Swansea University
Prifysgol Abertawe

Cronfa
Setting Research Free

**Cronfa - Swansea University Open Access Repository**

This is an author produced version of a paper published in:
*Environmental and Molecular Mutagenesis*

Cronfa URL for this paper:
http://cronfa.swan.ac.uk/Record/cronfa49200

**Paper:**
Dertinger, S., Kraynak, A., Wheeldon, R., Bernacki, D., Bryce, S., Hall, N., Bemis, J., Galloway, S., Escobar, P. et. al. (2019). Predictions of genotoxic potential, mode of action, molecular targets, and potency via a tiered multiflow® assay data analysis strategy. *Environmental and Molecular Mutagenesis*
http://dx.doi.org/10.1002/em.22274

This item is brought to you by Swansea University. Any person downloading material is agreeing to abide by the terms of the repository licence. Copies of full text items may be used or reproduced in any format or medium, without prior permission for personal research or study, educational or non-commercial purposes only. The copyright for any work remains with the original author unless otherwise specified. The full-text must not be sold in any format or medium without the formal permission of the copyright holder.

Permission for multiple reproductions should be obtained from the original author.

Authors are personally responsible for adhering to copyright and publisher restrictions when uploading content to the repository.

http://www.swansea.ac.uk/library/researchsupport/ris-support/

October 26, 2018

**TITLE:** Tiered Data Analysis Strategy for MultiFlow® DNA Damage Assay: Predictions of Genotoxic Potential and Mode of Action, Insights into Molecular Targets, and Potency Determinations

**AUTHORS:** Stephen D. Dertinger[1], Andrew R. Kraynak[2], Ryan P. Wheeldon[3], Derek T. Bernacki[1], Steven M. Bryce[1], Nikki Hall[1], Jeffrey C. Bemis[1], Sheila M. Galloway[2], Patricia A. Escobar[2], George Johnson[3]

**AFFILIATIONS:**

[1]Litron Laboratories, Rochester, NY, USA

[2]Merck & Company, Inc., West Point, PA, USA

[3]Institute of Life Science, Swansea University Medical School, Swansea University, SA2 8PP Wales, United Kingdom

**KEY WORDS:** TK6 cells, mode of action, multiplexed, γH2AX, p53, clastogen, aneugen, benchmark dose

**RUNNING TITLE:** Tiered Data Analysis Strategy

**ABSTRACT**

The *in vitro* MultiFlow® DNA Damage Assay multiplexes γH2AX, p53, phospho-histone H3, and polyploidization biomarkers into a single flow cytometric analysis [Bryce et al., 2016]. The current report describes a tiered, sequential data analysis strategy based on data generated from exposure of human TK6 cells to a previously described 85 chemical training set and a new pharmaceutical-centric test set (n=40). In each case, exposure was continuous over a range of closely spaced concentrations, and cell aliquots were removed for analysis following 4 and 24 hr of treatment. The data analyses described herein begins with a machine learning ensemble in addition to a rubric that considers fold increases in biomarkers against global evaluation factors (GEFs)—a positive finding in either is sufficient to make a genotoxic-positive call. In the case of a positive finding, tier 1 further considers the machine learning predictions and GEFs to classify the activity as clastogenic and/or aneugenic. Test set results demonstrated the generalizability of the first tier, as 35/40 (88%) concordance with *a priori* genotoxicity expectations was observed, and 21/24 (88%) of the chemicals identified as genotoxic were predicted to exhibit the expected mode of action (MoA). A second tier applies unsupervised hierarchical clustering to the biomarker response data, and these analyses were found to group certain chemicals, especially aneugens, according to their molecular targets. Finally, a third tier utilizes benchmark dose analyses and MultiFlow biomarker responses to rank genotoxic potency. The relevance of these rankings is supported by the strong agreement found between benchmark dose values derived from MultiFlow biomarkers compared to those generated from parallel *in vitro* micronucleus analyses. Collectively, the results suggest that a tiered MultiFlow data analysis pipeline is capable of rapidly and effectively identifying genotoxic hazards while providing additional information that is useful for modern risk assessments—MoA, molecular targets, and potency.

## INTRODUCTION

Our laboratories have pursued the development and validation of a multiplexed flow cytometric assay that combines information from several biomarkers relevant to DNA damage response pathways and aneuploidy induction [Bryce et al., 2014, 2016, 2017, 2018; Bernacki et al., 2016]. This so-called MultiFlow® DNA Damage Assay is formatted as an add-and-read test that efficiently prepares cells in microtiter plates for flow cytometric analysis. The biomarkers measured are: i) phosphorylation of H2AX at serine 139 (γH2AX) to detect DNA double strand breaks, ii) phosphorylation of histone H3 at serine 10 (p-H3) to identify mitotic cells, iii) nuclear p53 content as an indicator of p53 activation in response to DNA damage, iv) frequency of 8n+ cells to monitor polyploidization, and v) determination of nuclei counts to provide information about treatment-related cytotoxicity and cytostasis. Relative to individual, standard *in vitro* genotoxicity assays, an advantage of the MultiFlow method is that it goes beyond genotoxic hazard identification, since it is capable of distinguishing between clastogenic and aneugenic MoA [Bryce et al., 2016].

Given the multiplexed nature of the MultiFlow assay, the data analysis procedures used to synthesize and interpret biomarker responses have resembled pattern-recognition tools as opposed to parametric and non-parametric pair-wise tests that are commonly applied to traditional single endpoint genotoxicity assays. One published example of a MultiFlow data analysis strategy makes use of a series of global evaluation factors (GEFs) [Bryce et al., 2017]. This approach is based on cutoff response values that were derived for each biomarker and time point from data collected by 7 laboratories. To optimize agreement with *a priori* calls, a rubric was developed around the collection of cutoff values that categorizes chemicals as genotoxic or not, and if the former, whether the activity is clastogenic, aneugenic, or both. This approach was reported to exhibit good sensitivity and specificity across laboratories, and it provided reliable MoA information. However, an important caveat is that the initial report did not

evaluate the method's performance against chemicals that were outside of the training set, i.e., with an external test set that was not used to develop the GEFs and associated rubric.

Other data analysis strategies have made use of supervised machine learning tools. In this paradigm, mathematical algorithms were developed based on training set data where genotoxic potential and MoA are known. The labeled data provided a means to create models that could then be used to make predictions based on new biomarker response data that were not part of the training set. For instance, most recently, an ensemble of three machine learning algorithms consisting of logistic regression, random forest, and an artificial neural network has been described [Bryce et al., 2018]. In this case, a majority vote was used to make a final prediction about genotoxicity and genotoxic MoA. As with GEFs, this machine learning strategy also demonstrated good performance characteristics, but in this case in a more convincing fashion, as performance was maintained with an external test set of 103 chemicals.

Whereas there are advantages and disadvantages to the GEF and machine learning data analysis strategies, their use is not mutually exclusive, and there may be merit to using them in combination. The current experiments were therefore designed to extend our work with MultiFlow data analysis strategies by testing the performance of the GEF rubric and/or a machine learning ensemble using chemicals outside the training set. Furthermore, we investigated the utility of hierarchical clustering to group genotoxic chemicals with similar molecular targets, and evaluated the capacity of MultiFlow biomarker responses to provide genotoxicity potency ranking. For these investigations, MultiFlow data were generated from TK6 cells exposed to a diverse set of chemicals using a continuous treatment design (i.e., 24 hr), and in some cases these analyses were supplemented with *in vitro* micronucleus measurements. The results are discussed in terms of the performance and benefits of a sequential, tiered, high information content data analysis pipeline (see Figure 1).

