Regulatory Toxicology and Pharmacology 72 (2015) 17–25

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

A feasibility study: Can information collected to classify for mutagenicity be informative in predicting carcinogenicity?

Petko I. Petkov[a], Grace Patlewicz[b,*], Terry W. Schultz[c], Masamitsu Honma[d], Milen Todorov[a], Stefan Kotov[a], Sabcho D. Dimitrov[a], E. Maria Donner[b], Ovanes G. Mekenyan[a]

[a] Laboratory of Mathematical Chemistry (LMC), As. Zlatarov University, Bourgas, Bulgaria
[b] DuPont Haskell Global Centers for Health and Environmental Sciences, Newark, DE 19711, USA
[c] College of Veterinary Medicine, The University of Tennessee, Knoxville, TN 37996-4500, USA
[d] Division of Genetics and Mutagenesis, National Institute of Health Sciences, Tokyo, Japan

ARTICLE INFO

Article history:
Received 5 September 2014
Available online 16 March 2015

Keywords:
Carcinogenicity classification
Integrated Approaches to Testing and Assessment (IATA)
(Q)SAR
Adverse Outcome Pathway (AOP)

ABSTRACT

Carcinogenicity is a complex endpoint of high concern yet the rodent bioassay still used is costly to run in terms of time, money and animals. Therefore carcinogenicity has been the subject of many different efforts to both develop short-term tests and non-testing approaches capable of predicting genotoxic carcinogenic potential. In our previous publication (Mekenyan et al., 2012) we presented an *in vitro–in vivo* extrapolation workflow to help investigate the differences between *in vitro* and *in vivo* genotoxicity tests. The outcomes facilitated the development of new (Q)SAR models and for directing testing. Here we have refined this workflow by grouping specific tests together on the basis of their ability to detect DNA and/or protein damage at different levels of biological organization. This revised workflow, akin to an Integrated Approach to Testing and Assessment (IATA) informed by mechanistic understanding was helpful in rationalizing inconsistent study outcomes and categorizing a test set of carcinogens with mutagenicity data on the basis of regulatory mutagenicity classifications. Rodent genotoxic carcinogens were found to be correctly predicted with a high sensitivity (90–100%) and a low rate of false positives (3–10%). The insights derived are useful to consider when developing future (non-)testing approaches to address regulatory purposes.

1. Introduction

Carcinogenicity is a complex toxicological endpoint of high concern. At the same time the rodent bioassay currently employed to assess carcinogenic potential is costly to run in terms of time, money and number of animals. Therefore carcinogenicity has been the subject of many efforts to develop *in vitro* and *in vivo* short-term tests, specifically capable of predicting genotoxic carcinogenic potential. The available genotoxicity tests assess the potential of substances to cause cancer or heritable diseases in humans. The data generated is used in both the hazard identification and risk characterization of substances for regulatory and product stewardship purposes.

Hazard identification for genotoxicity mainly relies on *in vitro* studies determining mutagenicity of substances in bacteria and in mammalian cells following an initial review of existing literature and Structure Activity Relationship/Quantitative Structure Activity Relationship (SAR/QSAR) pre-screening. Effects such as DNA damage, formation of strand breaks or adducts are other helpful indicators for genotoxicity. *In vivo* studies are also used to evaluate genotoxic potential further and are typically conducted to put *in vitro* observations into perspective.

Given the many different modes of action for mutagenesis, a number of tests are needed to assess whether a chemical is genotoxic or not with any degree of confidence. When combined appropriately, positive results from mutagenicity tests can be used to predict carcinogenicity. Some modes of actions involved in the cancer initiation step (e.g., epigenetic DNA methylation) remain without experimental data support because no appropriate test systems for their identification have yet been developed. This can potentially bring some limitations to the currently employed strategies for predicting carcinogenesis. There have been a number of efforts to investigate strategies for evaluating mutagenicity both from the perspective of classifying a chemical as a mutagen or in directing further work in the assessment of carcinogenic potential (Zeiger, 1998; Kirkland et al., 2005, 2014; Cimino, 2006; Matthews

* Corresponding author at: EPA/ORD/NCCT, 109 T.W. Alexander Dr. B205-01, RTP, NC 27711, USA.
E-mail address: patlewig@hotmail.com (G. Patlewicz).

http://dx.doi.org/10.1016/j.yrtph.2015.03.003
0273-2300/© 2015 Elsevier Inc. All rights reserved.

et al., 2006; Benigni et al., 2010). These articles examined the most frequently used *in vitro* and/or *in vivo* genotoxicity assays for their capability to discriminate between rodent carcinogens and non-carcinogens. For the assessment of non-pharmaceuticals, genotoxicity assays have typically been used as part of a three tiered-testing approach, with Tier I *in vitro* testing followed by Tier II *in vivo* genotoxicity testing in somatic cells to determine the biological relevance of chemicals that are positive in the preceding *in vitro* tests. Tier III *in vivo* testing may comprise tests in gonadal cells as well as multigenerational tests. The most common genotoxicity testing batteries include assays that measure gene mutation (i.e., point mutations that affect single genes or blocks of genes), clastogenicity (i.e., structural chromosome aberrations), and aneuploidy (i.e., numerical chromosome aberrations) (Dearfield et al., 2002; Cimino, 2006). Indeed the US EPA's test battery is a three-tiered scheme (Cimino, 2006) where Tier I includes bacterial reverse mutation assays for gene mutations (e.g., Ames tests), Tier II, an *in vitro* mammalian cell gene mutation assay (e.g., mouse lymphoma test), and Tier III comprises either the *in vivo* bone marrow mammalian chromosome aberration or the *in vivo* erythrocyte micronucleus assay. Japan's National Institute of Health Sciences (NIHS), employs a very similar testing strategy. Whilst positive results concerning *in vitro* genotoxicity demonstrate an intrinsic genotoxic activity of a chemical, this is sometimes only observed under extreme culturing conditions or in the presence of high concurrent cytotoxicity and therefore may not be relevant for *in vivo* genotoxicity (Kirkland et al., 2005, 2006). As a result, a high number of "irrelevant positive" results detected by *in vitro* assays (especially chromosomal aberrations) appear not to be confirmed in follow-up *in vivo* assays (EFSA, 2011).

In an effort to improve predictivity, strategic testing has taken the form of Integrated Testing Strategies (ITSs) (Grindon et al., 2006; Combes et al., 2007; Kirkland et al., 2007a,b; Kirkland et al., 2011). The aim of an ITS is to maximize the use of all scientific relevant information and where possible, avoid the use of animal testing. The ITS described in the REACH Technical Guidance (ECHA, 2014) is a case in point.

