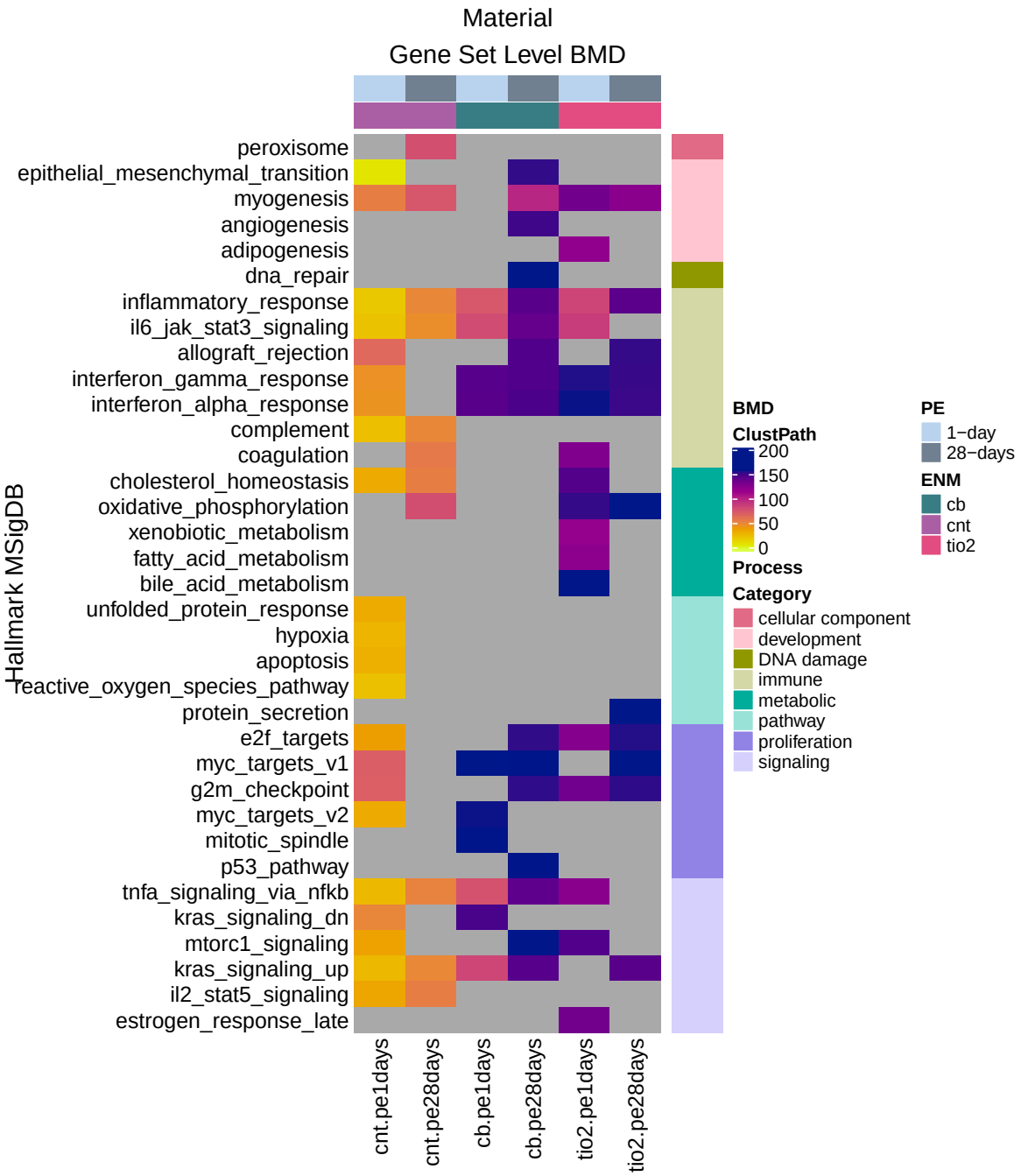
- Gene Set Enrichment Scores: In these heatmaps the rows indicate an enriched gene set and columns indicate a material exposure and post-exposure timepoint. The teal, purple, and pink colored bars at the top of the heatmap indicate the material type (carbon black, multi-walled carbon nanotube, and titanium dioxide, respectively). The light blue and slate blue colored bars at the top of the heatmap indicate post-exposure timepoints 1- and 28-days. The left heatmap displays the $-log_{10}(FishersFDR)$ enrichment scores and the right displays the $log(OddsRatio)$, where gray blocks indicate where a gene set is not enriched and darker the color indicates a higher enrichement score.

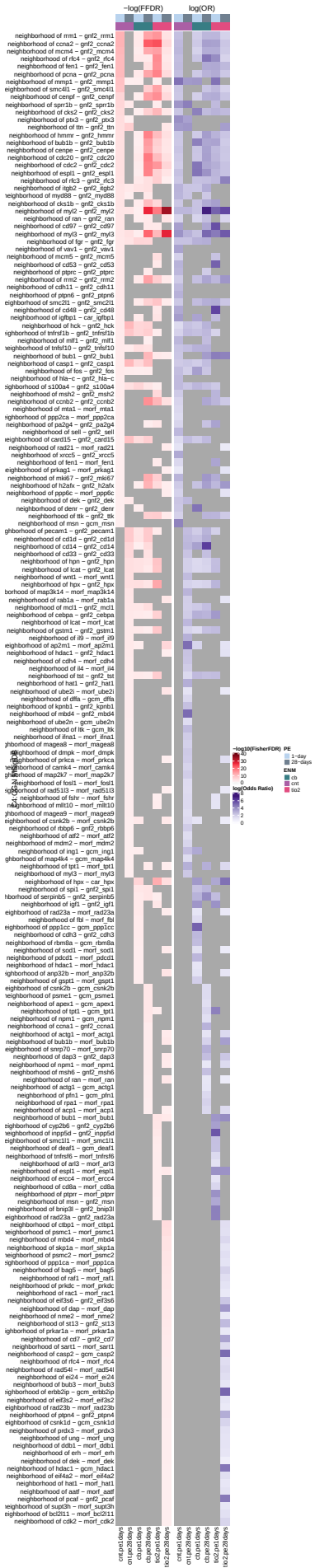
- Gene Set Level BMD Heatmaps: In these heatmaps the rows indicate an enriched gene set and columns indicate a material exposure and post-exposure timepoint. KEGG and MSigDB Hallmark gene sets are grouped by their respecitve process categories, see (21, 22, 23, 24, 25, 26, 27) and references there-in. The teal, purple, and pink colored bars at the top of the heatmap indicate the material type (carbon black, multi-walled carbon nanotube, and titanium dioxide, respectively). The light blue and slate blue colored bars at the top of the heatmap indicate post-exposure timepoints 1- and 28-days. Gray indicates where a gene set is not enriched for a material type and post-exposure timepoint. Gene set BMDs with bright yellow indicate a smaller benchmark dose estimate and dark purple indicate a bechmark dose close to the maximum experimental dose.

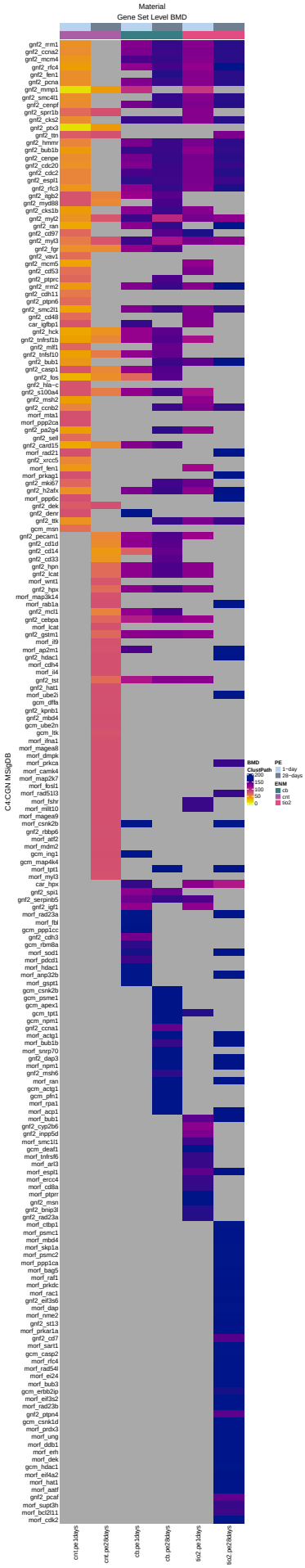

−log(FFDR)
log(OR)
KEGG Pathway
phagosome
peroxisome
lysosome
cell cycle
oocyte meiosis
p53 signaling pathway
focal adhesion
regulation of actin cytoskeleton
tight junction
cytokine−cytokine receptor interaction
jak−stat signaling pathway
ecm−receptor interaction
wnt signaling pathway
abc transporters
erbb signaling pathway
neuroactive ligand−receptor interaction
cell adhesion molecules
dna replication
proteasome
ribosome biogenesis in eukaryotes
aminoacyl−trna biosynthesis
rna transport
protein processing in endoplasmic reticulum
ubiquitin mediated proteolysis
ribosome
mismatch repair
spliceosome
nucleotide excision repair
homologous recombination
dilated cardiomyopathy
hypertrophic cardiomyopathy
rheumatoid arthritis
malaria
arrhythmogenic right ventricular cardiomyopathy
asthma
systemic lupus erythematosus
staphylococcus aureus infection
leishmaniasis
prostate cancer
endometrial cancer
graft−versus−host disease
type i diabetes mellitus
allograft rejection
huntington disease
primary immunodeficiency
parkinson disease
alzheimer disease
non−small cell lung cancer
pathways in cancer
amyotrophic lateral sclerosis
glycolysis / gluconeogenesis
n−glycan biosynthesis
metabolism of xenobiotics by cytochrome p450
drug metabolism − cytochrome p450
metabolic pathways
glutathione metabolism
oxidative phosphorylation
linoleic acid metabolism
drug metabolism − other enzymes
valine, leucine and isoleucine degradation
citrate cycle (tca cycle)
fatty acid degradation
beta−alanine metabolism
tryptophan metabolism
nitrogen metabolism
caffeine metabolism
chemokine signaling pathway
toll−like receptor signaling pathway
hematopoietic cell lineage
nod−like receptor signaling pathway
natural killer cell mediated cytotoxicity
b cell receptor signaling pathway
antigen processing and presentation
cytosolic dna−sensing pathway
olfactory transduction
osteoclast differentiation
complement and coagulation cascades
gastric acid secretion
progesterone−mediated oocyte maturation
vasopressin−regulated water reabsorption
cardiac muscle contraction
ppar signaling pathway
dorso−ventral axis formation
cnt.pe1days
cnt.pe28days
cb.pe1days
cb.pe28days
tio2.pe1days
tio2.pe28days
cnt.pe1days
cnt.pe28days
cb.pe1days
cb.pe28days
tio2.pe1days
tio2.pe28days
−log10(FisherFDR)
40
30
20
10
0
log(Odds Ratio)
8
6
4
2
0
Process
Category
Metabolism
Genetic Information Processing
Environmental Information Processing
Cellular Processes
Organismal Systems
Human Diseases
PE
1−day
28−days
ENM
cb
cnt
tio2


Material
Gene Set Level BMD
KEGG Pathway
phagosome
peroxisome
lysosome
cell cycle
oocyte meiosis
p53 signaling pathway
focal adhesion
regulation of actin cytoskeleton
tight junction
cytokine−cytokine receptor interaction
jak−stat signaling pathway
ecm−receptor interaction
wnt signaling pathway
abc transporters
erbb signaling pathway
neuroactive ligand−receptor interaction
cell adhesion molecules
dna replication
proteasome
ribosome biogenesis in eukaryotes
aminoacyl−trna biosynthesis
rna transport
protein processing in endoplasmic reticulum
ubiquitin mediated proteolysis
ribosome
mismatch repair
spliceosome
nucleotide excision repair
homologous recombination
dilated cardiomyopathy
hypertrophic cardiomyopathy
rheumatoid arthritis
malaria
arrhythmogenic right ventricular cardiomyopathy
asthma
systemic lupus erythematosus
staphylococcus aureus infection
leishmaniasis
prostate cancer
endometrial cancer
graft−versus−host disease
type i diabetes mellitus
allograft rejection
huntington disease
primary immunodeficiency
parkinson disease
alzheimer disease
non−small cell lung cancer
pathways in cancer
amyotrophic lateral sclerosis
glycolysis / gluconeogenesis
n−glycan biosynthesis
metabolism of xenobiotics by cytochrome p450
drug metabolism − cytochrome p450
metabolic pathways
glutathione metabolism
oxidative phosphorylation
linoleic acid metabolism
drug metabolism − other enzymes
valine, leucine and isoleucine degradation
citrate cycle (tca cycle)
fatty acid degradation
beta−alanine metabolism
tryptophan metabolism
nitrogen metabolism
caffeine metabolism
chemokine signaling pathway
toll−like receptor signaling pathway
hematopoietic cell lineage
nod−like receptor signaling pathway
natural killer cell mediated cytotoxicity
b cell receptor signaling pathway
antigen processing and presentation
cytosolic dna−sensing pathway
olfactory transduction
osteoclast differentiation
complement and coagulation cascades
gastric acid secretion
progesterone−mediated oocyte maturation
vasopressin−regulated water reabsorption
cardiac muscle contraction
ppar signaling pathway
dorso−ventral axis formation
cnt.pe1days
cnt.pe28days
cb.pe1days
cb.pe28days
tio2.pe1days
tio2.pe28days
BMD
ClustPath
200
150
100
50
0
Process
Category
Metabolism
Genetic Information Processing
Environmental Information Processing
Cellular Processes
Organismal Systems
Human Diseases
PE
1−day
28−days
ENM
cb
cnt
tio2


−log(FFDR)
log(OR)
Hallmark MSigDB
peroxisome
epithelial_mesenchymal_transition
myogenesis
angiogenesis
adipogenesis
dna_repair
inflammatory_response
il6_jak_stat3_signaling
allograft_rejection
interferon_gamma_response
interferon_alpha_response
complement
coagulation
cholesterol_homeostasis
oxidative_phosphorylation
xenobiotic_metabolism
fatty_acid_metabolism
bile_acid_metabolism
unfolded_protein_response
hypoxia
apoptosis
reactive_oxygen_species_pathway
protein_secretion
e2f_targets
myc_targets_v1
g2m_checkpoint
myc_targets_v2
mitotic_spindle
p53_pathway
tnfa_signaling_via_nfkb
kras_signaling_dn
mtorc1_signaling
kras_signaling_up
il2_stat5_signaling
estrogen_response_late
cnt.pe1days
cnt.pe28days
cb.pe1days
cb.pe28days
tio2.pe1days
tio2.pe28days
−log10(FisherFDR)
40
30
20
10
0
log(Odds Ratio)
8
6
4
2
0
Process
Category
cellular component
development
DNA damage
immune
metabolic
pathway
proliferation
signaling
PE
1−day
28−days
ENM
cb
cnt
tio2

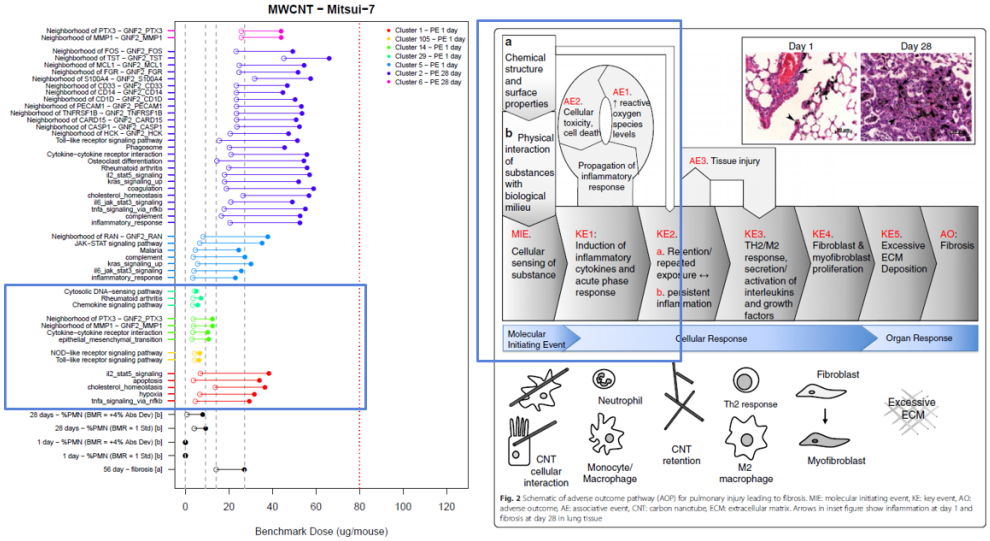

Material
Gene Set Level BMD
Hallmark MSigDB
peroxisome
epithelial_mesenchymal_transition
myogenesis
angiogenesis
adipogenesis
dna_repair
inflammatory_response
il6_jak_stat3_signaling
allograft_rejection
interferon_gamma_response
interferon_alpha_response
complement
coagulation
cholesterol_homeostasis
oxidative_phosphorylation
xenobiotic_metabolism
fatty_acid_metabolism
bile_acid_metabolism
unfolded_protein_response
hypoxia
apoptosis
reactive_oxygen_species_pathway
protein_secretion
e2f_targets
myc_targets_v1
g2m_checkpoint
myc_targets_v2
mitotic_spindle
p53_pathway
tnfa_signaling_via_nfkb
kras_signaling_dn
mtorc1_signaling
kras_signaling_up
il2_stat5_signaling
estrogen_response_late
cnt.pe1days
cnt.pe28days
cb.pe1days
cb.pe28days
tio2.pe1days
tio2.pe28days
BMD
ClustPath
200
150
100
50
0
Process
Category
cellular component
development
DNA damage
immune
metabolic
pathway
proliferation
signaling
PE
1−day
28−days
ENM
cb
cnt
tio2