## MATERIALS AND METHODS

*Chemicals*

The identities of 85 previously reported training set chemicals [Bryce et al., 2018] and a new set of pharmaceutical-centric test set chemicals (n=40), the source, and other information, are provided in Table I. Merck supplied 20 of the 40 test chemicals (coded) to Litron, and these were stored at -20°C until they were solubilized in dimethyl sulfoxide (DMSO), at which point they were refrozen at -20°C. Additional test set chemicals (n = 20) were selected by Litron scientists largely from the list recommended by Kirkland and colleagues for evaluating new genotoxicity tests [Kirkland et al., 2016]. Our *a priori* expectation regarding the *in vitro* mammalian cell genotoxicity potential for each of the 125 chemicals can be found in Table I. As explained in more detail below, the experiments reported herein occurred in the absence of an exogenous metabolic activation system. Thus, the *a priori* calls provided in Table I reflect expected genotoxicity assay results in the context of an S9-free mammalian assay system.

*Cell Culture and Treatments*

TK6 cells were purchased from ATCC® (cat. no. CRL-8015). Cells were grown in a humidified atmosphere at 37°C with 5% $CO_2$, and were maintained at or below 1 x $10^6$ cells/mL. The culture medium consisted of RPMI 1640 with 200 µg/mL sodium pyruvate (both from Sigma-Aldrich, St. Louis, MO), 200 µM L-glutamine, 50 units/mL penicillin and 50 µg/mL streptomycin (from Mediatech Inc., Manassas, VA), and 10% v/v heat-inactivated horse serum (Gibco®, a Thermo Fisher Scientific Company, Waltham, MA).

Chemicals selected by Litron scientists were tested using the same experimental design described previously [Bryce et al., 2016, 2017]. Briefly, treatments occurred in U-bottom 96 well plates, with 198 µL TK6 cell suspension (2 x $10^5$/mL) combined with 2 µL of DMSO-solubilized test chemical per well. The highest concentration tested was 1 mM, and the 19 additional concentrations were tested using a square root dilution scheme—that is, each concentration differed from the one above by a factor of 70.71%. In this manner a wide range of

concentrations were evaluated (i.e., nearly 3 orders of magnitude, 0.0014 to 1 mM). Each of the 20 concentrations was tested in a single well, whereas solvent was evaluated in 4 replicate wells. Upon addition of test chemical the plates were immediately incubated in a humidified atmosphere at 37°C with 5% $CO_2$ for 24 hr.

Merck-supplied chemicals were tested similarly, with the following exceptions. Preliminary dose-range finding experiments were used to generate 24 hr relative nuclei count (RNC) data for each chemical provided (*via* MultiFlow® — Cleaved PARP Kit, Litron Laboratories, Rochester, NY). Concentrations for the definitive experiment were chosen based on the RNC results with the intention to test at least one concentration that approached or slightly exceeded the MultiFlow assay's cytotoxicity limit, that is 80% reduction to RNC at 24 hr [Bryce et al., 2016]. There were two exceptions, 14n and 16p, compounds that were tested up to maximal feasible concentrations due to the low quantity of chemical that could be supplied (4.41 and 100 µM, respectively). For the definitive experiments, 10 concentrations of each chemical were tested in duplicate wells of a 96 well plate. As described above, the majority of chemicals were tested using a square root 2 dilution scheme. Based on data from preliminary dose-range finding experiments, some chemicals were tested using finer dilution schemes.

*MultiFlow Assay*

TK6 cells were prepared for analysis using reagents and instructions included in the MultiFlow® DNA Damage Kit — p53, γH2AX, Phospho-Histone H3 (Litron Laboratories, Rochester, NY). Components and preparation of the MultiFlow working solution have been described in detail previously [Bryce et al. 2016, 2017]. At the 4 and 24 hr sampling times, cells were resuspended with pipetting, then 25 µL were removed from each well and added to a new 96-well plate containing 50 µL/well of pre-aliquoted working MultiFlow reagent solution. Mixing was accomplished by pipetting the contents of each well several times. After incubation at room temperature for 30 min, samples were analyzed *via* flow cytometry.

Flow cytometric analysis was carried out using either a FACSCanto™ II flow cytometer equipped with a BD™ High Throughput Sampler or a Miltenyi Biotec MACSQuant® Analyzer 10 flow cytometer with integrated 96-well MiniSampler device. Stock photomultiplier tube detectors and associated optical filter sets were used to detect fluorescence emissions associated with the fluorochromes: FITC (detected in the FITC channel, to use BD instrument parlance), PE (PE channel), propidium iodide (PerCP-Cy5.5 channel), and Alexa Fluor® 647 (APC channel).

Representative bivariate graphs, gating logic, and position of regions were described in detail in earlier reports [Bryce et al., 2016, 2017; Bernacki et al., 2016]. Briefly, two biomarker measurements, γH2AX and p53, were based on the shift in median channel fluorescence intensity relative to same-plate solvent controls. Polyploidy and p-H3 biomarker measurements were based on their frequency among other nuclei. Nuclei to counting bead ratios were calculated for each sample, and these ratios were used to determine absolute nuclei counts (those with 2n and greater DNA-associated propidium iodide fluorescence). Nuclei counts were used to derive RNC, and %cytotoxicity was calculated as 100% minus %RNC at 24 hr.

*MultiFlow Data Analysis: Pre-Processing*

Data analyses described herein were restricted to those concentrations that did not exceed the MultiFlow assay's cytotoxicity limit, i.e., the top concentration of each chemical had to exhibit ≤ 80% reduction to RNC at the 24 hr time point. This has been described previously by Bryce and colleagues [2016, 2017]. The present report differs slightly, such that in addition to the 80% maximum cytotoxicity limit noted above, only two concentrations within the cytotoxicity range 70-80% were permitted. Finally, except for 14n and 16p as noted above, in the absence of excessive cytotoxicity the top concentration was 1 mM or the lowest precipitating concentration, whichever was lower.

For the GEF, machine learning, and benchmark dose analyses described below, 4 and 24 hr γH2AX, p53, and p-H3 measurements, and 24 hr polyploidy frequencies, were converted

to fold-change values by dividing them by the mean value associated with solvent-exposed cultures on the same plate (Microsoft Excel 2008, v12.3.6). This was performed for every test article concentration that was not excluded due to excessive cytotoxicity or other limits described above.

Unsupervised clustering analyses benefitted from several transformations. First, feature scaling (also known as unity-based normalization) was applied to every test article concentration to bring the values into the range 0 to 1 [Jayalakshmi and Santhakumaran, 2011]. Second, for each biomarker response and time point combination, fold-change values versus normalized concentration curves were used to generate an area under the curve (AUC) value. AUC provided a means of converting each biomarker dose-response relationship for every chemical into a single value. This was accomplished using Microsoft Excel *via* the trapezoidal rule as described at www.statisticshowto.com/how-to-find-the-area-under-a-curve-in-microsoft-excel. One (1) was subtracted from every biomarker's fold-change value before AUC calculations were made in order to set the no effect (baseline) value to zero. With this offset in place, AUC values were zero or nearly so in the case of no response, positive in the case of an increase, and negative in the case of a reduction. Also note that polyploid fold change values were transformed with the square root function, a processing step that converted this biomarker's dynamic range to one that more closely approximated that of the other biomarkers (found to be advantageous for artificial neural network models, see Bryce et al., 2018).

*MultiFlow Data Analysis: Global Evaluation Factors*

MultiFlow biomarker/time point combinations were compared to GEFs reported by Bryce and colleagues [2017]. GEFs for the three clastogen-responsive biomarkers 4 hr γH2AX, 4 hr p53, and 24 hr γH2AX, were 1.51-, 1.40-, and 2.11-fold, respectively; GEFs for the three aneugen-responsive biomarkers 4 hr p-H3, 24 hr p-H3, and 24 hr polyploidy, were 1.71-, 1.52-,

and 5.86-fold, respectively; and the GEF for the pan-genotoxicant (clastogen- and aneugen-responsive) biomarker, 24 hr p53, was 1.45-fold. Meeting or exceeding these interlaboratory-derived values identified a significant biomarker response at a particular time point. To synthesize the results of these multiple comparisons and to make judgments about genotoxic potential and MoA, the following rubric was applied. A genotoxic call with a clastogenic MoA required two successive concentrations to meet or exceed the GEF for at least two out of four clastogen-sensitive biomarkers: 4 hr γH2AX, 4 hr p53, 24 hr γH2AX, and 24 hr p53. A genotoxic call with an aneugenic MoA required two successive concentrations to meet or exceed the GEF for at least two out of four aneugen-sensitive biomarkers: 4 hr p-H3, 24 hr p-H3, 24 hr polyploidy, and 24 hr p53. In cases where both clastogen and aneugen call criteria were met, the call was genotoxic with a "mixed" MoA. When the above criteria were not met, the call was non-genotoxic under the test conditions.