Recently, Adverse Outcome Pathways (AOPs) which capture information on the causal links between a molecular initiating event, intermediate key events and an adverse outcome of regulatory concern have shown potential in providing a biological context to facilitate the development of mechanistically based Integrated Approaches for Testing and Assessment (IATAs) (which encompasses ITSs) for regulatory decision making (Ankley et al., 2010; Tollefsen et al., 2014). An IATA is a structured approach that integrates and weighs different types of data for the purposes of performing hazard identification, hazard characterization and/or safety assessment of a chemical or group of chemicals. Whilst there is a strong drive to develop AOPs that can be used to inform IATA, the OECD work programme being notable amongst these efforts, using AOPs in such a predictive capacity is still at an early stage of evolution. There are many practical challenges of gathering relevant data to derive and implement IATA and their elements for inclusion into tools, notably the OECD Toolbox.

Previously we introduced an *in vitro–in vivo* extrapolation workflow as a means of relating different short term genotoxicity tests together on the basis of their levels of biological organization. This so-called extrapolation workflow was used to facilitate the development of new genotoxicity models in the Tissue Metabolism Simulator (TIMES) platform and to help direct strategic testing (Mekenyan et al., 2012). Two (Q)SAR models, namely for *in vivo* genotoxicity in liver and *in vivo* micronucleus formation in bone marrow were developed. The exercise highlighted a number of practical issues notably the challenges of accounting for metabolic differences between *in vitro* and *in vivo* test systems (Mekenyan et al., 2012). The workflow developed was structured into 3 steps. Step one subdivided chemicals into positive or negative calls based on results from *in vitro* mutagenicity assays. Step two performed a similar categorization based on *in vivo* genotoxicity effects in liver, whilst step three was based on results from *in vivo* micronucleus formation in bone marrow (Mekenyan et al., 2012). The overall product was a five-level framework, where 3 concurrent negative results across the 3 levels of biological organization was denoted Level 1 and 3 concurrent positive results as Level 5.

Given recent efforts in developing AOPs and associated AOP-informed IATA particularly under the OECD work programme (see http://www.oecd.org/chemicalsafety/testing/adverse-outcome-pathways-molecular-screening-and-toxicogenomics.htm), this study re-evaluated the *in vitro–in vivo* extrapolation workflow by considering the mechanistic basis of each of the test systems. The intent was to create a mechanistically informed IATA where the elements comprised the different short-term tests grouped together on the basis of their test capability. The resulting IATA would then be used to predict the classifications of a test set of carcinogens in accordance with the Globally Harmonized System (GHS) categories for mutagenicity (United Nations, 2013). This exercise is to an extent complementary to one recently performed by Benigni et al. (2013) who investigated the use of assays measuring DNA reactivity (such as Ames) and cell transformation assays to classify carcinogens into International Agency for Research on Cancer (IARC) classes 1 and 2.

2. Materials and methods

A dataset of 162 chemicals gathered as part of the previous publication (Mekenyan et al., 2012) was relied upon during the initial part of this study. The data available for these substances were categorized by their respective test capability. The outcomes for the tests across the levels of biological organization were also reconsidered in light of the test capabilities. The studies were reviewed applying expert scientific knowledge and only studies with data that met the end-point specific criteria were included. If the pattern of data was equivocal (positive and negative result) for same chemicals, the positive data was accepted to be predominant (i.e., acceptance of the worst case scenario).

The test systems included in the dataset originated from the following study types:

- Bacterial reverse mutation test (Ames test) (OECD Test Guideline (TG) 471; Ames et al., 1973; Mortelmans and Zeiger, 2000).
- Mammalian chromosome aberration test (OECD TG 473; Dean and Danford, 1984).
- Mouse lymphoma thymidine kinase locus test (OECD TG 476; Clive et al., 1979; Clements, 2000).
- *In vivo* liver unscheduled DNA synthesis (UDS) (OECD TG 486).
- *In vivo* alkaline single-cell gel electrophoresis (comet) (draft OECD TG 489; Olive and Banath, 2006; Collins, 2004).
- Transgenic rodent gene mutation assay (OECD TG 488; Nohmi et al., 2000; Lambert et al., 2005).
- Mammalian bone marrow chromosome aberration test (OECD TG 475).
- Mammalian bone marrow micronucleus assay (OECD TG 474).
- Rodent dominant lethal test (OECD TG 478; Bateman, 1984; Green et al., 1985)).

To apply the resulting IATA in practice, an exercise to explore how well chemicals could be classified in accordance with the GHS categories for germ cell mutagenicity was undertaken. A test set of 107 unique chemicals were taken from the Istituto Superiore di Sanita, Carcinogen database (ISSCAN) version 4a as extracted from the QSAR Toolbox OECD version 3.2. This was supplemented by extracting overlapping genotoxicity/carcinogenicity data from other databases available in the OECD Toolbox v3.2. The ISSCAN database is actually freely available at the Istituto Superiore di Sanita's website: http://www.iss.it/ampp/dati/cont.php?id=233&lang=1&tipo=7.

The ISSCAN database includes carcinogenic potency information as well as summary carcinogenicity calls taken from the Carcinogenic Potency Database (CPDB), http://toxnet.nlm.nih.gov/cpdb. Potency is assessed based on a threshold dose (TD50) measured in rats and mice. The TD50, is the chronic dose-rate in mg/kg body weight/day for life to induce tumors in half of the test animals that would have remained tumor-free at zero dose. The TD50 value reported is the harmonic mean of the most potent TD50 values from each positive experiment in the species. In this study, the summary carcinogenicity result for a chemical with disconcordant outcomes (i.e., both negative and positive results) was taken as positive, i.e., a worst case scenario. It is noted that this simplification obviously ignores important sex and species differences that need to be considered on a case by case basis. For reasons of pragmatism and to ensure a broad coverage of test set chemicals, this worst case scenario was deemed appropriate in this preliminary IATA study.

GHS calls for the classification of chemicals based on germ cell mutagenicity. The most recent update of the GHS is provided in 2013, which is available from: http://www.unece.org/trans/danger/publi/ghs/ghs_rev05/05files_e.html. Under GHS, germ cell mutagens are classified in one of two categories based on a weight-of-evidence assessment (UN, 2013). The GHS guidance guides the combination of test systems, whether they be tests for genotoxicity or mutagenicity, for the purposes of predicting mutagenicity categories.

3. Results and discussion

3.1. Refinement of the in vitro–in vivo workflow informed by test capability

Our previous workflow (Mekenyan et al., 2012) subdivided chemicals as positive or negative on the basis of outcomes from *in vitro* mutagenicity assays regardless of their test origin. Subsequent subcategorizations were based on *in vivo* genotoxicity effects in the liver, followed by results in the *in vivo* bone marrow micronucleus test (MNT). Substances with inconsistent outcomes across the different biological levels were rationalized by considering factors such as *in vivo* only effects, including so-called substrate channeling, resulting in metabolic detoxification of chemicals, or by an ability to interact with proteins (or other biomolecules) whilst approaching the remote bone marrow tissue etc. Although these could be arguably justified on a case by case basis, one shortcoming was our default of relying on at least one positive mutagenicity test result regardless of Level. Fig. 1 illustrates the test systems and how they were originally subcategorized.

