National Toxicology Program

Peer Review of Draft NTP Approach to Genomic Dose-Response Modeling Expert Panel Meeting

October 23-25, 2017

National Institute of Environmental Health Sciences

Research Triangle Park, NC

Peer Review Report

Table of Contents

I. Attendees

Members in Attendance:

Carole Yauk (Panel Chair)
Lyle Burgoon
Rebecca Clewell
Ruili Huang
Kamin Johnson
Jorge Naciff
Setia Pramana
James Stevens
Fred Wright

NTP Board of Scientific Counselors Representative:

Katrina Waters (via webcast)

National Institute of Environmental Health Sciences (NIEHS) Staff:

Scott Auerbach
Mamta Behl
John Bucher
Pierre Bushel
Michael DeVito
Stephen Ferguson
Alison Harrill
Grace Kissling
David Malarkey
Arun Pandiri
Fred Parham
Shyamal Peddada
Rick Paules
Kristen Ryan
Keith Shockley
Nisha Sipes
Stephanie Smith-Roe
Ray Tice
Molly Vallant
Kristine Witt
Mary Wolfe

Federal Agencies

Stephen Edwards, US EPA
David Gerhold, NIH/NCATS
Jeff Gift, US EPA
Joshua Harrill, US EPA
Johanna Nyfeller, US EPA
Woodrow Setzer, US EPA
Russell Thomas, US EPA

Contract Staff to NIEHS

Canden Byrd, ICF
Logan Everett, Sciome
Cara Henning, ICF
Deepak Mav, Sciome
Sandra McBride, SSS
Jason Phillips, Sciome
Alex Sedykh, Sciome
Ruchir Shah, Sciome
Marjo Smith, SSS
Anna Stamatogiannakis, ICF
Dan Svoboda, Sciome

Public Attendees

Sorin Draghici, Wayne State University
Ernie Hood, Bridport Services
John House, NCSU
Andrew Williams, Health Canada

Day 1: October 23, 2017

II. Welcome, Introductions, and Background Information

The Peer Review of the Draft NTP Approach to Genomic Dose-Response Modeling Expert Panel Meeting met October 23-25, 2017 in Rodbell Auditorium, National Institute of Environmental Health Sciences (NIEHS), Research Triangle Park, North Carolina. Dr. Carole Yauk served as chair. The other Peer Review Panel members in attendance were Drs. Lyle Burgoon, Rebecca Clewell, Ruili Huang, Kamin Johnson, Jorge Naciff, Setia Pramana, James Stevens, and Fred Wright. Dr. Katrina Waters attended by webcast as the NTP Board of Scientific Counselors liaison. Interested public attended the meeting in person or watched the proceedings via webcast.

Dr. Yauk welcomed everyone to the meeting and asked all attendees to introduce themselves. Dr. Bucher welcomed participants, thanked Dr. Yauk for chairing the meeting and thanked the board members and staff for their work. Designated Federal Officer Dr. Mary Wolfe read the conflict of interest statement and asked panel members to sign updated Conflict of Interest forms.

Dr. Yauk introduced the scientific background behind the meeting, in which the expert panel would scrutinize the proposed NTP approach, vetting each step and helping the NTP by making recommendations on adoption or improvement of specific aspects of its proposal. She also described the format of the three-day meeting, culminating in the development of recommendations, which the panel would vote on.

Dr. Auerbach presented background information on the NTP proposed approach, acknowledged the parties who contributed to its development, and went over the panel's charge for the meeting. The overall goal for the proposed approach is to develop a biologically comprehensive, efficient assessment of test articles that can be used to estimate biological potency and highlight associations between transcriptomic changes and potential toxicological effects. Primary uses include development of biological potency estimates that can be used to identify screening-level exposure limits. Secondary uses include identification of potential toxicological effects, although the approach is not intended for traditional hazard identification.

III. Session I: Approaches to Genomic Dose-Response Analysis

A. Genomic Dose-Response: The Big Picture

Dr. Russell (Rusty) Thomas, Director of the US EPA National Center for Computational Toxicology, briefed the panel on the history of toxicogenomics and how it may now be integrated in a tiered 21st century toxicity testing framework. An important rationale for the inclusion of transcriptomics in Tox21/ToxCast is to expand the biological coverage beyond that in Tox21 and ToxCast, in an efficient and cost-effective way. He described current studies looking at whether bioactivity can be used as a conservative estimate of Point of Departure (POD). More than 300 chemicals with an *in vivo* POD have been assessed, and in 87%, the *in vitro* POD was seen to be protective. With the new Toxic Substances Control Act (TSCA) regulations, there is an increased potential to integrate

in vitro conclusions into regulatory decision-making applications. However, it will be necessary to arrive at consensus of the appropriate analysis approaches to derive meaning from transcriptomic data.

A.1. Questions for Clarification

Dr. Gerhold said that when he was in the pharmaceutical industry, there was criticism of toxicogenomic tests due to their lack of predictivity for toxicities and cancer. He asked Dr. Thomas what his view is of cancer prediction studies and whether it would be worthwhile to go back and look at pharmacokinetics — more qualitatively than quantitatively. Dr. Thomas said that the problems he is addressing today are more directed at the possibility of using toxicogenomics to identify a protective dose, not necessarily qualitatively predicting what the adverse effect would be. He felt that extending the analysis beyond the protective dose to identify adverse effects is not an efficient use of resources, since many of the environmental chemicals are promiscuous enough to result in numerous effects. However, extending toxicogenomics to identify modes of action for a subset of environmental chemicals that are more selective would be worthwhile.

Dr. Stevens supported Dr. Thomas's comments about not always trying to detect mode of action. He asked if the desire is to find chemicals with low potential for toxicity rather than identifying what chemicals would do if toxic. Dr. Thomas replied that in the environmental world, most people are more worried about detecting all chemicals with the potential to be toxic, and genomics is a first-tier approach. He noted that mode of action understanding would be important in certain cases, such as developmental toxicants. Dr. Stevens said that if potency in gene sets is to be determined, then gene sets are being interpreted as a surrogate for biological responses. Potency estimates may eliminate the gene set from being considered "adverse" without specifying which adverse effects are avoided. Employing this approach depends on the specificity and negative predictive value of the model. Dr. Thomas agreed and commented that in environmental applications, the goal is protection (with a tendency toward more conservative conclusions about toxicity) rather than predictivity. The overall specificity from a dose level is certainly important, he added.

Dr. Johnson asked whether Dr. Thomas had meant to state that the method did not work for receptor-mediated toxicity. Dr. Thomas replied that he had not explored the issue in depth, but there is some evidence that it does work to some degree for receptor-mediated effects as well.

Dr. Clewell said she was concerned with discussion of no-effect level without any consideration of mode of action. She asked how it could be decided what is the appropriate model without considering the different types of toxicities to be predicted or protected from. How can the appropriate biological space be covered in the testing system? For example, is the liver a sentinel tissue for *every* toxicity? Dr. Thomas agreed that general biological activity would not yield the most sensitive effect on a dose level 100% of the time. Defining the biological space that needs to be interrogated is still an open research question. He said that the community should take some comfort from the data he presented about conservative POD with ToxCast and Tox21.

Dr. Clewell appreciated the comment. Dr. Bucher noted that NTP has a series of studies underway in animals looking at the liver as a sentinel organ.

Dr. Naciff said it appeared that there is an assumption that any biological change described as a transcriptional change elicited by chemical exposure is an adverse event, and he did not agree with that concept. Dr. Thomas agreed that not all transcriptional perturbations are adverse; however, a reasonable percentage of environmental chemicals go from perturbation of the system to adversity in a fairly narrow potency range. Trying to sort out the non-adverse perturbations from the adverse perturbations is probably not the appropriate use of resources. He noted that he was not saying that biological activity itself identifies adversity, but in trying to be protective, at least there is the ability to draw a line in the sand and say that a certain dose is likely to be protective.

Referring to the international case study Dr. Thomas had described, Dr. Wright asked if he felt that an intermediate future should be sought in which transcriptomics-based PODs correlate highly with traditional approaches. Dr. Thomas replied that success would be more likely with short-term *in vivo* studies, but that a reasonable goal over the next few years would be to show that using the *in vitro* approaches can be protective.

Dr. Auerbach asked Dr. Thomas for his perspective on biological interpretation that can be derived from the NTP gene set approach. Dr. Thomas said that there is a change in the regulators' mindsets, moving from being focused on false predictivity to being comfortable with protection. He said discussions with scientists and regulators to that end have already begun. It is applicable to prioritization, and longer term, potentially, to screening level risk assessment. When protection is accepted, then the discussion can begin about using mode of action to lead to biological interpretation. He felt that it should be a linear process, first focusing on protection and then on biological interpretation in a staged approach. Dr. Auerbach noted that NTP will begin issuing 5-day reports with a gene set, with a lower confidence limit on the benchmark dose (BMD_L) and an upper confidence limit on the BMD (BMD_U), associated with a GO term; he asked whether the GO term should not be reported and instead the conclusions be presented based on "Gene Set 1" or "Gene Set 2". Dr. Thomas said that getting the regulators accustomed to the concept that the most sensitive pathway could be a surrogate of a conservative POD would be the first goal. Then, as knowledge and ability to interpret the meaning of the gene ontology (GO) biological process advance, interpretation will go forward. Thus, the regulators would gain the ability to see the biology that underlies the gene expression changes.

Dr. Draghici asked Dr. Thomas what he meant by "pathways." Referring to his slide labeled "Combined Correlation Between Cancer and Transcriptional PODs," he noted that in the case of the data shown, the lowest transcriptional benchmark dose (BMD) was at a pathway level. He said the data could be analyzed with several different pathway maps with similar results in terms of the relationship. Dr. Draghici asked if the aggregation shown illustrated the relationship between the genes or if it were just a combination of p values or fold changes for the genes in a set. Dr. Thomas explained that the data illustrated were based on having at least five genes that have a

transcriptional POD, that are perturbed in that pathway, with no enrichment or other statistical requirement, and showed the median BMD for that particular pathway aggregation.

B. Overview of the NC State Approach to Genomic Dose-Response Modeling

Dr. Fred Wright, the director of the NC State Bioinformatics Research Center, described the NC State approach for the panel.

In terms of statistical procedures, he noted the following for quality control in sequence-based transcriptomic technologies:

- Threshold individual genes based on expression level [removing transcripts with low counts or signal],
- Perform outlier checks, and
- Compare control samples to all other control samples.

Normalization is currently done per-experiment, e.g., using DESeq2 for sequence-based transcriptomics. Testing for statistical flags, NC State uses simple rank-based procedures. In multiple testing, there is false discovery control. Dose-response curve fitting is highly reliant on 4-parameter (Hill) logistic models, 3-parameter logistic models, or gain-loss models depending on the context and amount of data available.

Dr. Wright recounted the data pipeline used by his group: 1) count matrix generation, 2) count matrix quality control (QC) and normalization, 3) differential gene expression analysis, and 4) concentration response modeling and POD calculation. Regarding dose-response curve-fitting, he noted that:

- With lots of data, one can explore a large number of models,
- With few data points, a reduced number of models are explored,
- Nonparametric smoothing methods may work, but finding appropriate bandwidths may be difficult with little data,
- Most PODs involve interpolation, so different reasonable models often agree,
- For gene expression, there is a need to handle testing as well as estimation,
- Many approaches use one standard of deviation variation relative to controls to estimate PODs, so the POD is dependent on the experiment technology.

For discovery versus predictive pathway analysis, he said that final pathway-based PODs are based on minimum median pathway PODs, much like BMDExpress. He discussed the possible use of bootstrapping to quantify uncertainty at the per-gene level or for median pathway PODs. He described the use of ToxPi evaluations of pathway activity, which can be clustered for biological read-across.

B.1. Questions for Clarification/Panel Discussion

Regarding the one standard deviation issue, Dr. Yauk observed that the NTP proposed approach was closer to a hybrid model based on modeling rather than just the variability of controls. She asked Dr. Auerbach to clarify that point. He replied that the approach is to use the standard deviation in the model, not the standard deviation of the data. In

BMDExpress, a poorly fit feature will result in the BMR moving as a function of the fit to the curve. Estimation of POD does take into account the fit for the entirety of the curve.

Dr. Fred Parham said that the standard deviation that BMDExpress uses is not just the standard deviation of the control data, but is a standard deviation that is estimated from all of the data in the curve. This is only true if an assumption of constant variance is applied. If an assumption of non-constant variance is applied, then the BMR is based only on the standard deviation of the control group.

Dr. Thomas clarified that standard deviation in this instance is a source of uncertainty, not necessarily variability. He said there are different ways to characterize the uncertainty and variability to represent the BMR. Dr. Wright noted that as technology improves, those parameters would be expected to shrink somewhat, with a limit due to the underlying variability representing different animals.

Dr. Stevens asked what platform NTP is planning to use. Dr. Auerbach said multiple platforms would be used going forward, including Affymetrix, S1500, and RNASeq. As time goes on, it will shift more toward sequencing-based technologies. Dr. Stevens asked if the use of multiple platforms would hamper the analysis, increasing the degree of complexity and adding implementation problems to the overall plan. Dr. Auerbach said that when looking across multiple platforms, the POD answers tend to be fairly consistent. He described past NTP experience with the use of multiple platforms. Dr. Stevens asked that, if gene sets were to be used, would they be fully captured in the S1500 versus an RNASeq experiment. He noted that when it comes to potency determination, it's possible that there will need to be scaling or standardization across reduced representation platforms such as the S1500. Dr. Gerhold described the benefits of the various platforms. Dr. Thomas observed that a number of studies have shown gene biases regardless of platform used.

C. Overview of the US Army Approach to Genomic Dose-Response Modeling: Toxicogenomic Dose-Response Analysis to Inform Risk Assessment

Dr. Lyle Burgoon, leader of the Bioinformatics and Computational Toxicology Group at the US Army Engineer Research and Development Center, briefed the panel on his group's approach to genomic dose-response (GDR) modeling. He described preprocessing using Log2 transform and quantile normalization. He discussed (1) hypothesis-testing to identify probes associated with genes in adverse outcome pathway (AOP) networks of interest, employing Bayesian Region of Practical Equivalence (ROPE) and 95% highest density interval (HDI) analysis, and (2) screening for differentially expressed genes, employing the same methods to analyze only probes with at least 1.5x up/down regulation (in normal, as opposed to log-transformed, space). POD determination is based on monotonic dose-response. The group uses a tool called the Good Risk Assessment Values for Environmental Exposures (GRAVEE). They overlay data onto AOP pathway networks using AOPXplorer.

Dr. Burgoon provided details on the group's use of Bayesian analysis to identify differentially expressed probes and genes. He described Bayesian 95% HDI analysis in more depth, leading to a discussion of how uncertainty is derived. He discussed the

characteristics of parametric and non-parametric modeling. He illustrated the concepts he described through a case study of TNT exposure where gene expression was placed in a biological context. The process yielded a reference dose for TNT steatosis, using POD, IVIVE (in vitro – in vivo extrapolation), external dose, and uncertainty factors.

C.1. Questions for Clarification

Dr. Yauk asked when Bayesian HDI analysis is applied to identify differentially expressed genes, how many are yielded in the end, and how does it compare to the NTP's approach using analysis of variance (ANOVA) and fold change at p=.05? Dr. Burgoon said a comparison had shown there was not a large difference. He said the Bayesian approach is simpler, more transparent, and addresses the "p value crisis." It also allows continued study of chemicals using prior knowledge.

Dr. Johnson asked how the GRAVEE model compares to BMDExpress if extended to a POD number. Dr. Burgoon said that had not been done yet, but he noted that GRAVEE does not look at gene sets, which could produce some differences.

Dr. Peddada asked Dr. Burgoon how he is choosing priors and how he is implementing bootstrapping. Dr. Burgoon said they are doing the Bayesian analysis on a single dose, so priors are not based on a shape of the dose-response. They are filtering based on the following: as long as the criteria for being active in at least one dose are met, and a monotonic response is shown, it will be carried through for dose-response analysis. He described how priors are set for an individual dose, using a normal prior. With regard to bootstrapping, Dr. Burgoon said they randomly resample the curve itself.

Dr. Huang asked how GRAVEE handles non-sigmoidal curves. Dr. Burgoon said that for a linear curve, the lowest concentration or dose is used.

Dr. Auerbach asked what sort of run times would be involved in bootstrapping one thousand genes. Dr. Burgoon said it would be relatively fast (seconds as opposed to minutes).

Dr. Wright asked whether each gene is done separately or entire samples are considered during bootstrapping. Dr. Burgoon replied that they are done separately.

D. An Automated Method Identifies Dose-responsive Genes and Quantifies Points of Departure

Dr. David Gerhold of NIH/NCATS reported on his group's GDR modeling approach. He noted that this is the first time in history when rich, dense gene expression results can be investigated at a wide range of doses. It can be hoped that the new era will be more predictive of *in vivo* biology.

Dr. Gerhold offered the following principles/observations:

- A consensus on BMD/POD method and pathway decisions will facilitate cooperation among Tox21 members and consistent risk assessment.