−log(FFDR)
log(OR)
neighborhood of rrm1 – gnf2_rrm1
neighborhood of ccna2 – gnf2_ccna2
neighborhood of mcm4 – gnf2_mcm4
neighborhood of rfc4 – gnf2_rfc4
neighborhood of fen1 – gnf2_fen1
neighborhood of pcna – gnf2_pcna
neighborhood of mmp1 – gnf2_mmp1
neighborhood of smc4l1 – gnf2_smc4l1
neighborhood of cenpf – gnf2_cenpf
neighborhood of sprr1b – gnf2_sprr1b
neighborhood of cks2 – gnf2_cks2
neighborhood of ptx3 – gnf2_ptx3
neighborhood of ttn – gnf2_ttn
neighborhood of hmmr – gnf2_hmmr
neighborhood of bub1b – gnf2_bub1b
neighborhood of cenpe – gnf2_cenpe
neighborhood of cdc20 – gnf2_cdc20
neighborhood of cdc2 – gnf2_cdc2
neighborhood of espl1 – gnf2_espl1
neighborhood of rfc3 – gnf2_rfc3
neighborhood of itgb2 – gnf2_itgb2
neighborhood of myd88 – gnf2_myd88
neighborhood of cks1b – gnf2_cks1b
neighborhood of myl2 – gnf2_myl2
neighborhood of ran – gnf2_ran
neighborhood of cd97 – gnf2_cd97
neighborhood of myl3 – gnf2_myl3
neighborhood of fgr – gnf2_fgr
neighborhood of vav1 – gnf2_vav1
neighborhood of mcm5 – gnf2_mcm5
neighborhood of cd53 – gnf2_cd53
neighborhood of ptprc – gnf2_ptprc
neighborhood of rrm2 – gnf2_rrm2
neighborhood of cdh11 – gnf2_cdh11
neighborhood of ptpn6 – gnf2_ptpn6
eighborhood of smc2l1 – gnf2_smc2l1
neighborhood of cd48 – gnf2_cd48
neighborhood of igfbp1 – car_igfbp1
neighborhood of hck – gnf2_hck
ighborhood of tnfrsf1b – gnf2_tnfrsf1b
neighborhood of rtf1 – gnf2_rtf1
eighborhood of tnfsf10 – gnf2_tnfsf10
neighborhood of bub1 – gnf2_bub1
neighborhood of casp1 – gnf2_casp1
neighborhood of fos – gnf2_fos
neighborhood of hla−c – gnf2_hla−c
ighborhood of s100a4 – gnf2_s100a4
neighborhood of msh2 – gnf2_msh2
neighborhood of ccnb2 – gnf2_ccnb2
neighborhood of mta1 – morf_mta1
ighborhood of ppp2ca – morf_ppp2ca
neighborhood of pa2g4 – gnf2_pa2g4
neighborhood of sell – gnf2_sell
eighborhood of card15 – gnf2_card15
neighborhood of rad21 – morf_rad21
neighborhood of xrcc5 – gnf2_xrcc5
neighborhood of fen1 – morf_fen1
eighborhood of prkag1 – morf_prkag1
neighborhood of mki67 – gnf2_mki67
neighborhood of h2afx – gnf2_h2afx
neighborhood of ppp6c – morf_ppp6c
neighborhood of dek – gnf2_dek
neighborhood of denr – gnf2_denr
neighborhood of ttk – gnf2_ttk
neighborhood of msn – gcm_msn
ghborhood of pecam1 – gnf2_pecam1
neighborhood of cd1d – gnf2_cd1d
neighborhood of cd14 – gnf2_cd14
neighborhood of cd33 – gnf2_cd33
neighborhood of hpn – gnf2_hpn
neighborhood of lcat – gnf2_lcat
neighborhood of wnt1 – morf_wnt1
neighborhood of hpx – gnf2_hpx
borhood of map3k14 – morf_map3k14
neighborhood of rab1a – morf_rab1a
neighborhood of mct1 – gnf2_mct1
neighborhood of cebpa – gnf2_cebpa
neighborhood of lcat – morf_lcat
neighborhood of gstm1 – gnf2_gstm1
neighborhood of il9 – morf_il9
eighborhood of ap2m1 – morf_ap2m1
neighborhood of hdac1 – gnf2_hdac1
neighborhood of cdh4 – morf_cdh4
neighborhood of il4 – morf_il4
neighborhood of tst – gnf2_tst
neighborhood of hat1 – gnf2_hat1
neighborhood of ube2i – morf_ube2i
neighborhood of dffa – gcm_dffa
neighborhood of kpnb1 – gnf2_kpnb1
neighborhood of mbd4 – gnf2_mbd4
neighborhood of ube2n – gcm_ube2n
neighborhood of ltk – gcm_ltk
neighborhood of ifna1 – morf_ifna1
jhborhood of magea8 – morf_magea8
ighborhood of dmpk – morf_dmpk
ighborhood of prkca – morf_prkca
neighborhood of camk4 – morf_camk4
ghborhood of map2k7 – morf_map2k7
neighborhood of fost1 – morf_fost1
ighborhood of rad51l3 – morf_rad51l3
neighborhood of fshr – morf_fshr
neighborhood of mll10 – morf_mll10
jhborhood of magea9 – morf_magea9
eighborhood of csnk2b – morf_csnk2b
neighborhood of rbbp6 – gnf2_rbbp6
neighborhood of atf2 – morf_atf2
neighborhood of mdm2 – morf_mdm2
neighborhood of ing1 – gcm_ing1
ghborhood of map4k4 – gcm_map4k4
neighborhood of tpt1 – morf_tpt1
neighborhood of myl3 – morf_myl3
neighborhood of hpx – car_hpx
neighborhood of spi1 – gnf2_spi1
hborhood of serpinb5 – gnf2_serpinb5
neighborhood of igf1 – gnf2_igf1
eighborhood of rad23a – morf_rad23a
neighborhood of ltbl – morf_ltbl
eighborhood of ppp1cc – gcm_ppp1cc
neighborhood of cdh3 – gnf2_cdh3
neighborhood of rbm8a – gcm_rbm8a
neighborhood of sod1 – morf_sod1
neighborhood of pdcd1 – morf_pdcd1
neighborhood of hdac1 – morf_hdac1
ighborhood of anp32b – morf_anp32b
neighborhood of gspt1 – morf_gspt1
eighborhood of csnk2b – gcm_csnk2b
eighborhood of psme1 – gcm_psme1
neighborhood of apex1 – gcm_apex1
neighborhood of tpt1 – gcm_tpt1
neighborhood of npm1 – gcm_npm1
neighborhood of ccna1 – gnf2_ccna1
neighborhood of actg1 – morf_actg1
neighborhood of bub1b – morf_bub1b
eighborhood of snrp70 – morf_snrp70
neighborhood of dap3 – gnf2_dap3
neighborhood of npm1 – morf_npm1
neighborhood of msh6 – gnf2_msh6
neighborhood of ran – morf_ran
neighborhood of actg1 – gcm_actg1
neighborhood of pfn1 – gcm_pfn1
neighborhood of rpa1 – morf_rpa1
neighborhood of acp1 – morf_acp1
neighborhood of bub1 – morf_bub1
eighborhood of cyp2b6 – gnf2_cyp2b6
eighborhood of inpp5d – gnf2_inpp5d
eighborhood of smc1l1 – morf_smc1l1
neighborhood of deaf1 – gcm_deaf1
neighborhood of tnfrsf6 – morf_tnfrsf6
neighborhood of arl3 – morf_arl3
neighborhood of espl1 – morf_espl1
neighborhood of ercc4 – morf_ercc4
neighborhood of cd8a – morf_cd8a
neighborhood of ptprr – morf_ptprr
neighborhood of msn – gnf2_msn
neighborhood of bnip3l – gnf2_bnip3l
eighborhood of rad23a – gnf2_rad23a
neighborhood of ctbp1 – morf_ctbp1
eighborhood of psmc1 – morf_psmc1
neighborhood of mbd4 – morf_mbd4
neighborhood of skp1a – morf_skp1a
eighborhood of psmc2 – morf_psmc2
ighborhood of ppp1ca – morf_ppp1ca
neighborhood of bag5 – morf_bag5
neighborhood of raf1 – morf_raf1
neighborhood of prkdc – morf_prkdc
neighborhood of rac1 – morf_rac1
neighborhood of eif3s6 – gnf2_eif3s6
neighborhood of dap – morf_dap
neighborhood of nme2 – morf_nme2
neighborhood of st13 – gnf2_st13
ighborhood of prkar1a – morf_prkar1a
neighborhood of cd7 – gnf2_cd7
neighborhood of sart1 – morf_sart1
neighborhood of casp2 – gcm_casp2
neighborhood of rfc4 – morf_rfc4
neighborhood of rad54l – morf_rad54l
neighborhood of ei24 – morf_ei24
neighborhood of bub3 – morf_bub3
ighborhood of ertbb2ip – gcm_ertbb2ip
neighborhood of eif3s2 – morf_eif3s2
eighborhood of rad23b – morf_rad23b
neighborhood of ptpn4 – gnf2_ptpn4
eighborhood of csnk1d – gcm_csnk1d
neighborhood of prdx3 – morf_prdx3
neighborhood of ung – morf_ung
neighborhood of ddb1 – morf_ddb1
neighborhood of erh – morf_erh
neighborhood of dek – morf_dek
neighborhood of hdac1 – gcm_hdac1
neighborhood of eif4a2 – morf_eif4a2
neighborhood of hat1 – morf_hat1
neighborhood of aatf – morf_aatf
neighborhood of pcaf – gnf2_pcaf
leighborhood of supt3h – morf_supt3h
eighborhood of bcl2l11 – morf_bcl2l11
neighborhood of cdk2 – morf_cdk2
cnt.pe1day
cnt.pe28days
crt.pe1day
crt.pe28days
cb.pe1day
cb.pe28days
tio2.pe1day
tio2.pe28days
cnt.ex1day
cnt.ex28days
crt.ex1day
crt.ex28days
cb.ex1day
cb.ex28days
tio2.ex1day
tio2.ex28days
−log10(FisherFDR)
50
40
30
20
10
0
log(Odds Ratio)
10
8
6
4
2
0
PE
1−day
28−days
ENM
cb
cnt
tio2

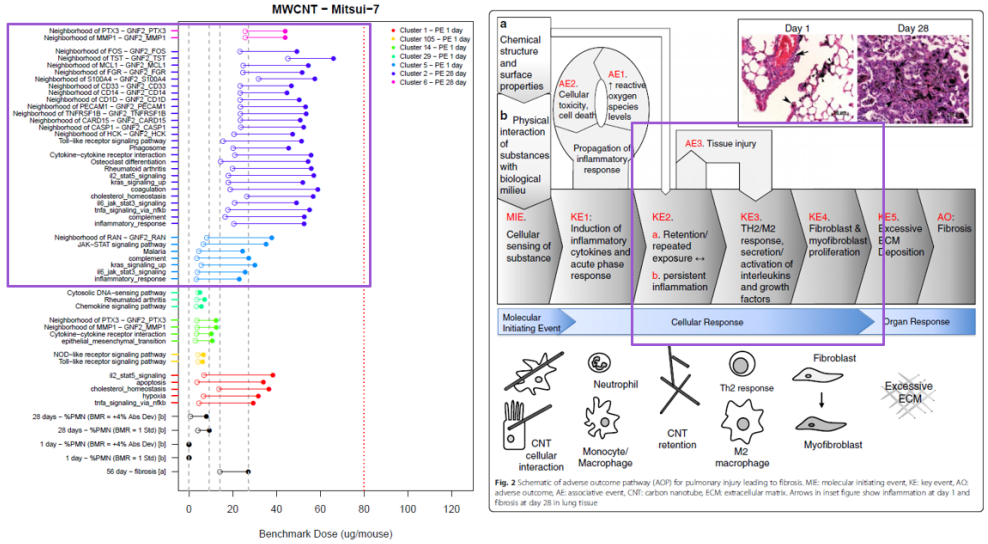

Material
Gene Set Level BMD
C4:CGN MSigDB
BMD
ClustPath
200
150
100
50
0
PE
1−day
28−days
ENM
cb
cnt
to2
cnt.pe.1days
cnt.pe.28days
cb.pe.1days
cb.pe.28days
to2.pe.1days
to2.pe.28days

7 Shared between Core Genes from ALOHA & Current Pulmonary Fibrosis Gene Sets

7.1 Comparison with Labib et al. (28) Pulmonary Fibrosis AOP



(a) PE 1-day



(b) PE 28-days

Figure 12: Comparison between ALOHA gene clusters enriched with the most biologically relevant gene sets at both time points after exposure to multi-walled carbon nanotubes from Guo et al. (11) (right) and Figure 2 from Labib et al. (28) illustrating their proposed pulmonary fibrosis adverse outcome pathway (AOP) (left). [Note: Figure 2 utilized here on the left is directly from Labib et al. (28) can be found in the original article written by Sarah Labib et al. published by Springer Nature in their journal *Particle and Fibre Toxicology* on March 15, 2016. The figure cited is under Creative Commons license.]

7.2 Comparison with Ehrhart et al. (29) Pulmonary Fibrosis AOP

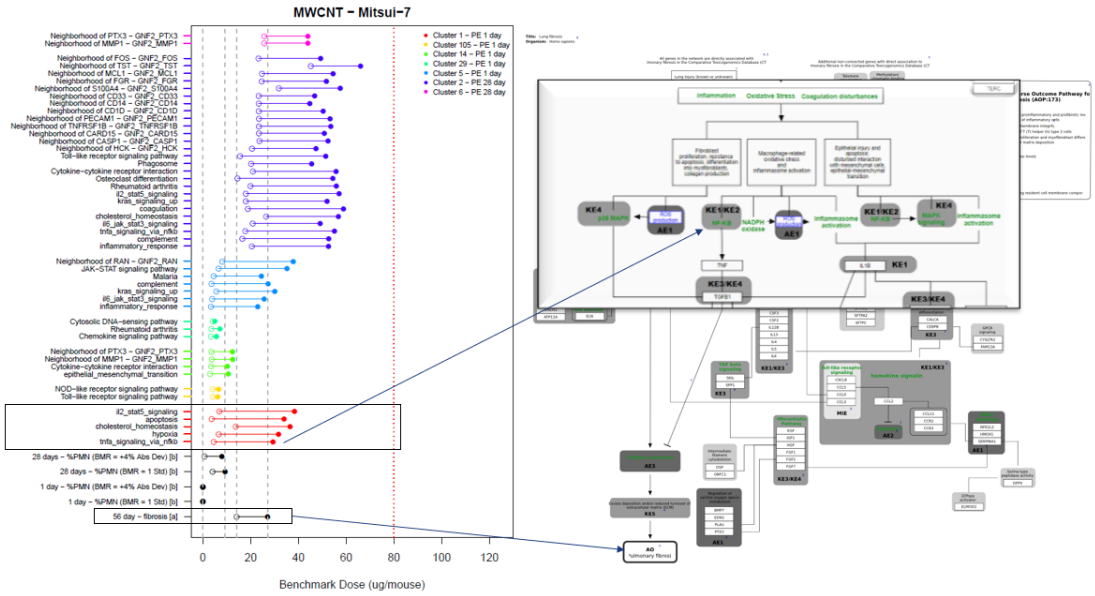


Figure 13: Comparison between ALOHA gene Cluster 1 enriched with the most biologically relevant gene sets 1-day post-exposure to multi-walled carbon nanotubes from Guo et al. (11) (right) and alignment with "$NF-\kappa B$" key events in the Lung Fibrosis (*Homo sapiens*) Wiki Pathway (29, 30, 31, 32, 33) (left).

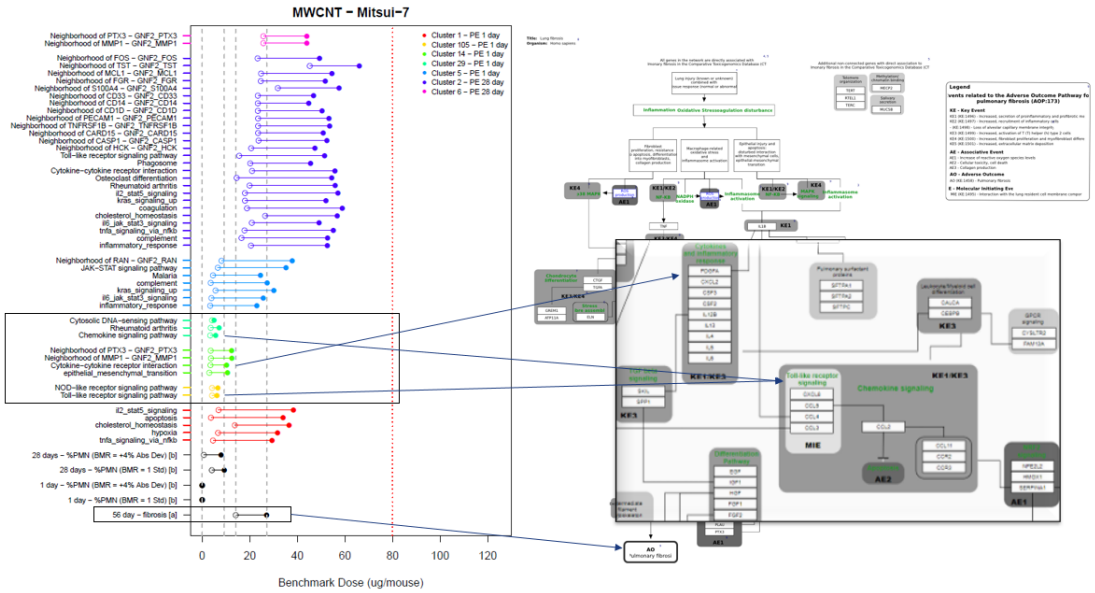


Figure 14: Comparison between ALOHA gene Clusters 105, 14, & 29 enriched with the most biologically relevant gene sets 1-day post-exposure to multi-walled carbon nanotubes from Guo et al. (11) (right) and alignment with "$Cytokineandinflammatoryresponse$" & "$Toll-likereceptorsignaling$" key events in the Lung Fibrosis (*Homo sapiens*) Wiki Pathway, see (29, 30, 31, 32, 33) and references therein (left).