At Level 1, an overall positive mutagenicity test call would be assigned based on at least one positive result, from either an Ames test with metabolic activation S9 (Ames-S9), an *in vitro* chromosomal aberration test (ivt CA) or a Mouse lymphoma tk assay (MLA). The same approach was used to assign *in vivo* mutagenicity calls at Levels 2 and 3. For example, a final mutagenicity call at Level 1 could be assigned as positive based on a negative Ames result and a positive ivt CA result. However, a chemical could still be assigned as negative at Level 2, because neither of the selected tests in Level 2 had a comparable test capability to that of the ivt CA test, which gave a positive result at Level 1. Accordingly the chemical could be positive *in vitro* but negative *in vivo* in the liver. For such cases we hypothesized that these chemicals were being detoxified in the liver, which may be practically implausible. The extent to which the *in vitro–in vivo* extrapolation workflow needed to be refined to address such shortcomings was considered.

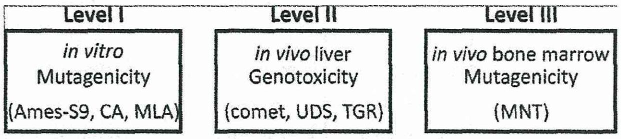

Fig. 1. Workflow adapted from Mekenyan et al. (2012).

Each of the assays reflected in the dataset was subcategorized in accordance with its presumed test capability. The revised categorizations are reflected in Fig. 2.

Level 1A for bacterial *in vitro* mutagenicity is represented by the Ames test with rat liver S9 exogenous metabolic activation (Ames-S9). This assay accounts for short length DNA damage (e.g., 2–3 nucleotides) only. Level 1B is characterized by *in vitro* mutagenicity as assessed in the mammalian chromosome aberration (ivt CA) test, which in general accounts for DNA and/or protein damage and the mouse lymphoma tk assay (MLA), which detects structural chromosome aberrations, aneuploidy, and recombination events (e.g., such as gene conversion) that result in loss of heterozygosity. Level 1 assays exhibit different though complementary test capabilities. Level 2 is subdivided into 3 groups, denoted by A, B and C. Group A is for *in vivo* genotoxicity in the liver as assessed in the comet assay and the unscheduled DNA synthesis (UDS) assay. The comet assay accounts for long length DNA damage (e.g., 20–30 nucleotides). The liver UDS test evaluates the role of DNA repair. Both are indicator tests measuring primary DNA damage. Group B is represented by the *in vivo* transgenic rodent mutation (TGR) assay, which detects point mutations. The TGR assay has a similar capability to the Ames test in that it takes into accounts the same type and extent of damage (i.e., short chain length DNA damage). Group C, *in vivo* mutagenicity assessed in the mammalian chromosome aberration test (iv CA), identifies clastogenic events (e.g., structural chromosome aberrations, aneuploidy). The *in vivo* mammalian chromosome aberration (iv CA) is similar in scope to the ivt CA test shown in Level 1B. Level 3 has only one category, *in vivo* mutagenesis and only one test, the bone marrow micronucleus test (MNT). This assay detects clastogenic activity in a site remote from the liver, which is the primary site of metabolic activation for mutagenic chemicals. The comet assay in Level 2A, the *in vivo* liver TGR in Level 2B and the iv CA in Level 2C are complementary to the MNT.

3.2. Mapping Ames test chemicals across the workflow

To investigate the utility of the revised workflow in practice, the original dataset of 162 chemicals were profiled and categorized by test type. Fig. 3 presents the results across the workflow for chemicals that are Level 1A (i.e., Ames test) positive.

Whilst nine chemicals could not be compared, the majority of *in vitro* Ames positive chemicals were found to be positive in the ivt CA test (70/78). Chemicals which were negative or not assessed in the ivt CA test were found to be mostly positive (14/17) in the liver *in vivo* comet and/or UDS test. This supports the expectation of consistent test capabilities between these Level 2A tests and the Ames test in Level 1A. Two different scenarios are proposed when comparing *in vitro* Ames and ivt CA positive chemicals with *in vivo* comet or UDS outcomes. Thirty-seven of the Level 1A and Level 1B positive chemicals are also positive in Level 2A assays. There is no Level 2B or Level 2C data to confirm the positive liver

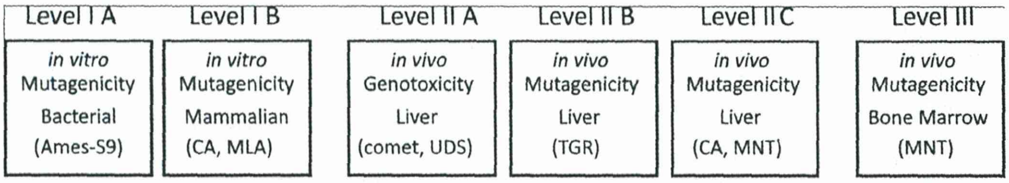
Level I A
Level I B
Level II A
Level II B
Level II C
Level III
in vitro Mutagenicity Bacterial (Ames-S9)
in vitro Mutagenicity Mammalian (CA, MLA)
in vivo Genotoxicity Liver (comet, UDS)
in vivo Mutagenicity Liver (TGR)
in vivo Mutagenicity Liver (CA, MNT)
in vivo Mutagenicity Bone Marrow (MNT)

Fig. 2. *In vitro–in vivo* workflow informed by test capability.