- Public BMDExpress 2.0 software and visualization tools are useful, although he suggests changes to the algorithm for identifying "significant genes."
- False positives need to be minimized. With 21,000 genes there is a multiplicity problem.
- The simplest model (most constrained) applicable to transcriptional regulation will minimize overfitting and minimize false positives.
- Conclusions close to the lowest dose are most difficult to interpret, since there is no information from lower doses and may lead to false positives.

Dr. Gerhold described several steps taken to optimize their algorithm without knowing "truth," including the use of 3 probe-sets per gene. He provided the statistical details of the NCATS POD method, with several examples illustrating each point. He noted that biphasic responses are a particular challenge, and described how they are treated.

He related a series of recommendations:

- Use a single model (Hill eq.) for consistency and improvement in performance. It is the simplest, most constrained model applicable to transcriptional regulation and minimizes overfitting and false positives. For biphasic responses where the model will tend to fit the response at lower doses, these low-dose events tend to be the ones of importance for dose-response conclusions.
- A trend test facilitates true/false positive decisions, especially at the lowest dose. The POD is conservative to minimize false positives. Once we make calls for each gene, potentially use BMD/POD number of standard deviations from control or minimum fold change to adjust stringency.
- A database is imperative to store data and experiment annotations. NCATS uses enterprise grade database storage. This allows central storage and search for all processed data.
- Suggests changing to a minimum change of 3 standard deviations from mean of vehicle controls, instead of using 2-fold. A standard deviation basis adjusts for noisy experiments/noisy genes.
- Suggest retesting these algorithms in cases where there are gene responses at the lowest dose tested, since it is expensive to test every chemical at low [nM] concentrations. A trend test could be helpful here as well.

D.1. Questions for Clarification/Panel Discussion

Dr. Pramana commented about the filtering methods that Dr. Gerhold described. Dr. Gerhold said that in the example Dr. Pramana was discussing, the fit was quite good, although it was counterintuitive. Dr. Thomas noted that in the NCATS experiments, there were 10-12 vehicle controls on each plate, and BMR was defined based on variability in those replicate controls, whereas in his group, particularly in the *in vivo* model, there is an animal-to-animal variability, as opposed to technical variability in the NCATS work. He recommended discussion of what represents technical variability versus true variability. Dr. Gerhold agreed that that is an important point. He asked Dr. Thomas what he would recommend doing with an experimental data set where there are changes at the lowest dose — whether they should be filtered out and thrown away,

or whether the POD should be defined as less than or equal to the lowest dose. He said that is why they do the trend test. Dr. Thomas said that was one approach, but a better experimental design would include an adequate floor. Without that, no amount of statistics can compensate. Dr. Thomas added that in his experience, the Hill model leads to as many spurious fits as the polynomial model, and the method for aggregating and interpreting those fits will help account for the poor fits. Dr. Gerhold said that in his experience, the Hill model does not mislead the way the polynomials do.

Dr. Wright agreed that he had not seen as many misleading fits from the Hill model but felt that other model fits may not differ that much with respect to the POD. He said it would be reasonable to consider some sort of constraint on the polynomial model to avoid multiple direction-changes in the curve.

Dr. Auerbach asked Dr. Gerhold what filter thresholds he used to identify the probes to be modeled and whether he had applied a fold-change cutoff. Dr. Gerhold said that at the very end of the experiment, a fold-change cutoff of 1.6 was applied. He said that would be subject to change depending on how noisy the data set is. Dr. Gerhold added that a T-test was used for each dose with a metric threshold value of .05, and a test for the trend for the slope of the curve with a threshold of p=.01.

E. Overview of the NTP Proposed Approach to Genomic Dose-Response Modeling

Dr. Auerbach discussed the “bigger picture” questions related to why the NTP has chosen the overall approach to GDR modeling that had been described, including (1) why a BMD approach was selected instead of the more traditional NOEL/LOEL approach and (2) why a statistical and effect size filter is used before performing modeling rather than letting the models alone determine what is responding to treatment.

Relating an overview of the proposed approach, he discussed study design in terms of many dose levels, limited biological replication, and select target organs or cell types. He described the BMDExpress 2.0 software, with its filtering features, dose-response model fitting features, and determination of gene set level potencies. He also talked about plans to include biological interpretation through a data and literature-driven curation of the gene sets to provide a baseline toxicological interpretation and contextualization of the active gene sets.

Dr. Auerbach said that the BMD approach was chosen based on the goal of accurately estimating the minimum biological potency as opposed to detecting hazards. In addition, the plan is to fit a diversity of features from a single study with an array of potencies with the hope of accurately estimating their PODs. NTP has concluded that some sort of filter is needed other than the global goodness of fit filter and the BMD/BMDL ratio. He detailed the logic behind the proposal to use the parametric models from the US EPA approach:

- They are validated.
- They are used broadly in risk assessment to derive potencies (BMD).

- Diversity of models allows for adequate fitting of a variety of dose-response patterns
- There is valuable documentation and guidance on how to use the models.

He also explained the decision to use gene sets versus individual genes:

- Gene sets have a better coverage of biological space,
- Gene sets better represent the underlying totality of the emergent properties at the cellular and tissue level, and
- Gene sets give better representation of the uncertainty in biological potency.

Part of the proposed approach is to attempt biological interpretation. Dr. Auerbach emphasized that it would not be used for traditional hazard labeling at this time.

E.1. Questions for Clarification

Dr. Naciff asked what would be done in the case of a gene set where one is up-regulated and one is down-regulated and whether the different responses would be weighted. Dr. Auerbach said that up to now, they have not weighted the different fits. Directionality of pathways is being looked at in a recent internal release of BMDExpress, but that has not been specifically examined in relation to an interpretation. Right now, they are simply looking at the potency values of response, with a median value being reported as the potency value for the pathway.

Dr. Stevens asked for confirmation that the approach is aimed at identifying biological responses and not at hazard identification. Dr. Auerbach confirmed that assertion. Dr. Stevens noted that conversations tend to trend toward hazard assessment quickly, so if the focus is solely biological response, a different context is created for discussion. Dr. Auerbach clarified that the analysis outcome is an empirically derived relationship and not a qualitative hazard determination. He also emphasized that biological interpretation should be differentiated from a hazard call. Dr. Stevens said that his impression is that NTP wants to identify all possible PODs, whether they are used in risk assessment or not, and Dr. Auerbach confirmed that statement. Dr. Stevens felt that by trying to define all possible PODs and also defining the toxicological significance of biological pathways, the approach crosses the boundary into risk assessment. He said he would like more clarity on the type of input being requested by NTP — whether the approach is an adequate way to define all possible PODs absent of any toxicological context or how to interpret the potencies and PODs as a starting point for future risk assessment. Dr. Auerbach replied that NTP would like input on the actual POD approach, but any insight about how to perform biological interpretation would be very helpful.

Dr. Gerhold asked Dr. Auerbach to define what is meant by “gene set” with respect to potency reporting. Dr. Auerbach replied that they are pre-defined gene sets, building out the Hallmark gene sets. They are curated gene sets which already carry biological associations, in most cases.

Dr. Clewell asked how the gene sets analyses would be used right now, in terms of both the *in vitro* and *in vivo* studies, and whether they would be used to make preliminary

decisions about chemicals. Dr. Auerbach said that both *in vitro* and *in vivo* studies would be used in parallel in order to build out the database. The intention of the *in vitro* screening data is to prioritize, using a margin of exposure-based approach, and subsequently to do an *in vivo* study. Dr. Clewell asked if the initial chemicals would be those already tested in Tox21 and ToxCast biomarker assays or new chemicals that are perhaps of interest in a regulatory sense. Dr. Auerbach replied that both would be done, in parallel, partially depending on events in the real world.

E.2. Public Comments (*ad hoc*)

Dr. Stephen Edwards from US EPA commented in support of biological interpretation. He mentioned the "80/20 rule," where one can get 80% of the way with 20% of the effort, with everything getting harder from there. He said that as the field moves into that last 20%, the understanding of biological effects will become more and more important. He said that preparing for that now will prepare for the future and he recommended establishing a common framework for interpreting data. He said that biological interpretation can make assumptions explicit and can make it obvious why a particular perturbation is not being called a hazard. Dr. Auerbach supported the statements.

Dr. Jeff Gift from US EPA clarified aspects of the EPA BMDS software and spoke to how a future version could address some the concerns raised by panel members. He said the models being used by the current version of BMDExpress could easily be adjusted to address specific problems. He suggested allowing a non-constant variance model where variance is modeled as a function of the response. Also, polynomial models could be constrained so that they are monotonic by specifying whether the parameters are strictly negative or strictly positive. In addition, the EPA's statistical workgroup is looking closely at Bayesian approaches for both model averaging and non-parametric modeling, and those features may be available within BMDS within a couple of years. Dr. Auerbach said that in the NTP approach, the ratio of the BMD_U and BMD_L are used to filter the data, and by dropping the polynomial 3 and making it a Hill model, the BMD_U may not be available. He wondered how the problem could be overcome.

E.3. Panel Discussion

Dr. Stevens said he was on board with the proposal. He wished to highlight the ambitious scope of the project. He encouraged focusing on clear goals.

Dr. Auerbach said he looked to the panel to help determine how much biological interpretation would be appropriate, among other questions.

Dr. Johnson was the reviewer for Session I, addressing the overall approach. He approved of the approach and said that if it is just used as a screening-level approach, a hazard would not need to be identified. He felt that whatever approach ends up being used, it should be data-driven, with the omic POD linked to the apical POD using appropriate data sets. Whatever method gives the best correlation is what should be used, certainly *in vivo*. In terms of pathways, again, a data-driven approach should be

used. From a risk assessment standpoint, as long as the outcome is protective, that is sufficient.

Dr. Yauk initiated the open panel discussion for Session I. She asked that the focus remain on the broader approach, as there would be ample time later in the meeting for more detailed discussion.

Dr. Burgoon said that with respect to screening, he noted that Dr. Auerbach had referred to chemical prioritization and interim exposure limits. He said that when he had tried to institute those elements at EPA, he was told that interim exposure limits need to be legally defensible. That could be an issue for the NTP approach also, as there could be an implication of injury resulting in a need for legal defensibility. Dr. Auerbach felt that rewording of that material may potentially be reasonable. He asked Dr. Burgoon what his avenue to finding success was and whether NTP is missing a specific detail that should be incorporated. Dr. Burgoon replied that where judicial review was possible, the major issues were whether the scientific community had adopted a particular method and whether it was reproducible. He said that for NTP it would depend on the gene sets and their plausibility. He noted that that would be a larger challenge, because peer reviewed publication in the literature would not be sufficient to demonstrate community agreement. Dr. Auerbach asked what a pathway definition or gene set definition would look like that would be plausible and defensible. Dr. Burgoon said the focus at the Army is on solidly supported pathways, building the case for biological plausibility. Using the scientific literature for support may help support the case, without the need to do extra studies.

Related to gene sets, Dr. Pramana noted that it is possible to do some clustering of different genes with different patterns, which could be helpful for interpretation and for finding the right gene sets. Dr. Auerbach mentioned that thanks to Dr. Stevens, NTP had acquired WGCNA-derived clusters from a large liver gene expression data set. He described the evaluation process NTP had undertaken with the data set. He said some had proposed pre-clustering every experiment, but the problem is that as the NTP approach goes forward, there is a desire to use relational analysis, so that with each new data set, the genes would cluster differently, creating a problem for relational analysis. He added that if pre-clustered sets derived from WGCNA were used, it would be a valid approach. The only challenge is the need to create huge data sets with every tissue and cell type to be studied.

Dr. Wright said it was his impression from the previous discussion that using methods that are agnostic about which pathways might be affected might not work as well in situations where there is a clear receptor-mediated response that might be highly toxic. If so, he asked whether there should be consideration of a domain of applicability in which the proposed procedures are expected to work and whether those areas should be documented. Dr. Auerbach replied that the question is reasonable, but the problem is that the behavior of the chemicals in diverse systems can exhibit highly offset potencies. Dr. Thomas commented that exploring the domain of applicability of the current models being applied would help augment the battery of approaches being used to identify PODs, whether for the more promiscuous chemicals or the more specific

ones. Dr. Wright asked if anyone had a counter-example where the approach would not work. Dr. Thomas described a situation in which BMDs were highly variable depending on the system affected. Dr. Clewell said that in transcriptomics, you would be presumably seeing a sensitive pathway POD that would not be seen in the ToxCast assays. It would come back to whether you are measuring the right tissue or the right model. You might not see activity in the liver, for example, because the agent affects a different tissue set.

Dr. Gerhold said there are two kinds of toxicants — those like drugs that are designed against a specific target and those that are used for industrial purposes that are not designed to hit a biological target. There are also two kinds of pathways, he noted. He felt that they should be put into separate bins and treated somewhat differently. He suggested two ways to make the process a bit easier and more objective. First, it helps to profile all of the genes and then assign them to pathways that are split into as many pathways as possible in a non-redundant way. He observed that the weakness of pathways is that they are usually defined in one particular organ or cell type, which is often not specified. It would be helpful if that was defined more carefully.

Dr. Stevens said he was concerned about all of biology being described as deterministic pathways, and he felt that biological systems do not actually work that way. He said that it is not possible to say that PODs for biological responses will be the only ones defined, completely divorcing it from hazard identification and risk assessment. He also suggested other methods could be useful. For example, he asked about doing dose-response modeling of Generalized Compressed Suffix Array (GSCA) scores. He asked if the scope of the project could be more broadly defined, beginning with the gene-level dose-response curve analysis and eventually moving into biological interpretation, giving the project a series of stopping points to analyze progress. Dr. Auerbach agreed that the biological interpretation approach needs a more delineated plan. He noted that the GDR idea has been in the literature for a while, and NTP now wants to design the specifics of the analytical method for a defined purpose. He said that for the primary purpose, biological potency methods are more easily designed, but biological interpretation with gene sets that may be changed could be problematic. The goal is to systematize the process, pre-defining and pre-approving all definitions, so that whenever a gene set comes up, it gets a canned interpretation with all citations needed. Dr. Stevens felt that the scope of what Dr. Auerbach described would be huge, with a high degree of difficulty to achieve scientific consensus on the gene sets and interpretation. He said that trying to do too many things all at once would prevent progress on the initial goal of getting the dose-response for biological effects nailed down and standardized. Dr. Auerbach replied by asking Dr. Stevens what approach and interpretation he would propose. Dr. Stevens said that as Dr. Auerbach had mentioned, all of the BMDs and PODs look about the same, regardless of which aggregation method is used. If that could be rigorously shown, then NTP could demonstrate that even if additional genes were added to a set, the method tends to smooth variability and get a better boundary in estimating the POD. He noted that that approach does not require biological interpretation and has a very clear endpoint. By example of where that approach would be challenging, Dr. Auerbach discussed a study of a purported pathway linked to diabetes that demonstrated a POD from MCHM,

although a biological interpretation of a link to diabetes is inappropriate. He wondered whether pathway names should not be reported. Dr. Johnson said they should be reported as a hypothesis.

Dr. Yauk said that the discussion thus far suggested that biological interpretation is absolutely critical but it may be more for the future than this particular project.

Adding to the discussion of biological interpretation, Dr. Bushel said that a data-driven approach involves feedback mechanisms. He felt that having some biological interpretation early in the development of the system and leveraging it would help to refine some of the pathways or gene sets, building in expert knowledge.

Dr. Naciff agreed that biological interpretation should be employed early in the process. He felt that guidance on biological activity would be important, particularly when there is no other data available. He also stressed the importance of putting everything in the context of the time of exposure.

Dr. Thomas said that it was appropriate to comment on the use of the BMD approaches to model both the dose-response behavior of the genes and gene sets. He also felt it would be appropriate to discuss the use of the BMD approaches to identify a dose below which biological effects are unlikely to occur. He said that both of those questions are appropriately within the scope of the project. However, predicting hazards and interpreting likely hazards associated with BMDs of specific pathways is out of scope, at least for now.

Dr. Auerbach discussed slides depicting GO data for four chemicals over time. He noted that the potency on the most sensitive pathway or gene set did not change.

Dr. Clewell was curious whether the 5-day studies would take bioaccumulation into account. She felt that a tiered decision tree might be in order to adjudge whether a particular agent would likely bioaccumulate; if so, extending or adding an additional time point might be necessary. Dr. DeVito opined that if a compound bioaccumulates, it simply should not be made; however, Dr. Gerhold pointed out that there are many drugs on the market today that are known to bioaccumulate, and are seen as safe. Dr. Stevens clarified they are safe in the context of a therapeutic use.

Dr. Auerbach noted that most laboratories that publish GDR analyses perform them at least somewhat akin to the proposed NTP approach. He asked the panelists to think outside what has been proposed. He wondered about the plausibility of doing a front-end redux, counting, and then fitting dichotomous models.