*MultiFlow Data Analysis: Machine Learning Ensemble*

The development and use of three machine learning models, multinomial logistic regression (LR), artificial neural network (ANN), and random forest (RF), was described in detail previously [Bryce et al., 2018]. Briefly, these various models utilize 4 and 24 hr MultiFlow data fold-change values and predict whether a chemical exhibits genotoxic activity or not, and if present whether the genotoxicity occurs *via* a clastogenic, aneugenic, or clastogenic and aneugenic MoA. Each model's output was synthesized into genotoxicity and MoA calls as follows. Genotoxic, with evidence for a clastogenic MoA, required two successive concentrations to exhibit clastogen probability scores ≥ 80%, or one concentration to exhibit a clastogen probability score ≥ 90%. Genotoxic, with evidence for an aneugen MoA, required two successive concentrations to exhibit aneugen probability scores ≥ 80%, or one concentration to exhibit an aneugen probability score ≥ 90%. Non-genotoxic was defined as the absence of two

successive concentrations exhibiting clastogen or aneugen probability scores ≥ 80%, and no one concentration exhibiting a clastogen or aneugen probability score ≥ 90%.

A majority vote ensemble considered the genotoxicity calls from each of the 3 modeling approaches as described above. A simple majority (2 out of 3) was necessary for a summary genotoxic call. For most chemicals, MoA predictions were found to be in agreement across models. In instances when models showed significant clastogen and aneugen probabilities, the chemical was considered genotoxic with evidence for a mixed MoA.

*MultiFlow Performance Assessments*

Training and test set chemicals were evaluated against *a priori* genotoxicity and MoA expectations. This was accomplished *via* a hybrid strategy that made use of both GEF and machine learning predictions. With this approach, an overall genotoxic call was made when *either* the GEF or the machine learning ensemble was positive (see Figure 1). We supplemented these analyses by evaluating the performance of the GEF rubric and machine learning ensemble on their own.

For each strategy described above, performance was assessed by determining the level of agreement between expected and observed genotoxicity calls. This was accomplished by calculating the percentage of chemicals correctly identified as being genotoxic or non-genotoxic. Furthermore, for those agents that were identified as genotoxic, the level of agreement between MoA calls was also made by calculating the percentage of compounds that showed expected MoA. In the several instances where *a priori* MoA was either difficult to define or hypothesized to be a mixed MoA, any genotoxic MoA prediction was considered correct. In cases where a presumably non-genotoxic chemical was identified as genotoxic, any/all associated MoA calls were considered incorrect.

*Unsupervised Clustering*

Chemicals that were identified as aneugens by the hybrid GEF and machine learning approach were evaluated using JMP software's unsupervised clustering platform (JMP, v12.0.1). As described above, the biomarker response data were first converted to AUC values, and when clustering aneugens, the following 7 biomarkers were used as variables: 4 hr γH2AX, p-H3 and p53, and 24 hr γH2AX, p-H3, p53 and 24 hr polyploidy. The analysis options were set as follows: clustering method = hierarchical; method for calculating distances between clusters = "Ward"; data as usual = "Standardize Data"; data visualization = "Dendrogram", with "two way clustering".

Chemicals identified as clastogens by the hybrid GEF and machine learning approach were evaluated in a similar manner. However, in this case, the 4 variables were utilized: 4 hr γH2AX, 4 hr p53, 24 hr γH2AX, and 24 hr p53.

*Benchmark Dose Analyses*

A subset of the reference genotoxic chemicals (n = 34) were evaluated for *in vitro* micronucleus (MN) formation using TK6 cells from the same treated cultures used in the MultiFlow assay. These analyses were conducted at the 24 hr time point, and were accomplished *via* flow cytometric analysis using *In Vitro* MicroFlow® Kit reagents (Litron Laboratories, Rochester, NY). These methods have been reported in detail elsewhere [Avlasevich *et al.,* 2006]. For the MN endpoint, concentrations were limited to those that resulted in ≤ 55% reduction to relative nuclei counts.

The Benchmark Dose (BMD) for continuous data is defined as the dose or exposure that results in a predetermined percent change (benchmark response, BMR) in the response rate of an adverse effect relative to existing background incidence, generally in the range of 1-10% increase in the background [MacGregor et al., 2015]. BMD analyses were performed for the subset of 34 chemicals with concurrent MultiFlow and MicroFlow data. Specifically, γH2AX, p-

H3, p53, and *in vitro* MN dose responses were evaluated using PROAST (v63.3). Values for Critical Effect Size (CES, in PROAST notation) of 0.5 (BMR 50%), or 1.0 (BMR 100%, in the case of *in vitro* MN compounds mitomycin C, 4-nitroquinoline 1-oxide, and topotecan) were used for covariate BMD analysis for the compounds. The resulting 95% Confidence Intervals (CI's) were used to represent the relative potency of the compound for the endpoint under study. After ranking the *in vitro* MN induction potency of each compound, the data were compared with 24 hr γH2AX and 24 hr p53 endpoints for the clastogen group of compounds, and 24 hr p-H3 and 24 hr p53 endpoints for the aneugen group of compounds. These correlations are represented in cross system plots on a double Log scale [Soeteman-Hernández et al, 2016; Bemis et al., 2016]. The analyses were conducted separately for clastogens (n = 21) and aneugens (n = 13).

## RESULTS AND DISCUSSION

### *Tier 1 Analyses: Training Set*

The 85 reference chemicals that comprise the training set were given *a priori* classifications in regard to their genotoxic potential, as well as their predominant genotoxic MoA, clastogenicity or aneugenicity (Table I). Results for several of these agents are presented in detail in order to describe prototypical response profiles, and to introduce a new data visualization tool. These examples should provide a useful background for interpreting the aggregate chemical results that are presented hereafter.

Thapsigargin is an inhibitor of the sarco/endoplasmic reticulum $Ca^{++}$ ATPase [Rogers et al., 1995]. A radar plot portrays each biomarker response and time point combination as a function of concentration (Figure 2a). As expected for a non-genotoxicant, no substantial increases in γH2AX, p-H3, p53 or polyploidization biomarkers were observed, despite that fact that it was tested to cytotoxic concentrations (71% cytotoxicity). Thus, it is not surprising that

neither the GEF rubric or any of the three machine learning models predicted genotoxicity (Table II).

Treatment of TK6 cells with the reference genotoxicant 4-nitroquinoline 1-oxide resulted in a prototypical clastogenic response profile (Figure 2b). The γH2AX biomarker was increased at 4 and 24 hr. Whereas p53 activation at the 24 hr time point is a pan-genotoxicity signal, activation at 4 hr, as observed here, is quite specific for clastogens [Bryce et al., 2014, 2016]. Additionally, 4-nitroquinoline 1-oxide did not increase polyploidization, and the p-H3 biomarker was reduced in a dose-dependent manner. Both of these observations provide additional evidence of clastogenic as opposed to aneugenic activity. As shown in Table II, the GEF rubric and all three of the machine learning models predicted genotoxicity, with a clastogenic MoA.

Mebendazole's aneugenicity has been attributed to microtubule binding [Laclette et al., 1980]. MultiFlow response data illustrate a typical tubulin binder-induced aneugenic response profile (Figure 2c). While anti-γH2AX-associated fluorescence did not increase at either time point and p53 translocation was not apparent at 4 hr, marked p53 responses were observed at 24 hr. Furthermore, robust increases in p-H3 positive events were induced by mebendazole, and this was accompanied by polyploidization. GEFs as well as the machine learning ensemble identified this compound as genotoxic, with evidence for an aneugenic MoA.