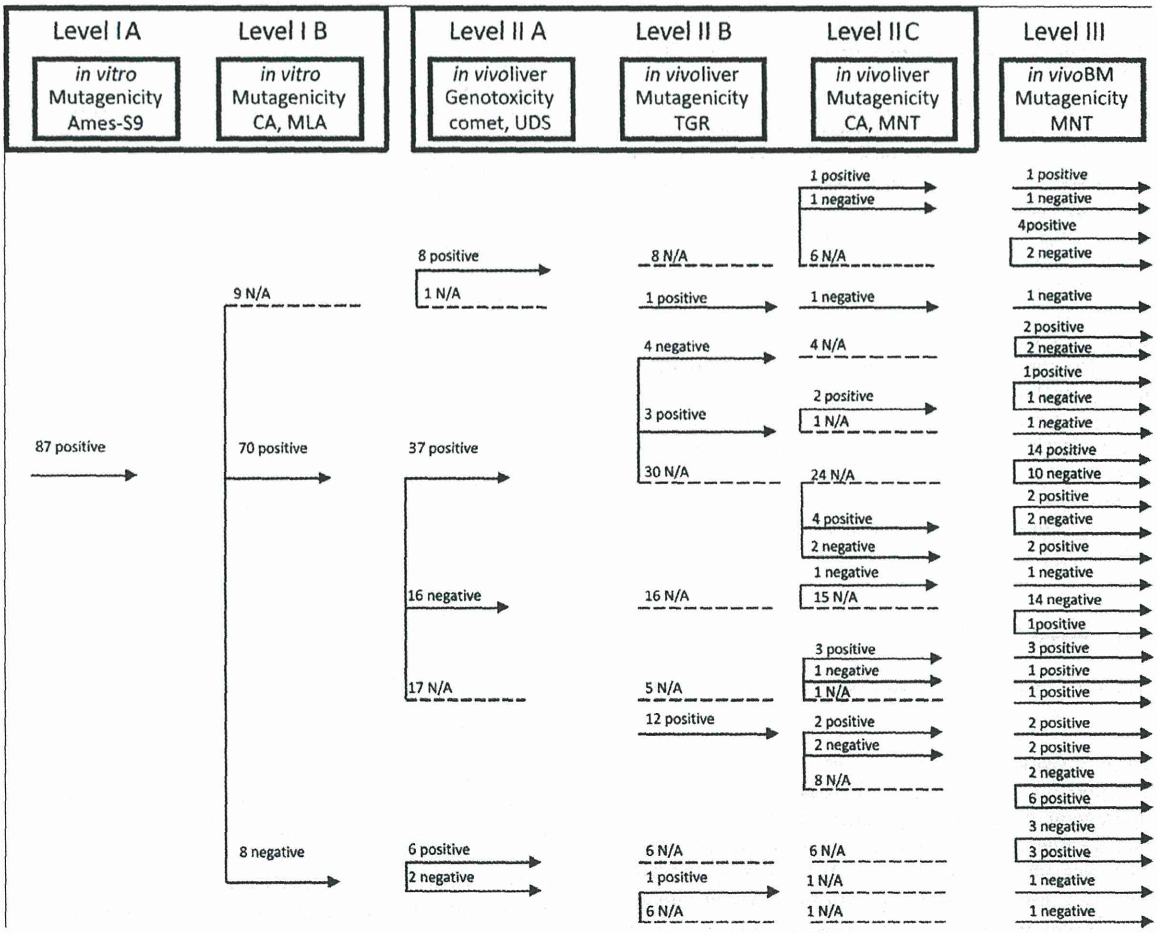

Fig. 3. Breakdown of the Ames positive test results across the workflow.

responses of these chemicals. Nevertheless, a number of these chemicals are negative in the *in vivo* MNT (i.e., Level 3).

For 17 chemicals where there is no data at Level 2A, a positive result at Level 1 is found with the liver *in vivo* TGR assay at Level 2B. This is an expected extrapolation based on the similar capability of the Ames and TGR tests. Sixteen Ames positive chemicals are negative at Level 2A. Whilst no data are available at Level 2B or 2C to corroborate the negative result at Level 2A, these chemicals are negative in the MNT (i.e., Level 3), suggestive of detoxification in the liver. *In vivo* detoxification can often account for the presence of *in vitro* positive and *in vivo* negative results for the same chemicals. Discrepancies between *in vitro* and *in vivo* results could also be explained by a number of other factors such as: non-physiological culture conditions, gender and species differences, cytotoxicity, incorrect route of administration, etc. Each of these specific experimental conditions may contribute to biologically irrelevant *in vitro* positive results (e.g., cytotoxicity) or indeed *in vivo* negative results. Given their complexity, these factors were not analyzed to identify *in vitro* or *in vivo* cytotoxic chemicals among the list of analyzed chemicals. Insufficient data were available to investigate the fate of the ivt CA positive chemicals in the *in vivo* CA test.

It is also important to recognize that *in vitro* cytotoxicity is not a singular phenomenon, it is a complex effect caused by a variety of mechanisms. The degree of cytotoxicity is often dependent on the method used for measuring it and the time at which the evaluation is performed. This is a significant weakness in the assessment of cytotoxicity, since the actual cytotoxicity may not be manifested for several hours after the time point at which the assessment is made, and yet the process of cytotoxicity could be well under way. Moreover, the relevance of cytotoxicity for the *in vivo* situation is highly questionable. On the other hand, *in vivo* cytotoxicity can result in tumor formation in rodents as result of persistent

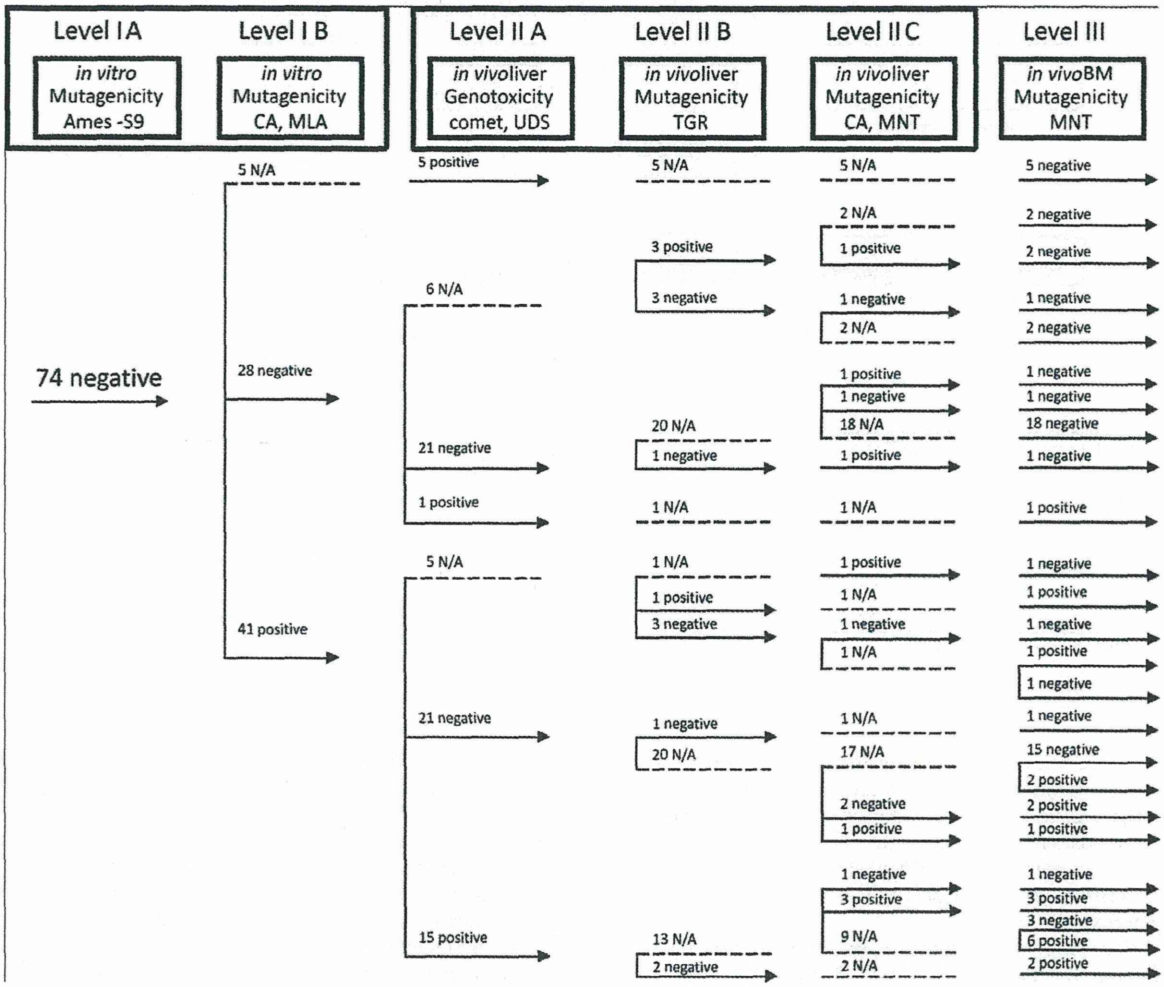

Fig. 4. Breakdown of the Ames negative test results across the workflow.