Dr. Gerhold said it made him uneasy to consider identifying a pathway that responds on a subset of genes and then incorporates all of the genes in that pathway; this practice might shift the median BMD upward. Higher and higher doses would result in finding more and more genes in the pathway, he observed. One solution would be to use the minimum number of genes that give you the redundancy to be confident in the dose-response conclusion, which in most cases is two. He recommended first testing and determining whether one is convinced that the pathway is responding to that chemical

and then perhaps taking the BMD or the POD of the second gene in the pathway. Dr. Auerbach agreed that PODs at higher doses would shift BMDs higher, because the median across the set increases. He said that defining a more systematic way of determining whether a gene set is active, and how that is reported, is going to be the subject of further discussion.

Dr. Yauk said it makes sense to work on groups of genes, given that much of the data over the past decade has regarded gene sets as opposed to individual genes.

Dr. Burgoon fully supported AOP and AOP network approaches. He said that the biggest challenge has been putting the networks together and getting community-wide support that conclusions can be drawn from them. He described his experience with the p53 pathway, illustrating the point that determining the utility of a pathway depends in part on the question being asked.

Regarding AOPs, Dr. Stevens asked how many AOP nodes are not captured with an equivalent GO biological process term. He said he supports AOPs but finds them data-poor. He related his experience where AOPs were not valid when the attempt was made to extrapolate them to a much larger chemical space. He observed that the domain of applicability has not been defined for AOPs. Dr. Burgoon agreed and said he has had issues with how AOPs have typically been compiled and applied, which is why his group has gone to Gene Ontology, to reactome, to Wiki Pathway, and others. He agreed that the domain of applicability with AOPs has been too small which is why his group has moved to diseases.

Dr. Yauk summarized the panel discussion. She said there seemed to be a good deal of support for the big picture of what the NTP has proposed. The one take-home message, she noted, is that perhaps defining the scope in a more detailed way would be useful, focusing more on the POD goal.

IV. Session II: Filtering of Measured Features

A. Some Pertinent Findings from MAQC Related to Reproducibility of Gene Expression

Dr. Bushel presented some pertinent findings from the MicroArray and Quality Control (MAQC) Consortium based on gene expression data and related to reproducibility.

He provided a brief overview of MAQC. It is an FDA-led, community-wide, crowd-sourced effort to assess technical performance and application of genomics technologies in regulatory settings, clinical applications and safety evaluation. It has evolved in four stages: MAQC-I, MAQC-II, MAQC-III/SEQC-I, and MAQC-IV/SEQC-II, which is the most recent iteration, dating from 2015. It addresses QC and reliable use of Whole Genome/Exome Sequencing and Targeted Gene Sequencing in clinical applications and regulatory science research. In his talk, he focused on discussing some of the findings from MAQC-I and MAQC-III/SEQC-I.

He described the reproducibility crisis in science today, with widespread failure (>70%) in reproduction of another scientist's experiments. MAQC-I addressed the issue in the

context of a gene expression microarray platform. Several factors affect reproducibility in toxicogenomics studies:

- Study design
- Platform
- Between and within study sites
- Data processing/Normalization
- Treatment effect

Dr. Bushel described an MAQC-I rat toxicogenomic study, assessing reproducibility between two sites. To determine concordance of differentially expressed genes (DEGs) between the sites, a formula was applied to yield the percent of overlapping genes (POG). There was agreement between the sites in terms of the reproducibility obtained from the T-test, but depending on the ranking, the concordance differed. Fold change value outperformed p-value significance. Coupling p-value with fold change improved reproducibility, within a site, between sites, and between microarray platforms over just using a p-value alone.

MAQC-III/SEQC-I focused on reproducibility between gene expression platforms – are DEGs detected on the microarray platform also detected on the mRNA-Seq platform? Dr. Bushel described a toxicogenomics study design using male Sprague-Dawley rats. Part of the results was from the use of a measurement called root mean squared distance (RMSD), computing the average RMSD for all pairs of biological replicates and compound treatments. As more low-expressed genes were added to the list of genes, the variability between biological replicates increased. The results were validated by qPCR. Genes above the median of expression displayed a high degree of concordance between microarray and RNA-Seq, while genes below the median did not. It was also shown that the strength of perturbation is linearly correlated with the treatment response, and that the more the system is perturbed, the higher the concordance. Concordance also increased at the pathway level.

Dr. Bushel's take-home messages were:

- Use a fold change threshold coupled with a p-value cut-off to balance biological meaningfulness with statistical significance
- Filter out low expressed genes
- Know your chemical's transcriptional strength (if at all possible)
- Pathways perform better than genes individually
- It is of interest to see if some of these findings can be extended to other transcriptomics platforms such as Tempo-Seq

B. NTP's Proposed Approach to Filtering Unresponsive Genes

Dr. Auerbach gave a short presentation on the NTP's proposed approach to filtering unresponsive genes, which is the first step of the proposed NTP analysis pipeline.

The proposal is to use a combined one-way ANOVA and fold change filter to remove unresponsive features. Dr. Auerbach explained the reasoning behind the choice. He

noted that using orthogonal filters is in alignment with the MAQC recommendations. Combining the two filter types maximizes reproducibility and minimizes false discovery. He said that thresholds will need to be determined for each platform and provided an example of how that was accomplished for Affymetrix 230 2.0 microarrays. He described the methods to be employed for permissiveness and noise elimination and methods to ensure reproducibility.

The approach uses a combined statistical and effect size threshold filter for responsive features. It is an empirical method for deriving statistical and effect size thresholds that considers noise reduction, permissiveness, and reproducibility.

B.1. Questions for Clarification

Dr. Pramana asked Dr. Auerbach about the null models used in reaching filtering conclusions and what the approach would be if there had been signal in the data. Dr. Auerbach said that the original reason was based on experiences during the Elk River study. Even with its weak signal chemicals, there was clearly a response in a small subset of genes, so the approach was taken to separate the signal from the noise. It has subsequently been applied to multiple weak signal sets. He noted that with the approach, there is clear biological plausibility with what is pulled from the data.

Dr. Wright noted that in his description, Dr. Auerbach had described a number of active gene sets, but never said that no genes were found under the "no multiple testing correction" element with the null data. Dr. Auerbach confirmed Dr. Wright's understanding. Dr. Auerbach said that "active gene set" was defined as a minimum of three genes with an adequate BMD fit. The gene set must be 5% populated, with a nominal Fisher's Exact Test value of $p<0.05$. Thus, there are multiple combined features.

Dr. Huang asked Dr. Auerbach if he had looked at false negative rates and whether truly responsive genes may have been filtered out in the process. Dr. Auerbach said there was a need for well-defined prototype compounds to be able to do so. Most of the prototype response genes have huge effect sizes, so in most cases they would not be filtered out and would not be good tests for false negatives.

B.2. Public Comments (*ad hoc*)

Dr. Draghici had some issues with leaving out a multiple testing correction step. He suggested that Dr. Auerbach reconsider. Dr. Auerbach noted that it was an integrated approach with many levels of filtering, resulting in being stricter than a traditional collection of gene sets after running a T-test. He said that the criteria of 3 genes and 5%, although arbitrary, results in being very restrictive in what is actually being reported. Dr. Draghici said he was concerned about false positives.

Dr. Peddada asked Dr. Auerbach to explain what he meant by "fold change." Dr. Auerbach said it was a maximum fold change at any dose level. Dr. Peddada expressed concern about Type 1 error. Dr. Auerbach replied that evaluating any one of the steps independently would be challenging, but all of the extra layers of filtering

should increase robustness. Dr. Peddada said he did not understand how false discovery would be prevented. Dr. Auerbach said that when results are reported, they would include the number of probes that came through, the number that were modeled, and the number that made it into active pathways.

Dr. Edwards addressed the exclusion of the multiple testing correction. He indicated that with the other filters being applied, he did not have concerns that this correction was excluded from the approach.

B.3. Panel Discussion

Dr. Wright was the first reviewer for Session II. He thought that it is possible for a series of weak predictors to collectively make a good prediction rule. Whether the NTP procedure is achieving that, he was not yet entirely sure. He noted agreement that some kind of filtering should be done. The proposal as written did not give every particular value that Dr. Auerbach had shown in his presentation; the lack of specificity in the proposal made it difficult to evaluate. He hoped there would be more database investigation of the particular thresholds used in the filters. He noted that when there is a large proportion of genes that are features that are truly dose responsive, even when using a multiple testing correction such as false discovery control, a large number of features may still be captured. Even an uncorrected procedure using $p<0.05$ may yield a smaller but still appreciable number of features. $P<0.05$ would also indicate a strong signal, and there would probably be a lot of true discoveries in the set. He noted that once thresholding has taken place so that only things above the expression threshold are considered, the mean-variance relationship is not carried forward. Then, applying the fold change filter does not account for variation. He said there are different types of p-values that could be used. With the ANOVA-based p-values for filtering, the doses are used as a categorical response, not using the ordering — there is no presumption about the order of anything in this case. Thus, Dr. Wright observed, sets make it through the filter whether or not they show a dose-dependent trend. He said the approach would suffer in power if there were few replicates per dose. An alternative would be to use a trend test or rank regression. He suggested testing the efficacy of various methods as data accrues to adjust the method if needed.

Dr. Auerbach appreciated many of the points raised by Dr. Wright, and said he is supportive of using test data to refine empirical approaches. He said that the newest version of BMDExpress includes a Williams' trend test and will also implement a non-parametric approach for identifying differentially expressed genes. He asked Dr. Peddada to describe that feature in more detail.

Dr. Huang was the second reviewer for the filtering approaches session. She said that a trend test to be used in filtering would help to reduce false positive/false negative rates. She recommended accounting for outliers in the approach since outliers could pass the ANOVA test, throwing off curve fitting. She felt that the selection of threshold values should be experiment-dependent, because different platforms could have different noise levels.

Dr. Auerbach replied that NTP is working with Sciome to implement curve-p for adjustment of monotonicity, which will also take into account some of the outliers and outlier effects. He said the hope is that it would be implemented in BMDExpress within the next couple of months. He added that two trend tests are also being implemented. Optimization will need to be performed for each platform and may even have to be done for each experiment, particularly now that costs are lower.

Dr. Burgoon read a statement on p-values from the American Statistical Association. "P-values do not measure the probability that the studied hypothesis is true or probability that the data were produced by random chance alone." He recommended that the p-value threshold be removed entirely and rely solely on fold change. He also endorsed dropping the Fisher's Exact Test.

Dr. Auerbach showed some data slides. He said he had not yet run the filtering approach in the absence of ANOVA and just with fold change. He compared results for a large data set when several different filters were applied.

Dr. Stevens approved of the NTP approach to filtering. He said he was skeptical of the 3-gene approach. Regarding the table Dr. Auerbach had shown where the number of GO biological processes covered after the filtering decreased, he assumed it was because the number of genes that make it through the filter was going down. Dr. Auerbach confirmed his impression. He asked if the analysis could be done controlling for the number of genes that make it through the filter as a covariate. He suggested that that could allow the filtering to be less stringent.

Dr. Auerbach said that there should be more truth as filtering becomes more rigorous. Dr. Stevens agreed as long as there was some method for pathway enrichment. He said it could be argued that overly restricting the data set would throw out too much positive signal. It is always a problem when working with individual genes.

Dr. Peddada asked why the fold change was 1.5. Dr. Auerbach replied that the data was actual, not simulated. They are null data from animals that were treated with the same vehicle. There is interindividual animal variability as well. Dr. Peddada observed that therefore it is not truly null data.

Dr. Wright commented on the microarray data. There is resulting sampling variability, so the fold change cutoff is a sample fold change; this implies it is a statistic, not a parameter, he observed. Dr. Auerbach agreed with the assessment.

Dr. Yauk noted that some on the panel favored false discovery rate adjustment, while others were comfortable with fold change only. She asked for comments on an alternative trend test rather than the ANOVA.

Dr. Gerhold reiterated Dr. Auerbach's point that a trend test does not accommodate non-monotonic changes, and Dr. Huang's approach uses a trend test at three consecutive points, which allows for non-monotonic changes.

Dr. Auerbach showed several data slides depicting the Williams' trend test versus ANOVA, with the eventual PODs.

Dr. Clewell asked which version of the Williams' trend test was used, whether it was or was not the one that allows for non-monotonicity. Dr. Auerbach said it was the one that does not allow for non-monotonicity. Dr. Clewell noted that it might be unnecessarily filtering some things out. She asked whether removing fold change would result in seeing more differences in the data Dr. Auerbach had shown.

Dr. Pramana suggested trying other trend tests, as his group has done. Dr. Auerbach asked, given that there is very little difference between the ANOVA and the trend test, how would adding a different trend test impact the results? Dr. Pramana said he would expect significant differences between the Williams' test and the multiple contrast test (MCT) that would allow understanding of the pattern of the dose-response in each gene. Dr. Auerbach said that was a great suggestion.

Dr. Auerbach asked the statisticians present whether there is a threshold they were comfortable with. Dr. Gerhold suggested using estimation of the false discovery rate. Dr. Yauk asked for comment on going with a fold change-only approach. Dr. Burgoon replied that that is his group's approach with polymerase chain reaction (PCR) arrays. He added that he is not generally supportive of the false discovery approach because his group does not use p-values; instead, they do the ROPE analysis. Dr. Thomas said a null data set would be useful to try to limit false positives, and he favored exploring that and other data-driven approaches.

Dr. Wright noted that the software includes all datasets that pass the filter including any using the polynomial model, with the polynomial-2 chosen the largest fraction of the time. Dr. Auerbach said that it depends on the data set. Dr. Wright said that a smaller model would be favored if it has the same fit because of the AIC criteria, resulting in a "soft preference" for monotonic trends at the modeling step.

Dr. Yauk asked the panelists to comment on the pitfalls of the trend test and whether there was opposition to adding a trend test to the proposed NTP approach.

Dr. Burgoon said he did not see an advantage to adding a trend test.

Dr. Gerhold said that he would favor the use of a 3-point trend test and provided his reasoning for that conclusion. Dr. Wright said he saw the ANOVA and the trend test as being at two extremes, with these extremes not resulting in different transcriptional PODs. He saw Dr. Gerhold's method as being in between, so he argued that method also likely would not result in different transcriptional PODs. Dr. Gerhold replied that the trend test disqualifies many changes that would have been called positives, which are probably false positives. Dr. Wright speculated that using Dr. Gerhold's approach with the NTP approach would not result in very different PODs. Dr. Gerhold said he had several examples where an opposite call would be made as a result of applying the trend test to the data. He added that he was disturbed by the fact that there were 250 PODs called both by his group's approach and the BMDExpress approach, but that there were 23 genes called to go in opposite directions. Dr. Wright clarified that the draft NTP approach has filtering and curve fitting in two steps, and the current discussion was only on filtering, but is moving toward curve fitting. Dr. Gerhold

explained that his group's approach does the ANOVA test and then the trend test for each set of three data points, and then the data are fit to a curve to get the POD.

Dr. Peddada described use of the trend test within the draft NTP approach. Dr. Auerbach added that when the data is modeled with the trend test versus the ANOVA, a high number of genes come through the analysis, but only a small fraction are behaving differently.

Dr. Clewell wondered whether with so many corrections going on, some might be affecting different parts of the process more than others. She proposed relying less stringently on filters and allowing more genes through the process; if the assumption were correct that the pathways are what drive the POD, then it may not matter so much if individual genes are showing different patterns. She said she would prefer the trend test to the ANOVA. Dr. Auerbach said he would take a look at her suggestion.

Dr. Burgoon felt that the trend test and ANOVA were not contributing to the filtering as much as the fold change. Thus, he endorsed fold change. Dr. Auerbach counter-argued that the idea of using a purely statistical approach was problematic. "Is there any reason to think that using the effect size filter, or fold change, would be out-competed by one of these other methods?" he asked. Dr. Wright said there are situations where taking the logarithm of data is a variance-stabilizing transformation. In that scenario, "fold change is everything," he observed. Thus, fold change can be looked at as a shrinkage procedure.

Dr. Gerhold said he agreed with the MAQC conclusion that fold change works better than T-test or ANOVA, but before accepting it for gene counting, it should be tested.

Dr. Yauk observed that Dr. Auerbach would require additional experiments to look at some of the different approaches that had been proposed. She raised the issue of regulatory acceptance.

Regarding the comparison between trend test and ANOVA, Dr. Thomas asked what the benchmark of comparison would be. He offered several examples of how this benchmark could be established.

Dr. Gerhold said the best method he could think of to arrive at a benchmark in these experiments is to look for concordance across the three runs of the same experiment. Dr. Auerbach asked him what he would consider to be good concordance in genomic space, where it tends to be lower than in other spaces because so many features are being tested. Dr. Gerhold said that if the same genes or most of the same genes were called in all three runs, and were always up or always down for a given gene in all three runs, it would be considered to be working well.

Dr. Stevens said it seemed that the whole discussion was centering on a requirement of dose-response in individual genes, which are then filtered through to find biological processes. He asked if would not be just as reasonable to flip that approach on its head and find the biological processes that are enriched in multiple experiments, and then determine how many show dose-response relationships. He said that approach seems

valid, when what is ultimately being sought is whether there was really any difference in the number of pathways, gene sets, or GO biological processes that were hit that yield a maximum value to define PODs.