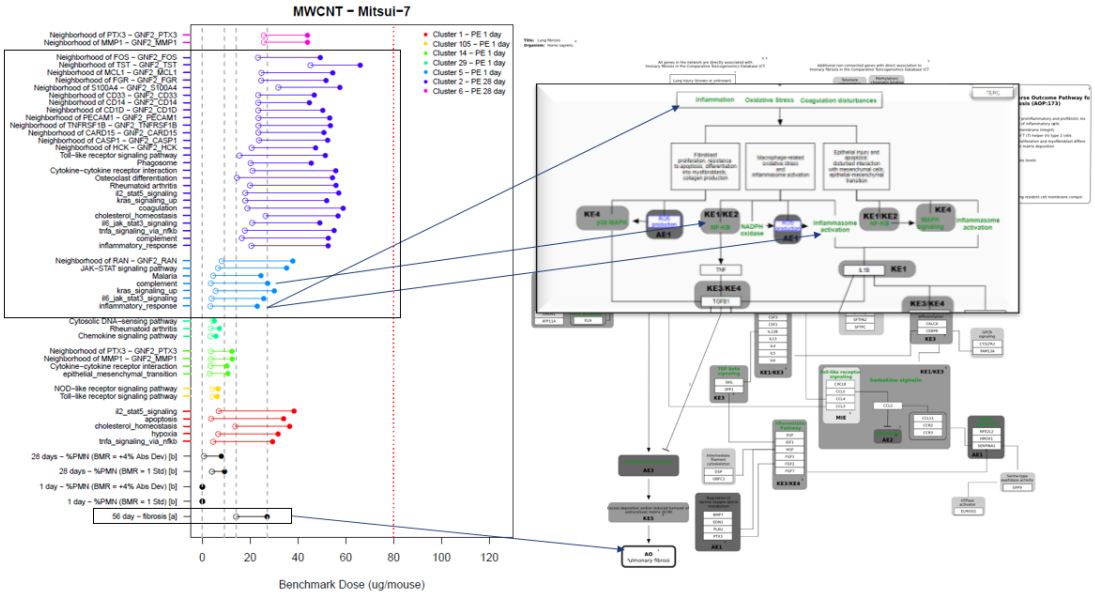


Figure 15: Comparison between ALOHA gene Cluster 5 enriched with the most biologically relevant gene sets 1-day post-exposure to multi-walled carbon nanotubes from Guo et al. (11) (right) and alignment with "$Inflammation$", "$NF-\kappa B$", & "$Inflammasomeactivation$" key events in the Lung Fibrosis (*Homo sapiens*) Wiki Pathway, see (29, 30, 31, 32, 33) and references therein (left).

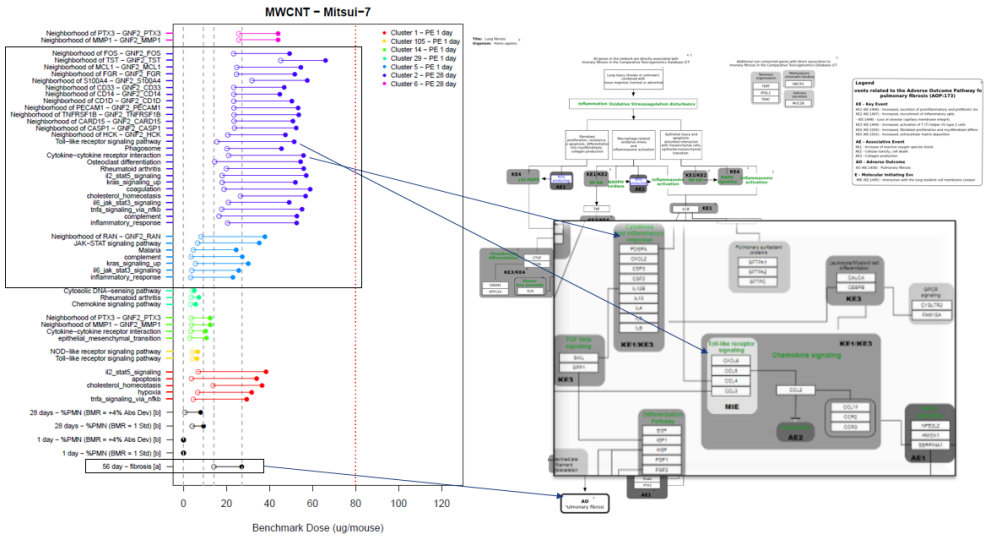


Figure 16: Comparison between ALOHA gene Cluster 2 enriched with the most biologically relevant gene sets 28-days post-exposure to multi-walled carbon nanotubes from Guo et al. (11) (right) and alignment with "$Cytokineandinflammatoryresponse$" & "$Toll-likereceptorsignaling$" key events in the Lung Fibrosis (*Homo sapiens*) Wiki Pathway (29, 30, 31, 32, 33) (left).

7.3 Core Genes of Most Sensitive Gene Sets

PE Timepoint	Enriched Gene Set	ALOHA Core Genes
1-day	Epithelial Mesenchymal Transition	Tnfrsf12a, Cxcl3, Mmp3, Tnc, Il6, Lox, Cxcl2, Timp1, Cxcl1, Ptx3, Cxcl5, Spp1, Serpine1
1-day	Cytosolic DNA sensing Pathway	Ccl12, Il6, Lox, Ptx3, Cxcl5, Serpine1
1-day	Neighborhood of PTX3	Il1b, Il6, Ccl4, Il33
28-days	Il6 Jak Stat3 Signaling	Ccl7, Cd44, Il1b, Tnf, Cxcl3, Cd14, Csf2, Tlr2, Ccr1, Cxcl10, Pik3r5, Il6, Osmr, Cxcl2, Tnfrsf1b, Csf2rb2, Cxcl1
28-days	Chemokine Signaling Pathway	Ptgs1, Ccl12, Il6, Cxcl5
28-days	Neighborhood of MMP1	Ccl7, Cxcr1, Ccl8, Ccl9, Cxcl10, Ccl12, Ccl17, Cxcl1, Cxcl5

Table 19: Core genes from the most sensitive enriched gene sets at 1- and 28-days after exposure to **multiwalled carbon nanotubes** (11). Gene names in dark green are also found in the pulmonary fibrotic model in Brass et al. (34).

PE Timepoint	Enriched Gene Set	ALOHA Core Genes
1-day	Inflammatory response	Csf3r, Ccl7, Il1b, Nfkbia, Ffar2, Cd14, Cxcl10, Ccl12, Ccl17, Timp1, Cxcl9, Cxcl5
1-day	Toll-like receptor signaling pathway	Cxcr2, Ccl7, Nfkbia, Ccl8, Ccl11, Ccl9, Cxcl10, Ccl12, Ccl17, Ccl4, Cxcl1, Cxcl9, Ccl2, Cxcl5
1-day	Neighborhood of FOS	Klf4, Nfkbia, Ms4a6d, Dusp1
28-days	Myogenesis	Cfd, Tnnt2, Mb, Myl4, Tcap, Csrp3, Ckmt2, Myl7, Sln, Fabp3, Hrc, Tnnc1, Actc1, Actn2, Cox7a1
28-days	Hypertrophic cardiomyopathy	Myh6, Tnnt2, Actc1, Tnni3
28-days	Neighborhood of MYL2	Myh6, Tnnt2, Mb, Tcap, Csrp3, Ckmt2, Myl7, Fabp3, Tnnc1, Myoz2, Actc1, Actn2, Cox7a1, Tnni3

Table 20: Core genes from the most sensitive enriched gene sets at 1- and 28-days after exposure to **carbon black** (20).

PE Timepoint	Enriched Gene Set	ALOHA Core Genes
1-day	Inflammatory response	Ccl7, Il1b, Cd14, Cxcl10, Il6, Ccl17, Timp1, Rgs1, Cxcl5, Mefv, Ccl20
1-day	Toll-like receptor signaling Pathway	Ccl7, Cxcl10, Ccl3, Ccl17, Ccl4, Cxcl1, Ccl2, Cxcl5, Ccl20
1-day	Neighborhood of MMP1	Il1b, Il6, Cxcl5
28-days	Myogenesis	Cfd, Tnnt2, Pgam2, Ckm, Mb, Myl4, Tcap, Csrp3, Myl3, Myom2, Ckmt2, Cox6a2, Myl7, Sln, Fabp3, Hrc, Myh4, Myl1, Tnnc1, Actc1, Actn2, Cox7a1, Eno3, Mybpc3, Sgcg
28-days	Drug Metabolism	Upp1, Cyp2a4, Cyp2a5
28-days	Neighborhood of HPX	Igfbp1, C8a, Fgg, Aldob, Dpys

Table 21: Core genes from the most sensitive enriched gene sets at 1- and 28-days after exposure to **titanium dioxide** (13).

References

[1] Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, et al. Ncbi geo: archive for functional genomics data sets—10 years on. *Nucleic acids research*, 39(suppl_1):D1005–D1010, 2010.

[2] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research*, 41 (D1):D991–D995, 2012.

[3] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[4] National Institute of Environmental Health Sciences NIEHS. Chemical effects in biological systems, Apr 2020. URL `https://manticore.niehs.nih.gov/cebssearch/`.

[5] National Institute of Occupational Safety & Health NIOSH. Nioshtic-2 publications search, Dec 2020. URL `https://www2a.cdc.gov/nioshtic-2/advsearch2.asp`.

[6] Andrew Williams and Sabina Halappanavar. Application of biclustering of gene expression data and gene set enrichment analysis methods to identify potentially disease causing nanomaterials. *Beilstein journal of nanotechnology*, 6(1):2438–2448, 2015.

[7] Andrew Williams and Sabina Halappanavar. Application of bi-clustering of gene expression data and gene set enrichment analysis methods to identify potentially disease causing nanomaterials. *Data in brief*, 15:933–940, 2017.

[8] RStudio Team and others. Rstudio: integrated development for r. 42:14, 2015. URL `http://www.rstudio.com`.

[9] R Core Team. R: A language and environment for statistical computing, 2018. URL `https://www.R-project.org/`.

[10] Gerhard Schöfl. *reutils: Talk to the NCBI EUtils*, 2016. URL `https://CRAN.R-project.org/package=reutils`. R package version 0.2.3.

[11] Nancy L Guo, Ying-Wooi Wan, James Denvir, Dale W Porter, Maricica Pacurari, Michael G Wolfarth, Vincent Castranova, and Yong Qian. Multiwalled carbon nanotube-induced gene signatures in the mouse lung: potential predictive value for human lung cancer risk and prognosis. *Journal of Toxicology and Environmental Health, Part A*, 75(18):1129–1153, 2012.

[12] Julie A Bourdon, Anne T Saber, Nicklas R Jacobsen, Keld A Jensen, Anne M Madsen, Jacob S Lamson, Håkan Wallin, Peter Møller, Steffen Loft, Carole L Yauk, et al. Carbon black nanoparticle instillation induces sustained inflammation and genotoxicity in mouse lung and liver. *Particle and fibre toxicology*, 9(1):1–14, 2012.

[13] Mainul Husain, Anne T Saber, Charles Guo, Nicklas R Jacobsen, Keld A Jensen, Carole L Yauk, Andrew Williams, Ulla Vogel, Hakan Wallin, and Sabina Halappanavar. Pulmonary instillation of low doses of titanium dioxide nanoparticles in mice leads to particle retention and gene expression changes in the absence of inflammation. *Toxicology and applied pharmacology*, 269(3):250–262, 2013.

[14] Edilberto Bermudez, James B Mangum, Brian A Wong, Bahman Asgharian, Paul M Hext, David B Warheit, and Jeffrey I Everitt. Pulmonary responses of mice, rats, and hamsters to subchronic inhalation of ultrafine titanium dioxide particles. *Toxicological sciences*, 77(2): 347–357, 2004.

[15] Alison Elder, Robert Gelein, Jacob N Finkelstein, Kevin E Driscoll, Jack Harkema, and Gunter Oberdorster. Effects of subchronically inhaled carbon black in three species. i. retention kinetics, lung inflammation, and histopathology. *Toxicological Sciences*, 88(2):614–629, 2005.

[16] Ankit Rohatgi. Webplotdigitizer: Version 4.4, 2020. URL `https://automeris.io/WebPlotDigitizer`.

[17] Anne Thoustrup Saber, Keld Alstrup Jensen, Nicklas Raun Jacobsen, Renie Birkedal, Lone Mikkelsen, Peter Møller, Steffen Loft, Håkan Wallin, and Ulla Vogel. Inflammatory and genotoxic effects of nanoparticles designed for inclusion in paints and lacquers. *Nanotoxicology*, 6 (5):453–471, 2012.

[18] Dale W Porter, Ann F Hubbs, Robert R Mercer, Nianqiang Wu, Michael G Wolfarth, Krishnan Sriram, Stephen Leonard, Lori Battelli, Diane Schwegler-Berry, Sherry Friend, et al. Mouse pulmonary dose-and time course-responses induced by exposure to multi-walled carbon nanotubes. *Toxicology*, 269(2-3):136–147, 2010.

[19] Robert R Mercer, Ann F Hubbs, James F Scabilloni, Liying Wang, Lori A Battelli, Sherri Friend, Vincent Castranova, and Dale W Porter. Pulmonary fibrotic response to aspiration of multi-walled carbon nanotubes. *Particle and fibre toxicology*, 8(1):1–12, 2011.

[20] Julie A Bourdon, Sabina Halappanavar, Anne T Saber, Nicklas R Jacobsen, Andrew Williams, Håkan Wallin, Ulla Vogel, and Carole L Yauk. Hepatic and pulmonary toxicogenomic profiles in mice intratracheally instilled with carbon black nanoparticles reveal pulmonary inflammation, acute phase response, and alterations in lipid homeostasis. *Toxicological sciences*, 127(2):474–484, 2012.

[21] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[22] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1): D353–D361, 2017.

[23] Minoru Kanehisa, Yoko Sato, Miho Furumichi, Kanae Morishima, and Mao Tanabe. New approach for understanding genome variations in kegg. *Nucleic acids research*, 47(D1):D590–D595, 2019.

[24] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[25] Arthur Liberzon. A description of the molecular signatures database (msigdb) web site. In *Stem Cell Transcriptional Networks*, pages 153–160. Springer, 2014.

[26] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6):417–425, 2015.

[27] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[28] Sarah Labib, Andrew Williams, Carole L Yauk, Jake K Nikota, Håkan Wallin, Ulla Vogel, and Sabina Halappanavar. Nano-risk science: application of toxicogenomics in an adverse outcome pathway framework for risk assessment of multi-walled carbon nanotubes. *Particle and fibre toxicology*, 13(1):15, 2015.

[29] Friederike Ehrhart, Penny Nymark, Linda Rieswijk, Kristina Hanspers, Egon Willighagen, and Jonathan Melius. Lung fibrosis (homo sapiens), 2018. URL `https://www.wikipathways.org/index.php/Pathway:WP3624`.

[30] Marvin Martens, Ammar Ammar, Anders Riutta, Andra Waagmeester, Denise N Slenter, Kristina Hanspers, Ryan A. Miller, Daniela Digles, Elisson N Lopes, Friederike Ehrhart, Lauren J Dupuis, Laurent A Winckers, Susan L Coort, Egon L Willighagen, Chris T Evelo, Alexander R Pico, and Martina Kutmon. WikiPathways: connecting communities. *Nucleic Acids Research*, 49(D1):D613–D621, 11 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa1024. URL `https://doi.org/10.1093/nar/gkaa1024`.

[31] Thomas Kelder, Martijn P. van Iersel, Kristina Hanspers, Martina Kutmon, Bruce R. Conklin, Chris T. Evelo, and Alexander R. Pico. WikiPathways: building research communities on biological pathways. *Nucleic Acids Research*, 40(D1):D1301–D1307, 11 2011. ISSN 0305-1048. doi: 10.1093/nar/gkr1074. URL `https://doi.org/10.1093/nar/gkr1074`.

[32] Martina Kutmon, Anders Riutta, Nuno Nunes, Kristina Hanspers, Egon L. Willighagen, Anwesha Bohler, Jonathan Mélius, Andra Waagmeester, Sravanthi R. Sinha, Ryan Miller, Susan L. Coort, Elisa Cirillo, Bart Smeets, Chris T. Evelo, and Alexander R. Pico. WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Research*, 44(D1):D488–D494, 10 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv1024. URL `https://doi.org/10.1093/nar/gkv1024`.

[33] Denise N Slenter, Martina Kutmon, Kristina Hanspers, Anders Riutta, Jacob Windsor, Nuno Nunes, Jonathan Mélius, Elisa Cirillo, Susan L Coort, Daniela Digles, Friederike Ehrhart, Pieter Giesbertz, Marianthi Kalafati, Marvin Martens, Ryan Miller, Kozo Nishida, Linda Rieswijk, Andra Waagmeester, Lars M T Eijssen, Chris T Evelo, Alexander R Pico, and Egon L Willighagen. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, 46(D1):D661–D667, 11 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx1064. URL `https://doi.org/10.1093/nar/gkx1064`.