toxic effects that create an intracellular environment disturbing (or inducing) cell proliferation and ultimately leading to tumor formation.

The outcomes of Ames negative chemicals across the levels of biological organization were also reconsidered (Fig. 4).

Only 28 out of the 74 Level 1A negative chemicals were also negative at Level 1B. Of these 28 compounds, six were not assessed at Level 2A, whilst 21 out of 22 were negative at Levels 2A and 3. In contrast, 41 out of the 74 Level 1A negative chemicals are positive at Level 1B. Such results may be expected, as a positive CA test indicates either DNA or protein damage, whereas the Ames test only assesses DNA damage. Analyzing the *in vivo* fate of the 41 Level 1A negative and Level 1B positive chemicals, gives rise to two possible scenarios. Twenty-one of the Level 1B positive chemicals are negative at Level 2A, the same result as in Level 1A. When factoring in test capabilities, it becomes evident that these chemicals should not be considered as being detoxified in the liver.

Interestingly, there are 15 Level 1A negative chemicals which are positive at both Level 1B and Level 2A. Although there is a similarity in the test capabilities between the comet and Ames test, the extent of DNA damage between the 2 test systems is different. No data is available to analyze the fate of these chemicals at Level 2B but based on the positive data at Level 3, these chemicals remain positive across all *in vivo* assays.

Taking into account test capabilities appears to be helpful in interpreting different test outcomes and in designing strategic testing strategies.

3.3. Classification of chemicals into mutagenicity categories

In terms of applying the insights derived from the revised workflow in practice, an investigation was undertaken to determine its utility in predicting mutagenicity classification and labeling categories under GHS.

Under GHS, two categories are defined for germ cell mutagens, Category 1A, Category 1B and Category 2. Chemicals known to induce heritable mutations in germ cells of humans based on positive evidence from human epidemiological studies are defined as Category 1A; whereas, Category 1B chemicals are those which should be regarded as if they induce heritable mutations in the germ cells of humans on the basis of results of mammalian studies. Since it is extremely difficult to obtain reliable information from studies on the incidence of mutations in human populations, or on possible increases in their frequencies, for the present investigation, chemicals known to induce or regarded as inducing heritable mutations in germ cells in humans were combined into a single category (Category 1). Thus, to place a substance in Category 1,

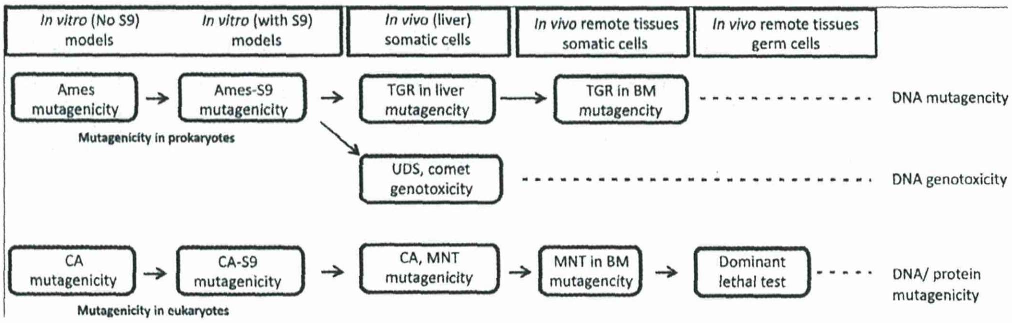

Fig. 5. IATA workflow for predicting mutagenicity categories.

Table 1
Performance of individual assays for predicting rodent carcinogenicity.

	In vitro					In vivo liver			In vivo BM	In vivo germ cells
Mutagenicity tests	Ames –S9	Ames +S9	CA –S9	CA +S9	MLA +S9	UDS, comet	TGR	MNT, CA	MNT	RDLT
#Chem	478	344	217	63	89	126	54	65	290	111
Carcinogenicity prediction										
Sensitivity, %	81	75	68	59	77	90	94	90	75	82
False positives, %	55	49	46	59	78	67	74	70	71	74

Note: RDLT = rodent dominant lethal test.

positive evidence from human or mammalian germ cell mutations (such as the dominant lethal test) would be needed.

For Category 2, positive *in vitro* mutagenicity results should be confirmed by positive *in vivo* somatic cell mutagenicity or genotoxicity results in mammals. In other words, Category 2 could be based on the combination of positive results from both mutagenicity and genotoxicity tests.

An additional category, Category 3, was proposed in this study to complement the existing GHS classification. The intent was not to formally expand the GHS classification; but to help discriminate between tests that assessed genotoxicity vs. mutagenicity[1]. Thus, chemicals for which there is positive data from an *in vivo* somatic cell mutagenicity test, with support from an *in vitro* mutagenicity test, are considered to be Category 2 mutagens. The new Category 3 is proposed to account for the situation where positive *in vivo* genotoxicity data (e.g., comet or UDS) are used in lieu of positive *in vivo* mutagenicity (e.g., TGR or MN) data yet supported by positive *in vitro* mutagenicity (e.g., Ames or ivt CA) data.

To account for the need for germ cell information, the IATA was extended to capture a component characterizing the dominant lethal test (DLT) (Fig. 5).

Chemicals positive in the dominant lethal test are classified as Category 1 mutagens without the need for additional positive mutagenicity data. This is due to the fact that, according to the GHS classification germ cells, mutagenicity is considered the apical endpoint. The role of germ cell tests in the regulations of different countries is still under discussion. US EPA, US FDA, Canada, UK, EU, and Japan place germ cell tests in Tier 2 and Tier 3. India and Australia only have Tier 1 tests and germ cell tests are not specifically included. According to the US FDA, US EPA and the European Food Safety Authority (EFSA), chemicals which are positive in somatic cells tests would normally be assumed to reach the germ cells and hence be considered as germ cell mutagens (EFSA, 2011). Chemicals negative in somatic cells tests would be assumed to be negative in germ cells. Thus, the conclusion is that routine testing for genotoxicity/mutagenicity in germ cells is not necessary. NIHS, Japan considers germ cell tests helpful in resolving conflicting situations where there are strong mutagens in somatic cells which are not carcinogenic. In these cases, the recommendation is that germ cell tests should be conducted in order to assess whether the mutagenic effect could eventually cause heritable diseases.