Dr. Thomas expressed skepticism that the approach Dr. Stevens proposed would be any better than the others under discussion. Dr. Burgoon agreed with Dr. Stevens. He said he liked the intuitiveness of the approach, because it narrows down the space. He discussed the conflict in predictivity between *in vitro* data and human data, with the need to concentrate on valid human predictivity. Dr. Auerbach said that it is possible that getting the truth on a quantitative level may actually be using C max or C steady state levels, of which about 500 are curated in Drug Matrix and available to be used for validation.

Dr. Yauk brought up the question of what goes into the calculation of fold change or ANOVA test to begin with – i.e., how much signal do you need to begin to analyze a gene. Dr. Wright reiterated that he used overall average gene expression level across all doses including controls, using as few as but no less than five counts. Dr. Rick Paules asked Dr. Wright for clarification of his remarks regarding a 5-count average. Dr. Paules said the question was important because at really low levels, there is heightened noise. Dr. Wright explained that they used a value as low as 5 because their q-value approach, which estimates proportion of false discoveries, told them that they actually were not paying a penalty going from 10 down to 5.

Dr. Gerhold agreed about using an average but said his group had been using a higher average, because going as low as 5 or 10 the ANOVA test would not be passed. He said they use 20 or 30 as an average minimum.

Dr. Bushel noted that in MAQC, it did not matter what type of preprocessing was done; the trend still remained the same.

Dr. Stevens said he was concerned about the direction of the discussion, in that it might provide NTP with too much input on too many different tests. When choosing different tests that don't have any fundamental difference in the number of biological responses, it might not be worthwhile to test all the different methods. He argued the test of the filter should be whether the filter yields biological responses as a function of the gene sets.

Dr. Yauk recapped the day's materials and discussions. She noted that there appeared to be much support for the general proposed NTP approach.

Day 2: October 24, 2017

V. Session III: Fitting Features to Dose-Response Models

A. Interpreting the Results of EPA Dose-Response Models

Dr. Jeff Gift from the US EPA National Center for Environmental Assessment briefed the panel on the EPA's approach to dose-response modeling, which served as a model for the proposed NTP approach.

He described the EPA's Benchmark Dose Technical Guidance, which took 12 years to formulate, and was published in 2012, as well as other pertinent EPA documents. He went over the key steps in EPA's BMD analysis:

- Benchmark Response
- Model Selection
- Model Fit
- BMDLs
- Akaike Information Criterion (AIC)

He noted that BMD software (BMDS) can analyze continuous data. The preferred approach is to use a BMR that corresponds to a level of change representing a minimal biologically significant response, such as a 10% decrease in body weight. In the absence of biological consideration, a BMR of a change in the mean equal to one control standard deviation from the control mean is recommended. In some cases, use of different BMRs is supported.

Dr. Gift provided details about the process of continuous model selection and continuous model forms, as well as restriction of model parameters. He described methods to determine whether the model fits the data, including variance and goodness-of-fit tests. He discussed the "sufficiently close" concept, which can vary depending on the needs of the assessment, but generally should not be more than 3-fold.

- If BMDLs are not sufficiently close, EPA recommends picking the model with the lowest BMDL.
- If BMDLs are sufficiently close, EPA recommends selecting the model with the lowest AIC.
- If multiple models have the same AIC, EPA recommends combining BMDLs.

Dr. Gift made a side-by-side comparison between the NTP proposed GDR modeling approach and the analogous EPA approach, covering the various steps in the process. His conclusions were:

- BMD modeling for traditional continuous endpoints using the "best method" approach has been used for more than 20 years and the methods are well-defined.
- Alternative methods are being researched to address model uncertainty (e.g., model averaging) and provide more accurate modeling results.
- BMDExpress leverages BMDS model executables to extend methods to alternative endpoints (i.e., gene expression)
- BMDExpress modeling and model selection criteria are generally consistent with EPA methods in areas of overlapping purpose.
- BMDExpress is well-positioned to adapt to updates to BMD modeling approaches (i.e., adoption of model averaging)

A.1. Questions for Clarification

Dr. Wright asked about how BMDExpress accounts for the fact that the control dose (0) cannot be expressed on a log scale. Dr. Gift said that the dose is not log-transformed during the analysis (only in visualization) and the EPA models allow for the assumption that the response distribution is lognormal. Dr. Auerbach agreed that in order to visualize the data, there cannot be a zero value on the log dose axis and confirmed that the modeling is not done on log-dose.

Dr. Pramana said that using log dose would generate a different pattern than dose, with what is observed perhaps not being the actual fitting than would be seen using the dose. Dr. Auerbach discussed some of the background thinking behind using log dose for visualization, including allowing better visualization of responses at low doses.

Dr. Wright noted that his group, in contrast to the EPA method, converts dose to a log scale first. Dr. Auerbach asked how the zero-dose control is handled. Dr. Wright replied that they choose a log scale value that is lower than the smallest dose by the average gap in the other doses. Dr. Thomas asked if they were fitting the fold change or had transformed the normalized count data. Dr. Wright explained that the y axis is log scale of whatever the normalized expression data are.

B. Fitting Curves Using Non-Parametric Approaches

Dr. Keith Shockley from NIEHS described methods of non-parametric curve-fitting. He said that he would compare and contrast non-parametric with parametric modeling.

He discussed parametric models, which are specified by parameters and require considerable specification. They are not typically very flexible and do not use observed data to make predictions. Non-parametric models are much more flexible and do use observed data to make predictions. Non-parametric models do have parameters but are able to follow and describe a dose-response pattern more reliably than parametric models.

He related the pros and cons of parametric models:

Pros:

- Reduce unknown (and possibly complicated) function $f(x)$ to a simple form with few parameters
- Can produce consistent results when the curve fits the data well
- May have familiar and useful parameters

Cons:

- A pre-specified parametric model may not fit the data well
- Carry distributional assumptions (e.g., normality)
- Different parametric models may produce different BMD estimates, reflecting model uncertainty

- Model averaging can be helpful when true function is not on edge of model averaging space

He then went over the pros and cons of non-parametric models:

Pros:

- Makes fewer assumptions about $f(x)$
- Uses the data to learn about the potential shape of $f(x)$
- Should fit the data very well

Cons:

- Parameters may not be readily interpretable
- Carry distributional assumptions
- May be computationally intensive
- May not be as familiar as parametric approaches

He described approaches to estimating POD from fitted curves, illustrating three cases: a Hill equation model and AC_{10} parameter, B-spline and concentration curve that crosses a response threshold, and polynomial interpolation and entropy-based POD. He presented data comparing and contrasting the three approaches.

In summary, Dr. Shockley noted that:

- Parametric modeling requires pre-specifying the model but is more familiar and may have interpretable parameters.
- Non-parametric modeling is more flexible but may be less familiar and may not have readily interpretable parameters.
- Simulation studies and repeatability of experimental results can be used to evaluate the performance of proposed modeling approaches.

B.1. Questions for Clarification

Dr. Yauk asked Dr. Shockley to comment on how the requirements for study designs might differ between parametric and non-parametric models. He replied that there is in fact a large difference, the biggest being the limitation in the number of doses being used in GDR modeling. He said that with some of the smoothing or non-parametric curve-fitting methods, more doses are needed to accurately track the dose-response curve. There is also the issue of number of replicates, since increasing the number of replicates often means decreasing the number of doses.

Dr. Gift described a paper in the literature that tried to address the question. He brought up the issue of data cloud. He said that parametric modelers are sometimes criticized for over-parameterization. He asked if there was such a thing for non-parametric modeling. Dr. Shockley said that non-parametric models could be thought of as assuming an infinite number of parameters, with parameters being a function of the observed data. With non-parametric methods, the data is being used to guide predictions.

Dr. Pramana asked Dr. Shockley about finding the best bandwidth for not over-fitting the data. Dr. Shockley replied that there are specifications for non-parametric modeling, with the user specifying, for example, the degree of the polynomial and the bandwidth — a tuning parameter to be worked out. Dr. Pramana asked if that would need to be done for all genes. Dr. Shockley said an automated approach might need to be developed, or a few genes could be chosen followed by evaluation using simulations or comparison to real data. Dr. Pramana asked how to obtain the purity of non-monotonic trends. Dr. Shockley said that often dose-response relationships are presented as a sigmoidal curve, looking at a specific parameter such as an AC_{10} or AC_{50}. But how would the POD be obtained when the Hill equation is not appropriate? Using a non-parametric approach, he observed, you could first look for the place where the response first goes outside the detection band or where the change in entropy is maximal. Dr. Pramana and Dr. Shockley continued to discuss fitting on the log scale versus fitting on the linear arithmetic scale.

Dr. Wright asked whether in the model Dr. Shockley had shown, the bandwidth would be insensitive to the scale of the dose. Dr. Shockley said that the Hill model is sensitive to the scale. He added that the form would look very different when fitting a Hill model on a linear scale versus a log scale. Dr. Wright clarified that his question was for non-parametric techniques, which can also differ depending on the choice. Dr. Shockley agreed, noting that it could be done both ways.

Dr. Naciff asked Dr. Shockley what his recommendation would be. Dr. Shockley replied that there are pros and cons with both approaches and it will depend on what the model is to be used for. He recommended evaluating different approaches using simulation studies and testing reproducibility on real data to find the approach that might be the most suitable.

Dr. Gift asked if there was a way with non-parametric modeling to determine when one is relying too much on the data. Dr. Shockley again advocated simulation studies to arrive at a more realistic sense of what precision and bias in the estimates really are.

Dr. Thomas asked how many risk assessments internationally have used a non-parametric model to arrive at POD. Dr. Burgoon argued that non-parametric modeling is used widely in fields of data science today and that perhaps it would not be appropriate to look to the “extremely conservative” risk assessment community to determine whether or not a method is accepted. He said other communities’ practices should be considered, where non-parametric modeling is in widespread use. Dr. Thomas said his point was that if the goal was acceptance by regulators, a practical path forward between non-parametric and more traditional parametric approaches would be advisable. Dr. Burgoon agreed and noted that he was not advising NTP to move into non-parametric right away.

C. NTP’s Proposed Approach to Curve Fitting and Determination of Feature Potency

Dr. Auerbach introduced the second step of the proposed NTP analysis pipeline: the application of the continuous parametric models to fit dose-response curves to the

responsive features that pass the fold change and ANOVA filter. The fits are then used to identify potency (BMD) values for each of the features.

Features are fit to 9 parametric continuous models, derived from the EPA's BMDS software: Hill, power, linear, poly2, poly3, and exponential 2, 3, 4, and 5. The BMR chosen is 1.349xSD. This assumes constant variance and is the standard deviation at control of the model. It approximates a 10% shift in the area under the normal distribution. Once the models are fit, a two-step process is employed for best model selection. From the best fit model, a BMD, BMD_L and BMD_U are determined.

He explained why parametric models are used, why 9 models were used, and why a 2-step model selection was employed. He also explained the choice of a BMR of 1.349xSD.

C.1. Questions for Clarification

Dr. Clewell asked whether there would be a specified process for what is a preferred model or technique when interpreting results from the 9 models. Dr. Auerbach replied that the intention is to constrain the uncertainty. He noted that all of the 9 models are fit to each data set, and then the best fit is determined. Restricting to one model (e.g., the Hill model) may result in a loss of information — if a BMD_U cannot be calculated from that model, the feature is removed from the downstream analysis. In reality, the feature may actually be responding, but it is not perfectly fit at the top end with a Hill model, and therefore that piece of information is lost. The biggest challenge is with weak signal data when no models adequately fit. Dr. Gift said that the EPA would say that lacking prior information suggesting that one model is better than another, trying as many models as possible would be the best approach. He said his suggestion with regard to polynomial models was not to exclude the unrestricted polynomial models but to add the option of including restricted polynomial models. Dr. Auerbach asked him if he was recommending dropping the unconstrained and adding the constrained. Dr. Gift said that was not the recommendation. He said they should be kept for certain situations.

Dr. Naciff asked whether Dr. Auerbach meant that for 1,000 genes, for example, every single modeling process would need to be run for every single gene. Dr. Auerbach said that that is the way the system is currently set up. Dr. Naciff asked if the genes should be grouped first by gene set, and then modeled by gene set. Dr. Auerbach asked if he meant clustering by general dose-response shape, and Dr. Naciff confirmed that that was what he meant. Dr. Auerbach said it was a reasonable suggestion.

Dr. Burgoon asked Dr. Auerbach to elaborate on the 10% shift representing a BMR. He felt that that approach looked very conservative and that it was difficult to comprehend how that would represent a biologically significant injury. Dr. Auerbach said it was always a challenge with BMRs when there is no prior knowledge. Dr. Burgoon asked why a 1.5-fold change is not used, since that is what traditionally is used in the genomics literature. Dr. Thomas said the 10% shift was in accordance with convention in the BMD literature. Dr. Burgoon argued that the distribution is assumed to be normal, but that assumption probably is not based on the actual data in-hand. That decreases

the justification because if the data are not normal, the change cannot be characterized as a 10% shift.

Dr. Shockley commented that the determination that 1.5-fold is biologically meaningful is very gene and context-dependent, and part of the trend toward 1.5 was driven by improved ability to measure precisely due to technological advances and not necessarily due to the biological meaning of the response.

C.2. Public Comments

Dr. Gerhold said he had at least five examples in his talk of places where the polynomial fit yields a questionable conclusion in BMDExpress. He wanted to encourage use of models that approximate the way gene expression regulation actually occurs in cells. He noted that to justify a polynomial fit, three events would have to occur: (1) at a low dose, the chemical would cause signal transduction resulting in expression of the gene, (2) at a slightly higher dose, the chemical would cause another signal transduction response to override the response at the lower dose and reduce expression, and (3) then a third event would override the previous ones. He said he had never seen an example of that happening as a function of dose.

Dr. Auerbach didn't think there is belief in the biology of that fit, but it is being used from a purely utilitarian standpoint, because fitting a Hill model results in basically the same BMD value. He said he was open to the possibility of dropping the poly3 model.

Dr. Gerhold asked about the possibility of up-calls and down-calls. Dr. Auerbach said that the directionality may change in certain cases, but in most the directionality is what would be seen in a Hill model and the BMD models are what would be expected.

Dr. DeVito commented that fitting many different models to a monotonic dose-response curve would yield essentially the same BMD. He noted that not every chemical or every gene should have the same BMR, but the biology is not known well enough to make that decision tree for every gene. He said that doing so in this automated pipeline approach would be misleading and would lead to mistaken conclusions. He recommended thinking of it as a screen of a screen, flagging the need to go deeper.

C.3. Panel Discussion

Dr. Yauk asked panel members to concentrate on both the advantages and disadvantages of any alternative methods they would suggest and to focus on the big picture related to which dose-response models the NTP should be using and how the results should be interpreted and used.

Dr. Pramana was the first panel reviewer. He suggested that a trend test should be included in fitting for filtering. He also suggested addition of asymmetric dose-response models such as a Gompertz model or Richard model. He encouraged clarification as to whether modeling was performed on log dose or dose. He endorsed the concept of using model averaging. He did not like non-parametric approaches, citing the need to apply that approach for perhaps thousands of genes. He suggested using the control

dose as the starting point for searching for the BMR. He felt that AIC would not be adequate for nested models.

Dr. Auerbach appreciated the suggestion to add more models, particularly ones that would potentially have a logical fit to capture some component of biology that may have otherwise been missed. He wished to defer to what EPA considers to be best practices in this instance. Regarding the AIC versus the nested chi square, both options are available, and the choice seems to make very little difference in the eventual data.

Dr. Pramana returned to the concept of clustering, with the fitting based on the shape of the dose-response relationship. There would be several options for how to perform it, he noted and would make fitting the model faster. He discussed issues with non-linear modeling, including difficulty in convergence. He said the filtering was most important, because once the unnecessary features are filtered out, the model fitting can proceed more quickly. Dr. Auerbach appreciated the suggestion of identifying the basic structure of the response, which can be done with the Origene output. When input into the model-fit software, models could be preselected that match the dose-response shape.

Drs. Huang, Wright, and Clewell were the panel reviewers for this topic. Dr. Yauk called on each of them to share their thoughts.

Dr. Huang said her concern with the polynomial model is that if there is something that can only be fit to a polynomial, are the results reliable? She said that in cases where something could not be fit to a Hill model, the curves are very noisy. She noted that polynomials tend to be very sensitive to outliers and suggested masking outliers. She suggested adding manual assessment of curve-fitting to the BMDExpress software to be able to manually reject some of the implausible fits. Dr. Auerbach felt that the suggestion for human intervention was good. He noted that he had run the experiment of dropping the poly3 model from consideration and re-running some the data, and the median BMD value for the POD was nearly the same. He pledged to look more deeply at the issue.

Dr. Bucher commented on Dr. Huang's suggestion of introducing a human element to manipulate the data. He said that question would need to be included in any data challenge issued regarding the proposed NTP approach. Dr. Huang said that based on her group's experience, human eyes are still the best judge of an appropriate fit.