Crizotinib is another aneugen that is instructive for several reasons. Crizotinib is a potent inhibitor of c-Met and ALK (anaplastic lymphoma kinase), with cell-based assay IC50 values in the low nM range [Awad and Shaw, 2014]. Even so, there is evidence that the agent's *in vitro* aneugenic activity may be related to off-target effects on aurora kinase(s) [Kong et al., 2018]. Data presented in Figure 2d support this view, as it generated response profiles that are similar to several confirmed aurora kinase inhibitors tested in the MultiFlow assay (e.g., ZM-447439 and tozasertib). As with many tubulin binders, p53 activation and polyploidization were observed at the 24 hr time point. In the case of this kinase inhibitor, polyploidization was especially robust,

and was evident well before the assay's cytotoxicity limit was reached (i.e., 8-fold increase in polyploidy at 59% cytotoxicity). Unlike tubulin binders, the proportion of p-H3-positive events was not elevated. Rather, at the highest concentrations tested, severe decreases were observed. These observations are consistent with aurora kinase inhibition, as this activity would be expected to repress serine 10 phosphorylation of histone H3 on mitotic chromosomes [Crosio et al., 2002]. Despite the response profile being quite different than spindle poisons, both the GEF rubric as well as all three of the machine learning models identified crizotinib as genotoxic, with evidence for an aneugenic MoA (Table II).

Results from the tier 1 data analyses are presented for all 85 training set chemicals in Table II. For the combined GEF plus machine learning ensemble, the concordance between *a priori* expected and observed genotoxicity calls was 99%. For those agents with a genotoxic call, the agreement with expected MoA was 98%. In both cases, the one mischaracterized agent was imatinib mesylate (identified as a clastogen). Table II also provides performance metrics for the machine learning ensemble and GEFs used in isolation. The most obvious difference between the two is that the former was effective for both genotoxicity calls and MoA predictions (at least with a training set size of 85 chemicals), while the GEF rubric showed a lower level of agreement between expected and observed genotoxic activity calls (i.e., 93% concordance), especially for clastogens. Note: Supplemental file 1a-c provides Manhattan-type plots that show machine learning probabilities for each of the 85 chemicals at every concentration tested.

*Tier 1 Analyses: Test Set*

With promising results evident for 85 training set chemicals, work with compounds that were not used to devise the GEF rubric or the machine learning models were tested in the MultiFlow assay. The results from tier 1 analyses are presented in Table III. For this set of 40 diverse chemicals, the combined GEF plus machine learning ensemble resulted in 88%

agreement between expected and observed genotoxicity calls.

Three suspected genotoxicants were not identified as such: 6f, 13m, and 14n. While 6f was an anticipated aneugen, it was not observed to affect any of the aneugen-sensitive biomarkers, despite the fact that analyses included concentrations that induced up to 63.8% cytotoxicity. Compound 14n was also classified *a priori* as aneugenic, and in this case only one aneugen biomarker was slightly induced: 4 hr p-H3 was increased by 1.39-fold at the highest concentration tested, 4.41 µM. This false negative result for 14n should be qualified to some degree, since cytotoxicity at the highest feasible concentration tested was 48.7%, well below the assay's cytotoxicity limit of 80%. The third false negative result, 13m, is also noteworthy. Whereas 6f and 14n showed slight to nil biomarker responses, 13m caused robust increases that exceeded biomarker GEFs for 4 hr p-H3 and 4 hr γH2AX across several consecutive concentrations, as well as 24 hr polyploidy at the highest concentration (Figure 3a). This response profile was not observed in the 85 chemical training set, and consequentially the GEF rubric was not developed with this in mind, and the machine learning models have no experience with this pattern.

Tier 1 mischaracterized two non-genotoxicants as genotoxic: 2b and 12L. In the case of 2b (a.k.a., sodium diethyldithiocarbamate trihydrate), it should be noted that this compound has been shown to induce cytogenetic damage in both CHO and TK6 cells [Hilliard et al., 1998; Galloway et al., 1998; Greenwood et al., 2004], and DNA double strand breaks in rat hepatocytes [Storer et al., 1996], but only at concentrations deemed overly cytotoxic by current testing standards. There are at least two biologically plausible causes for indirect effects leading to *in vitro* DNA damage: diethyldithiocarbamate chelates copper and zinc, and it is a potent inhibitor of superoxide dismutase [Heikkila et al., 1976; Nicotera et al., 1989].

Of the chemicals identified as genotoxic, tier 1 analyses were also used to predict their genotoxic MoA. As shown in Table III, 88% agreement was observed between expected and observed calls. One compound, 16p, showed mixed activities, as both clastogen and aneugen

biomarker responses were detected. This was an expected result, as 16p has an azobenzimidazole structure that was previously observed to induce premature centromere separation at metaphase in addition to induction of micronuclei and structural aberrations. MultiFlow biomarker results for this atypical agent are shown in Figure 3b. The three chemicals with misidentified MoA included the aneugen call for ciprofloxacin, a fluoroquinoline class antibiotic that was expected to exhibit clastogenic activity based on its reported topoisomerase II inhibitor activity, and the two *a priori* non-genotoxicants discussed above (i.e., 2b and 12L; both identified as clastogens).

Overall, the high concordance values speak to the generalizability of the combined GEF and machine learning ensemble to detect genotoxicants, and to furthermore provide an indication of genotoxic MoA. As with the training set, GEF and machine learning were also considered in isolation. Similar to the training set, these analyses suggest the use of the current GEF rubric alone is somewhat suboptimal, as agreement between expected and observed genotoxicity calls fell to 75%, a result that is largely attributable to false negative calls. Supplemental file 2a-c provides Manhattan-type plots that show machine learning probabilities for each of the 40 test set chemicals at every concentration evaluated.

*Tier 2 Analyses*

A set of 21 *a priori* aneugens and mixed MoA chemicals that were identified as such in tier 1 analyses were evaluated *via* unsupervised hierarchical clustering using 4 and 24 hr MultiFlow biomarker data that were each converted to a single AUC value. The resulting groupings are presented in Figure 4 in the form of a two dimensional dendrogram. The clade denoted “TB” was entirely comprised of tubulin binders. Note that whereas the exact mechanism of test agent 17q is not known, it is a benzimidazole-containing structure and therefore expected to have tubulin-binding properties. The other clear grouping is denoted “KI”, a clade that included each of the presumptive mitotic kinase inhibitors that were tested: AMG

900, crizotinib, tozasertib, hesperadin, ZM-447439, and 10j.

The set of 46 *a priori* clastogens that were identified as such in tier 1 analyses were also evaluated *via* unsupervised clustering using the 4 clastogen-responsive biomarkers. The results are shown in Figure 5. For this set of diverse clastogens, it is less obvious that clusters formed around different molecular targets. That said, the clade identified as "TI" was highly enriched for topoisomerase inhibitors (6/8), and the "C-L" grouping was enriched for DNA cross-linking agents (5/9).

Taken together, a second tier that consists of unsupervised hierarchical clustering appears to complement genotoxic potential and MoA analyses, as it provides useful information about likely molecular targets. This is especially true in the case of delineating aneugens that target mitotic kinases versus those that interfere with tubulin polymerization.

*Tier 3 Analyses*

BMD metrics served as a basis for tier 3, analyses that were conducted to determine whether MultiFlow biomarker(s) could provide a reliable indication of chemicals' genotoxic potency. The advantage of using BMD-derived potency metrics has been previously discussed by Soeteman-Hernández and colleagues [2015, 2016]. As shown in Figures 6 and 7, the BMDs in the MultiFlow endpoints were plotted against micronucleus response BMDs on a double-log scale. As opposed to representing correlation with a numerical coefficient value, a linear relationship with intercept zero equals a straight line in a double-log plot. Therefore, two lines with unity slope have been drawn on each correlation plot in such a manner that the majority of the BMD confidence intervals are encompassed between the lines. The distribution of BMD positions within the two lines show approximate linearity, differing by a proportionality constant. Furthermore, the vertical distance between the two lines translates into an uncertainty margin given by the estimation of a BMD on the *y* axis based on a specified BMD on the *x* axis, and vice versa. The uncertainty margin is used as a measure of correlation between two endpoints.