[34] David M Brass, Ivana V Yang, Marcus P Kennedy, Gregory S Whitehead, Holly Rutledge, Lauranell H Burch, and David A Schwartz. Fibroproliferation in lps-induced airway remodeling and bleomycin-induced fibrosis share common patterns of gene expression. *Immunogenetics*, 60(7):353–369, 2008.

Chapter 4

ALOHA R Packages and R Package Tutorials

The following chapter contains documentation of the R packages and their functions, which were part of the software development to execute the ALOHA method and perform comparative analyses. In addition to the R package documents there are also short tutorials of how to utilize the functions with other datasets for further analyses.

Package 'temposeqData'

April 19, 2021

Type Package

Title TempO-Seq Data Processing Functions

Version 0.1.1

Date 2021-01-13

Author Sarah E. Davidson B.S.

Maintainer Sarah E. Davidson B.S. <davidsss@mail.uc.edu>

Description This package contains functions to prepare TempO-Seq data for analysis.

License GPL (>= 2)

Encoding UTF-8

LazyData true

RoxygenNote 6.1.1

Suggests knitr,
rmarkdown

VignetteBuilder knitr

R topics documented:

`temposeqNorm` *Log2(Count per Million + 1) Normalization of TempO-Seq Data*

Description

Log2(Count per Million + 1) Normalization of TempO-Seq Data

Usage

```
temposeqNorm(rawcountdata)
```

Arguments

`rawcountdata`	Data-frame of raw read count data from TempO-Seq gene expression measurement.

Value

tl2dat - Normalized TempO-Seq data, log2 trasformed (count per million + 1) Data.

Index

'temposeqData' Vignette

Sarah E. Davidson B.S. (PhD Candidate)

2021-01-12

Objective

In this vignette we demonstrate how to normalize RAW gene expression count from the S1500+ TempOSeq platform devleoped by Mav et al. (2018) with count per million (CPM) normalization function in the `temposeqData` package. The goal is to normalize RAW transcript counts to $log_2(CPM + 1)$ scale.

Load the R Package

First thing first is to load the R package.

```
library(temposeqData)
```

Load the Data

Now that we have the R package let us get into our normalization example. Here, we will use an example dataset from Ramaiahgari et al. (2019) to demonstrate the normalization process.

Let us load in the example dataset with raw read counts obtained from the TempO-Seq platform.

```
data("bap-2d-run2-plate1")
```

We can do a quick investigation of the data structure. Since the dataset we are loading is a `data.frame` containing dose-response information we will evaluate the dimensions of the data and structure in the first 6 columns and first 6 rows.

```
dim(bap_2d_run2_plate1) # check the dimensions
#> [1] 2978   42
bap_2d_run2_plate1[1:6,1:6] # look at first 6 rows and first 6 columns
#>           X2D_RG_PLAT_1_102116_C23 X2D_RG_PLAT_1_102116_D23
#> Conc_nM                          0                        0
#> ABCC5_26                        77                      195
#> ABCG1_37                         2                        0
#> ADAT1_113                        2                       11
#> APOE_358                       366                     1004
#> ARNT2_448                       14                       38
#>           X2D_RG_PLAT_1_102116_E23 X2D_RG_PLAT_1_102116_F23
#> Conc_nM                          0                        0
#> ABCC5_26                        61                       96
#> ABCG1_37                         0                        0
#> ADAT1_113                        0                        9
#> APOE_358                       385                      558
#> ARNT2_448                        7                       62
```

```
#>           X2D_RG_PLAT_1_102116_G23 X2D_RG_PLAT_1_102116_H23
#> Conc_nM                          0                        0
#> ABCC5_26                        75                      215
#> ABCG1_37                         0                        1
#> ADAT1_113                        4                       16
#> APOE_358                       354                     1576
#> ARNT2_448                       26                       51
```

Data Preparation

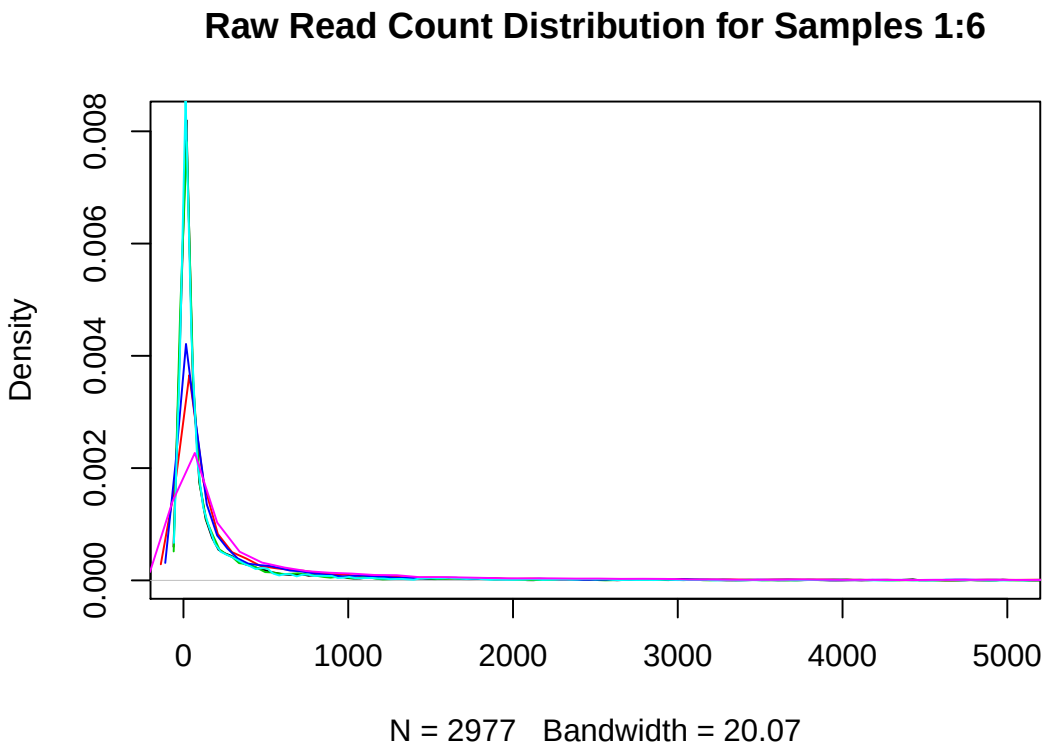
Now to normalize the raw read counts for each probe we need to first remove the row containing doses.

```
doses <- bap_2d_run2_plate1[1,] # store the first row with doses for later
raw_reads_only <- bap_2d_run2_plate1[-1,] # remove the first row
```

The following lines of code can be used to ensure the dose row is removed. However, we leave this as an exercise for the reader.

```
dim(raw_reads_only) # check the dimensions
raw_reads_only[1:6,1:6] # look at first 6 rows and first 6 columns
```

Evaluate a Raw Read Count Sample



CPM Normalization

Now that the dataset only contains raw read counts we can move on to the normalization step.

Raw read counts in samples are converted to count per million in the `temposeqNorm()` function using the following equation:

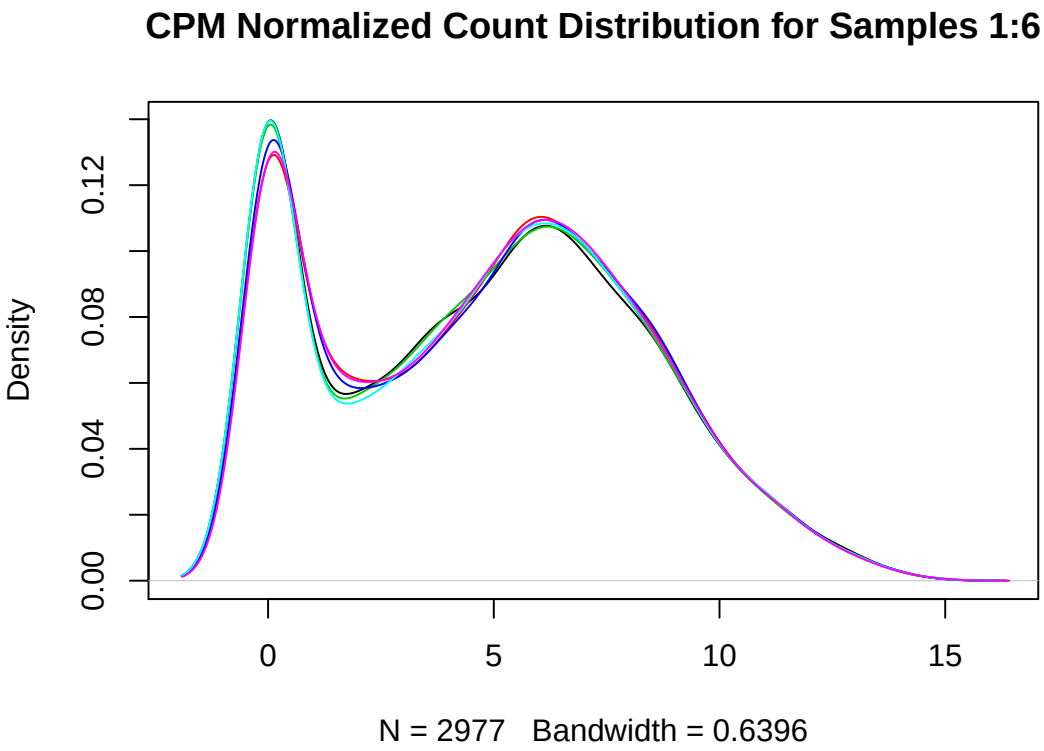
$$s_j^* = \frac{1 \times 10^6}{\{\sum_{i=1}^{n=2977} s_{ij}\}} \times s_j \tag{1}$$

where we assume s_j is the j^{th} sample, s_{ij} is the observed raw transcript read count for gene i in sample j, and s_j^* is the CPM normalized sample. After CPM normalization the function also adds 1 to the new counts and transforms it to a log_2 scale.

To perform this normalization for all samples in the dataset we use the following command from our package.

```
norm_reads_only <- temposeqNorm(raw_reads_only) # normalize data
```

Evaluate a Normalized Count Samples



That looks much better. Now that the samples are normalized to a similar distribution we can then put the dose-response dataset back together and save it to the desired directory using the `save()` or `write.table()` function.

All we need to do is add the doses object back to the dataset such that we have the same dose-response matrix form we had when we started.

```
norm_drdata <- rbind(doses,norm_reads_only) # combine the data back together
dim(norm_drdata) # check the dimensions
#> [1] 2978   42
```

Now the data is ready to be used in the analyses of your choice to answer your research question(s). Any further data processing is dependent upon the reader's own research objectives.

References

Mav, Deepak, Ruchir R Shah, Brian E Howard, Scott S Auerbach, Pierre R Bushel, Jennifer B Collins, David L Gerhold, et al. 2018. "A Hybrid Gene Selection Approach to Create the S1500+ Targeted Gene Sets for Use in High-Throughput Transcriptomics." *PloS One* 13 (2). Public Library of Science San Francisco, CA USA: e0191105.

Ramaiahgari, Sreenivasa C, Scott S Auerbach, Trey O Saddler, Julie R Rice, Paul E Dunlap, Nisha S Sipes, Michael J DeVito, et al. 2019. "The Power of Resolution: Contextualized Understanding of Biological Responses to Liver Injury Chemicals Using High-Throughput Transcriptomics and Benchmark Concentration Modeling." *Toxicological Sciences* 169 (2). Oxford University Press: 553–66.

Package 'lxsplineBMD'

April 19, 2021

Type Package

Title Benchmark Dose Estimation for Bayesian Local Extrema Spline Dose-Response Modeling Package

Version 0.1.2

Date 2021-01-13

Author Sarah E. Davidson

Maintainer Sarah E. Davidson <davidsss@mail.uc.edu>

Description Add-on package for 'lxsplines' package by Wheeler et al. (2017). Functions in this package are meant to evaluate BLX splines for dose-response modeling, convert into smooth curves, and estimate benchmark doses (BMDs) for biological responses (primarily gene express at this point).

License GPL (>= 2)

Imports lxsplines, scales, dplyr, stringr, magrittr, numbers, snow

Encoding UTF-8

LazyData true

RoxygenNote 6.1.1

Suggests knitr, rmarkdown

VignetteBuilder knitr

R topics documented:

`batchSAVE_fit` *Batch (Chunk) Lx-Spline Fit Object Save Function.*

Description

This function divides a large list of BLX spline fit objects into smaller lists and saves them into a specified directory. Sometime the list of BLX spline fits on multiple responses is too large to be parallelized over in an time or storage efficient manner.

Usage

```
batchSAVE_fit(fit_list, no_cores, strictlyLess = TRUE,
  out_dir = getwd())
```

Arguments

`fit_list`	List of lx-spline fit output object (i.e. lxfitS object) to be saved in chunks for batch conversion.
`no_cores`	The number of cores being used to parallelize over.

strictlyLess	Logical indicator if the maximum number of cores (and chunks) should be the largest divisor less than or equal to the number of cores specified. (Default = TRUE) (NOTE: FLASE => Total number of chunks should be equal to the number of specified cores.)
out_dir	The directory where the chunks of the lx-spline fit output are to be stored.

Value

Print statements about contaning information about the chunks of lx-spline fit objects.

`batch_bmdMCMC2.0` *Batch Processing of Estimating Postior Distribution of Benchmark Doses (BMDs) for Multiple Responses - Standard Deviation BMR.*

Description

This is a function estimates the BMD estimates for BLX spline fits on a list of multiple responses (genes) with a specified standard deveiation change from the control response as the benchmark response.

Usage

```
batch_bmdMCMC2.0(fit_list, lxEST_list, gene = NULL, ox, bmr = 1,
  anyBMD = T)
```

Arguments

fit_list	A list of BLX spline fit objects.
lxEST_list	A list of smoothed BLX spline fit objects.
gene	A vector of responses (genes) in the list to evaluate. (Default setting is 'NULL' and will evaluate all responses in the fit list).
ox	Sequence of numeric values for smoothing spline converted from (-0.5,0.5) to the original dose scale.
bmr	Specified standard deviation change from control to use as the benchmark response for estimating the BMD.
anyBMD	Logical indicator whether the BMD can be anywhere in the dose range (TRUE) or must be in the low dose range between dose=0 and than the first change point (FALSE). (Default setting is TRUE).
ignore_tilda_min	Logical indicator whether the first "bump" in the tilde ('~') shaped dose-response curves should be ignored. (Default setting is TRUE).

Details

Input for the fit and lxEST list should be the same. The 'ignore_tilda_min' option is not included here and it is assumed that the first bump of tilde-shaped is an artifact of BLX spline fitting and does not necessarily reflect biological responses.

Value

batch_BMD_est - A list of BLX BMD objects for specified genes.

See Also

'bmdMCMC2.1'

`batch_bmdMCMC2.0_foc` *Batch Processing of BMD Estimation with Fold-of-Control BMR.*

Description

This function estimate the posterior BMD distribution for BLX splines on a set of responses.

Usage

```
batch_bmdMCMC2.0_foc(fit_list, lxEST_list, gene = NULL, ox, foc = 2,
  anyBMD = T)
```

Arguments

`fit_list`	List of BLX spline fits on multiple responses.
`lxEST_list`	List of smoothed BLX spline fits on multiple responses.
`gene`	Single gene name (response) or list of specific genes.
`ox`	Sequence of numeric values for smoothing spline converted from (-0.5,0.5) to the original dose scale.
`foc`	Specified fold of control level to use as the benchmark response for estimating the BMD. (Default Fold-of-control setting is (FOC) 2).
`anyBMD`	Logical indicator whether the BMD can be anywhere in the dose range (TRUE) or must be in the low dose range between dose=0 and than the first change point (FALSE). (Default setting is TRUE).

Value

List of posterior BMD estimate objects for the set of responses.

batch_lxEST *Batch (Chunk) Spline to Polynomial Function.*

Description

Batch (Chunk) Spline to Polynomial Function.

Usage

```
batch_lxEST(fit_list, knots, tx)
```

Arguments

fit_list	The list of lx-spline fit output object (i.e. lxfitS object) being converted from splines to polynomial.
knots	The knot set (splines) specified lxfitS to obtain the fit.
tx	Sequence of numeric values from -0.5 to 0.5 (recommended seq(-0.5,0.5,length.out = N), where N is the number of values to be part of the sequence).

Value

lxEst - The matrix of average of polynomial fits across MCMC samples, from burn-in sample to last MCMC sample, for all elements in the batch set.

x.Pred - The list of matrices containing predicted polynomial fit for MCMC samples after burn-in sample, for all elements in the batch set.

knots - The knot set specified for the lxfitS to obtain the fit.

tx - Sequence of numeric values from -0.5 to 0.5 used for the conversion to the smooth polynomial.

---

blxbmd_hist *Bayesian Local Spline MCMC BMD Estimate Histogram Plotting Function.*

Description

Bayesian Local Spline MCMC BMD Estimate Histogram Plotting Function.

Usage

```
blxbmd_hist(bmdest, quantiles = c(0.05, 0.5, 0.95),
  up_color = "indianred1", down_color = "skyblue", q_color = "black",
  ...)
```

Arguments

bmdest	The bmd estimates for all MCMC samples after the burn-in period.
quantiles	Quantile values fof MCMC samples after the burn-in period (Default: c(0.05,0.5,0.95)).
up_color	Color for the upregulation bmd estimate histogram (Default: "indianred1").
down_color	Color for the downregulation bmd estimate histogram (Default: "skyblue").
...	Additional graphics arguments to be passed to the main plotting window.

Value

Generates a histogram of the posterior distribution of the up and down regulated BMD estimates.

---

blxbmd_hist2.0 *Bayesian Local Spline BMD Histogram Plotting Function for Up & Down Regulated BMDs.*

Description

This function generates histogram plots of the BMD posterior distributions.