Dr. Wright felt that the issue of whether to fit models on the original scale and then just display on a log-scale versus performing fits on log-transformed data should at least be investigated. He said the ideas under discussion largely came down to a continuum ranging from a few-parameter view that things should be a monotone to a high-parameter view that allows the data to specify the shape, with that extreme represented by non-parametric methods. His impression was that non-parametric methods penalize for the number of parameters but do so implicitly through cross-validation. He preferred models that do not allow more than one change in direction. He said that there should be a provision to strip off third-degree polynomials if the fit changed direction twice or changed direction between doses. He felt that the convergence criteria for the Hill model might be a bit too stringent. Dr. Auerbach agreed with Dr. Wright's comments on

the poly3 models and felt that there was probably a simple way to deal with them. He said NTP must work with EPA to deal with the convergence issues.

Dr. Clewell said that she had not understood why there were so many models, but it seemed to come down to achieving a balance between adopting tools the regulatory community has accepted and moving forward with some of the newer, better approaches. She said that as the field advances, many people are finding advantages to using some of the non-parametric approaches, and making them available to the community would be useful. She felt that if parametric models were going to be used, fewer models should be included. She did not see why high doses would be allowed to define the curve, when the low dose will likely define the POD. She said she has done manual curation but that it is not practical to do so in genomics where there are so many genes and so many models. She asked Dr. Thomas to comment on why the models were limited in ToxCast and whether those considerations might be applicable to the NTP proposed approach.

Dr. Thomas replied that the ToxCast pipeline was developed with common practices in pharmacology dose-response in mind. For GDR data, the methods instead seek to move toward acceptance in risk assessment. So the effort was undertaken to see if benchmark dose modeling could be applied to genomic data, resulting in a learning process over the past ten years. Currently, there are both the more pharmacological ToxCast approaches and the benchmark dose approaches in the genomic realm, and the community is beginning to use a lot of the genomic approaches in the *in vitro* screening paradigm. He said that at some point, there will need to be a decision about which approach to take. Dr. Clewell agreed.

Dr. Stevens noted that even though the way of modeling a dose-response curve for any given gene is clearly critically important, the larger question is whether changing the methodology would signal any systematic error in the aggregated biological pathway POD. Would reducing the number of models result in a significant systematic error? He echoed the concern about introducing human intervention because it could introduce bias into the data set.

Dr. Auerbach showed some data slides illustrating using a BMR of 1.349xSD versus 1xSD, which showed a small shift in the POD values. He also showed data depicting using the Hill model versus using all 9 models. He discussed loss of information as a result of weak signal data.

Dr. Burgoon said he felt the panel was becoming focused on issues that did not affect the goals of the approach. He noted that he does not look at his model fits and only does so when a decision is called for. He uses his pipeline as a screen, similar to the proposed application of the NTP approach.

Dr. Gift agreed with Dr. Burgoon's assessment. He said EPA has tried to be a bit more prescriptive with respect to documentation, suggesting the most appropriate approach. He said they have tried to be more flexible in the software and suggested NTP do the same with respect to BMDExpress. He noted that what the software allows you to do may be different, and more flexible, compared to the guidance. Regarding the nested

testing approach, he said that EPA had always used the AIC approach for selecting models, but they did use the European nested testing approach to implement their exponential models.

Dr. Johnson agreed that some of the details that had been discussed probably did not have a large effect on the results of the approach. He felt that the process should be benchmarked based on concordance to the apical effect, and the method should be tweaked if there are ways to improve concordance.

Dr. Gift cited a 2015 paper by Slob and Setzer in which they stated their belief that Hill and exponential models alone were sufficient to handle all continuous data, at least in the toxicological realm.

Dr. Auerbach said that if the number of models used is reduced, inevitably some of them would be forced to a different fit, and the uncertainty around that fit would be greater. Thus, would the thresholds be relaxed, perhaps on the BMD_L/BMD ratio, or the BMD_U/BMD_L ratio? Would that be a reasonable course of action? As models are dropped, there is not as good a fit to the data, and so the uncertainty is greater. For weak signal chemicals, all potential information is lost. He said that as the number of models is reduced, the range of the signal that comes through is reduced. EPA uses stricter standards based upon modeling one specific endpoint for setting risk values with potentially very significant economic impact, whereas the NTP is proposing more of a screening level-based approach, with somewhat looser standards and more tolerance for uncertainty. Dr. Gift said that EPA has found in simulation tests that the more models thrown into the suite, bias is improved even compared to the Hill model alone. The model averaging approach yields additional improvement.

Dr. Burgoon said it was proof positive that NTP should not drop the models and implement model averaging in the future. He said that what NTP has proposed is appropriate for screening applications. Dropping models would produce very good fits for some genes but would increase uncertainty and cause some low signal genes to be lost from the analysis. Responding to a query from Dr. Auerbach, Dr. Burgoon described some of his group's risk assessment studies on three chemicals.

Dr. Stevens said that the NTP work would set a precedent as to methodology used in risk assessment applications. If the current approach is proposed as a pilot, to be evaluated for future improvement to move into risk assessment, that would be reasonable. He urged NTP to be careful about referring to the approach as being "good enough for screening," rather than "sufficient to pilot and do iterative learning to see where we end up for risk assessment."

VI. Session IV: Gene Set-Level Potencies

A. When Is a Pathway Changed?

Dr. Sorin Draghici from Wayne State University addressed the panel. He emphasized that a gene set is not a pathway. He described existing approaches, including classical approaches and Gene Set Enrichment Analysis, in which all of the genes are ranked

based on the correlation to the phenotype. He described the limitations of the gene set approach, illustrated by the example of the insulin signaling pathway.

Dr. Draghici illustrated statistical methods to analyze pathway perturbation, including approaches to validate pathway analysis methods. His recommendations were:

- Use all knowledge available, i.e., use pathways not gene sets if possible,
- Use methods that can assess pathway impact based on the topology and calculate significance based on resampling (e.g., impact analysis), not simple enrichment,
- Use methods that can identify putative mechanisms based on known pathway topology, and
- Take into consideration and eliminate individual pathway bias.

A.1. Questions for Clarification

After mentioning that mitochondrial damage is often involved in diseases such as Alzheimer's and Parkinson's, Dr. Stevens asked Dr. Draghici why pathways should be used if what the pathways are called is less accurate than the Gene Ontology (GO) term. Dr. Draghici replied that Dr. Stevens's point that mitochondria damage is an important part of the disease phenomenon is well taken, but he felt that with knowledge evolving, pathway analysis would still be useful. He said that GO analysis has other issues that render GO less useful than pathways. Dr. Stevens agreed that every approach has its problems, but noted that the problem with pathway is that it assumes that the functional coupling between nodes is correct. He felt that the pathways would migrate substantially as more information is collected. GO also has its problems, but it does not imply causality, it only implies functionality. He said that using only pathways, which are a poor representation of the complexity of biology, would lead to a conclusion that pathways are not accurate enough to give a true reflection of what operates more as a highly stochastic system. Dr. Draghici largely agreed with Dr. Stevens but noted that GO has also been through many revisions. He felt that the pathways approach will also evolve in time. He noted that he was not advocating that it should be used alone but that it should be used together with GO and the other existing tools.

Dr. Wright said that from a purely statistical perspective, the use of gene sets as opposed to pathways might offer some benefit in terms of averaging. In the quest to make the correspondence between transcriptional POD and apical POD as tight as possible, he asked Dr. Draghici if paying more attention to pathway structure could lead to improvement, or would it lead in a different direction? Dr. Draghici said it was a very interesting question, and the answer would depend on the research goal. If the goal is simply to seek a threshold and answers to binary questions such as whether a certain chemical is dangerously toxic, then GO analysis only would suffice. However, if more information is being sought, understanding how the various genes interact would be crucial. Dr. Auerbach explained that Dr. Draghici's presentation was illustrating some of the tools that would be needed as there is more evolution toward risk assessment.

Dr. Edwards said he liked what Dr. Draghici had done with pathways. He noted that a possible compromise would be to use the pathways to identify the gene sets that might

be informative of that pathway, which would allow use of gene expression data to identify the groups of genes or profiles. It would be a hybrid approach that would lead to improvement in identifying informative gene sets and would allow for detection of phenomena that may be important to a pathway but are not transcriptionally regulated.

B. Deriving Points of Departure Using Toxicogenomics for Chemical Risk Assessment

Dr. Andrew Williams, a biostatistician from Health Canada, briefed the panel on strategies for deriving PODs using toxicogenomics for chemical risk assessment. He presented data comparing traditional approaches and toxicogenomics to inform mode of action and PODs in a risk assessment drinking water study. The approaches in the study reached similar conclusions. He plotted the study on Dr. Thomas' graph of temporal concordance of apical and transcriptional PODs. Another study he depicted revealed that the more stringent the filtering, the more the BMD distribution was driven down. His group also compared eleven different approaches to group genes for derivation of a POD to apical PODs for previously published work on six chemicals sampled at four time points. They found the eleven different approaches gave PODs that were all within 10-fold of apical PODs. He described how his group analyzed the data presented in the Dunnick et al 2017 paper. For one chemical, one of the methods yielded a 13-fold range, and would probably not be used. For the other chemical, the methods yielded PODs within 10-fold of each other. In playing with the thresholds, they found that as more information was included, the median BMDs increased. However, the range was not large. He endorsed the use of gene sets.

He cited another 2017 Thomas paper which stated that BMD values from GSEA-identified genes and the most sensitive biologically enriched pathways were shown to be good predictors of the most sensitive apical response BMD values. Dr. Williams described other case studies illustrating his points.

He recommended the following:

- Use of significant gene sets, pathways and/or signatures over individual genes.
- Modeling composite scores
    - Modeling the GSEA score
    - First principle component
    - Cumulative Expression Differences
- Filtering
    - Significant gene sets, pathways and/or signatures
    - MAQC (unadjusted p-value <0.05 and 1.5-fold change cut-off)

B.1. Questions for Clarification

Dr. Auerbach asked Dr. Williams which of the methods consistently produced the most sensitive outcome. Dr. Williams replied that it was the lowest pathway BMD, with a 5-gene filter.

Dr. Clewell noted that regardless of what conditions had been chosen, the methods were within a factor of 10, which Dr. Williams confirmed. Dr. Clewell asked if the same sort of analysis was performed with different types of gene sets, and Dr. Williams replied that the gene set definitions were not changed. He noted that Dr. Auerbach had data where he had looked at changing the different databases and had found that, with one exception, they were not very different. Dr. Auerbach added that generally, what determines how low a gene set goes is determined by coverage of biological space.

Dr. Stevens asked if an analysis performed just based on differential gene expression would yield the same answer as an analysis performed with pathway analysis? If so, one could just map differentially expressed genes and look for the enrichment. He said the percentage of differentially expressed genes has a very high effect size. Dr. Auerbach replied that one could use very strict, stringent standards for filtering genes, which would yield perfect curves, and then the lowest one could be chosen. That way, it would be based on perfectly-fit data. That approach would avoid much of the biological interpretation. Dr. Stevens pointed out they could then look at a database (e.g., GO biological processes) and find ones that showed a significant enrichment by some criteria and ask how often they deviate from the BMD calculation just based on the percentage of differentially expressed genes. His expectation would be that since they are a subset of the larger set, they would probably be within a reasonable approximation of what would be seen in the larger set, unless there is a unique pathway that is exquisitely sensitive to a particular biological event.

C. NTP's Proposed Approach to Estimating Gene Set Level Potencies

Dr. Auerbach presented the third step in the proposed NTP analysis pipeline, describing how fitted features are transformed into genes, parsed into pre-defined gene sets, and used to determine gene set potency values.

He noted that for a feature to be considered for gene set analysis, it must pass an additional set of filters to ensure the adequacy of the curve fits. Specifically, for a feature to be considered, its best model must:

- Have converged BMD, BMD_L and BMD_U values
- Not map to more than one gene
- Not have a BMD > highest dose
- Have a nominal global goodness of fit p-value >0.0001
- $BMD_L/BMD_U < 40$

He described the derivation of the global goodness of fit p-value, how gene sets are identified as active, and how median BMD values are estimated. A gene set must contain at least 3 genes, be at least 5% populated, and be enriched by Fisher's Exact Test with a $p<0.05$.

C.1. Questions for Clarification

Dr. Yauk asked Dr. Auerbach to discuss how error is represented on the gene sets. He replied that the BMD_L will be provided, but the focus will be on reporting just the BMD,

using the philosophy of picking the pathway with the lowest BMD value. Dr. Yauk clarified that she meant to inquire about representing the confidence intervals of the gene BMDs. Dr. Auerbach said that when looking at the distribution in a gene set, it is looking at the median value within the gene set, within the range of the most sensitive pathway.

Dr. Naciff asked, given that there may be upregulated genes and downregulated genes within a single gene set, how is the most sensitive gene set defined? Dr. Auerbach said that at this point, the directionality is not being considered.

Dr. Stevens asked Dr. Auerbach to define if the approach were agnostic or agglomerative. Dr. Auerbach replied that it is a partial agglomerative. There are pre-defined sets, and a feature must pass all the filter criteria to end up in one of the gene sets. Each gene set then provides one agglomerative BMD value, he added.

Dr. Wright asked about the example Dr. Auerbach had shown. In it, the single gene that had the median BMD also had middle BMD_L and BMD_U. For the actual calculation of the BMD_L for the pathway, he wanted to know if it used the median of the BMD_L column, because it would not necessarily match to the same gene. Dr. Auerbach agreed that sometimes it would not and indicated it is using the median of the BMD_L column rather than using the BMD_L associated with the median BMD. Dr. Wright asked Dr. Auerbach if there were recommendations about how to handle the gene sets in situations where there may be extrapolation from a smaller set of genes that are interrogated. Dr. Auerbach said that the optimization still needs to be worked out for that component. They discussed use of the S1500+ gene set.

C.2. Public Comments (*ad hoc*)

Dr. Peddada asked Dr. Auerbach if use of a weighted average had been considered. Dr. Auerbach said that there is a weighted calculation incorporated into BMDExpress that weights the fit quality, and that has been considered as the approach metric (as opposed to the currently-proposed metric of median). It has not been explored much, but is definitely a possibility, he added.

Dr. Gift said that in circumstances where the dose-response is shallow, he could imagine that the BMD_U would be many-fold higher than the BMD, with the BMD_L lower than the BMD. In such a scenario, the 40-fold BMD_L/BMD range would be skewed by the very high BMD_U. He recommended that Dr. Auerbach look at the sensitivity of using a BMD/BMDL ratio as opposed to a BMD_U/BMD_L ratio. Dr. Auerbach mentioned that he had looked at that in some of the datasets, and the two different methods yielded virtually identical results.

Dr. Bushel asked Dr. Auerbach what is done if there are several significant gene sets, when the BMD is determined from the most significantly enriched pathway. How is the median value determined when the estimates are fairly close to each other? Dr. Auerbach replied that each gene set is ranked by its median BMD, with the one with the lowest median BMD being reported.

C.3. Panel Discussion

Dr. Stevens was the first panel reviewer. In terms of strengths, he felt that linking the estimates of BMD from the genes to the biology is critical. There are standard methods proposed for deriving gene set significance, and the redundancy problem has been addressed. There are also limitations, he said. He felt that the redundancy problem was being created by the way NTP was going about it. He would have been more comfortable with some examples to work through as part of the proposal. He said that to judge how adequate the approach is, it would need to be seen relative to some other methodologies. He could not separate the NTP proposed approach from the Hallmark gene sets. He felt that it was inappropriate to say that a biological effect is being adequately estimated without specifying the biological effect. He suggested that by applying the method to different types of approaches such as pathway, GO, or Hallmark, one can ask whether one arrives at the same conclusion, one can make arguments about what is being missed, and one can suggest whether they represent all of the biology needed. Reducing redundancy too much may oversimplify complexity, he observed. He felt that it would have been good to run some of the alternative approaches to generate preliminary data. It does make sense to perform the calculations on existing data sets rather than launching new data sets from new experiments, he said. He found the charge question regarding "comment on how a gene set-level POD should be determined" asked panelists to exceed the scope of the project. Assigning a mode of action to a gene set implies risk assessment, he noted.

He encouraged NTP to take on something other than nuclear hormone receptors, such as, for example, fibrosis or cholestasis, conditions offering non-receptor-mediated, complex biology. There is a risk that some people will not want to go with just the biology and will want to look at the genes. Asking such questions will depend on whether there is a disconnect between the PODs of the individual genes versus the method being employed. He felt that that question had been answered adequately. Dr. Auerbach said some examples would be added to the document. He said that a comparison of methods would be conducted using the S1500+ data and would be included. He agreed that the Hallmarks are limited, and the question becomes whether to go through the curation effort when there are so many other efforts to curate gene sets. He noted that the beauty of the Hallmark gene sets is that they have already gone through a degree of redux. Dr. Stevens agreed and suggested NTP should not get bogged down in doing a lot of curation. However, he noted, the Hallmark gene sets are pretty new, whereas GO is well-established. He likes GO better, because it only implies function, not cause and effect relationships, leaving cause and effect determination to risk assessment. He wondered how well the Hallmark gene sets would be accepted and whether they would be seen as encompassing the scope of biology toxicologists are interested in as opposed to being more focused on therapeutics. Dr. Auerbach replied that they are the latter and represented a bridge into a toxicological interpretation. The gene set can selectively be built out as necessary, he observed.