For the aneugens, when comparing MN induction to p53 responses, the cross system plots show good correlation, with the majority of the compounds located between the two lines (Figure 6). Taking the microtubule binder nocodazole as an example, the horizontal dashed line intersections with the sloped dashed lines may be considered as the respective upper and lower bounds of the uncertainty range for the *in vitro* MN endpoint. The intercepts of approximately -3 and -1 on the Log scale correspond to lower and upper bounds of $10^{-3}$ = 0.001 and $10^{-1}$ = 0.1 μM, respectively. Hence, the *in vitro* MN BMD for nocodazole is estimated to lie between 0.001 and 0.1 μM considering an uncertainty margin of approximately 1 Log. In fact, the *in vitro* MN potency for nocodazole in the dataset represented in Figure 6 has both BMDL and BMDU either side of -2 Log, and hence within the estimated potency of -3 Log and -1 Log estimated from the p53 response. The MN vs. 24 hr p-H3 system plot also indicates the BMDs for the majority of compounds in both systems are proportionally related (Figure 6), however the two lines are drawn further apart than the MN vs. 24 hr p53 system (i.e., 2 logs versus 1 log). In both cases, MN vs. p53 and MN vs. p-H3, the data are randomly scattered with good correlation.

For the clastogens, good correlation is observed for MN vs. γH2AX and MN vs. p53, with data randomly scattered between the two diagonal lines of the unity slopes, with distances of approximately 3 Log, and 2 Log respectively for each system (Figure 7).

The correlations observed here are consistent with those of other genotoxicity endpoints which have been compared using similar methodologies. Bemis and colleagues [2016] obtained an uncertainty margin of approximately 1.5 Log when comparing the *in vitro* MN responses against *in vivo* MN responses for a group of 7 clastogens. Similarly, Soeteman-Hernández et al. [2015] assessed the ability to predict *in vivo* MN potency from *in vitro* MN data. BMD confidence intervals span 2 orders of magnitude, with *in vivo* BMD confidence intervals generally smaller than those from *in vitro* studies.

*Conclusions*

The MultiFlow DNA Damage Assay's ability to predict chemicals' *in vitro* genotoxic potential and MoA was demonstrated with an external test set of 40 largely pharmaceutical-centric compounds. Whereas the GEF and associated rubric exhibited high specificity and accurate MoA predictions, it provided lower sensitivity to detect genotoxicants relative to a machine learning ensemble. A hybrid strategy whereby GEFs and machine learning are used to make calls appears to be advantageous. This approach should allow for the identification of most genotoxicants while training set data are still being expanded. Furthermore, even as data used to build prediction algorithms become more extensive, concurrent use of the GEF rubric represents a safety net of sorts, as it is capable of highlighting biomarker response patterns that the machine learning model(s) may not have encountered. In this respect, the hybrid strategy should be useful to novice laboratories, as well as established groups as they begin investigating new chemical spaces that have not been tested, or that are currently underrepresented in the training set. The compound 13m is a useful example. Although the GEF rubric did not classify the novel response profile as genotoxic, the fact that three biomarkers were elevated over their respective GEFs serves to suggest that the machine learning algorithms require additional training in this chemical space if the responses are indeed caused by *bona fide* genotoxic activity that needs to be reliably detected.

Unsupervised clustering is able to group certain genotoxicants with the same or similar molecular targets based on multifactorial biomarker response patterns. This was especially successful with aneugens that were clustered into tubulin binder and kinase inhibitor groups. While these analyses do not offer proof of molecular targets, they do represent a powerful hypothesis-generating tool, one that could be used to efficiently design the necessary follow-up test(s) aimed at directly and conclusively identifying molecular target(s) responsible for *in vitro* genotoxicity

With respect to the BMD analyses reported herein, the strong correlation of MultiFlow

biomarkers to *bona fide* genomic damage in the form of MN provides assurances of the relevance of the new assay's endpoints. Furthermore, the correlations suggest that potency determinations based on MultiFlow endpoints, at least on a rank-order basis, are likely comparable to those derived from the MN assay. This bolsters the use case whereby the constellation of MultiFlow assay biomarkers serve as a reliable genotoxicity screening tool that is predictive of *in vitro* MN formation, with the benefit of providing more mechanistic information. Finally, dose-response analyses such as these are worth pursuing further because they reflect the paradigm shift that has been transitioning genotoxicity away from a simple binary yes/no characteristic to a quantitative metric that has the potential to better inform risk assessments as margin of exposure and other toxicological principles can be considered [Pottenger and Gollapudi, 2009, 2010; Gollapudi et al., 2013; Johnson et al., 2014; MacGregor et al., 2015a,b; Dearfield et al., 2017].

**REFERENCES**

Attia SM, Aleisa AM, Bakheet SA, Al-Yahya AA, Al-Rejaie SS, Ashour AE, Al-Shabanah OA. 2009. Molecular cytogenetic evaluation of the mechanism of micronucleus formation induced by camptothecin, topotecan, and irinotecan. Environ Mol Mutagen 50:145-151.

Avlasevich SL, Bryce SM, Cairns SE, Dertinger SD. 2006. In vitro micronucleus scoring by flow cytometry: differential staining of micronuclei versus apoptotic and necrotic chromatin enhances assay reliability. Environ Mol Mutagen 47:56-66.

Awad MM, Shaw AT. 2014. ALK inhibitors in non-small cell lung cancer: Crizotinib and beyond. Clin Adv Hematol Oncol 12:429-439.

Aydemir N, Bilaloğlu R. 2003. Genotoxicity of two anticancer drugs, gemcitabine and topotecan, in mouse bone marrow in vivo. Mutat Res 537:43-51.

Bemis JC, Willis JW, Bryce SM, Torous DK, Dertinger SD, Slob W. 2016. Comparison of in vitro and in vivo clastogenic potency based on benchmark dose analysis of flow cytometric micronucleus data. Mutagenesis 31:277–285.

Bernacki DT, Bryce SM, Bemis JC, Kirkland D, Dertinger SD. 2016. γH2AX and p53 responses in TK6 cells discriminate promutagens and nongenotoxicant in the presence of rat liver S9. Environ Mol Mutagen 57:546-558.

Bryce SM, Bemis JC, Mereness JA, Spellman RA, Moss J, Dickinson D, Schuler MJ, Dertinger SD. 2014. Interpreting in vitro micronucleus positive results: simple biomarker matrix discriminates clastogens, aneugens, and misleading positive agents. Environ Mol Mutagen 55:542–555.

Bryce SM, Bernacki DT, Bemis JC, Dertinger SD. 2016. Genotoxic mode of action predictions from a multiplexed flow cytometric assay and a machine learning approach. Environ Mol Mutagen 57:171-189.

Bryce SM, Bernacki DT, Bemis JC et al. 2017. Interlaboratory evaluation of a multiplexed high information content in vitro genotoxicity assay. Environ Mol Mutagen 58:146-161.

Bryce SM, Bernacki DT, Smith-Roe SL, Witt KL, Bemis JC, Dertinger SD. 2018. Investigating the generalizability of the MultiFlow® DNA damage assay and several machine learning models with a set of 103 diverse test chemicals. Toxicol Sci 162:146-166.

Cheung JR, Dickinson DA, Moss J, Schuler MJ, Spellman RA, Heard PL. Histone markers identify the mode of action for compounds positive in the TK6 micronucleus assay. Mutat Res 777:7-16.

Chinnasamy N, Rafferty JA, Hickson I, Ashby J, Tinwell H, Margison GP, Dexter TM, Fairbairn LJ. 1997. O6-benzylguanine potentiates the in vivo toxicity and clastogenicity of temozolomide and BCNU in mouse bone marrow. Blood 89:1566-73.