Chemicals having positive *in vitro* mutagenicity and positive *in vivo* mutagenicity data are classified as Category 2 mutagens e.g., a chemical found to be positive in the *in vitro* Ames and *in vivo* liver TGR tests in Fig. 5 is considered a Category 2 mutagen. In this case, defining a Category 2 mutagen is based on a combination of tests with the same test capability (i.e., DNA damage). However, different combinations of an *in vitro* and *in vivo* mutagenicity test could also be relied upon to classify a chemical as a Category 2 mutagen. Examples of Category 2 mutagens based on tests with the same capabilities include: a combination of *in vitro* Ames and *in vivo* TGR in bone marrow, an ivt CA and *in vivo* liver CA, or an ivt CA and *in vivo* MNT. Classification as a Category 2 mutagen could also be made on the basis of tests with different capabilities such as a combination of an *in vitro* Ames and *in vivo* liver CA, an *in vitro* Ames and *in vivo* MNT, an ivt CA and *in vivo* TGR, or an ivt CA and *in vivo* TGR in bone marrow.

Indeed the European Chemical Agency's (ECHA) ITS relies on a combination of assays with different capabilities (ECHA, 2014). For example, a mutagenicity classification could be assigned based on a combination of *in vitro* Ames and/or ivt CA and *in vivo* MNT. In this case, the combination of an *in vitro* Ames and *in vivo* MNT outcome would be preferable for a Category 2 mutagen classification. In contrast, NIHS's workflow for assessing impurities of pharmaceuticals requires combinations of assays with the same capabilities – a combination of *in vitro* Ames and *in vivo* liver TGR tests.

There are advantages and disadvantages in basing a classification on a combination of tests with similar or different capabilities. The advantage of a mutagenicity classification based on a combination of tests with the same capabilities is the high predictive certainty across mutagenicity pathways, in contrast, tests with different capabilities offers an expanded domain.

[1] GHS categorizes tests as mutagenicity or genotoxicity tests.

Category 3 mutagens as proposed could be determined using a combination of *in vitro* and *in vivo* assays with the same or differing capabilities. For example, a combination of an *in vitro* Ames and *in vivo* comet (or UDS) result may define a Category 3 mutagen by accounting for DNA-based genotoxicity outcomes. On the other hand, a combination of an ivt CA and *in vivo* comet (or UDS) result could define a Category 3 mutagen by accounting for DNA- and protein-based mutagenicity effects.

The above analysis describes the general working hypothesis of defining mutagenicity categories as combinations of positive results of mutagenicity tests. The requirement for mutagenicity tests covering a range of assays restricted the number and breadth of chemical classes addressed in this work. Thus, specific chemical classes such as the α,β-unsaturated aldehydes which are well known to be positive *in vitro* but negative *in vivo* were not included in this analysis. Unfortunately, such chemical classes were not available in the set of 107 unique chemicals used in this investigation.

3.4. Performance of mutagenicity categories for predicting rodent carcinogenicity

Using the test set of data, the performance of individual *in vitro* and *in vivo* assays for predicting rodent carcinogens was first examined before investigating combinations of assays based on either their similar or different test capabilities. Table 1 reflects the performance of the individual assays in predicting rodent carcinogenicity.

Most of the mutagenicity assays (e.g., Ames, comet, MNT, RDL) exhibited a high sensitivity to rodent carcinogenicity and a fairer performance in terms of specificity provided by *in vitro* Ames and ivt CA tests only. The net result of this investigation is a high sensitivity at the expense of a high rate of false positive carcinogenic predictions. The generated mutagenicity data are expected to be more reliable if all mutagenicity tests are conducted properly and cytotoxicity is accounted for. This will not only increase the sensitivity of the individual mutagenicity tests in predicting rodent carcinogenicity but also will increase their specificity (i.e., by eliminating would also be expected to reduce the number of false positives in the scheme for carcinogenicity prediction).

The high rate of false positive results was discussed during a workshop organized by European Centre for Validation of Alternative Methods (ECVAM) (see EFSA, 2011). A need for better guidance on the likely mechanisms resulting in positive results considered irrelevant to humans and their supporting evidence was identified.

A combination of two (or more) genotoxicity tests is a commonly used approach to increase predictivity of rodent carcinogenicity (Kirkland et al., 2005, 2006). If a combination of assays is based on the presumption that only a single positive genotoxicity result is considered as evidence that the substance is carcinogenic, then the sensitivity of the prediction increases. However, the specificity of such combinations will decrease because many positive predictions will be obtained for non-carcinogens. To reduce the number of false positives, we proposed an approach in which carcinogenicity is only assigned if positive results are found in both (or more) mutagenicity tests simultaneously. This approach is expected to significantly increase the confidence of a correct carcinogenicity prediction at the expense of a low rate of false positives. Such a combination of two assays is embedded in the definition of above presented GHS mutagenicity classification. The combinations of assays as defined by mutagenicity categories could be justified by a mechanistic rationale underlying the extrapolation workflow we have introduced. In the present investigation, we examined the performance of the three mutagenicity categories for predicting rodent carcinogenicity.

Table 2
Performance of the mutagenicity categories based on tests with the same capability.

Combination of *in vitro* and *in vivo* tests	Defined category mutagens	Sensitivity to carcinogens, (%)	Rate of false positive carcinogens, (%)	Total # chemicals
Ames and comet	Category 3	94 (31/33)	6 (2/33)	33
Ames and TGR	Category 2	100 (16/16)	–	16
ivt CA and CA	Category 2	100 (13/13)	–	13
ivt CA and MNT	Category 2	92 (36/39)	8 (3/39)	39
RDLT	Category 1	82 (37/45)	18 (8/45)	45

Initially, we examined the relationship between rodent carcinogenicity and mutagenicity categories, accounting for same test capability. Thus, the workflow used to define Category 3 mutagens derived by combining positive *in vitro* Ames and *in vivo* comet data and then relating it to rodent carcinogenicity (Table 2).

33 chemicals were found to overlap between defined Category 3 mutagens and carcinogenicity. Obviously, relationships between Category 3 mutagens and rodent carcinogens indicated a very good performance in terms of sensitivity. Only 2 out of the 33 chemicals were found to be positive in Category 3 and negative according to observed carcinogenicity. Such a small deviation could be expected given the limited domain of Category 3 mutagens (i.e., not all positive comet data indicating genotoxicity will ultimately result in mutations). Although there is a limited number of overlapping chemicals, the net result of this investigation is a low rate (6%) of false positive carcinogenicity predictions.