Dr. Burgoon was the second panel reviewer. He started his comments with the third charge question, "Should the pathway be associated with a known toxicity mode of action?" He wondered if that approach was going down the road toward risk

assessment. As to whether the most sensitive or most enriched pathway should be used, he said he leaned toward sensitive, coming from his public health perspective. He noted that he takes issue with the use of the Fisher's Exact Test for determining significant gene set enrichment. He said the evidence shows it is not needed. He addressed the issue of p-value and his experience he had never seen it approach the proposed threshold in gene expression data, although he had fewer issues with the model fit approach than he did when he had first written his preliminary comments. Dr. Auerbach asked Dr. Burgoon what sort of thresholds on the global goodness of fit p-value he would suggest. Dr. Burgoon answered that he is more comfortable with 0.1 or 0.05.

Dr. Auerbach asked what alternative would be suggested to replace the Fisher's Exact Test. Dr. Burgoon observed that his group does not use it and generally uses 3 or 5 genes, as long as they are in the pathway.

Dr. Wright said that even though he had been critical of using the Fisher's Exact Test in other contexts, he felt compelled to defend it. He noted that if enrichment is to be performed for higher reproducibility, then resampling methods should be considered. He said that if an overall pathway BMD_L is desired, bootstrapping could be used to determine a value that can be trusted. Regarding the goodness of fit testing, he wondered if an R^2 value could be used instead of a p value. In the case of an enormous sample size, genes will be lost, having failed the goodness of fit test. He said he liked the median as a measure of central tendency for PODs. Regarding the issue that things might be missed with the Hallmark gene set approach, he wondered if the minimum median could simultaneously be used, while keeping the concept that a number of genes are being affected by the chemical, albeit at a very low concentration. Dr. Auerbach responded that the resampling-based approach seemed reasonable, and he would try it. He agreed that an R^2 value might be more effective than filtering through fits. He agreed that the Hallmark gene sets do have a coverage issue at this point, but capturing a secondary metric might be a reasonable approach.

Dr. Naciff asked Dr. Auerbach how the most significant genes were selected. Dr. Auerbach said that once the analysis has been performed, there is a collection of BMD values for each feature. The BMD values are ranked to determine the most sensitive 10. Then they compare the median BMD value for the lowest 10 and the BMD value for the most sensitive pathway; if they are notably different, then it is possible that some biology may not have been captured in the gene sets.

Dr. Draghici said he would definitely use re-sampling instead of the Fisher's Exact Test. He noted that biological outcomes are needed, and gene-set level potencies and biological interpretations overlap. He asked if it is really necessary to establish a set of numbers called potencies of the gene sets, or is it better to go directly to the biological interpretation. What is really important is whether a particular substance is dangerous or not, and at what level. Regarding the issue of 3 genes versus 2 genes, Dr. Auerbach speculated that it is a rare occurrence when two critical genes respond while other genes in the pathway do not. He said he was fine with the re-sampling approaches. He noted that the challenge is the overarching goal of being either protective or predictive.

He said the goal is to be protective right now, and he cautioned about letting biological interpretation drive the initiative at present. In the future, he noted, the direction would be to evaluate the use of GDR for risk assessment.

Dr. Bucher said that in a GDR modeling effort, it does not matter if a substance is toxic or not. What matters is (1) the distance between the dose that causes biological activity and (2) environmental conditions at which humans would be exposed to the chemical along with the kinetics for the chemical to get to a tissue of interest.

Dr. Stevens thought it should not be hard to get people to buy in that it's a conservative approach, per biological responsiveness. He recommended that NTP drop the 3-gene requirement. The 5% populated with included statistical significance would be sufficient, he felt. If the 3 genes filter were dropped and the filter metric looking at the significance for enrichment were increased, he wondered whether it would result in just as good a filter without penalizing small gene sets for having fewer genes. Dr. Auerbach asked if that filter would imply one gene would qualify in some of the smaller gene sets; i.e., would one gene pass that threshold and result in a pathway? Dr. Stevens said that it could, but then it would be a matter of what is done in the risk assessment with the information. It is unlikely that the risk assessment would be made based on just the one single gene, without other accompanying information.

Dr. Draghici said he would recommend keeping the 3-gene requirement. In his experience with GO analysis, every time there are very few genes in a set, the statistics are not reliable.

Dr. Wright favored the 3-gene minimum. His concern overall was more statistical than biological. He noted that if testing was to be done, then re-sampling could include permutation testing; if the goal is confidence levels, bootstrapping must be done.

Dr. Johnson agreed that the number of genes should be more than one.

Dr. Stevens retracted his suggestion regarding fewer than 3 genes.

Dr. Yauk asked whether 3 genes were enough, or might 5 be better.

Dr. Williams agreed with the 3-gene requirement, particularly with the other conditions and constraints being included.

Dr. Stevens noted that the 3-gene requirement is arbitrary. He questioned cases where 3 genes does not constitute 5%. He speculated that dropping the 3-gene requirement would make no difference if the 5% cutoff is in place. Dr. Auerbach said that in small gene sets, it does make a difference. He noted that in GO, there are gene sets of 2, and if the 5% condition replaces the 3 genes condition, there would be many single-gene GO terms.

Dr. Wright commented that when dealing with a restricted set of genes such as the S1500+ set, there can be gene sets for the entire transcriptome that have 15 or 20 members, of which only 5 or fewer are actually present among the analyzed sets. He recommended maintaining awareness of the effects of filters on these restricted sets.

Dr. Auerbach acknowledged that 3 genes and 5% was arbitrary but noted that 60 genes would also be arbitrary. He said any suggestions of another approach would be appreciated.

Dr. Stevens said he had concerns about the S1500+ not having been tested in this kind of approach. He understood its original rationale, but costs are coming down so rapidly, it is practical to analyze the entire transcriptome. He questioned how much biology is captured in the S1500+. Dr. Auerbach said he understands the concern but there is an extrapolation engine that will fit curves to the entire genome. Dr. Stevens said that putting the S1500+ set in against the entire genome would result in finding enhanced enrichment because it has been selected for things that are biologically significant. Dr. Auerbach agreed but noted that all of the enrichment is adjusted against what was measured on the platform.

VII. Session V: Study Design

A. Improving Study Designs for Quantifying Biological Potency with Genomics Data

Dr. R. Woodrow Setzer from the National Center for Computational Toxicology briefed the panel on strategies for improving study designs for quantifying biological potency with genomics data.

In this instance, “design” signifies the number of dose or concentration groups, what concentrations to use, and how to distribute replicates among doses. Dr. Setzer pointed out that resource and structural constraints would limit some or all of those elements. He described the design considerations of GDR features:

- Most curves are likely to be sigmoidal (approximated by a Hill model) but can be non-monotonic, mainly at high doses.
- Thousands of endpoints (genes) are tested, which implies a more complicated decision compared with a chronic bioassay
- For a chemical, the design should function well over the full range of:
    - gene-specific potencies (e.g., BMDs)
    - gene-specific dose-response shapes (e.g., power parameter, limiting fold change)

Dr. Setzer presented and discussed a variety of conceptual tools for evaluating experimental design, from statistical theory and from simulations. He described the classical toxicology design, with modifications for dose-response and BMD estimation. He also detailed the optimal design for Hill dose-response. His conclusions included:

- Focusing on the dose considerations leads to designs with more dose levels and fewer replicates per dose.
- Practical designs will have multiple dose levels, log-spaced, and evenly weighted.
- Dose spacing should depend on the range of doses and the steepness of the curve.

- The lower end of the dose range is critical for risk assessment, and the dose design will depend on weighing considerations of coverage of low doses, dose spacing, and cost.
- Both simulation and theory should jointly inform designs used.

A.1. Questions for Clarification

Dr. Johnson asked Dr. Setzer if he would choose doses around the POD if it was known ahead of time. Dr. Setzer pointed out that the POD is not generally know prior to the experiment and instead forms the motivation for doing the experiment. Dr. Johnson noted that typically molecules do not demonstrate toxicity at doses less than 1 mg/kg/day. With that in mind, he proposed to Dr. Setzer a hypothetical study design. He asked if Dr. Setzer would choose doses such as 1 or 10 or 100 mg/kg/day up to the maximum tolerated dose (MTD). Dr. Setzer asked for further information regarding the POD. Dr. Johnson replied it would be the BMD based on 10% change from the mean of the control group. Dr. Setzer said he probably would not select those doses. He said he would do log-spaced doses, would consider resources regarding how much replication could be used, and would use 6-10 dose groups.

Dr. Wright cited the data Dr. Setzer had presented on 44 chemicals. Dr. Wright speculated that at this point Dr. Setzer might have feelings about how the experiments could have been done differently or how an experiment would be designed if there was no prior information about the chemical but using lessons learned from the 44 chemicals. He asked Dr. Setzer if his recommended approach boiled down to using the traditional standard practice but shifting toward using more doses and fewer replicates. Dr. Setzer confirmed that that was his sense. However, he added, it may well be that log-spacing is not the right scale.

B. NTP's Proposed Approach to Study Design for Genomic Dose-Response Modeling

Dr. Auerbach presented NTP's proposed approach to study design for GDR modeling.

He noted that:

- Traditional toxicity assessments are designed/powered for pairwise statistical analysis with the goal of identifying No Observed Effect Levels.
- That approach is often not conducive to applying a dose-response modeling approach such as BMD.
- For GDR studies, NTP proposes to use a BMD-focused study design:
    - More dose levels and fewer biological replicates
    - Example design: 10-12 dose levels, 3 biological replicates/dose group

This proposed study design will allow for better coverage of the numerous dose-response relationships in each study, more confident fits of the data, and greater certainty in the BMD estimates for the features.

In vivo studies would use male Sprague Dawley rats, would last for 5 days (i.e., 5 doses, 1 per day), would target the liver and expert-selected target organs, and the top dose selection would be the 5-day MTD.

For *in vitro* studies, the parameters would be human cells, with the sex determined by availability. Duration would be determined by experts, with a goal to employ a timepoint that maximizes response to the test article. The cell types would be commonly used organotypic cultures and would look across multiple organs, covering a broad array of biological space. The top dose would be one that clearly challenges the cell system to produce a response to the test article. It was proposed that a lethal concentration 20 (LC20) would be used as the top dose.

B.1. Questions for Clarification

Dr. Clewell asked Dr. Auerbach to elaborate on plans for an MTD preliminary *in vivo* study in terms of the number of doses and replicates and whether any other phenotypic endpoints would be considered. Dr. Auerbach said the intention had been to follow the protocol done by the Iconix Group, which created Drug Matrix: evaluating the LD_{50}, typically taking a 50% reduction in the LD_{50}, and dosing with small decrements in doses to see where there is an effect. It is important to make sure that there is a dose with a clear toxicological effect to ensure there is a clear response at the top dose.

B.2. Public Comments (*ad hoc*)

Dr. Paules asked Dr. Setzer about the lessons learned he had presented with data on the 44 chemicals. He wondered if Dr. Setzer had bracketed the doses around an AC_{50} or an LC_{20}, or whether he had given all of the chemicals the same dose range and dose spacing. Dr. Setzer noted that it was a pilot study for a much larger high-throughput design. The dosing started at 100 micromolar as long as it was soluble, and then it was half-log spacing down from that, for every chemical. Dr. Paules asked how well that had worked with the 44 chemicals. Dr. Setzer said the data was still being analyzed, but from the standpoint of potentially missing some dose-response trends, it looked like a few had been missed. He said that analysis for very potent genes has not yet been completed.

Dr. Paules said that the depth of the reads and the number of reads can contribute to how much information is gained. He noted that in the S1500+ gene set, which is comprised of approximately 3000 genes, the target is an average read of about 500 mapable reads per gene, or about 1.5 million reads per sample. He said that the EPA approach is about 5 million reads per sample, which are then attenuated down to about 3 million reads per sample. Using RNASeq, that would be an equivalent to 30-50 million reads per sample. Going lower would compromise the ability to interpret some of the responses. He asked Dr. Setzer if he had any experience with that issue. Dr. Setzer said his group starts with about 20,000 genes in the gene set. They prefilter before doing any analysis, only including genes for any chemical where the average count is at least 5. That generally throws out almost half of the genes.

Dr. Alison Harrill from NTP commented on points that had been brought up in discussions regarding the *in vitro* and *in vivo* study designs. First of all, she said, the chemical being studied needs to be taken into account, potentially including the compound's pharmacokinetics since they might influence temporal measurements. She also asked the panel to consider the inclusion of the female sex in some investigations.

Dr. Shockley noted that with 3 replicates at each concentration, there is effectively a much larger sample in total and more power. Dr. Bucher said that works well for continuous data but not for some of the pathology and incidence data that is used.

B.3. Panel Discussion

Dr. Clewell was the first panel reviewer for study design. She said it was important when considering study design to talk about purpose and it was unclear what the purposes of the different studies were in the NTP proposed approach. Apparently, the *in vivo* studies are more about trying to set a preliminary POD that can be used for prioritization of further testing, she noted, as opposed to building the database and using the testing as a pilot to improve methods. She felt that the *in vitro* side was more about developing methods, tools, and technologies and beginning to answer the question of how to move the field toward *in vitro* approaches. For the *in vivo* studies, she said she agreed with the idea of reducing the number of animals per dose while increasing the number of doses. She endorsed the half-log scale and the idea of 3 replicates. She recommended considering extra replicates for the control group.

She wondered why the study was limited to the male rat, while recognizing that the ability to compare to historical male data would lend overall power to the conclusions. However, she did not believe that deficiencies with historical study design should be propagated. She also pointed out that the question of developmental toxicity had not been addressed. Using females would allow the ability to compare fetal toxic effect with maternal effect. That would be a good first step toward understanding developmental toxicity. She suggested using pharmacokinetic (PK) data, or at least PK simulations, to predict the time to reach steady state internal concentrations. She said that in general, she is a proponent of bringing metabolism and dosimetry into consideration earlier before decisions are made. She agreed with using the liver as a sentinel tissue, but if this approach is going to be a pilot for how to proceed in the future, efforts should not stop with just the liver and expected target tissue. She suggested looking at the kidneys, for example.

Regarding the *in vitro* aspects of the approach, she asked how it would be decided what is the appropriate model. She thought it would be great to start with the HepaRG model, but it would be good to establish some collaborations with people using other liver models, including primary models, and then compare those results to the *in vivo* results to see how organotypic the cells need to be to get the gene. She suggested setting a goal for the tissues thought to be most important to begin with along with a plan for developing them, depending on whatever technology is developed. So, for example, if the kidney was considered to be a prime target to develop next, it should be put out to the community to allow people to start developing it. She said she would like

to see some discussion about what those targets might be—certainly neurotoxicity would be high on the list.

She endorsed incorporating metabolism much earlier in the process. She suggested using our understanding of metabolism early on to define ranges of exposure, or chemical concentration, that are likely to be seen *in vivo* to benchmark those around expected *in vivo* response. One of the most important conversations moving forward, she observed, will be about how to ensure that *in vitro* dosing is consistent with or even relevant to the *in vivo* system.

Dr. Auerbach responded to Dr. Clewell's comments. He agreed that there is a need to better define the purpose of the project in the document. He said NTP had been thinking about the issue surrounding the sole use of male rats in the experiments. He felt that using male rats and female mice might be a compromise, although that would double the cost of a screening-level study. When the proposed NTP approach was first put out as a screening-level assessment of potency, the original focus was in fact on male rats because it was not meant to be a guideline-level type of assessment. One question would be what the added value of adding females would be from a potency standpoint. In most cases, the offset in potency from male to female, at least within the same species, would not be significant. In terms of adding development studies, an animal welfare issue arises.

Dr. Auerbach mentioned that Dr. DeVito was conducting some comparison studies looking at developmental exposures and 5-day exposures, to see how far off the estimates of potency are. Dr. Auerbach noted that in most cases, in his experience, the cancer bioassay pathology is typically more sensitive than the developmental toxicities, although there certainly can be some exceptions. He said that he would accept female-only studies and observed that the focus on males results largely from legacy work, with most of the databases built on male rats. He felt that Dr. Clewell's suggestion to collect more organs than just the liver was not unreasonable, although pathological assessment might not be valuable in a 5-day study.

Dr. Clewell added that perhaps it could be crowd-sourced, being aware that her suggestion would entail spending much more money. She pointed out that it would save money in the long term to collect the samples now. She said that many would be interested to know the minimum number of tissues needed to have a good idea of general systemic toxicity.

Dr. Auerbach said it would be important to have a suite of models when looking for potential offset potencies across many tissues. He cited work currently being done to develop "organs in a dish" models. Regarding metabolism, he said that the chemistry is very expensive, and developing methods to do kinetics would double the cost of the screening studies. It would depend on the target, he added. He agreed with Dr. Clewell's assertion that the animal samples should be taken when the internal concentrations reach steady state. He pushed back a bit on the concept of human exposure-relevant dosing. He said dosing in that range could be done, but if higher dosing is not used, little or nothing would be seen in terms of biological potency.