Usage

```
blxbmd_hist2.0(bmdest, dose_range, id_name = NULL, quant = c(0.05,
  0.95), down_color = "skyblue", up_color = "indianred",
  q_color = "black", ...)
```

Arguments

bmdest	BLX bmd estimate object.
dose_range	The dose range of the original data.
id_name	Name to add to the title of the histogram plot.
quant	Quantiles for the BMDL and BMDU.
down_color	Color for the down-regulated BMD estimates.
up_color	Color for the up-regulated BMD estimates.
q_color	Color for the quantile lines.
...	Additional plotting parameters.

Value

A stacked plot of the down and up regulated BMD estimates.

`blxplot` *Bayesian Local Spline Plotting Function.*

Description

Bayesian Local Spline Plotting Function.

Usage

```
blxplot(x, y, fit, color = "red", ...)
```

Arguments

`x`	The original x variable input (independent variable, e.g. dose).
`y`	The original y variable input (dependent variable, e.g. gene response).
`fit`	The lxfitS output object.
`color`	Specified color to distinguish lx-spline curve fit from the original data (Default: color = "red").
`...`	Additional graphics arguments to be passed to the main plotting window.

Value

Plot of the original data and the lx-spline curve fit (in an alternative color).

`blxsmoothplot` *Bayesian Local Spline Plotting Function with Smoothed Spline.*

Description

This function produces a plot of observed response data and the smooth BLX spline DR curve.

Usage

```
blxsmoothplot(x, y, fit, fxfit, batch = F, gene = NULL,
  colors = c("red", "blue"), ...)
```

Arguments

`x`	The original x variable input (independent variable, e.g. dose).
`y`	The original y variable input (dependent variable, e.g. gene response).
`fit`	The lxfitS output object.
`fxfit`	The spline2poly output object, gives you the smoothed spline.
`...`	Additional graphics arguments to be passed to the main plotting window.
`color`	Specified color to distinguish lx-spline curve fit from the original data (Default: color = "red").

Value

Plot of the original data, the lx-spline curve fit (in an alternative color), and the smoothed curve spline fit.

`bmdF` *LX-spline BMD estimation function for continuous endpoints.*

Description

Benchmark dose estimation equation for continuous endpoints (Crump K.S., 1995) to be passed to bmd estimation functions utilizing uniroot.

Usage

```
bmdF(x, f, bmr, sd, increasing = TRUE)
```

Arguments

`x`	A numeric value corresponding to a possible dose value in the dose range being tested.
`f`	The splinefun object transforming polynomial prediction to smoothed spline function/equation.
`bmr`	Specified benchmark response.
`sd`	The standard deviation of samples for BMD estimation, more specifically sqrt(1/tau).
`increasing`	Logical argument if the dose-response trend is initially increasing or decresing. (Default setting is TRUE, increasing.)

Value

The difference between the estimated response at the benchmark dose and the proposed dose value.

`bmdF_foc` *LX-spline BMD estimation function for continuous endpoints.*

Description

Benchmark dose estimation equation for continuous endpoints using fold of control (FOC) to be passed to bmd estimation functions utilizing uniroot.

Usage

```
bmdF_foc(x, f, foc, increasing = TRUE)
```

Arguments

x	A numeric value corresponding to a possible dose value in the dose range being tested.
f	The splinefun object transforming polynomial prediction to smoothed spline function/equation.
foc	Specified fold of control (FOC).
increasing	Logical argument if the dose-response trend is initially increasing or decresing. (Default setting is TRUE, increasing.)

Value

The difference between the estimated response at the benchmark dose and the proposed dose value.

`bmdMCMC2.0` *(OLDER VERSION NOT RECOMMENDED)Standard Deviation BMR Benchmark Dose (BMD) Estimation for all posterior fit samples from the MCMC.*

Description

This is a function estimates the BMD estimates for BLX spline fits with a specified standard deveiation change from the control response as the benchmark response.

Usage

```
bmdMCMC2.0(fit, lxEST, ox, bmr = 1, anyBMD = T)
```

Arguments

fit	BLX spline fit object.
lxEST	Smoothed BLX spline fit object.
ox	Sequence of numeric values for smoothing spline converted from (-0.5,0.5) to the original dose scale.
bmr	Specified standard deviation change from control to use as the benchmark response for estimating the BMD.
anyBMD	Logical indicator whether the BMD can be anywhere in the dose range (TRUE) or must be in the low dose range between dose=0 and than the first change point (FALSE). (Default setting is TRUE).

Value

BMD_est - Matrix of numeric values containing the down and up regulated BMD estimate (columns) for each posterior sample (rows).

bmdMCMC2.1 *Standard Deviation BMR Benchmark Dose (BMD) Estimation for all posterior fit samples from the MCMC.*

Description

This is a function estimates the BMD estimates for BLX spline fits with a specified standard deveiation change from the control response as the benchmark response.

Usage

```
bmdMCMC2.1(fit, lxEST, ox, bmr, anyBMD = T, ignore_tilda_min = T)
```

Arguments

fit	BLX spline fit object.
lxEST	Smoothed BLX spline fit object.
ox	Sequence of numeric values for smoothing spline converted from (-0.5,0.5) to the original dose scale.
bmr	Specified standard deviation change from control to use as the benchmark response for estimating the BMD.
anyBMD	Logical indicator whether the BMD can be anywhere in the dose range (TRUE) or must be in the low dose range between dose=0 and than the first change point (FALSE). (Default setting is TRUE).
ignore_tilda_min	Logical indicator whether the first "bump" in the tilde (‘~‘) shaped dose-response curves should be ignored. (Default setting is TRUE).

Value

bmd - Matrix of numeric values containing the down and up regulated BMD estimate (columns) for each posterior sample (rows).

fd - Numeric vector of the fold of control benchmark responses for each posterior sample.

f0 - Numeric vector of the background response (i.e. response in the control group).

`bmdMCMC2.1_foc` *Fold of Control (FOC) Benchmark Dose (BMD) Estimation for all posterior fit samples from the MCMC.*

Description

This is a function estimates the BMD estimates for BLX spline fits with a specified fold-of-control as the benchmark response.

Usage

```
bmdMCMC2.1_foc(fit, lxEST, ox, foc, anyBMD = T, ignore_tilda_min = T)
```

Arguments

`fit`	BLX spline fit object.
`lxEST`	Smoothed BLX spline fit object.
`ox`	Sequence of numeric values for smoothing spline converted from (-0.5,0.5) to the original dose scale.
`foc`	Specified fold of control level to use as the benchmark response for estimating the BMD.
`anyBMD`	Logical indicator whether the BMD can be anywhere in the dose range (TRUE) or must be in the low dose range between dose=0 and than the first change point (FALSE). (Default setting is TRUE).
`ignore_tilda_min`	Logical indicator whether the first "bump" in the tilde (‘~‘) shaped dose-response curves should be ignored. (Default setting is TRUE).

Value

bmd - Matrix of numeric values containing the down and up regulated BMD estimate (columns) for each posterior sample (rows).

fd - Numeric vector of the fold of control benchmark responses for each posterior sample.

f0 - Numeric vector of the background response (i.e. response in the control group).

`both_bmd` *Standard Deviation Benchmark Dose (BMD) Estimation for a single fit (KEEP TILDE MIN).*

Description

This function estimates the up and down regulation benchmark doses for a single fit (e.g. single BLX spline MCMC posterior fit sample). For tilde-shaped dose-respones curves the first bump in the low dose region is not ignored.

Usage

```
both_bmd(fd, x, f, bmr, sd, a1, a2, aext, anyBMD)
```

Arguments

`fd`	The fold of control benchmark response (BMR).
`x`	Sequence of points in the dose range used to estimate the BMD.
`f`	BLX spline function for the dose-response curve.
`bmr`	The specified standard deviation change from control response.
`sd`	Standard deviation of response.
`a1`	Changepoint parameter alpha 1.
`a2`	Changepoint parameter alpha 2.
`aext`	The number of changepoint parameters for the BLX spline.
`anyBMD`	Logical indicator of whether the BMD can be anywhere in the dose range (TRUE) or must be in the low dose range between dose=0 and than the first change point (FALSE).

Value

bmd - Vector of numeric values for down and up regulated BMD estimates.

fd - Numeric value of the fold of control benchmark response.

f0 - Numeric value of the background response (i.e. response in the control group).

`both_bmd2.0` *Standard Deviation Benchmark Dose (BMD) Estimation for a single fit (IGNORE TILDE MIN).*

Description

This function estimates the up and down regulation benchmark doses for a single fit (e.g. single BLX spline MCMC posterior fit sample). For tilde-shaped dose-respones curves the first bump in the low dose region is ignored as an artifact of modeling.

Usage

```
both_bmd2.0(fd, x, f, bmr, sd, a1, a2, aext, anyBMD)
```

Arguments

`fd`	The fold of control benchmark response (BMR).
`x`	Sequence of points in the dose range used to estimate the BMD.
`f`	BLX spline function for the dose-response curve.
`bmr`	The specified standard deviation change from control response.
`sd`	Standard deviation of response.
`a1`	Changepoint parameter alpha 1.
`a2`	Changepoint parameter alpha 2.
`aext`	The number of changepoint parameters for the BLX spline.
`anyBMD`	Logical indicator of whether the BMD can be anywhere in the dose range (TRUE) or must be in the low dose range between dose=0 and than the first change point (FALSE).

Value

bmd - Vector of numeric values for down and up regulated BMD estimates.

fd - Numeric value of the fold of control benchmark response.

f0 - Numeric value of the background response (i.e. response in the control group).

`both_bmd2.0_foc` *Fold of Control (FOC) Benchmark Dose (BMD) Estimation for a single fit (IGNORE TILDE MIN).*

Description

This function estimates the up and down regulation benchmark doses for a single fit (e.g. single BLX spline MCMC posterior fit sample). For tilde-shaped dose-respones curves the first bump in the low dose region is ignored an artifact of model fitting.

Usage

```
both_bmd2.0_foc(fd, x, f, foc, a1, a2, aext, anyBMD)
```

Arguments

`fd`	The fold of control benchmark response (BMR).
`x`	Sequence of points in the dose range used to estimate the BMD.
`f`	BLX spline function for the dose-response curve.
`foc`	The specified fold of control change from the control response.
`a1`	Changepoint parameter alpha 1.
`a2`	Changepoint parameter alpha 2.
`aext`	The number of changepoint parameters for the BLX spline.
`anyBMD`	Logical indicator of whether the BMD can be anywhere in the dose range (TRUE) or must be in the low dose range between dose=0 and than the first change point (FALSE).

Value

bmd - Vector of numeric values for down and up regulated BMD estimates.

fd - Numeric value of the fold of control benchmark response.

f0 - Numeric value of the background response (i.e. response in the control group).

`both_bmd_foc` *Fold of Control (FOC) Benchmark Dose (BMD) Estimation for a single fit (KEEP TILDE MIN).*

Description

This function estimates the up and down regulation benchmark doses for a single fit (e.g. single BLX spline MCMC posterior fit sample). For tilde-shaped dose-respones curves the first bump in the low dose region is not ignored.

Usage

```
both_bmd_foc(fd, x, f, foc, a1, a2, aext, anyBMD)
```

Arguments

fd	The fold of control benchmark response (BMR).
x	Sequence of points in the dose range used to estimate the BMD.
f	BLX spline function for the dose-response curve.
foc	The specified fold of control change from the control response.
a1	Changepoint parameter alpha 1.
a2	Changepoint parameter alpha 2.
aext	The number of changepoint parameters for the BLX spline.
anyBMD	Logical indicator of whether the BMD can be anywhere in the dose range (TRUE) or must be in the low dose range between dose=0 and than the first change point (FALSE).

Value

bmd - Vector of numeric values for down and up regulated BMD estimates.

fd - Numeric value of the fold of control benchmark response.

f0 - Numeric value of the background response (i.e. response in the control group).

---

constant_set *Shape Constant Set Function.*

Description

Shape Constant Set Function.

Usage

```
constant_set(fit, extrema)
```

Arguments

fit	The lx-spline fit output object (i.e. lxfitS object).
extrema	The number of extrema allowed in the lxspline fit (i.e. number of possible change points).

Value

M - The multiplicative constant to convert the spline to a polynomial.

degree_find *Spline Degree Finder Function*

Description

Spline Degree Finder Function

Usage

```
degree_find(fit)
```

Arguments

fit	The lx-spline fit output object (i.e. lxfitS object).

Value

Degree of the spline to be used to convert the spline to a polynomial. Called by lxEST() and batch_lxEST().

lxEST *LX-spline function estimation*

Description

LX-spline function estimation

Usage

```
lxEST(fit, knots, tx)
```

Arguments

fit	The lx-spline fit output object (i.e. lxfitS object).
knots	The knot set (splines) specified lxfitS to obtain the fit.
tx	Sequence of numeric values from -0.5 to 0.5 (recommended seq(-0.5,0.5,length.out = N), where N is the number of values to be part of the sequence).

Value

lxEst - Average of polynomial fits across MCMC samples, from burn-in sample to last MCMC sample.

x.Pred - The predicted polynomial fit for MCMC samples after burn-in sample.

knots - The knot set specified for the lxfitS to obtain the fit.

tx - Sequence of numeric values from -0.5 to 0.5 used for the conversion to the smooth polynomial.

NRVar_filter_list *Non-reponse Variability Filter List Generator Function.*

Description

This is a function to find probes/genes that are un-responsive due to read-depth limitations, where many of the dose groups all of their responses are zero-read counts, for a given dose group for TempO-Seq data.

Usage

```
NRVar_filter_list(tdrData)
```

Arguments

tdrData	Column form of the dose-response data (dose and probes in columns, replicates in rows).

Value

toss - Character vector of probes/genes that are un-reponsive due to read depth limitations.

keep - Character vector of probes/genes that pass minimum set read depth requirements.

zeros_threshold - Numeric value of number of 0 reads are allowable for a probe (i.e. total number of replicates minus number of dose groups).

tt - Matrix of criterion variables. For one probe/gene (i.e. row); mm = overall median value of reads, vv = overall variability of reads, zz = number of zero reads.

predShape1.0 *Shape Prediction Function.*

Description

Shape Prediction Function.

Usage

```
predShape1.0(fit, shapeBFout)
```

Arguments

fit	The lx-spline fit output object (i.e. lxfitS object).
shapeBFout	The output object from the bayes' factor shape calculation function.

Value

predshape - The predicted curve shape based off of the bayes' factor shape results.

predShape1.2 *Shape Prediction Function for BLX Spline Dose-Response Curves 1.2.*

Description

This function predicts the shape of the dose-response curve using the BLX spline fit object and the Bayes Factors testing shape options.

Usage

```
predShape1.2(fit, shapeBFout)
```

Arguments

fit	BLX spline fit object (i.e. lxfitS object).
shapeBFout	Shape Bayes Factor Output object.

Value

predshape - The predicted curve shape based off of the bayes' factor shape results.

quick_lxEST *Quick LX-spline Conversion to Smooth Curve and estimation of quantiles.*

Description

Quick LX-spline Conversion to Smooth Curve and estimation of quantiles.

Usage

```
quick_lxEST(fit, knots, no.points = 200, quantiles = c(0.05, 0.25, 0.5,
  0.75, 0.95))
```

Arguments

fit	The lx-spline fit object (i.e. output from lxfitS).
no.points	The number of points to consider in the dose (x-value) range.
quantiles	The quantile values to use for the posterior distribution estimates.
knot	The spline knot values used for fitting the BLX spline curve.

Value

lxEST_quant - The smooth lxEST fit quantile bounds.

knots - The spline knots used for fittilg the BLX spline curve.

tx - The sequence of numbers used to estimate the smooth curve.

`sa1plot` *Posterior Trace Plotting Function for the 'a1' Changepoint Parameters.*

Description

This function generates a trace plot of the a1 BLX spline changepoint parameter over all MCMC iterations after burn-in.

Usage

```
sa1plot(fit, col = "red", ...)
```

Arguments

`fit`	The BLX spline fit object.
`col`	The color for the a1 changepoint parameter. (Default setting is "red" for a1.)
`...`	Additional plotting arguments for the abline.

Details

The plot generated from this function can be used to evaluate the convergence of the BLX spline and the location of the a1 changepoint parameter.

Value

Generates a trace plot for the a1 changepoint parameter.

`sa2plot` *Posterior Trace Plotting Function for the 'a2' Changepoint Parameters.*

Description

This function generates a trace plot of the a2 BLX spline changepoint parameter over all MCMC iterations after burn-in.

Usage

```
sa2plot(fit, col = "red", ...)
```

Arguments

`fit`	The BLX spline fit object.
`col`	The color for the a2 changepoint parameter. (Default setting is "red" for a2.)
`...`	Additional plotting arguments for the abline.

Details

The plot generated from this function can be used to evaluate the convergence of the BLX spline and the location of the a2 changepoint parameter.

Value

Generates a trace plot for the a2 changepoint parameter.

---

sa_both_plot *Posterior Trace Plotting Function for Changepoint Parameters.*

Description

This function generates a trace plot of the BLX spline changepoint parameters over all MCMC iterations after burn-in.

Usage

```
sa_both_plot(fit, sa_cols = c("red", "blue"), line_col = "black", ...)
```

Arguments

`fit`	The BLX spline fit object.
`sa_cols`	Vector of colors to indicate the changepoint parameters (a1 and a2). (Default setting is "red" for a1 and "blue" for a2.)
`line_col`	The color of the lines for the knot locations. Here the knot locations are assumed to be at -0.5 and 0.5.
`...`	Additional plotting arguments for the abline.

Details

The plot generated from this function can be used to evaluate the convergence of the BLX spline and the location of the changepoint parameters.