In the next analysis, we examined the relationship between Category 2 mutagens, defined by *in vitro* Ames and *in vivo* TGR and rodent carcinogenicity. Small numbers of chemicals (16 chemicals) were found overlapping between Category 2 mutagens and carcinogenicity. All Category 2 mutagens were found to be positive carcinogens and thus providing a performance of 100% in terms of sensitivity.

A similar investigation was used to relate mutagenicity categories defined based on test accounting for DNA and/or protein damage along the categorization workflow. Carcinogens were correctly predicted based on a combination of ivt CA and *in vivo* CA (or MN) tests. Here, all chemicals belonging to Category 2 mutagens based on a combination of positive data in tests with the same capability were found to be carcinogens.

The relationship between Category 2 mutagens, based on positive ivt CA and *in vivo* MNT data and carcinogenicity, also indicated very high sensitivity (92%). Only 3 out of 39 (8%) chemicals were found to be mutagenic based on positive ivt CA and *in vivo* MNT data and observed to be negative in carcinogenicity tests. This data inconsistency could be due to factors such as cytotoxicity. In-depth analysis of the experimental conditions of both ivt CA and *in vivo* MNT tests would be needed to put the two positive results into perspective. Nonetheless, the small rate of false positive carcinogens indicates that chemicals which are positive at the same time in two mutagenicity tests sharing a similar capability for detecting chromosome breakages correlated more than 90% of the time with rodent carcinogenicity.

Category 1 is defined based on positive *in vivo* data in the DLT. Thus, the relationship between Category 1 mutagens and carcinogenicity is based on 45 overlapping chemicals. Interestingly, 37 out of 45 (82%) Category 1 chemicals, based on a single positive *in vivo* result in DLT, are observed to also be carcinogens. Only 8 out of 45 (18%) chemicals appear to be mutagens and could possibly elicit heritable disease but are not carcinogenic. However, most germ-cell mutagens to date have been identified using only male rats, thereby efficiently precluding comparisons of gender differences when comparing DLT and rodent carcinogenicity results.

Table 3
Performance of the mutagenicity categories based on tests with different capabilities for predicting positive rodent carcinogenicity.

Combination of *in vitro* and *in vivo* tests	Defined category mutagens	Sensitivity to positive carcinogenicity, (%)	Rate of false positive carcinogens, (%)	Total # chemicals
Ames and liver CA/MNT	Category 2	100 (12/12)	–	12
Ames and MNT	Category 2	95 (37/39)	5 (2/39)	39
Ames, ivt CA and MNT	Category 2	97 (29/30)	3 (1/30)	30
ivt CA and comet/UDS	Category 3	91 (21/23)	9 (2/23)	23

Table 4
Distribution of 107 unique chemicals across different chemical classes.

#	Chemical classes	# Chemicals	# False positives
1	Aromatic amines	14	1
2	N-nitrosamines	11	–
3	Nitroaromatics	8	2
4	Organochlorines	8	–
5	Azoarenes	6	–
6	Nitrogen mustards	5	–
7	Heteroaromatics	5	4
8	Aziridines	4	–
9	Alkylsulfonates	4	–
10	Epoxides	4	–
11	Phosphate(thiophosphate) esters	4	1
12	Phenols	3	2
13	Acetanilide derivatives	2	1
14	Aliphatic amines	1	1
15	Miscellaneous	28	–

However, despite this discrepancy, the small number of false positive predictions indicated that most positive DLT chemicals will be also carcinogens.

A similar analysis was performed to investigate the relationship between Category 2 and 3 mutagens and rodent carcinogenicity by defining the categories based on a combination of tests with different capabilities. For example, *in vitro* Ames and *in vivo* MNT are used to define Category 2 and ivt CA and *in vivo* comet to define Category 3 (Table 3).

The result of this investigation is almost the same as that observed when accounting for same test capabilities – sensitivity of 90–100% at the expense of a relatively small number (3–10%) of false positive carcinogenicity predictions. Based on the limited databases used to relate mutagenicity and carcinogenicity that was available for study, the ultimate conclusion about the advantage of using tests with same or different capabilities could not be conclusively established.

Prediction of non-carcinogens was also investigated based on a combination of negative results from the same couple of assays used to define the three mutagenicity categories. In all cases, a combination of two negative mutagenicity results was found to be insufficient for predicting non-carcinogens. The high rate of false negatives (about 50%) is presumably due to the fact that many different mechanisms, not necessarily related to genotoxicity, are involved in cancer formations. This investigation will be expanded in a separate paper.

In summary, combining results from *in vitro* and *in vivo* tests with similar capabilities appears to improve predictivity principally by reducing the number of false positives. On the other hand, the tests used, whether singly or in combination, have a high rate of false negatives with respect to predicting carcinogenicity. Whilst the numbers of chemicals tested in all the assays necessary to demonstrate the proposed assay couplings is limited, the dramatic drop in false positives, as compared to individual assay is noteworthy. 107 unique chemicals were used to relate mutagenicity categories with carcinogenicity. Some of these chemicals existed in two of more mutagenicity categories at the same time, thus resulting in 250 outcomes. It should be noted that, some of the genotoxicity tests (ivt CA, *in vitro* comet) used in this approach are prone to false positive results due to cytotoxicity. Hence, when a limited number of chemicals are used to relate mutagenicity to carcinogenicity, the issue of cytotoxicity may not be evident. However, if a large database is screened, a peer data review will be needed in an attempt to eliminate cytotoxic chemicals from the list of false positives. The distribution of the 107 unique chemicals across different chemical classes is presented in Table 4.

The most pronounced chemical classes are: aromatic amines, N-nitrosamines, nitroaromatics, organochlorines, azoarenes, etc. The exhaustive list of false positives includes: 6-mercaptopurine, chloramphenicol, pyrimethamine, methotrexate, bisphenol A, 1,2-diamino-4-nitrobenzene, acetaminophen, m-phenylenediamine, cyclohexylamine, phenol, thiabendazole and chlorpyrifos. Most of these chemicals are mutagenic only in germ cells (i.e., DLT) which flag for heritable diseases but which do not culminate in tumor formation. It is still difficult to explain lack of carcinogenicity of the other compounds (acetaminophen, phenol, thiabendazole and chlorpyrifos) which are somatic cell mutagens.

4. Conclusions

An *in vitro–in vivo* extrapolation workflow for genotoxicity previously developed was refined on the basis of test capabilities. Reevaluating the dataset of 162 chemicals through this revised workflow addressed some of the shortcomings identified when interpreting study outcomes. This had implications for the manner in which the original TIMES *in vivo* models were developed and the insights were used to make subsequent refinements. The revised workflow was extended in an effort to investigate its practical utility in predicting GHS mutagenicity categories for rodent carcinogens. Evaluating the performance of various combinations of test systems accounting for their similarity or differences in test capability showed little variation. This could of course have been biased on account of the limited test set of 250 mutagenicity calls available for study.