Dr. Clewell responded that chemical-specific chemistry would not necessarily need to be performed to bring metabolism into the approach. She agreed that the higher doses are necessary *in vitro*, beyond human-relevant exposures. She said she was advocating doing IVIVE and kinetics ahead of time to help inform *in vivo* work.

Dr. Yauk asked about the time point issue and whether signal would be lost by waiting until 24 hours after exposure. She felt that all on the panel were prepared to accept the concept of more doses and fewer replicates for this application.

Dr. Stevens said that if the intent is *in vivo* screening studies, a 4- or 5-day study would work, even though it would miss the point in an adaptive response where there would be the maximum gene expression. If the idea is both a way to screen and to pilot for risk assessment, he did not see how a cause-and-effect relationship could be established without having multiple time points. He cited practices in pharmaceutical work. He felt that the value of time points in getting to real risk assessment was under-appreciated in the NTP document. Dr. Auerbach noted that tolerance for change in effect is lower in pharmaceutical experiments than in chemical risk assessment. Dr. Stevens agreed but said the more important value is to establish cause-and-effect relationships in a risk assessment. If there is only one time point, it would be difficult to perform that type of risk assessment, he observed.

Dr. Bucher said that given the fact that most of the environmental toxicology risk assessment values are set on 90-day studies with phenotypic outcomes, the next phase of the research means making the linkage Dr. Stevens described. Dr. Stevens said he was reluctant to assume that a future direction would be pursued. He said the panel had been asked to evaluate a document that laid out a strategy without adequately providing the context of applying the strategy. He said he would feel more comfortable if the issue he had described was acknowledged and taken into account in the document.

Dr. Naciff seconded the prior comments about time, particularly if there is a desire to eventually use the approach for risk assessment. If so, he said, more time points would definitely be needed. He noted that for new chemicals when there is a dose range-finding study, both sexes should be included using the findings to determine if there is a difference in gender; if there is, then the more sensitive sex should be chosen for the next step.

Dr. Johnson said he felt that the tissue selection is appropriate. He advocated collecting all of the organs at necropsy to save them as contingency sentinels. Blood should also be collected and saved for possible metabolite analysis later. He noted that as a screening study, animal use and welfare should be considered, minimizing the experiment to the extent possible while answering the relevant questions. He said he liked the idea of using modeling to predict *in vivo* internal dose or some *in vitro* concentrations. He described his group's experimental design approach to establish a maximum tolerated dose (MTD).

Dr. Clewell asked why a study would need to be repeated after establishing an MTD and why it would not suffice in and of itself. Dr. Auerbach replied that the dosing study

is used to try to find a no-biological-effect level and would not work for estimating biological potency.

Dr. Auerbach asked Dr. Johnson to elaborate on his description of arriving at dose levels. Dr. Johnson said the important consideration was not the number of doses, but the number of animals chosen for the entire study.

Recapping the day's proceedings, Dr. Yauk said that the sense of the panel was a wish to see non-parametric models introduced as soon as possible, and that they should be integrated with BMDExpress to help build confidence and experience. There was reluctance to reduce the number of models. Overall, the panel seemed relatively happy with the approach, including the BMR, she noted. The objective is to look at the dose at which biological effects occur, with the foresight that someday, there will be movement toward risk assessment. It seemed that most panel members favored dropping the Fisher's Exact Test, she observed. The panel seemed focused primarily on gene sets or gene groups or pathways versus individual genes. Generally, 3 genes seemed acceptable to the group. Panel members agreed with the concept of more doses and fewer replicates. They agreed that both sexes should be included, and endorsed the idea of a pilot study to identify the more sensitive sex. PK data should be considered early in dose setting, and more consideration should be given to the time point selection. The group discussed the opportunity for others to work on tissues that NTP may not initially use, with primary unused tissues being banked.

Day 3: October 25, 2017

VIII. Session VI: Biological Interpretation

A. Using the AOP Framework to Aid in Gene Set Identification

Dr. Edwards presented material to the panel on how to use the AOP framework to aid in gene set identification.

He said it was important to appreciate that AOPs are not just the end game, but are a framework for organizing whatever information is on hand. In the use of AOPs to connect toxicity pathways to regulatory endpoints, he noted that the key element to remember with AOPs is that the community is trying to be systematic about how to gather that information and organize and translate that information to other people.

He depicted the factors determining predictivity of early key events. He said that one advantage of the AOP framework is that it is transparent about when there is a lack of data at any particular stage. He noted that quantitative understanding can exist at different levels, but that it is not necessary to use the AOP information to tie the dose-response to an apical endpoint. The same approach holds true for modifying factors.

He depicted the flow from toxicants through toxicity pathways to regulatory endpoints and emphasized that the AOP provides a scaffold for all data, including high-throughput toxicology, GDR, more traditional toxicology data, and epidemiology. He said it would be important to start thinking about how such different types of data can be treated in a similar manner within the AOP framework.

Regarding the criticism that AOPs are too simple and do not represent the complexity inherent in biology, Dr. Edwards described AOP networks and how these will emerge as key event are entered into the databases by multiple users. The AOPs will intersect, the AOPs will form networks, and more complex biology will emerge out of this piece-by-piece approach. He said that his key point was to emphasize that AOPs are useful no matter how much information you have. However, they are more robust given more information. More people adding key events will increase confidence in the entire process, he noted. He described ongoing work to automate the generation of computationally-predicted AOPs (cpAOPs) and provided examples of those studies to automate extraction of subnetworks, including a depiction of data illustrating a cpAOP network for fatty liver disease extracted from ToxCast, CTD, and TG-Gates data.

Dr. Edwards said that the end goal for the work is to determine whether large data sets can be programmatically mined to find key event while keeping track of the individual genes and the individual pathways.

A.1. Questions for Clarification

Dr. Bucher asked Dr. Edwards how the Bradford Hill guidelines for causality are incorporated into the development of the AOPs. Dr. Edwards replied that for the OECD-endorsed AOPs, the Bradford Hill considerations are the basis for the evidence evaluation, although they were modified slightly because AOPs are chemically agnostic. He went into more detail about the role of key event relationships in the AOP framework. Dr. Bucher observed that the fact that the approach emphasizes doing genomic assessments on a developing lesion over time will assist in the agnostic data mining aspects of AOP development. Dr. Edwards confirmed that assertion.

Dr. Auerbach asked Dr. Edwards whether it was true that in order to build out AOPs adequately, there would need to be much phenotypic anchoring in studies, or at least a small subset in different tissues. Dr. Edwards confirmed that would be needed eventually, but even hypothetical associations can be labeled as AOPs, albeit with very low confidence. Such an AOP would still be useful, but by the same token, if there is concern about the confidence, the AOP should continue to be built out. Over time as more empirical data are available, there will be increased confidence regarding the areas of the most concern.

Dr. Burgoon asked if it would be possible for people to just enter key events into the wiki and not create an AOP wholesale. Dr. Edwards said that there is a button on every page to create key events, key event relationships, and stressors, with nothing needing to be tied to an AOP. The hope is that people will contribute what they know, and an AOP can be assembled later.

Dr. Stevens asked if there are unbiased ways to say how far the strategy can be extended, in terms of knowing when the applicability domain has been exceeded by the attempt to extract information and apply it to a new molecule that may be more complex. Dr. Edwards addressed the question in detail. He and Dr. Stevens discussed the concept of species applicability related to AOPs.

Dr. Johnson agreed that there is much data in the liver and asked about the current ability to predict hazard in an organ like the liver that has a lot of data. Dr. Edwards replied that today, there is more work to do to understand the existing data. He noted that the further down the road to risk assessment, the more scrutiny there will be. However, he added, he believes the community is close to being able to say something about what the downstream outcomes would be.

Dr. Bucher observed that the approach seemed to hinge on ability to measure the probabilities between each of the steps in the AOP, specifically, evaluating how many times in a particular data set the association is seen versus how many times in a comparable data set it is not. That would allow pursuing the predictive road in a more confident way, he said. Dr. Edwards agreed and stated that weighted edges are being built in computational work, generating an iterative process of evaluating edge weights as opposed to simply the structure of the network.

B. Application of Weighted Gene Co-Expression Network Analysis (WGCNA) to Dose-Response Analysis

Dr. Stevens from Eli Lilly briefed the panel on techniques to improve interpretation of nonclinical results using modularity to reduce complexity without loss of biological information.

He discussed multi-scale complexity, depicting the concept of multi-scale modeling of pathophysiology. His talk addressed in detail:

- Modeling biological complexity
    - The modular nature of complex systems
    - Leveraging modular systems models using gene expression data
    - Translating gene expression data into biological understanding
    - Reducing redundancy in MSiqDB
    - Knowing what we don't know
- Understanding molecular pathogenesis
    - Correlating expression modules with pathology
    - Closing the loop from transcription factor to pathogenesis
    - Predicting adaptive vs. progressive responses
    - Closing the loop on transcriptional control
- Applications of WGCNA to dose-response analysis
    - Separating injury signals from tissue stereotypic response
    - Perturbing network in culture
    - Translation to human

Dr. Stevens reiterated his concern about the incorporation of time into the NTP model.

B.1. Questions for Clarification

Regarding the data Dr. Stevens had presented, Dr. Bucher said it had been noticed that it dealt with a homogeneous population, and NTP was recently interested in understanding variable susceptibility issues using tools such as the Diversity Outbred

mouse. He wondered if there would be value in testing the concept Dr. Stevens had presented in such a system to tease out whether the initial events are occurring in all of the population. This would help determine if the conversion to a pathological condition is more a function of the genetic susceptibility of the animal rather than inherent ability of the initial event to create a pathology. Dr. Stevens said that his group is actually doing an experiment right now along the lines Dr. Bucher described. He added that it is an idea he would love to pursue further. Dr. Bucher said that the chemicals that fail in predicting POD for a later apical outcome in the 5-day studies could be put into a longer-term study with a sequential genome assessment.

Dr. Johnson said he understood and appreciated that the modules change over time but asked if the dose-response relationship overall changes over time. Dr. Stevens said that the slope of the curve changes over time with the adaptive response, while the slope of the curve increases with the progressive response. He noted that after one day of exposure, the system is most likely to show a shift from initial conditions. At 29 days, the dose-response behavior remains, but the slope is much shallower. Dr. Johnson asked whether Dr. Stevens thought the POD changes over time. He said he had suspected that it did, but based on data presented by Dr. Auerbach, apparently it did not. Drs. Stevens, Johnson, Naciff, and Auerbach discussed the time series issues in detail.

Dr. Clewell said her group has also seen gene signatures at later time points and higher doses, resembling what Dr. Stevens had characterized as a tissue stereotypic effect, converging into a phenotypic response that is not dependent on the initial molecular event. Earlier in time and lower in dose, the patterns emerge that are more specific to the chemical itself. Dr. Stevens said he felt that there was value in the approach he presented and that he was in the process of suggesting that it be added into the toolbox of how to use gene expression data. He noted that the effort is highly transparent, and by putting the tools into the public domain, it is designed to encourage community acceptance. This acceptance will ultimately aid eventual regulatory acceptance.

Dr. Edwards said that Dr. Stevens could interpret the events he had described as an AOP. Dr. Stevens replied that that would meet the dose and time elements of Bradford Hill, but it would be unlikely to be used by others. Dr. Edwards reiterated that AOPs are all along the continuum of development and may not be used, but nonetheless they all offer value as a construct. Dr. Stevens said he would like to see a movement toward using Bayesian network theory to assign a probability to the edge.

C. NTP's Proposed Approach to Biological Interpretation of Genomic Dose-Response Results

Dr. Auerbach described NTP's plans for developing a standardized biological interpretation of GDR data, which he described as the least developed component of the proposed analysis pipeline. Historically, biological interpretation of data has been done by using gene ontologies or pathways and, in certain cases, signatures or co-regulation modules. The challenges with many of those gene sets are redundancy, incomplete coverage of gene space, and only partial congruency with gene expression data.

He described the advantages of the Hallmark gene sets:

- Limited redundancy
- High percentage of each gene set is regulated at the level of transcript abundance
- Empirically validated/curated
- Challenges include limited gene coverage and no toxicological interpretation
- NTP wishes to develop Hallmarks+

Expanding the Hallmark gene sets would overcome its limitations, using a data- and literature-driven approach. The Hallmark discovery process began by performing WGCNA on approximately 130,000 microarray samples from GEO, which were curated by Sciome as part of the S1500+ gene selection process. Once the Hallmark gene sets have been saturated and NTP feels there is adequate coverage of the transcriptome, the gene sets will need to be curated (using Illumina Correlation Engine) in a way that facilitates interpretation of GDR data. Dr. Auerbach provided an example annotation to help illustrate the concept.

In a bid to be provocative, Dr. Auerbach showed a possible term that would depict the elements of the approach based on a particular gene set, with its GO definition. He challenged the panel as to what NTP could conclude based on the data.

Dr. Stevens said that in the case shown (response to DNA damage stimulus), p53 would be a special case, where there should be a fold cutoff. Dr. Auerbach agreed that at some level, in some instances, an automated interpretation would not be appropriate. Dr. Clewell said that the example illustrated her concern overall with using the existing annotated pathways for gene expression sets. She would want the first step, prioritization, to be more centered around an agnostic approach. This would help build the overall understanding so that the data can be used toward understanding mode of action and forming conclusions about hazard. She felt that the NTP approach should not address biological interpretation.

Dr. Naciff responded to Dr. Auerbach's question and said that NTP essentially cannot conclude anything but can mark the results as a point for further studies.

In conclusion, Dr. Auerbach asked for the panel's input regarding whether NTP should include biological interpretation as part of its approach.

C.1. Public Comments (*ad hoc*)

Dr. Mav from Sciome suggested that Dr. Auerbach include area under the curve for the best fit in his example. Dr. Auerbach clarified that in the example, 14% of genes made it all the way through the analysis and into the gene set; he believed that it was substantially populated.

Dr. Harrill said there needs to be some careful thinking about how the pathways are presented in reporting, because, especially in the *in vivo* studies, it is very difficult to separate some of the *in vivo* biology from the gene expression data. She said there had

been considerable discussion on how to integrate the *in vivo* data and the toxicogenomics tools.

Regarding Dr. Auerbach's example slide, Dr. Draghici understood there would be a predefined set of genes for which numbers would be reported, as shown. He thought it might be a bit sub-optimal to use a predefined set of genes, whether Hallmark or enhanced Hallmark. He suggested that another alternative might be to perform the analysis in GO, starting with the lowest terms and calculating a p-value, etc, which would give a custom level of abstraction, providing the most knowledge and understanding for the particular compound. Dr. Auerbach thought it was an excellent suggestion.

Dr. Bushel agreed that it would be very informative to have ancillary information other than the gene expression data. He wondered if it would be possible to use the annotations to describe BMDs, and if it would be possible to include some of the ancillary toxicological information to help bring context to the biology and allow more informative results.

C.2. Panel Discussion on Biological Interpretation

Dr. Yauk opened the panel discussion, noting that this was the least developed part of the NTP's proposal, and also the trickiest. It is an area where NTP will substantially benefit from the panel's input for future research. She asked what the panel's thoughts were on the approach the NTP should be taking.

Dr. Naciff was the panel reviewer for the biological interpretation section. He felt that one of the strengths is the filtering step. He did not agree with not taking p-value cutoff into account, because the results would include thousands of genes with a fold change above 1.5 at a given dose. Overall, he liked that step, especially for interpreting what the gene expression set means in the context of toxicity or biological activity in the cell. He said it is a critical step and one of the strengths of the proposed approach. The second strength is the use of the gene sets. It is a given that more granularity is needed in what exactly every single gene change means in the context of very specific pathways. It is a strength to reduce redundancies to better define what would be the most sensitive pathways, but that is not achievable with the 50 Hallmark genes. Finding the specific sets of genes based on toxicity will be great given the purpose of NTP's work, but an interpretation of biological activity is not needed; providing the BMD is most important. He said biological interpretation will be necessary to understand what happens when the BMD is exceeded. He agreed with previous remarks that the time element is needed in the context of biological interpretation of changes. Additionally, it should be shown that the BMD doesn't change across times, although that may not work for all chemistries. He suggested more examples in the manuscript to aid understandability. Regarding the *in vitro* systems, he wondered how many cell types would be needed to understand the biology. He noted that in his industry at present, there is much demand for *in vitro* testing.

Dr. Auerbach said he had been thinking about how to phenotypically anchor *in vitro* studies. It would need to be a parallelogram approach that relates to the AOP, linking

gene expression to AOP and using organotypic cultures that reflect human biology. There would ultimately need to be linkages to key event processes, as Dr. Edwards had described, with mapping of the gene sets being used to key events. He said that Dr. Naciff's point about time was a good one. He asked, however, at what point there is enough data to conclude the approach is valid and for the panelists to nominate any compounds they think would break the paradigm. He said that need for a p-value would be voted upon, and noted that the panel was setting historical precedent in many aspects. He agreed that more granularity is needed with respect to the Hallmarks.