Value

Generates a trace plot with both changepoint parameters.

`shapeBF` *Bayes Factor Calculation for Possible Dose-Response Curve Shapes Function.*

Description

Bayes Factor Calculation for Possible Dose-Response Curve Shapes Function.

Usage

```
shapeBF(fit)
```

Arguments

`fit`	The lx-spline fit output object (i.e. lxfitS object).

Value

BF_jk - The vector of resulting Bayes' Factors calculated for each shape comparison.

jkMat - The matrix of model comparison tests for reference.

evidtest - Logical matrix if the strength of evidence is sufficient for model HA or HB.

`shapeBF1.0` *Shape Bayes Factor Calculation Function 1.0.*

Description

Shape Bayes Factor Calculation Function 1.0.

Usage

```
shapeBF1.0(fit, BFsiglevel = 3.2)
```

Arguments

`fit`	BLX spline fit object (i.e. lxfitS object).
`BFsiglevel`	Significant Bayes Factor Level. (Default: The default value is 3.2, based on Jeffereys, H. (1998) & Kass, R. E., & Raftery, A. E. (1995).)

Value

BF_jk - The calculated Bayes Factor values.

jkMat - The matrix of shape comparisons (Vales from 1:6, each indicating a single dose-response curve shape.)

evidtest - The evidence test matrix with detailed test information and logical information (T/F) if there is sufficient data for passing significance level.

spline2poly *LX-splines to Polynomial Model f(x).*

Description

LX-splines to Polynomial Model f(x).

Usage

```
spline2poly(fit, n, CONST, cbx)
```

Arguments

fit	The lx-spline fit output object (i.e. lxfitS object).
n	The number of testpoints to construct splines over dose range.
CONST	The constant to multiply with the transformed spline matrix.
cbx	Spline array on the values covering the range of -0.5 to 0.5.
iter	MCMC iteration to convert from splines to polynomial (i.e. f(x)).

Value

tp The polynomial constructed from the splines for all MCMC iterations after burn-in.

ZR_finder *Zero-reponse (zero-read count) Finder Function.*

Description

This is a function to find the zero-read counts for a given dose group for TempO-Seq data.

Usage

```
ZR_finder(tdrData, dose.group)
```

Arguments

tdrData	Column form of the dose-response data (dose and probes in columns, replicates in rows).
dose.group	Dose Group value of interest.

Value

Logical output (T or F) if all of the responses for the specified dose group are "0". (All responses = 0, then TRUE; o.w. FALSE).

`ZUR_filter_list` *Zero-finder for Unreponsive probe Filter List Generator Function.*

Description

This is a function to find probes/genes that are un-responsive due to read-depth limitations, where many of the dose groups all of their responses are zero-read counts, for a given dose group for TempO-Seq data.

Usage

```
ZUR_filter_list(dose.groups, tdrData, unresp.threshold = NULL,
  cytotox.mask = FALSE, cytotox.dose = NULL)
```

Arguments

`dose.groups`	Vector of all unique dose groups in dose-response dataset.
`tdrData`	Column form of the dose-response data (dose and probes in columns, replicates in rows).
`unresp.threshold`	Number of dose groups that if reached deems probe/gene as un-reponsive due to limited read-depth. [Default (NULL): less than or equal to 3 dose groups => threshold=1; o.w. even number of dose groups => threshold=dosegroups/2, odd number of dose groups => threshold=(dosegroups/2)+0.5.]

Value

toss - Character vector of probes/genes that are un-reponsive due to read depth limitations.

keep - Character vector of probes/genes that pass minimum set read depth requirements.

Index

lxsplineBMD

Sarah E. Davidson B.S. (PhD Candidate)

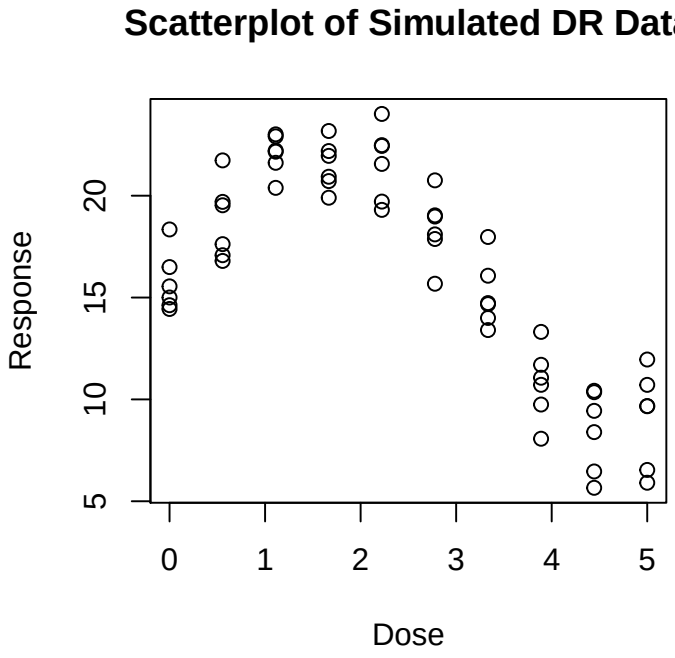
```
library(lxsplineBMD)
```

Objective

In this vignette we demonstrate how to estimate the a dose-response (DR) curve with BLX spline models developed by Wheeler, Dunson, and Herring (2017) specifying two possible changepoints ($H = 2$). Then to use the resulting curve estimate to estimate both the up- and down-regulated benchmark dose (BMD) estimates. Here, we will use a simulated DR data example for one response we would like to estimate. At the end we provide some comments on how this package can be applied to genomic DR data, and some guidance for gene expression datasets obtained from different technologies.

Data

```
set.seed(01132021)
x <- rep(seq(0,5,length.out = 10),each = 6) # dose
y <- 7*sin(0.3*pi*x) + 15 + rnorm(length(x),mean = 0,sd = 2) # response
```



BLX Spline Modeling

Now that we have our observed data we want to estimate it with the BLX spline.

Here, we will assume two changepoints and a knot set of size three located at the baseline, midpoint, and maximum of the dose range (i.e. $t = \{0, 0.5, 1\}$). Nonparametric models are sensitive to the number and placement of knots, which may influence the estimated dose-response curves. For dose-response modeling, we suggest the using the provided knot set or to use the following guidelines in :

1. The knot set size should be less than the total number of dose groups, but at least of size three.
2. There should be a knot located at the baseline (i.e. $dose = 0$) and the maximum dose.
3. Knots should be located between observed dose groups.
4. Knots should be equally-spaced in the dose range, unless if there is sufficient reason to have the knots located elsewhere for a more appropriate fit to the observed data.

We refer the reader to Wheeler, Dunson, and Herring (2017); Meyer and others (2008); and references therein for further details about shape-constrained splines and knot set selection.

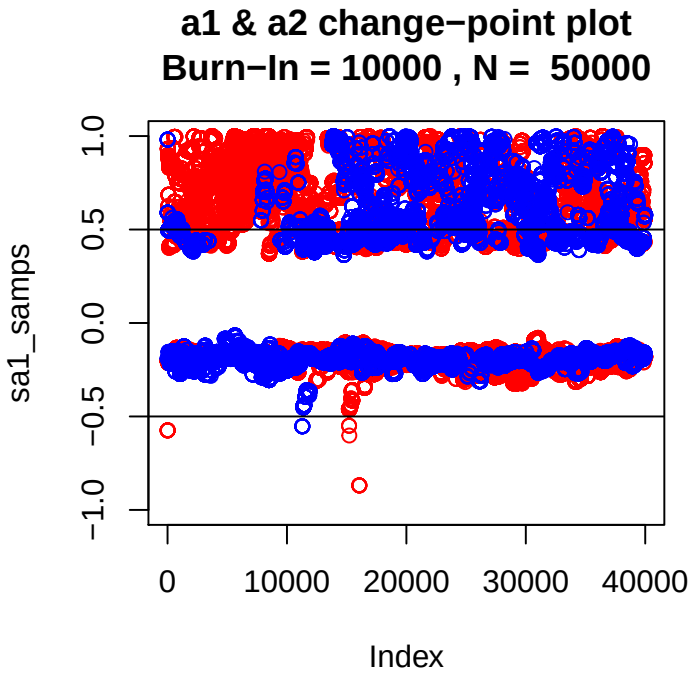
The following lines of code provide the set up and execution of BLX spline model fitting.

```
library(lxsplines) # load the lxsplines package
knotME <- c(0,0.5,1) # knot set

## fit blx spline with 2 extrema and assume initial increase in response ##
fitME <- lxfitS(x = x,y = y,splines = knotME,isIncrease = T,mEXTREMA = 2)
```

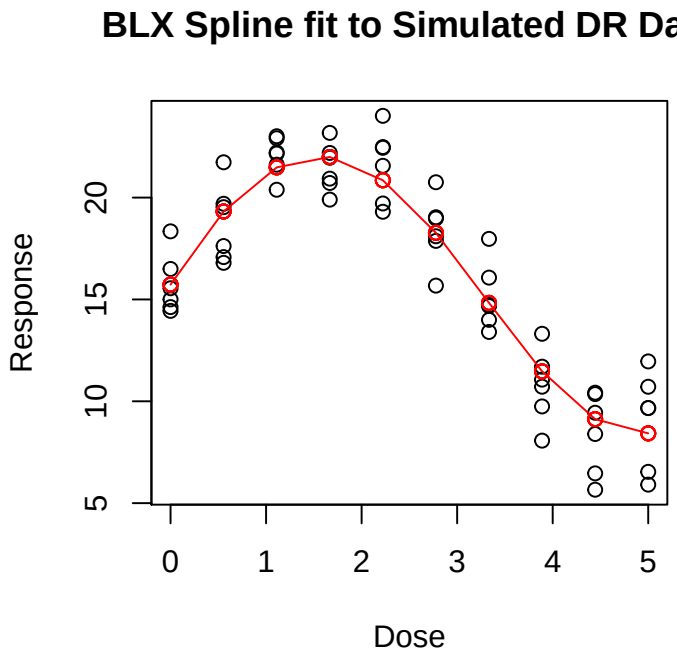
Evaluation of the posterior distribution of changepoint parameters can help us to determine convergence of the to ensure good mixing and convergence.

```
sa_both_plot(fit = fitME) # posterior samples of change point parameters after burn-in
```



The posterior samples indicate good mixing and convergence in the MCMC chain. So, now let us look at the curve fit to the data.

```
blxplot(x = x,y = y,fit = fitME,
        xlab = "Dose",ylab = "Response",
        main = "BLX Spline fit to Simulated DR Data")
```



BLX Spline to Polynomial Conversion

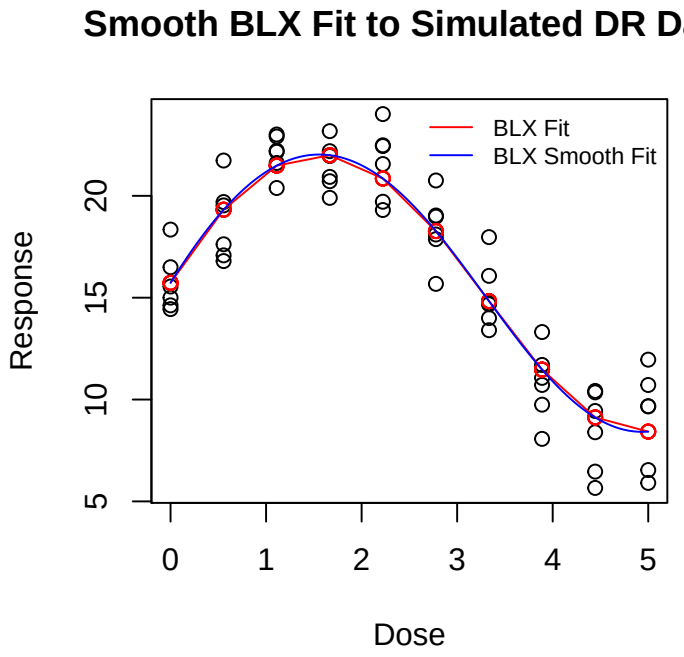
Now that we have the estimated DR curve we need to convert the current spline fit from estimates at each dose group to a smooth polynomial curve we can use to estimate the benchmark dose.

We will use 200 points in the dose region to estimate a smooth polynomial. For the conversion these points need to be on the scale of the model space, from $[-0.5, 0.5]$, **not** in the original dose-resposne scale.

```
tx   <- seq(-0.5,0.5,length.out = 200)
fxME <- lxEST(fit = fitME,knots = knotME,tx = tx)
```

The following plot show the BLX spline fit and the polynomial BLX fit overlayed on the simulated DR data.

```
blxsmoothplot(x = x,y = y,fit = fitME,fxfit = fxME,
              xlab = "Dose",ylab = "Response",
              main = "Smooth BLX Fit to Simulated DR Data")
legend(legend = c("BLX Fit","BLX Smooth Fit"),col = c("red","blue"),
       "topright",bty = "n",lty = "solid",cex = 0.75)
```



Benchmark Dose Estimation

The smoothed BLX curve estimate for our simulated data is ready to estimate the posterior distribution of BMD estimates for our simulated DR data.

First we need to transform the 200 points from the $[-0.5, 0.5]$ range to the original dose range. We are going to use the R package `scales` to efficiently convert our doses.

```
library(scales) # load the scaling package
ox <- rescale(tx,to = range(x),from = range(tx)) # transform dose scale for BMD estimation
```

We also need to specify our desired benchmark response (BMR). Here, we are use 1 standard deviation change in the baseline (i.e. control) response as our level of risk. Additionally, we need to specify whether it makes sense to allow the BMD estimate to be anywhere in the dose-respones range or to restrict it to be in the low-dose region (i.e. BMDs less than the location of the first changepoint).

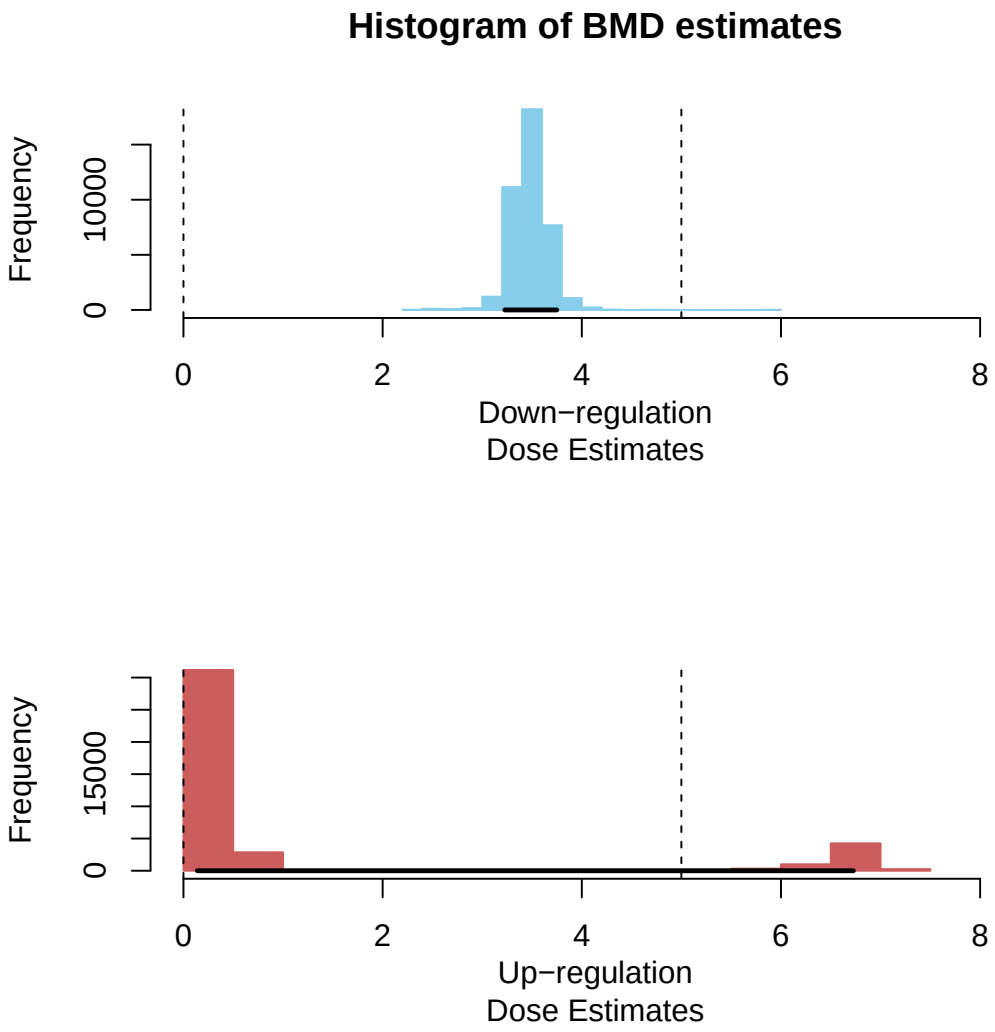
```
bmrME  <- 1 # 1 std. dev. change from control
anyBMD <- TRUE # allow the BMD to be anywhere
```

Now we can estimate the BMD for our simulated DR data.

```
bmdME <- bmdMCMC2.1(fit = fitME,lxEST = fxME,ox = ox,bmr = bmrME,anyBMD = anyBMD)
```

Let us look at the posterior distribution of BMD estimates.

```
blxbmd_hist2.0(bmdest = bmdME,dose_range = range(x))
```



We can see that the BMD for down-regulation is the most stable and reliable estimate. Similar results can then be used with additional supporting evidence to help inform the risk assessment at hand.

Genomic DR Application Examples

For genomic DR datasets, it is not advised to use the individual modeling response functions as this would be inefficient in terms of time. Instead the `lxsplineBMD` package contains functions for analyzing genes as a batch.