Cimino MC. 2006. Comparative overview of current international strategies and guidelines for genetic toxicology testing for regulatory purposes. Environ Mol Mutagen 47:362–390.

Cojocel C, Novotny L, Vachalkova A. 2006. Mutagenic and carcinogenic potential of menadione. Neoplasma 53:316-323.

Crosio C, Fimia GM, Loury R, Kimura M, Okano Y, Zhou H, Sen S, Allis CD, Sassone-Corsi P. 2002. Mitotic phosphorylation of histone H3: spatio-temporal regulation by mammalian Aurora kinases. Mol Cell Biol 22:874-885.

Curry PT, Kropko ML, Garvin JR, Fiedler RD, Theiss JC. 1996. In vitro induction of micronuclei and chromosome aberrations by quinolones: possible mechanisms. Mutat Res 352:143-150.

Dearfield KL, Gollapudi BB, Bemis JC et al. 2017. Next generation testing strategy for assessment of genomic damage: A conceptual framework and considerations. Environ Mol Mutagen 58:264-283.

de Graaf AO, van den Heuvel LP, Dijkman HB, de Abreu RA, Birkenkamp KU, de White T, van der Reijden BA, Smeitink JA, Jansen JH. 2004. Bcl-2 prevents loss of mitochondria in CCCP-induced apoptosis. Exp Cell Res 299:533-540.

Degrassi F, De Salvia R, Berghella L. 1993. The production of chromosomal alterations by β-lapachone, an activator of topoisomerase I. Mutat Res 288:263-267.

DeMarini DM, Brock KH, Doerr CL, Moore MM. 1987. Mutagenicity and clastogenicity of teniposide (VM-26) in L5178Y/TK +/- 3.7.2C mouse lymphoma cells. Mutat Res 187:141-149.

Dertinger SD, Phonethepswath S, Weller P, Avlasevich S, Torous DK, Mereness JA, Bryce SM, Bemis JC, Bell S, Portugal S, *et al.* 2011. Interlaboratory *Pig-a* gene mutation assay trial: Studies of 1,3-propane sultone with immunomagnetic enrichment of mutant erythrocytes. Environ Mol Mutagen 52:748–755.

Dertinger SD, Phonethepswath S, Avlasevich SL, Torous DK, Mereness J, Bryce SM, Bemis JC, Bell S, Weller P, MacGregor JT. 2012. Efficient monitoring of in vivo *Pig-a* gene mutation and chromosomal damage: Summary of 7 published studies and results from

11 new reference compounds. Toxicol Sci 130:328-348.

Diaz D, Scott A, Carmichael P, Shi W, Costales C. 2007. Evaluation of an automated in vitro micronucleus assay in CHO-K1 cells. Mutat Res. 630:1-13.

Floxin [package insert]. Raritan, NJ: Ortho-McNeil, 2008. http://www.accessdata.fda.gov/drugsatfda_docs/label/2008/019735s059lbl.pdf. Accessed October 12, 2018.

Futami T, Miyagishi M, Taira K. 2005. Identification of a network involved in thapsigargin-induced apoptosis using a library of small interfering RNA expression vectors. J Biol Chem 280:826-831.

Galloway SM, Aardema MJ, Ishidate Jr M, Ivett JL, Kirkland DJ, Morita T, Mosesso P, Sofuni T. 1994. Report from working group on *in vitro* tests for chromosomal aberrations. Mutat Res 312:241-261.

Galloway SM, Miller JE, Armstrong MJ, Bean CL, Skopek TR, Nichols WW. 1998. DNA synthesis inhibition as an indirect mechanism of chromosome aberrations: Comparison of DNA-reactive and non-DNA-reactive clastogens. Mutat Res 400:169-186.

Garcia-Canton C, Anadon A, Meredith C. 2013. Assessment of the in vitro γH2AX assay by high content screening as a novel genotoxicity test. Mutat Res 757:158-166.

Gewirtz DA. 1999. A critical evaluation of the mechanisms of action proposed for the antitumor effects of the anthracycline antibiotics adriamycin and daunorubicin. Biochem Pharmacol 57**:**727-741.

Gleevec [package insert]. East Hanover, NJ: Novartis, 2001. http://www.accessdata.fda.gov/drugsatfda_docs/label/2008/021588s024lbl.pdf. Accessed October 12, 2018.

Glover TW, Berger C, Coyle J, Echo B. 1984. DNA polymerase alpha inhibition by aphidicolin induces gaps and breaks at common fragile sites in human chromosomes. Hum Genet 67:136–142.

Gocke E, Bürgin H, Müller L, Pfister T. 2009. Literature review on the genotoxicity, reproductive toxicity, and carcinogenicity of ethyl methanesulfonate. Toxicol Lett 190:254-265.

Gollapudi BB, Johnson GE, Hernández LG et al. 2013. Quantitative approaches for assessing dose-response relationships in genetic toxicology studies. Environ Mol Mutagen 54:8-18.

Gollapudi P, Hasegawa LS, Eastmond DA. 2014. A comparative study of the aneugenic and polyploidy-inducing effects of fisetin and two model Aurora kinase inhibitors. Mutat Res 767:37-43.

Greenwood SK, Hill RB, Sun JT, Armstrong MJ, Johnson TE, Gara JP, Galloway SM. 2004. Population doubling: A simple and more accurate estimation of cell growth suppression in the in vitro assay for chromosomal aberrations that reduces irrelevant positive results. Environ Mol Mutagen 43:36-44.

Gulati DK, Witt K, Anderson B, Zeiger E, Shelby MD. 1989. Chromosome aberration and sister chromatid exchange tests in Chinese Hamster Ovary cells in vitro III: Results with 27 chemicals. Environ Mol Mutagen 13:133-193.

Han C, Nam MK, Park HJ, Seong YM, Kang S, Rhim H. 2008. Tunicamycin-induced ER stress upregulates the expression of mitochondrial HtrA2 and promotes apoptosis through the cytosolic release of HtrA2. J Microbiol Biotechnol 18:1197-1202.

Henderson L, Fedyk J, Windebank S, Smith M. 1993. Induction of micronuclei in rat bone marrow and peripheral blood following acute and subchronic administrate of azathioprine. Mutat Res 291:79-85.

Hilliard C, Armstrong M, Bradt C, Hill R, Greenwood S, Galloway S. 1998. Chromosome aberrations in vitro related to cytotoxicity of non-mutagenic chemicals. Environ Mol Mutagen 31:316-326.

IARC monograph, Clofibrate: http://monographs.iarc.fr/ENG/Monographs/vol66/mono66-17.pdf. Accessed October 12, 2018.

Jayalakshmi T, Santhakumaran A. 2011. Statistical normalization and back propagation for

classification. International Journal of Computer Theory and Engineering 3:89-93.

Johnson GE, Soeteman-Hernández LG, Gollapudi BB et al. 2014. Derivation of point of departure (PoD) estimates in genetic toxicology studies and their potential applications in risk assessment. Environ Mol Mutagen 55:609-23.

Hauf S, Cole RW, LaTerra S, Zimmer C, Schnapp G, Walter R, Heckel A, van Meel J, Rieder CL, Peters JM. 2003. The small molecule Hesperadin reveals a role for Aurora B in correcting kinetochore-microtubule attachment and in maintaining the spindle assembly checkpoint. J Cell Biol 161:281-294.

Heard PL, Rubitski EE, Spellman RA, Schuler MJ. 2013. Phenolphthalein induces centrosome amplification and tubulin depolymerization in vitro. Environ Mol Mutagen 54:308-316.

Heikkila RE, Cabbat FS, Cohen G. 1976. In vivo inhibition of Superoxide dismutase in mice by diethyldithiocarbamate. J Biol Chem 251:2182-2185.

Hernández LG, van Benthem J, Johnson GE. 2013. A mode-of action approach for the identification of genotoxic carcinogens. PLOS ONE 8:e64532. doi:10.1371/journal.pone.0064532.