Dr. Stevens encouraged NTP not to provide biological interpretation as part of the approach. He said that one of the ways to implement the approach poorly is to try to do too much, and going to the Hallmark gene sets implicitly takes on the issue of causality. He noted that NTP can do everything intended with GO biological processes, which would allow the scientific community to buy into the fact that genomics would be used to arrive at a POD, capturing to the extent possible the universe of biological processes. He would recommend not doing biological interpretation, because it would stretch NTP too far.

Dr. Auerbach noted that in the example he had shown, the gene set has a name associated with GO biological processes. He asked Dr. Stevens if there should be no NTP interpretation; that simply the GO definition should be put in, and that would be all. Dr. Stevens replied that GO terms give you function, and pathways give you cause-and-effect relationships. He felt that getting into cause-and-effect relationships would move toward risk assessment, and in that vein, it would rapidly become less of a data-driven discussion of whether the technology works, and more the domain of the political, contextual, biased arguments of how risk assessment is done, which would hinder NTP from making progress in using genomics to arrive at a very defined application. Dr. Auerbach asked Dr. Stevens about the example he had depicted. Dr. Stevens said he did not think that onus could be put on the panel. He felt that the NTP approach should be limited to what NTP could support from a data-driven perspective. Adding cause-and-effect would invite controversy and hinder NTP's effort.

Dr. Bucher asked for comments from other panelists.

Dr. Wright said that some GO definitions do include descriptions (such as DNA damage), so they may not be any less potentially explosive than Hallmark sets. However, the Hallmark set currently only covers about 20% of the transcriptional space. Any of the sets that are proposed would most likely give similar POD calculations as long as they span the biological space well enough. Dr. Auerbach clarified that the Hallmark+ is supposed to be an expansion of the Hallmark set—more sets, not adding genes to the existing sets, and covering greater biological space.

Dr. Naciff asked Dr. Auerbach to elaborate on the purpose and goal of NTP moving into BMD analysis using gene expression. Dr. Auerbach said that first and foremost, it is a biological potency exercise. Dr. Stevens felt it would be more accurate to say that an abstraction of biology comes out the other end, and like any abstraction, it is open to interpretation. Dr. Edwards felt that no matter what is done, there would be biological interpretation, which in his opinion is a critical step. He felt that context would need to

be provided, because otherwise observers would do so for their own purposes. He felt that there would be a constellation of processes for a given chemical that may be indicative of something that an individual GO annotation might not reflect. He recommended thinking in terms of what context can be provided now based on the data at hand, what context could be provided by bringing in some of the existing data sets, and what context could be provided by additional experimentation.

Dr. Burgoon said he did not have a problem with Dr. Auerbach's example, because it is a statement of fact, not a synthesis that came from NTP. He suggested adding a statement stating that it is a screening-level assessment and the purposes for which it should be used. That would absolve NTP from responsibility if anyone abuses the information.

Dr. Clewell said she liked the option offered by Dr. Edwards about adding more information on pathways, such as the 5 or 10 most sensitive pathways, which would add context. She said that adding that information would acknowledge that there are many things going on. She noted that the BMDs will cluster and agreed with Dr. Burgoon's suggestion to disclaim any attempt at risk assessment.

Dr. Johnson agreed about making a blanket statement noting that the work is screening-level only, and is not to be equated with an adverse effect.

Dr. Yauk said that Health Canada had looked at the bottom 5 pathways, which could be shown as well. Dr. Auerbach said that the report being developed for the first chemical would include between 10-25 pathways.

IX. Finalization of Panel Recommendations and Voting

A. Panel Discussion and Panel Recommendations

Dr. Yauk explained the format of the meeting's final session. For each of the six sessions of the meeting, the NTP's approach and possible panel recommendations would be individually projected and discussed by the panel, with opportunity for panelists to propose revisions. After any proposed changes were incorporated by consensus into the recommendations, the panel would vote on their recommendations for revision to NTP's approach. The chair would call for a motion and second, followed by the panel voting by a show of hands. Panel members voting "no" or abstaining would be asked to explain their actions.

Dr. Stevens said he would prefer a "yes, but" option in the voting, as he remained uncomfortable with the implementation plan, or lack thereof. He said there would be a great deal of work for NTP following the panel's input, and that simply passing the recommendations would imply a blanket go-ahead. Dr. Bucher noted that Dr. Stevens's qualifications would be captured in the meeting minutes as an important part of the record.

For each of the 6 sessions, elements of NTP's proposed approach were projected along with possible recommendations by the panel, which had been identified based upon

discussion during the meeting. The panel had opportunity to propose revision to the possible recommendations after which the chair called for a motion and vote.

A.1. Overall Approach

Proposed Approach

- Implement filtering
- Perform benchmark dose (BMD) modeling
- Define gene sets
- Report potency

Possible Recommendations

Scope:

Clarify the scope of the objectives to include use of BMD approaches to:

- Model the dose-response behavior of genes and gene sets
- Identify a dose below which biological and toxicological effects are unlikely to occur
- The design is sufficient at this time to evaluate its future application to risk assessment

Out of scope:

- Limit the toxicological interpretation of effects

Context of use:

- Screening and prioritization
- Interim point of departure (POD)

Time points:

- Specify how the approach will consider changes in dose-response relationships across different time points and how it will accommodate bioaccumulative substances

Other:

- Add examples to document to illustrate the method and test approach on existing datasets
- Include more details about objectives to discern objectives of *in vivo* and *in vitro* studies in approach

Dr. Yauk opened discussion on possible recommendations.

Dr. Johnson suggested adding “and toxicological” to the second bullet point under Scope. The panel concurred.

In the same bullet point, Dr. Stevens proposed changing “unlikely” to “likely.” Dr. Clewell opposed the suggestion, and explained her reasoning. Dr. Stevens withdrew the proposal.

In the third bullet point under “Scope,” Dr. Stevens felt there would not be enough information to evaluate application to risk assessment and suggested changing the word to “consider.” The panel concurred.

Dr. Johnson moved to adopt the updated recommendations. Dr. Stevens seconded. The panel voted 8 yes, 0 no, 0 abstain to recommend the following revisions to NTP’s proposed overall approach:

Recommendations

Scope:

Clarify the scope of the objectives to include use of BMD approaches to:

- Model the dose-response behavior of genes and gene sets
- Identify a dose below which biological and toxicological effects are unlikely to occur
- The design is sufficient at this time to consider its future application to risk assessment

Out of scope:

- Limit the toxicological interpretation of effects

Context of use:

- Screening and prioritization
- Interim point of departure (POD)

Time points:

- Specify how the approach will consider changes in dose-response relationships across different time points and how it will accommodate bioaccumulative substances

Other:

- Add examples to document to illustrate the method and test approach on existing datasets
- Include more details about objectives to discern objectives of *in vivo* and *in vitro* studies in approach

A.2. Filtering Measured Features

Proposed Approach

- Empirical approach maximizing permissiveness, noise reduction, and reproducibility. Details:
    - ANOVA p-value <0.05
    - Fold change >1.5
    - No multiple testing correction

Possible Recommendations

Test whether trend tests should be incorporated in initial database filter:

- Traditional Williams' test
- Williams' test variation that allows it to be used with non-monotonic data
- The design is sufficient at this time to evaluate its future application to RA
- Eliminate statistical tests for filter (ANOVA filter)
- Filter based on fold change as proposed
- Customize specific filter parameters for different platforms or experiments
- Begin to introduce nonparametric tests

The panel debated the elements of the segment at length, returning to several of the points raised in the earlier discussion of filtering.

Dr. Burgoon opposed any use of p-value. Dr. Wright opposed setting filter thresholds, as did Dr. Stevens. Ultimately, the panel changed the recommendations to specifically oppose the proposed approach, instead offering an alternative, as seen in the final version below. Dr. Wright moved to accept the revised version, Dr. Johnson seconded the motion, and the panel voted in favor, 8 yes, 0 no, 0 abstain.

Recommendations

- Do not use proposed approach. Instead, customize specific filter parameters and tests for different platforms or experiments, with the goal to enhance reproducibility of results
- Begin to introduce nonparametric tests

A.3. Fitting Features to Dose-Response Curves

Proposed Approach

- Features are fit to 9 parametric continuous models
- Benchmark response (BMR) = 1.349 x SD of controls
- 2-step process for best model selection [nested chi square and Akaike information criterion (AIC)]
- From the best fitting model a BMD, BMD_L and BMD_U is determined (BMD_L = BMD lower confidence limit and BMD_U = BMD upper bound)

Possible Recommendations

- Use parametric models as proposed; consider additional parametric models when available

- Introduce nonparametric models into BMDExpress to build confidence and experience
- Eliminate polynomial 3 model from consideration
    - Constrain parameters of polynomial 3 model to eliminate direction changes
- Specify explicitly whether the model-fitting approach uses dose or log-dose and investigate the effects of each
- During filtering stage, determine shape of response to pre-select a model for fitting
- Consider using model averaging to take into account model uncertainty as approach moves toward risk assessment

Dr. Wright suggested adding "multiple" to the reference to direction changes. Dr. Burgoon suggested striking the sentence beginning, "During filtering stage, etc." Dr. Wright and Dr. Huang suggested changing the polynomial 3 reference to simply, "Constrain parameters of polynomial models to eliminate multiple direction changes." The panel also determined a change to the first recommendation, as reflected below. Dr. Stevens moved to accept the revised version, Dr. Clewell seconded the motion, and the panel voted in favor of the motion, 8 yes, 0 no, 0 abstain.

Recommendations

- Use the parametric models proposed; consider additional parametric models when available
- Introduce nonparametric models into BMDExpress to build confidence and experience
- Constrain parameters of polynomial models to eliminate multiple direction changes
- Specify explicitly whether the model-fitting approach uses dose or log-dose and investigate the effects of each
- Consider using model averaging to take into account model uncertainty as approach moves toward risk assessment

A.4. Gene Set-Level Potencies

Proposed Approach

- Fit p-value threshold >0.0001
- BMD_U/BMD_L ratio threshold of <40
- Threshold for "active" gene sets
    - 3 genes, 5% populated, and Fisher Exact Test p-value <0.05
- Determining potency of a gene set: median and mean BMD

Possible Recommendations

- Eliminate use of Fisher Exact Test and enrichment testing
- To consider for future — apply resampling methods, such as permutation methods or bootstrapping

- Don't rely on individual genes, use groups of genes
- When estimating gene set potency, use weighted average instead of median of individual gene BMDs to capture variability
- Consider higher curve fit p-value >0.0001
    - Alternative: Use R2 value instead of a global goodness-of-fit p-value
- Use an alternative gene set to the Hallmark gene set (e.g., GO) that covers broader biological space
- Investigate the use of bootstrapping to determine confidence intervals on gene set

Dr. Stevens said that "Use an alternative gene set etc." did not belong in this segment, because it is agnostic to what set of genes is being used. He said it should be in the final segment. Dr. Burgoon and Dr. Wright suggested changes to the bullet referring to Fisher's Exact Test and enrichment testing. The panel debated how to refer to enrichment testing. See below for the final wording agreed upon by the panel.

The panelists agreed to strike the sentence beginning, "Don't rely…" They agreed to add "in addition to" to the sentence referring to R^2 value.

Dr. Stevens moved to accept the revised version. Dr. Burgoon seconded. The panel voted in favor of the motion, 8 yes, 0 no, 0 abstain.

Recommendations

- Eliminate use of Fisher Exact Test and investigate other methods, such as resampling, to perform enrichment testing
- When estimating gene set potency, use weighted average instead of median of individual gene BMDs to capture variability
- Consider higher curve fit p-value >0.0001
    - Alternative: Use R2 value instead of or in addition to a global goodness-of-fit p-value
- Investigate the use of bootstrapping to determine confidence intervals on gene set

A.5. Study Design

Proposed Approach

- BMD-centric design
    - In vivo parameters
        - Male Sprague Dawley rats, 6-8 weeks of age
        - 5 day repeat dose
        - Liver and other expert-selected organs
        - Use of a 5-day maximum tolerated dose (MTD)
    - In vitro parameters
        - Human cell lines, sex based on availability
        - Expert-determination of duration
        - Organotypic culture

        - Top dose selection: LC20 (20% reduction in cell viability relative to control)
- 10 to 12 dose levels, 3 replicates/dose group

Possible Recommendations

- Consider study design as 1st phase of larger effort to inform genomic-based risk assessment
- Include multiple time points
- Use pharmacokinetic predictions to determine steady-state timescale for duration determination and time point selection
- Include additional replicates in control group
- Include females in *in vivo* studies for studies where range-finding studies find differences between sexes
- Expand organ list beyond liver to top 3 endpoints [liver toxicity, kidney toxicity, and lung (inhalation), neurotoxicity]; collect all organs and blood to save for potential analysis later
- Incorporate metabolic considerations in study design in both *in vivo* and *in vitro*

Dr. Yauk suggested striking the phrase "collect all organs etc." Dr. Stevens suggested adding "Expand organ *collection* list*…"* Dr. Clewell suggested adding "…for future testing" to that bullet.

Dr. Naciff and Dr. Wright suggested adding a bullet, "Consider including additional replicates in the control group." Regarding the bullet referring to time points, Dr. Stevens suggested a more nuanced phrase (as seen below), including reference to an earlier time point as well as a reference to piloting for risk assessment.

After considerable discussion, the panel agreed to amend the sentence regarding including females, changing to "Use most sensitive sex, etc." while adding a sub-bullet regarding range-finding studies.

Dr. Stevens moved to accept the revised version. Dr. Johnson seconded. The panel voted in favor of the motion, 8 yes, 0 no, 0 abstain.

Recommendations

- Consider study design as 1st phase of larger effort to inform genomic-based risk assessment
- Include an earlier time point to the 5-day study design as a pilot for application to risk assessment
- Use pharmacokinetic predictions to determine steady-state timescale for duration determination and time point selection
- Consider including additional replicates in the control group
- Use most sensitive sex in *in vivo* studies
    - Range-finding studies can be used to find differences between sexes
- Expand organ collection list beyond liver to top 3 endpoints [kidney toxicity, and lung (inhalation), neurotoxicity]; for future testing
- Incorporate metabolic considerations in study design in both *in vivo* and *in vitro*

A.6. Biological Interpretation

Proposed Approach

- Expand and curate hallmark datasets to provide a toxicological and mechanistic interpretation that is species and organ/tissue specific. Expand:
    - Mine the GEO database to identify co-regulated gene sets not currently captured in the Hallmark gene sets
    - Mine existing phenotypic-anchored signatures such as those contained in the DrugMatrix database and those from the published literature
    - Remine MSigDB and CPDB in manner similar to what was done to create the Hallmark gene sets to identify additional sets that may have been overlooked

Possible Recommendations

- Focus proposal on identifying biologically responsive dose and not hazards
- Work toward linking biological effects to toxicological effects

Dr. Stevens voiced his intention to vote no on the segment in its initial version. He advocated a simpler proposed approach, with the expansion to be considered when there was adequate time, staff and other resources. Dr. Burgoon agreed. After considerable discussion, the committee decided to leave the Proposed Approach as is, but employ the same stratagem they had used in the second segment, Filtering Measure Features, negating the Proposed Approach. The panel then added several other recommendations, as reflected in the final version below.

Recommendations

- Do not use the proposed approach at this time
- Use an existing curated data set to produce a functioning pipeline
- Focus proposal on identifying biologically responsive dose and not hazards
- With release of data, include a statement that this is a screening assessment
- Report the lowest gene set and its name; list the bottom 5-10 gene sets; do not interpret further
    - Release all data publically
- Consider proposed approach at a later time with evaluation and comparison with more traditional gene sets

Dr. Stevens moved to vote on the segment. Dr. Johnson seconded the motion. The panel voted to approve the segment, 7 yes, 1 no, 0 abstain.

Dr. Stevens was the no vote. He explained that he did not agree with the stated Proposed Approach, and said he felt that an alternative approach stated around the existing curated sets would be simpler.

A.7. Next Steps

Dr. Auerbach briefly described the next steps in the process. There will be a meeting report distributed for review to ensure that panelists' comments were captured

accurately. When the review process is complete, the chair will sign the report and it will be posted to the NTP website. During this review process, panelists cannot add content to what was stated in the public meeting but could correct any inaccuracies.

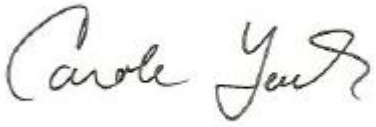
Dr. Bucher thanked participants on behalf of NTP. He said that the meeting would be seen as a milestone in the development of genomic-based risk assessment modeling.

Dr. Auerbach thanked the panelists for their input.

Dr. Yauk adjourned the meeting at 1:00 pm, October 25, 2017.

X. Approval of the Peer Review Report by the Chair of the Peer Review Panel

This peer review report has been read and approved by the Chair of the Peer Review of the Draft NTP Approach to Genomic Dose-Response Modeling Expert Panel Meeting.



Carole Yauk., Ph.D.

Date: Jan 8, 2018