NOTE: For large datasets, we suggest the batch processing functions be coupled with parallel computing on a high performance computing (HPC) server for computational efficiency.

Below is an example dataset from Ramaiahgari et al. (2019) that we leave for the reader as an exercise.

```
## R Packages ##
library(lxsplines)
library(lxsplineBMD)
library(scales)
```

```
library(tictoc)
## Data ##
data("bap-2d-run1-plate1-normalized_log10dose")
dim(bap_2d_run1_plate1_norm_log10dose)
## Data Prep ##
dose    <- t(bap_2d_run1_plate1_norm_log10dose[1,])
tdrData <- t(bap_2d_run1_plate1_norm_log10dose)
## Minimum Read Count (MRC) Filtering ##
geneFilter <- NRVar_filter_list(tdrData = tdrData)

length(geneFilter$toss) # number of probes failing MRC criteria
length(geneFilter$keep) # number of probes passing MRC criteria
## Choose Subset of Passing Probes for Example ##
set.seed(01132021)
mygenes <- sample(geneFilter$keep,size = 10,replace = F)
## BLX Spline Fitting ##
# parameters #
knotME <- c(0,0.5,1) # knot locations
aext   <- 2 # number of extrema
dirME  <- TRUE
# fit BLX spline #
tic() # start timer
fitMEBatch <- lapply(mygenes, function(g){
  res <- try(lxfitS(x = dose,y = tdrData[,g],
                    splines = knotME,
                    isIncrease = dirME,
                    mEXTREMA = aext))
})
toc() # end timer and report
## Spline 2 Polynomial Conversion ##
# points set-up #
tx <- seq(-0.5,0.5,length.out = 200)
ox <- rescale(tx,to = range(dose),from = range(tx))
# BLX fit conversion #
tic() # start timer
fxMEBatch  <- batch_lxEST(fit_list = fitMEBatch,knots = knotME,tx = tx)
toc() # end timer and report
## BMD Estimation ##
# bmd estimation set-up #
bmr <- 1 # 1 std. dev. change from control
foc <- 1.5 # 1.5 fold of control change in expression
anyBMD <- TRUE # allow the BMD to be anywhere in the dose range
# estimate BMDs with Standard Deviation Change from Control #
tic() # start timer
bmdMEBatch <- batch_bmdMCMC2.0(fit_list = fitMEBatch,lxEST_list = fxMEBatch,
                               gene = mygenes,bmr = bmr,ox = ox,anyBMD = anyBMD)
toc() # end timer and report
# estimate BMDs with Fold-of-Control #
tic() # start timer
bmdMEBatch <- batch_bmdMCMC2.0_foc(fit_list = fitMEBatch,lxEST_list = fxMEBatch,
                                   gene = mygenes,foc = foc,ox = ox,anyBMD = anyBMD)
toc() # end timer and report
```

Additional Guidelines & Suggestions: When evaluating genomic DR datasets with BLX splines it is

important to ensure the expression values are in the positive model space to remove issues with the dispersion matrix and ensure a model fit to genes/probes passing MRC criteria. Normalized and log-transformed expression value spaces do not always meet the positivity standard and it may vary between different technologies. To mitigat these issues we provide the following guidance:

1. For TempOSeq or other transcript count platforms, with log_2 normalized expression values "floored" at 0, we suggest leaving the data as is.
2. For microarray platforms, with log_2 normalized expression values centered around 0, we suggest adding some constant value (e.g. 10) to all observed expressions such that there are no negative expression values.

For additional guidance we refer the reader to the literature, including Davidson et al. (2021); NTP and others (2018) and references therein.

Dependency Package Note: To obtain the `lxsplines` R package contact the developer Dr. Matthew Wheeler (matt.wheeler@nih.gov) for tarball and further details.

References

Davidson, Sarah E, Matthew W Wheeler, Scott S Auerbach, Siva Sivaganesan, and Mario Medvedovic. 2021. "ALOHA: Aggregated Local Extrema Splines for High-Throughput Dose-Response Data Analyses." Unpublished.

Meyer, Mary C, and others. 2008. "Inference Using Shape-Restricted Regression Splines." *The Annals of Applied Statistics* 2 (3). Institute of Mathematical Statistics: 1013–33.

NTP, National Toxicology Program, and others. 2018. "NTP Research Report on National Toxicology Program Approach to Genomic Dose-Response Modeling: Research Report 5." National Toxicology Program.

Ramaiahgari, Sreenivasa C, Scott S Auerbach, Trey O Saddler, Julie R Rice, Paul E Dunlap, Nisha S Sipes, Michael J DeVito, et al. 2019. "The Power of Resolution: Contextualized Understanding of Biological Responses to Liver Injury Chemicals Using High-Throughput Transcriptomics and Benchmark Concentration Modeling." *Toxicological Sciences* 169 (2). Oxford University Press: 553–66.

Wheeler, Matthew W, David B Dunson, and Amy H Herring. 2017. "Bayesian Local Extremum Splines." *Biometrika* 104 (4). Oxford University Press: 939–52.

Package 'clusterBLXspline'

April 19, 2021

Type Package

Title Bayesian Local Extrema Spline Models Clustering Package

Version 0.1.1

Author Sarah E. Davidson

Maintainer Sarah E. Davidson <davidsss@mail.uc.edu>

Description Package that contains functions to prepare data from bayesian local extrema spline model fitting to cluster together responses.

License GPL (>= 2)

Imports stringr, stringi, dplyr, magrittr

Encoding UTF-8

LazyData true

RoxygenNote 7.1.1

Suggests knitr, rmarkdown

VignetteBuilder knitr

R topics documented:

`alphasMCMC_avg` *Posterior mean of Change-point parameter (alpha) Samples*

Description

This function obtains the posterior mean estimates for both alpha parameters after discarding the burn-in samples.

Usage

```
alphasMCMC_avg(fit)
```

Arguments

`fit`	BLX spline fit object.

Value

aAVG - A vector of posterior mean estimates for changepoint parameters.

`avgBestParamFit_probes2df` *Best Parametric Fit Estimated CSIMM Data Frame Creation Function.*

Description

This function takes the original dose response data, estimated parameters from best model, and probe information to create a matrix for CSIMM clustering.

Usage

```
avgBestParamFit_probes2df(data, fit_params, probe_info, pid = "Probe",
  join_id = "Array.Probe", gid = Category.Component, gsep = ";",
  center_expr = F, filt = F)
```

Arguments

`data`	The original dose-response infromation set up by columns as Dose, Gene 1, Gene 2, ..., Gene n.
`fit_params`	The list of parameter estimates for the best parametric model fits.
`probe_info`	A matrix with probe and gene mapping information.
`pid`	Column name for 'Probe ID'. (Default is "Probe".)

join_id	Column name for joining fit parameter data and probe information matrix. (Default is "Array.Probe".)
gid	Column name for the EntrezID gene information used for Cluster Enrichment Analyses (CLEAN). (Default is Category.Component. NOTE: Do not quote this input variable.)
gsep	The character designating if there are multiple gene symbols or EntrezIDs for a given entry. (Default is ";".)
center_expr	Logical input indicating whether to center the expression distribution. (Default is FALSE.)
filt	Logical input indicating whether to filter out extraneous probe with either unknown Entrez/Gene IDs (listed as NA) and/or multiple. (Defaul is FALSE.)

Value

A data.frame with first two columns containing gene information (EntrezID and Gene/Probe identifier) and estimated responses from best parametric model for each experimental dose group to cluster genes with CSIMM.

---

avgBLX_probes2df *BLX Spline Fit Estimated CSIMM Data Frame Creation Function.*

Description

This function takes the original dose response data, BLX spline function, and probe information to create a matrix for CSIMM clustering.

Usage

```
avgBLX_probes2df(data, fit_fncs, probe_info, pid = "Probe",
  join_id = "Array.Probe", gid = Category.Component, gsep = ";",
  center_expr = F, filt = F, conds_probe_info = F)
```

Arguments

data	The original dose-response infromation set up by columns as Dose, Gene 1, Gene 2, ..., Gene n.
fit_fncs	The list of BLX spline fit function objects.
probe_info	A matrix with probe and gene mapping information.
pid	Column name for 'Probe ID'. (Default is "Probe".)
join_id	Column name for joining fit parameter data and probe information matrix. (Default is "Array.Probe".)
gid	Column name for the EntrezID gene information used for Cluster Enrichment Analyses (CLEAN). (Default is Category.Component. NOTE: Do not quote this input variable.)

gsep	The character designating if there are multiple gene symbols or EntrezIDs for a given entry. (Default is ";".)
center_expr	Logical input indicating whether to center the expression distribution. (Default is FALSE.)
filt	Logical input indicating whether to filter out extraneous probe with either unknown Entrez/Gene IDs (listed as NA) and/or multiple. (Defaul is FALSE.)
conds_probe_info	Logical inpur indicating whether to filter out probes/genes that are duplicated. (Default is FALSE.)

Value

A data.frame with first two columns containing gene information (EntrezID and Gene/Probe identifier) and estimated responses from BLX spline fit for each experimental dose group to cluster genes with CSIMM.

`avgDR_probes2df` *Original DR Data Average Response CSIMM Data Frame Creation Function.*

Description

This function takes the average response for each dose group from original dose response data and probe information to create a matrix for CSIMM clustering.

Usage

```
avgDR_probes2df(data, probe_info, pid = "Probe",
  join_id = "Array.Probe", did = Conc_nM, gid = Category.Component,
  gsep = ";", center_expr = F, filt = F, conds_probe_info = F)
```

Arguments

data	Transposed dose-response dataset. First column contains doses followed by columns of probe/gene reponses.
probe_info	Dataset containing mapping from probe to gene.
pid	Pattern for Probe level information (this should not be a case sensitive pattern).
join_id	Variable name for left-join of clustering data (left) and probe/gene mapping data (right).
did	Variable name for dose (should not be quoted - passed to dplyr processes).
gid	Variable name for gene level (should not be quoted - passed to dplyr processes).
gsep	Pattern separating multiple mappings for probe/gene information.
center_expr	Should the probe/gene response across all samples be centered? (Default == FALSE).
filt	Should the probe/gene responses be filtered to have only 1 probe:1 gene/component, and no Missing (NA) gene/component? (Default == FALSE)

Value

A data.frame with first two columns containing gene information (EntrezID and Gene/Probe identifier) and average responses from original samples in each experimental dose group to cluster genes with CSIMM.

`bap_2d_run1_plate1_norm_log10dose` *log2(CPM+1) Normalized S1500+ TempO-Seq Dataset for normalization demonstration.*

Description

Normalized and log-transformed transcription counts after exposure of liver cells to Benzo-[a]-pyrene at various doses. This dataset is of size 2978 by 42. Row 1 contains doses for each sample, probes in rows 2-2978, and columns are observed samples.

Usage

```
data("bap-2d-run1-plate1-normalized_log10dose")
```

Format

An object of class `"data.frame"`.

Source

NTP-CEBS Database

References

Ramaiahgari S.C.et al. (2019) Toxicological Sciences, Volume 169, Issue 2, Pages 553–566. (Pub Med)

`betasMCMC_avg` *Posterior mean of Spline Coefficient parameters (beta) Samples*

Description

This function obtains the posterior mean estimates for both beta parameters after discarding the burn-in samples.

Usage

```
betasMCMC_avg(fit)
```

Arguments

fit	BLX spline fit object.

Value

bAVG - A vector of posterior mean estimates for spline coefficient parameters.

---

blx_paramavg *BLX Spline Posterior Mean Estimate CSIMM Data Frame Creation Function.*

Description

This function takes the posterior mean estimates for BLX spline parameters and probe information to create a matrix for CSIMM clustering.

Usage

```
blx_paramavg(fits, probe_info, pid = "Probe", join_id = "Array.Probe",
  gid = Category.Component, gsep = ";", filt = F,
  conds_probe_info = F)
```

Arguments

fits	The list of BLX spline fit function objects.
probe_info	Dataset containing mapping from probe to gene.
pid	Pattern for Probe level information (this should not be a case sensitive pattern).
join_id	Variable name for left-join of clustering data (left) and probe/gene mapping data (right).
gid	Variable name for gene level (should not be quoted - passed to dplyr processes).
gsep	Pattern separating multiple mappings for probe/gene information.
filt	Should the probe/gene responses be filtered to have only 1 probe:1 gene/component, and no Missing (NA) gene/component? (Default == FALSE)
conds_probe_info	Logical inpur indicating whether to filter out probes/genes that are duplicated. (Default is FALSE.)

Value

parameter_matrix - The matrix containing posterior mean estimates for BLX spline parameters.

a_est - A character string indicating the matrix contains posterior mean changepoint estimates.

b_est - A character string indicating the matrix contains posterior mean spline coefficient estimates.

blx_paramdat *BLX Spline Posterior Sample Replicate CSIMM Data Frame Creation Function.*

Description

This function takes equally-spaced posterior distribution samples for BLX spline parameters and probe information to create a matrix for CSIMM clustering.

Usage

```
blx_paramdat(fits, probe_info, a_reps = 30, b_reps = 30, incr = T,
  pid = "Probe", join_id = "Array.Probe", gid = Category.Component,
  gsep = ";", filt = F, conds_probe_info = F)
```

Arguments

fits	The list of BLX spline fit function objects.
probe_info	Dataset containing mapping from probe to gene.
a_reps	The number of equally-spaced posterior samples. (Default is 30.)
b_reps	The number of equally-spaced posterior samples. (Default is 30.)
incr	Logical input whether to take the samples in an increasing order starting from burn-in samples (TRUE) or decreasing starting from the final samples (FALSE). (Default is TRUE.)
pid	Pattern for Probe level information (this should not be a case sensitive pattern).
join_id	Variable name for left-join of clustering data (left) and probe/gene mapping data (right).
gid	Variable name for gene level (should not be quoted - passed to dplyr processes).
gsep	Pattern separating multiple mappings for probe/gene information.
filt	Should the probe/gene responses be filtered to have only 1 probe:1 gene/component, and no Missing (NA) gene/component? (Default == FALSE)
conds_probe_info	Logical inpur indicating whether to filter out probes/genes that are duplicated. (Default is FALSE.)

Value

parameter_matrix - The matrix containing posterior mean estimates for BLX spline parameters.

a_samp_idx - The vector of posterior sample indicies for changepoint parameter.

b_samp_idx - The vector of posterior sample indicies for spline coefficient parameters.

`blx_paramdist` *BLX Spline Posterior Quantile Estimate CSIMM Data Frame Creation Function.*

Description

This function takes the quantiles for BLX spline parameters and probe information to create a matrix for CSIMM clustering.

Usage

```
blx_paramdist(fits, probe_info, a_quants = c(0.05, 0.25, 0.5, 0.75,
  0.95), b_quants = c(0.05, 0.25, 0.5, 0.75, 0.95), pid = "Probe",
  join_id = "Array.Probe", gid = Category.Component, gsep = ";",
  filt = F, conds_probe_info = F)
```

Arguments

`fits`	The list of BLX spline fit function objects.
`probe_info`	Dataset containing mapping from probe to gene.
`a_quants`	Quantiles for the posterior distribution estimates of BLX change-point parameters.
`b_quants`	Qunatiles for the posterior distribution estimates of BLX spline coefficient parameters.
`pid`	Pattern for Probe level information (this should not be a case sensitive pattern).
`join_id`	Variable name for left-join of clustering data (left) and probe/gene mapping data (right).
`gid`	Variable name for gene level (should not be quoted - passed to dplyr processes).
`gsep`	Pattern separating multiple mappings for probe/gene information.
`filt`	Should the probe/gene responses be filtered to have only 1 probe:1 gene/component, and no Missing (NA) gene/component? (Default == FALSE)
`conds_probe_info`	Logical inpur indicating whether to filter out probes/genes that are duplicated. (Default is FALSE.)

Value

parameter_matrix - The matrix containing posterior estimates for BLX spline parameters.

a_quants - The quantiles for the changepoint posterior estimates.

b_quants - The quantiles for the spline coefficient posterior estimates.

`condense_multiprobe2gene` *Multiple Probe Measurement Averaging (Condensation) Function.*

Description

This function takes a prepared CSIMM dataframe and condenses probes mapping to the same EntrezID or Category.Component by averaging their expressions.

Usage

```
condense_multiprobe2gene(csdata, probe_info,
  join_id = "Category.Component", gid = Category.Component,
  aid = NULL, center_expr = F)
```

Arguments

`csdata`	CSIMM dataframe.
`probe_info`	Dataset containing mapping from probe to gene.
`join_id`	Variable name for left-join of clustering data (left) and probe/gene mapping data (right).
`gid`	Variable name for gene level (should not be quoted - passed to dplyr processes).
`aid`	A placeholder variable to keep in the probe infromation dataframe. (Default is NULL, and the place holder will be indicies 1:nrows(dataset).)
`center_expr`	Logical variable whether the values in the resulting csdata dataframe.

Value

A data.frame with probes mapping to the same EntrezID or Category.Component averaged together.

`context_lenfind` *Context Length Finding Function*

Description

This function takes a BLX parameter data matrix and finds the context information for gene clustering using parameter values in CSIMM.

Usage

```
context_lenfind(blx_param_df)
```

Arguments

`blx_param_df`	CSIMM BLX parameter data matrix/frame. First two columns contain probe/gene information followed by response data.

Value

param_nContexts - The total number of contexts based on parameters.

param_contextLengths - The number of samples in each context.

`context_lenfind2` *Context Length Identification Function.*

Description

This function calculates the number of contexts and samples in each context to inform the CSIMM clustering algorithm.

Usage

```
context_lenfind2(data)
```

Arguments

`data`	BLX CSIMM prepared data (list or data.frame). First two columns contain probe/gene information followed by response data.