Keshava C, Keshava N, Whong WZ, Nath J, Ong TM. 1998. Inhibition of methotrexate-induced chromosomal damage by folinic acid in V79 cells. Mutat Res 397:221-228.

Kimura A, Miyata A, Honma M. 2013. A combination of in vitro comet assay and micronucleus test using human lymphoblastoid TK6 cells. Mutagenesis 28:583-90.

Kirkland D, Kasper P, Müller L, Corvi R, Speit G. 2008. Recommended lists of genotoxic and non-genotoxic chemicals for assessment of the performance of new or improved genotoxicity tests: a follow-up to an ECVAM workshop. Mutat Res 653:99-108.

Kirkland D, Kasper P, Martus H-J, Müller L, van Benthem J, Madia F, Corvi R. 2016. Updated recommended lists of genotoxic and non-genotoxic chemicals for assessment of the performance of new or improved genotoxicity tests. Mutat Res 795:7-30.

Klein CB, King AA. 2007. Genistein genotoxicity: Critical considerations of in vitro exposure dose. Toxicol Appl Pharmacol 224:1-11.

Krishna G, Urda G, Tefera W, Lalwani ND, Theiss J. 1995. Simultaneous evaluation of dexamethasone-induced apoptosis and micronuclei in rat primary spleen cell cultures. Mutat Res. 332:1-8.

Kondo Y, Honda S, Nakajima M, Miyahana K, Hayashi M, Shinagawa Y, Sato S, Inoue K, Nito S, Ariyuki F. 1992. Micronucleus test with vincristine sulfate and colchicine in peripheral blood reticulocytes of mice using acridine orange supravital staining. Mutat Res 278:187-191.

Kong Y, Bender A, Yan A. 2018. Identification of novel aurora kinase A (AURKA) inhibitors via hierarchical ligand-based virtual screening. J Chem Inf Model 58:36-47.

Kurihara D, Matsunaga S, Kawabe A, Fujimoto S, Noda M, Uchiyama S, Fukui K. 2006. Aurora kinase is required for chromosome segregation in tobacco BY-2 cells. Plant J 48:572-580.

Laclette JP, Guerra G, Zetina C. 1980. Inhibition of tubulin polymerization by mebendazole. Biochem Biophys Res Commun 92:417-423.

Li Y, Luan Y, Qi X, Li M, Gong L, Xue X, Wu X, Wu Y, Chen M, Xing G, Yao J, Ren J. 2010. Emodin triggers DNA double-strand breaks by stabilizing topoisomerase II-DNA cleavage complexes and by inhibiting ATP hydrolysis of topoisomerase II. Toxicol Sci 118:435-443.

Lidoderm [package insert]. Chad's Ford, PA: Endo Pharmaceuticals, 2004. http://www.accessdata.fda.gov/drugsatfda_docs/label/2005/020612s007lbl.pdf. Accessed October 12, 2018.

Lotz AS, Havla JB, Richter E, Frölich K, Staudenmaier R, Hagen R, Kleinsasser NH. 2009. Cytotoxic and genotoxic effects of matrices for cartilage tissue engineering. Toxicol Lett 190:128-33.

Lu P-Z, Lai C-Y, Chan W-H. 2008. Caffeine induces cell death via activation of apoptotic signal and inactivation of survival signal in human osteoblasts. Int J Mol Sci 9:698-718.

Lynparza [package insert]. Wilmington, DE: AstraZeneca Pharmaceuticals LP, 2014. http://www.accessdata.fda.gov/drugsatfda_docs/label/2014/206162lbl.pdf. Accessed October 12, 2018.

MacGregor JT, Frötschl R, White PA et al. 2015a. IWGT report on quantitative approaches to genotoxicity risk assessment I. Methods and metrics for defining exposure-response relationships and points of departure (PoDs). Mutat Res 783:55-65.

MacGregor JT, Frötschl R, White PA et al. 2015b. IWGT report on quantitative approaches to genotoxicity risk assessment II. Use of point-of-departure (PoD) metrics in defining acceptable exposure limits and assessing human risk. Mutat Res 783:66-78.

Martelli A, Allavena A, Campart GB, Canonero R, Ghia M, Mattioli F, Mereto E, Robbiano L, Brambilla G. 1995. In vitro and in vivo testing of hydralazine genotoxicity. J Pharmacol Exp Ther 273:113-120.

Matsushima T, Hayashi M, Matsuoka A, Ishidate M Jr, Miura KF, Shimizu H, Suzuki Y, Morimoto K, Ogura H, Mure K, Koshi K, Sofuni T. 1999. Validation study of the in vitro micronucleus test in a Chinese hamster lung cell line (CHL/IU). Mutagenesis 14:569-80.

Mevacor [package insert]. Whitehouse Station, NJ: Merck & Co., Inc., 2012. http://www.accessdata.fda.gov/drugsatfda_docs/label/2012/019643s085lbl.pdf. Accessed October 12, 2018.

Moon JL, Kim SY, Shin SW, Park J-W. 2012. Regulation of brefeldin A-induced ER stress and apoptosis by mitochondrial NADP+ -dependent isocitrate dehydrogenase. *Biochem* Biophys Res Commun 417:760-764.

Muehlbauer PA, Spellman RA, Gunther WC, Sanok KE, Wiersch CJ, O'Lone SD, Dobo KL, Schuler MJ. 2008. Improving dose selection and identification of aneugens in the in vitro chromosome aberration test by integration of flow cytometry-based methods. Environ

Mol Mutagen 49:318-327.

National Toxiclogy Program Report, 6-Thioguanine: https://ntp.niehs.nih.gov/testing/status/agents/ts-m890081.html. Accessed October 12, 2018.

National Toxicology Program Report, Sodium Dodecyl Sulfate: http://ntp.niehs.nih.gov/testing/status/agents/ts-10604-g.html. Accessed October 12, 2018.

Nicotera TM, Notaro J, Notaro S, Schumer J, Sandberg AA. 1989. Elevated superoxide dismutase in Bloom Syndrom: A genetic condition of oxidative stress. Cancer Res 49:5239-5243.

Nikolova T, Dvorak M, Jung F, Adam I, Krämer E, Gerhold-Ay A, Kaina B. 2014. γH2AX assay for genotoxic and nongenotoxic agents: Comparison of H2AX phosphorylation with cell death response. Toxicol Sci 140:103-117.

Oliver J, Meunier JR, Awogi T, Elhajouji A, Ouldelhkim MC, Bichet N, Thybaud V, Lorenzon G, Marzin D, Lorge E. 2006. SFTG international collaborative study on in vitro micronucleus test V. Using L5178Y cells. Mutat Res 607:125-152.

Parry EM, Parry JM, Corso C, Doherty A, Haddad F, Hermine TF, Johnson G, Kayani M, Quick E, Warr T, Williamson J. 2002. Detection and characterization of mechanism of action of aneugenic chemicals. Mutagenesis 17:509-521.

Paulsson B, Kotova N, Grawé J, Henderson A, Granath F, Golding B, Törnqvist M. 2003. Induction of micronuclei in mouse and rat by glycidamide, genotoxic metabolite of acrylamide. Mutat Res 535:15-24.

Paxil [package insert]. Research Triangle Park, NC: GlaxoSmithKline, 2011. http://www.accessdata.fda.gov/drugsatfda_docs/label/2011/020031s058s066,020710s022s030lbl.pdf. Accessed October 12, 2018.

Payton M, Bush TL, Chung G, Ziegler B, Eden P, McElroy P, Ross S, Cee VJ, Deak HL,

Hodous BL, Nguyen HN et al. 2010. Preclinical evaluation of AMG 900, a novel potent and highly selective pan-aurora kinase inhibitor with activity in taxane-resistant tumor cell lines. Cancer Res 70:9846-9854.

Pepcid [package insert]. Whitehouse Station, NJ: Merck & Co, Inc., 2011. http://www.accessdata.fda.gov/drugsatfda_docs/label/2011/019462s037lbl.pdf. Accessed October 12, 2018.

Pottenger LH and Gollapudi BB. 2009. A case for a new paradigm in genetic toxicology testing. Mutat Res 678:148-151.

Pottenger LH and Gollapudi BB. 2010. Genotoxicity testing: Moving beyond qualitative “screen and bin” approach towards characterization of dose-response and thresholds. Environ Molec Mutagen 51:792-799.

Rogers TB, Inesi G, Wade R, Lederer WJ. 1995. Use of thapsigargin to study Ca2+ homeostasis in cardiac cells. Biosci Rep 15:341-349.

Rosefort C, Fauth E, Zanki H. 2004. Micronuclei induced by aneugens and clastogens in mononucleate and binucleate cells using the cytokinesis block assay. Mutagenesis 19:277-284.

Schuler M, Muehlbauer P, Guzzie P, Eastmond DA. 1999. Noscapine hydrochloride disrupts the mitotic spindle in mammalian cells and induces aneuploidy as well as polyploidy in cultured human lymphocytes. Mutagenesis 14:51-56.

Smart DJ, Halicka HD, Schmuck G, Traganos F, Darzynkiewicz Z, Williams GM. 2008. Assessment of DNA double-strand breaks and gammaH2AX induced by the topoisomerase II poisons etoposide and mitoxantrone. Mutat Res 641:43-47.

Soeteman-Hernández LG, Fellows MD, Johnson GE, Slob W. 2015. Correlation of in vivo versus in vitro benchmark doses (BMDs) derived from micronucleus test data: Proof of concept study. Toxicol Sci 148:355-367.

Soeteman-Hernández LG, Johnson GE, Slob W. 2016. Estimating the carcinogenic potency of chemicals from the in vivo micronucleus test. Mutagenesis 31:347-58.

Sprycel® (dasatinib) [package insert] Bristol-Myers Squibb Co, Princeton, NJ. (2010). Available at: https://www.accessdata.fda.gov/drugsatfda_docs/label/2010/021986s7s8lbl.pdf. Accessed October 12, 2018.

Storer RD, McKelvey TW, Kraynaka AR, Elia MC, Barnum JE, Harmon LS, Nichols WW, DeLuca JG. 1996. Revalidation of the in vitro alkaline elution/hepatocyte assay for DNA damage: improved criteria for assessment of cytotoxicity and genotoxicity and results for 81 compounds. Mutat Res 368:59-101.

Tagrisso™ (Osimertinib) [package insert] AstraZeneca Pharmaceuticals LP, Wilmington, DE. 2012. Available at: https://www.accessdata.fda.gov/drugsatfda_docs/label/2015/208065s000lbl.pdf. Accessed October 12, 2018.

Tayama S, Nakagawa Y. 2001. Cytogenetic effects of propyl gallate in CHO-K1 cells. Mutat Res 498:117-127.

Tweats DJ, Johnson GE, Scandale I, Whitwell J, Evans DB. 2016. Genotoxicity of flubendazole and its metabolites in vitro and the impact of a new formulation on in vivo aneugenicity. Mutagenesis 31:309-321.

Van Hummelen P, Elhajouji A, Kirsch-Volders M. 1995. Clastogenic and aneugenic effects of three benzimidazole derivates in the in vitro micronucleus test using human lymphocytes. Mutagenesis 10:23-29.

Verdoodt B, Decordier I, Geleyns K, Cunha M, Cundari E, Kirsch-Volders M. 1999. Induction of polyploidy and apoptosis after exposure to high concentrations of the spindle poison nocodazole.

Youngblom JH, Wiencke JK, Wolff S. 1989. Inhibition of the adaptive response of human lymphocytes to very low doses of ionizing radiation by the protein synthesis inhibitor

cycloheximide. Mutat Res 227:257-261.

Zerit [packet insert] Bristol-Myers Squibb Virology, Princeton, NJ, 2002. http://www.accessdata.fda.gov/drugsatfda_docs/label/2002/20412S017.pdf. Accessed October 12, 2018.

**FIGURE LEGENDS**

**Figure 1.** Flow chart representing a tiered MultiFlow assay data analysis pipeline. With this strategy chemicals are evaluated for their genotoxic potential and genotoxic mode of action (tier 1), insights into molecular target are provided by unsupervised clustering (tier 2), and finally potency metrics are generated (tier 3).

**Figure 2.** Radar plots show MultiFlow assay data for seven biomarker/time point combinations and for each of four chemicals: thapsigargin, 4-nitroquinoline 1-oxide (4NQO), mebendazole, and crizotinib. The biomarker data are expressed as fold-increase over mean solvent control on the same plate, and each chemical concentration appears as a different colored line. The top-most endpoint (24 hr p53, at 12 o'clock) is a pan-genotoxic biomarker, whereas the biomarkers arranged on the right side of the graph are responsive to clastogens and those arranged on the left are responsive to aneugens.

**Figure 3.** Radar plots show MultiFlow assay data for seven biomarker/time point combinations and for each of two chemicals: Merck-supplied test compounds 13m and 16p. The biomarker data are expressed as fold-increase over mean solvent control on the same plate, and each chemical concentration appears as a different colored line. Same format as Figure 2.

**Figure 4.** Unsupervised clustering results are shown as a two dimensional dendrogram for 21 chemicals that were identified as exhibiting aneugenic activity. As described in Materials and Methods, each biomarker dose response was converted to an area under the curve for this analysis. The abbreviations TB (tubulin binder) and KI (kinase inhibitor) are used to denote clades with chemicals that are known to exhibit these activities.

**Figure 5.** Unsupervised clustering results are shown as a two dimensional dendrogram for 46

chemicals that were identified as exhibiting clastogenic activity. As described in Materials and Methods, each biomarker dose response was converted to an area under the curve for this analysis. The abbreviations TI (topoisomerase inhibitor) and C-L (cross-linker) are used to denote clades that are enriched for chemicals known to exhibit these activities.

**Figure 6.** Left panel: BMD analyses of aneugen compounds represented in cross system plots with BMD50 confidence intervals for *in vitro* MN against BMD50 24 hr p-H3 responses in TK6 cells, with both *x* and *y* axes representing Log10 concentration of compounds in μM. The dashed parallel lines are drawn in such a way that encompasses most of the confidence intervals. Compound 'car' falls outside the trend with unbound confidence intervals in the 24 hr p-H3 endpoint. Right panel: BMD50 confidence intervals for *in vitro* MN against BMD50 24 hr p53 responses in TK6 cells, with both *x* and *y* axes representing Log10 concentration of compounds in μM. Dashed parallel lines encompassing most of the BMDs, similarly to the left panel correlation plot. Compounds 'gli' and 'des' lie outside the general observed trend, with unbound upper confidences interval in the 24 hr p53 endpoint. Dashed horizontal lines obtain the uncertainty range with corresponding circles intercept with the *x* axis predicting the BMD50 for *in vitro* MN response. See Table I for compound abbreviations. Abbreviation: BMD, Benchmark Dose. MN, micronucleus.

**Figure 7.** Left Panel: BMD analyses of clastogen compounds represented in cross system plots with BMD50 confidence intervals for *in vitro* MN (with the exception of mmc, nqo and top) versus BMD50 24hr H2AX responses in TK6 cells, with both *x* and *y* axes representing Log10 concentration of compounds in μM. The dashed parallel lines are drawn in such a way that encompasses all of the confidence intervals. Right Panel: BMD50 confidence intervals for *in vitro* MN (with the exception of mmc, nqo and top) versus BMD50 24hr p53 responses in TK6

cells, with both *x* and *y* axes representing Log10 concentration of compounds in µM. Dashed parallel lines encompassing most of the BMDs. Compound ola lies outside the general observed trend, with an unbound upper confidence interval in the p53 endpoint. Compound cis displays an unbound upper confidence interval in the p53 endpoint. See Table I for compound abbreviations. Abbreviation: BMD, Benchmark Dose. MN, micronucleus.