Value

nContexts - The total number of contexts.

contextLengths - The number of samples in each context.

`GMPC` *Gene Parametric Model Predicted Response Calculation.*

Description

This function take a vector of dose values (within the dose range) and best fit model parameter estimates to estimate the responses at the provided dose values.

Usage

```
GMPC(x, params)
```

Arguments

x	A numeric vector of dose values (within the dose range) for which responses will be estimated.
params	The parameter value matrix (nrow = 1) for the best fit model.

Value

A numeric vector of estimated responses for doses in vector 'x'.

---

origDR_probes2df *Original Dose Response Expression Data Prep for CSIMM clustering.*

Description

This function takes the orignal dose-response data frame and probe information to generate a dataframe for CSIMM gene clustering.

Usage

```
origDR_probes2df(data, probe_info, pid = "Probe",
  join_id = "Array.Probe", gid = Category.Component, gsep = ";",
  center_expr = F, filt = F, conds_probe_info = F)
```

Arguments

data	Transposed dose-response dataset. First column contains doses followed by columns of probe/gene reponses.
probe_info	Dataset containing mapping from probe to gene.
pid	Pattern for Probe level information (this should not be a case sensitive pattern).
join_id	Variable name for left-join of clustering data (left) and probe/gene mapping data (right).
gid	Variable name for gene level (should not be quoted - passed to dplyr processes).
gsep	Pattern separating multiple mappings for probe/gene information.
center_expr	Should the probe/gene response across all samples be centered? (Default == FALSE).
filt	Should the probe/gene responses be filtered to have only 1 probe:1 gene/component, and no Missing (NA) gene/component? (Default == FALSE)
conds_probe_info	Should duplicated probe/gene responses be removed? (Default == FALSE).
did	Variable name for dose (should not be quoted - passed to dplyr processes).

Value

A data.frame with first two columns containing gene information (EntrezID and Gene/Probe identifier) and normalized expression data to cluster genes with CSIMM.

`parameter_est_extraction` *Parameter estimate extraction function.*

Description

This function takes model fitting output from BMDExpress and extracts the parameter estimates for the best fitting model.

Usage

```
parameter_est_extraction(bmd_expr2_dat)
```

Arguments

`bmd_expr2_dat`	Matrix or Dataframe BMDExpress model fit output data. (Data must contain a variable "Probe ID" with gene information and "Best Model" and/or "Best Poly" variables containing the best model information.)

Value

tmodelest - The list of parameter estimates for the best fit model of each gene in the BMDExpress output provided.

Index

clusterBLXspline

Sarah E. Davidson B.S. (PhD Candidate)

```
library(clusterBLXspline)
```

Objective

In this tutorial we aim to show how to use the functions in the `clusterBLXspline` package to prepare BLX spline estimated response data for clustering with the the context-specific infinite mixture model (CSIMM) gene clustering algorithm developed by Liu et al. (2006). In addition to the BLX spline estimated response data we also provide an outline of how to utilize the other functions provided in the `clusterBLXspline` package to obtain a CSIMM input object with the observed normalized expression data, parametric model fit data, and BLX spline parameter data.

Load in the Data

For our tutorial we will use a dose-response dataset obtained by Ramaiahgari et al. (2019) as our example.

```
data("bap-2d-run1-plate1-normalized_log10dose")
```

BLX Spline Modeling

As mentioned in the objectives the primary example shown in this tutorial focuses on the use of clustering BLX estimated response data as input for CSIMM, which is the basis of the ALOHA method developed by Davidson et al. (2021). This section provides a baseline of obtaining BLX spline fits for each probe. If you are already familiar with this process we refer you to the next section.

First, we will set the example dataset to format into a CSIMM input data object. To do this we load the packages necessary for BLX spline modeling on probes, format the observed normalized expression data loaded previously.

```
## R Packages ##
library(lxsplines)
library(lxsplineBMD)
#>
#> Attaching package: 'lxsplineBMD'
#> The following object is masked _by_ '.GlobalEnv':
#>
#>     bap_2d_run1_plate1_norm_log10dose
#> The following object is masked from 'package:clusterBLXspline':
#>
#>     bap_2d_run1_plate1_norm_log10dose
## Data Prep ##
tdrData    <- t(bap_2d_run1_plate1_norm_log10dose)
geneFilter <- NRVar_filter_list(tdrData = tdrData)
```

Next, we are going to randomly select a subset of probes passing the minimum read criteria (n = 20) to serve as our demonstration.

```
## Subset ##
set.seed(01152021)
mygene     <- sample(x = geneFilter$keep,size = 20,replace = F)
mygene
#>  [1] "CDCA4_1183"    "AHR_23752"     "HOXA5_3080"    "GZMA_2853"
#>  [5] "MMADHC_10852"  "AQP1_26514"    "HPR_3094"      "SOX4_6686"
#>  [9] "HTR5A_24873"   "CHMP4A_1304"   "GDPD5_2628"    "LDHA_3766"
#> [13] "RAB1A_5660"    "GAPDH_2587"    "CPOX_28209"    "PMM2_5224"
#> [17] "COL1A1_1466"   "SMARCD2_6579"  "SLC30A3_15096" "DDX3Y_29022"
```

Once we have the subset of genes we have to get the BLX spline fits for each of the probes in out demonstration dataset. We leave this excercise for the reader.

```
## BLX Spline Fitting ##
knotME     <- c(0,0.5,1)
fitME      <- lapply(mygene,function(g){
  res <- try(lxfitS(x = tdrData[,1],y = tdrData[,g],
                    splines = knotME,
                    mEXTREMA = 2,
                    isIncrease = T))
  return(res)
})
names(fitME) <- mygene
```

Create the CSIMM Input Matrix

To create the matrix that will be the input data for the CSIMM gene clustering algorithm we first need to load the BLX spline fit data and the probe/gene mapping information.

```
data("mygene_fitME") # BLX spline fit file
data("Probe_File_Human_S1500+_Probe_to_Entrez_Gene") # probe to gene mapping
```

Second we need to create the BLX function list object.

```
# obtain the blx functions #
blxFitFnc <- lapply(fitME,function(g){splinefun(tdrData[,1],g$lmEst)})
```

Lastly, we can put together the CSIMM input matrix. We need to input the original dose-response dataset, the BLX fit function object, probe to EntrezID (i.e. Category Component) mapping, and logical statements for centering and filtering. Centering the data prior to clustering can help with convergence of CSIMM in obtaining gene clusters and help the overall clustering process. Filtering set to be `TRUE` is typically meant to filter out any promiscuous probes, i.e. probes mapping to multiple EntrezID's. In the random selection of probes there are not promiscuous probes so we will use `FALSE`.

```
avgBLXData <- clusterBLXspline::avgBLX_probes2df(
  data = tdrData, # dose-response data
  fit_fncs = blxFitFnc, # BLX spline fit data
  probe_info = probe2entrezID, # probe map
  center = T, # center responses
  filt = F)
```

You as the reader can view the resulting data using the following command.

```
View(avgBLXData)
```

The last piece of information necessary for performing the gene clustering with CSIMM is to obtain the context information. We do this using the 'context

```
avgBLXData_yc <- context_lenfind2(avgBLXData)
avgBLXData_yc
#> $nContexts
#> [1] 11
#>
#> $contextLengths
#>  [1] 1 1 1 1 1 1 1 1 1 1 1
```

Run CSIMM

We have all of the information we need to run the gene clustering using CSIMM. The following code illustrates the execution of the CSIMM clustering algorithm on the example data created in this tutorial. Here, we specify the MCMC (Markov Chain Monte Carlo) process to run for 50,000 iteration and discard the first 10,000 iterations in the chain as burn-in (or warm-up).

```
library(gimmR) # load in the gimmR package
myfilename <- "clusterBLX_demo_out"
csimm_out_avgBLXData <- gimmR::runGimmNPosthoc(
  tableData = avgBLXData, # data to cluster
  dataFile = myfilename, # base name for output files
  M = ncol(avgBLXData)-2, # number of sample columns
  # DO NOT include probe columns
  T = nrow(avgBLXData) # number of probes/genes to cluster
  nContexts = avgBLXData_yc$nContexts, # number of contexts
  contextLengths = avgBLXData_yc$contextLengths, # number of samples in each context
  burnIn = 10000,
  nIter = 50000)
```

We refer the reader to the *dependency package note* at the end of this tutorial and Liu et al. (2006) for more detailed information about the CSIMM clustering algorithm.

Other CSIMM Input Data Options

There are several other options for CSIMM input matrix options the following two examples provide the most common uses. (1) CSIMM input matrix for clustering on the normalized expression data. (2) CSIMM input matrix for clustering on the average normlized expression data for each context (or dose/concentration group).

```
# Original CSIMM Clustering Matrix by Dose Group #
clusterBLXspline::origDR_probes2df(data = tdrData,probe_info = probe2entrezID)
# CSIMM Clustering Matrix on Average Response per Dose Group #
avgDRData <- tdrData %>%
             as.data.frame() %>% # convert the matrix to a data.frame
             dplyr::group_by(Conc_nM) %>% # group by dose/concentration
             dplyr::summarise_each(funs = mean) %>% # obtain the mean for each column
             dplyr::ungroup() %>% # ungroup the data
             as.matrix() # convert back to a matrix
clusterBLXspline::avgDR_probes2df(data = avgDRData,probe_info = probe2entrezID)
```

Dependency Package Note: To obtain the `gimmR` package and additional information we refer the reader to gimmR webpage for download and UserManual_gimmR further details.

References

Davidson, Sarah E., Matthew W. Wheeler, Scott S. Auerbach, Siva Sivaganesan, and Mario Medvedovic. 2021. "ALOHA: Aggregated Local Extrema Splines for High-Throughput Dose-Response Analysis." *bioRxiv*. doi:10.1101/2021.03.29.437588.

Liu, Xiangdong, Siva Sivaganesan, Ka Yee Yeung, Junhai Guo, Roger Eugene Bumgarner, and Mario Medvedovic. 2006. "Context-Specific Infinite Mixtures for Clustering Gene Expression Profiles Across Diverse Microarray Dataset." *Bioinformatics* 22 (14). Oxford University Press: 1737–44.

Ramaiahgari, Sreenivasa C, Scott S Auerbach, Trey O Saddler, Julie R Rice, Paul E Dunlap, Nisha S Sipes, Michael J DeVito, et al. 2019. "The Power of Resolution: Contextualized Understanding of Biological Responses to Liver Injury Chemicals Using High-Throughput Transcriptomics and Benchmark Concentration Modeling." *Toxicological Sciences* 169 (2). Oxford University Press: 553–66.

Chapter 5

Conclusion

The goal of this dissertation was to develop a novel genomic dose-response (GDR) modeling approach which overcomes limitations in the current approach. Evaluation of genomic dose-response is currently limited by the reliance upon parametric models used in traditional risk assessment and the reliance on *a prior* gene sets to investigate the underlying molecular mechanisms. Current gene expression modeling approaches utilize parametric approaches which do not account for cases where cytotoxic doses induce non-monotonic curves (i.e. umbrella or J-shaped) Hsieh et al. (1). Additionally, one is not able to evaluate the underlying correlation structure of co-expressed genes. Here, we demonstrated the utility of an approach that leverages flexible modeling coupled with probabilistic gene clustering to identify co-expressed genes sharing a common dose-response pattern and estimate a benchmark dose linked with their underlying biology.

In Chapter 2, we discuss the developed GDR modeling approach Aggregated Local Extrema Splines for High-throughput Dose-response Analysis (ALOHA) and its comparison to currently established methods. ALOHA consists of two major components. First, individual genes are fit with a shaped-constrained non-parametric spline model, i.e. Bayesian local extrema splines (BLX) (2) with a fixed knot set, to evaluate the underlying dose-response trend. These fits are then used to estimate a gene-level benchmark dose estimate (BMD) and determined whether they are reliable for biological interpretation. Second, the individual BLX spline fits are clustered using context-specific infinite mixture models (CSIMM) (3) to probabilistically cluster genes having similar dose-response patterns. We showed the flexibility of dose-response modeling in ALOHA retains genes with non-monotonic trends while filtering out probes with little to no biological activity. Additionally, in cases where ALOHA outperforms BMDExpress (4, 5, 6) in the predictive ability of unobserved data, it does so by several orders of magnitude while remaining consistent in the gene level benchmark dose estimates. Comparing the clustering results of ALOHA with CSIMM we demonstrated our approach obtains gene clusters with more closely related dose-response patterns while maintaining a similar level of functional coherence. Coupling the flexible modeling with the evaluation of coordinated dose-response patterns in ALOHA allowed us to identify biologically relevant clusters (i.e. enriched for gene sets related to molecular processes linked with toxicity) while obtaining potency estimates similar to those from the currently established approach BMDExpress. Biological relevance of the resulting BMD estimates were shown by utilizing the following three criteria. First, there is a high level of functional coherence in gene clusters, i.e. genes grouped together also tend to be members of the same gene sets/pathways being evaluated (7). Second, the gene clusters are able to capture gene sets also enriched using standard enrichment methods. Finally, ALOHA produces similar if not more coherent gene set level BMDs from core genes. ALOHA unlike other current approaches more appropriately estimates the underlying change in gene expression with the use of more flexible models. ALOHA then leverages these fits to evaluate the coordinated dose-response patterns among genes to capture cellular processes related to adverse health outcomes and estimate an interim point-of-departure

for chemicals with little to no *in vivo* toxicological data.

In Chapter 3, we discuss the application of ALOHA to a set of publicly available gene expression data sets on lung tissue of mice after exposure to either multi-walled carbon nanotubes (MWCNT), carbon black (CB), and nanotitanium dioxide (TiO_2). Here, we compared the resulting enriched gene sets, and their corresponding BMD estimates, between the three materials to identify similarities and differences in perturbed biological pathways and level of potency. We were able to demonstrate clear differences in not only the potency of the three materials but also differences in the biological processes they perturbed. MWCNT was by far the most potent at both 1- and 28-days post-exposure having the smallest gene set level BMDs. CB was slightly more potent than TiO_2, but our results did not show a substantial difference in their potency. Additionally, we showed all three materials impact immune and signaling responses, but there were some key differences in the biological processes that were perturbed (e.g. enrichment of the p53 pathway only for CB at 28-days post-exposure). The biological pathways in the most biologically relevant and sensitive gene clusters were representative of key events in the development of adverse outcomes also found in the results of other toxicogenomic studies on these materials and/or were consistent with previously proposed adverse outcome pathways. Within the material, we were able to demonstrate the gene set level BMD estimates obtained from ALOHA were overlapped with the most robust *in vivo* point of departure (i.e. BMD/BMDL where the *in vivo* endpoint model fit was reliable – not saturated, hockey-stick shaped, unresponsive, or produces only estimates near zero – and LOAEL/NOAEL otherwise). Carbon black was the only material where the most sensitive gene set BMD/BMDL interval estimates were slightly higher for all but one of the *in vivo* interval estimates. This may be the result of a relatively small number of dose groups included, the spacing of experimental dose groups, and/or the resulting model fits to the data. Thus, these results show the applicability of our results to estimate a baseline point-of-departure and evaluate the underlying biological mechanisms that may contribute to the development of adverse health effects. These methods can be used in preliminary risk assessments to rapidly evaluate chemicals with little to no *in vivo* toxicology data, particularly when applying them to *in vitro* data. The results from these analyses may be used to estimate an initial point-of-departure, prioritize chemicals based on potency and activation of molecular mechanisms linked to adverse outcomes, and generate mechanistic hypotheses.

Overall, in this dissertation, ALOHA is shown to enable the simultaneous investigation of the correlation structure of co-expressed genes with similar dose-response patterns and leverage these genes to estimate cohesive benchmark dose estimates that are competitive with currently accepted approaches. This adds to the literature by proposing a new genomic dose-response modeling approach that accounts for the biological phenomenon at the gene expression level and may provide insights into molecular mechanisms. These contributions help support the overall objective of utilizing high-throughput data, such as gene expression, to develop preliminary guidance to prevent chemical exposures that lead to

adverse health effects in exposed individuals.

References

[1] Jui-Hua Hsieh, Alexander Sedykh, Ruili Huang, Menghang Xia, and Raymond R Tice. A data analysis pipeline accounting for artifacts in tox21 quantitative high-throughput screening assays. *Journal of biomolecular screening*, 20(7):887–897, 2015.

[2] MW Wheeler, DB Dunson, and AH Herring. Bayesian local extremum splines. *Biometrika*, 104(4):939–952, 2017.

[3] Xiangdong Liu, Siva Sivaganesan, Ka Yee Yeung, Junhai Guo, Roger Eugene Bumgarner, and Mario Medvedovic. Context-specific infinite mixtures for clustering gene expression profiles across diverse microarray dataset. *Bioinformatics*, 22(14):1737–1744, 2006.

[4] Scott S. Auerbach. BMDExpress 2.3, 2017. URL `https://github.com/auerbachs/BMDExpress-2/wiki`.

[5] Jason R Phillips, Daniel L Svoboda, Arpit Tandon, Shyam Patel, Alex Sedykh, Deepak Mav, Byron Kuo, Carole L Yauk, Longlong Yang, Russell S Thomas, et al. Bmdexpress 2: enhanced transcriptomic dose-response analysis workflow. *Bioinformatics*, 35(10):1780–1782, 2019.

[6] Longlong Yang, Bruce C Allen, and Russell S Thomas. Bmdexpress: a software tool for the benchmark dose analyses of genomic data. *BMC genomics*, 8(1):387, 2007.

[7] Johannes M Freudenberg, Vineet K Joshi, Zhen Hu, and Mario Medvedovic. Clean: Clustering enrichment analysis. *BMC bioinformatics*, 10(1): 234, 2009.