Future work will include extending the workflow to incorporate outcomes from available TIMES models and implementing the workflow into a practical IATA as a software tool for systematic use. Such a tool would facilitate a guided weight of evidence assessment for rodent carcinogens which are also genotoxic on the basis of available experiment data coupled with predicted outcomes from TIMES models. Extending the chemical dataset to include chemicals from the chemical classes of heteroaromatics, nitroaromatics, phenols and aromatic amines is of high priority. Efforts to account for test sex and species differences will also be investigated.

Conflicts of interest statement

At the time of this work, Grace Patlewicz and Maria Donner were employees of DuPont and received no other compensation but their salary for this work. DuPont has a license agreement with the Laboratory of Mathematical Chemistry (LMC) to access their suite of software models and help fund their development.

Acknowledgments

This work is dedicated to Dr. Gilman D. Veith given his efforts in the area of predictive carcinogenicity. We also wish to thank Drs.

Takeshi Morita, Shuichi Hamada, Akihiro Wakata, Masayuki Mishima, and Jiro Maniwa from the National Institute of Health Sciences, Japan who collected and peer reviewed genotoxicology data used in this study, in particular, data from the *in vivo* micronucleus tests.

References

Ames, B.N., Durston, W.E., Yamasaki, E., Lee, F.D., 1973. Carcinogens are mutagens: a simple test system combining liver homogenates for activation and bacteria for detection. Proc. Natl. Acad. Sci. U.S.A. 70, 2281–2285.

Ankley, G.T., Bennett, R.S., Erickson, R.J., Hoff, D.J., Hornung, M.W., Johnson, R.D., Mount, D.R., Nichols, J.W., Russom, C.L., Schmieder, P.K., Serrrano, J.A., Tietge, J.E., Villeneuve, D.L., 2010. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Environ. Toxicol. Chem. 29, 730–741.

Bateman, A.J., 1984. The dominant lethal assay in the male mouse. In: Kilbey, B.J., Legator, M., Nichols, W., Ramel, C. (Eds.), Handbook of Mutagenicity Test Procedures, second ed. Elsevier Scientific, Amsterdam, pp. 471–484.

Benigni, R., Bossa, C., Worth, A., 2010. Structural analysis and predictive value of the rodent *in vivo* micronucleus assay results. Mutagenesis 25, 335–341.

Benigni, R., Bossa, C., Battistelli, C.L., Tcheremenskaia, O., 2013. IARC classes 1 and 2 carcinogens are successfully identified by an alternative strategy that detects DNA-reactivity and cell transformation ability of chemicals. Mutat. Res. 758, 56–61.

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

Clements, J., 2000. The mouse lymphoma assay. Mutat. Res. 455, 97–110.

Clive, D., Johnson, K.O., Spector, J.F., Batson, A.G., Brown, M.M., 1979. Validation and characterization of the L5178Y/TK+/− mouse lymphoma mutagen assay system. Mutat. Res. 59, 61–108.

Collins, A.R., 2004. The comet assay for DNA damage and repair: principles, applications, and limitations. Mol. Biotechnol. 26, 249–261.

Combes, R., Grindon, C., Cronin, M.T., Roberts, D.W., Garrod, J., 2007. Proposed integrated decision-tree testing strategies for mutagenicity and carcinogenicity in relation to the EU REACH legislation. Altern. Lab. Anim. 35, 267–287.

Dean, B.J., Danford, N., 1984. Assays for the detection of chemically-induced chromosome damage in cultured mammalian cells. In: Venitt, S., Parry, J.M. (Eds.), Mutagenicity Testing – A Practical Approach. IRL Press, Oxford, pp. 187–232.

Dearfield, K.L., Cimino, M.C., Mc Carroll, N.E., Mauer, I., Valcovic, L.R., 2002. Genotoxicity risk assessment: a proposed classification strategy. Mutat. Res. 521, 121–135.

European Chemicals Agency (ECHA), 2014. Guidance on information requirements and Chemical Safety Assessment Chapter R.7a: Endpoint specific guidance, Version 3.0. ISBN: 978-92-9244-749-6.

European Food Safety Authority (EFSA), 2011. Scientific opinion on genotoxicity testing strategies applicable to food and feed safety assessment. EFSA J. 9, 2379–2448.

Green, S., Auletta, A., Fabricant, J., Kapp, R., Manandhar, M., Sheu, C.J., Springer, J., Whitfield, B., 1985. Current status of bioassays in genetic toxicology-the dominant lethal assay. A report of the U.S. Environmental Protection Agency Gene-Tox Program. Mutat. Res. 154, 49–67.

Grindon, C., Combes, R., Cronin, M.T., Roberts, D.W., Garrod, J.F., 2006. Integrated testing strategies for use in the EU REACH system. Altern. Lab. Anim. 34, 407–427.

Lambert, I.B., Singer, T.M., Boucher, S.E., Douglas, G.R., 2005. Detailed review of transgenic rodent mutation assays. Mutat. Res. 590, 1–280.

Kirkland, D., Aardema, M., Henderson, L., Müller, L., 2005. Evaluation of the ability of a battery of 3 *in vitro* genotoxicity tests to discriminate rodent carcinogens and non-carcinogens. I. Sensitivity, specificity and relative predictivity. Mutat. Res. 584, 1–256.

Kirkland, D., Aardema, M., Müller, L., Hayashi, M., 2006. Evaluation of the ability of a battery of three *in vitro* genotoxicity tests to discriminate rodent carcinogens and non-carcinogens II. Further analysis of mammalian cell results, relative predictivity and tumour profiles. Mutat. Res. 608, 29–42.

Kirkland, D., Pfuhler, S., Tweats, D., Aardema, M., Corvi, R., Darroudi, F., Elhajouji, A., Glatt, H., Hastwell, P., Hayashi, M., Kasper, P., Kirchner, S., Lynch, A., Marzin, D., Maurici, D., Meunier, J.R., Müller, L., Nohynek, G., Parry, J., Parry, E., Thybaud, V., Tice, R., van Benthem, J., Vanparys, P., White, P., 2007a. How to reduce false positive results when undertaking *in vitro* genotoxicity testing and thus avoid unnecessary follow-up animal tests: report of an ECVAM workshop. Mutat. Res. 628, 31–55.

Kirkland, D.J., Aardema, M., Banduhn, N., Carmichael, P., Fautz, R., Meunier, J.R., Pfuhler, S., 2007b. *In vitro* approaches to develop weight of evidence (WoE) and mode of action (MoA) discussions with positive *in vitro* genotoxicity results. Mutagenesis 22, 161–175.

Kirkland, D., Reeve, L., Gatehouse, D., Vanparys, P., 2011. A core *in vitro* genotoxicity battery comprising the Ames test plus the *in vitro* micronucleus test is sufficient to detect rodent carcinogens and *in vivo* genotoxins. Mutat. Res. 721, 27–73.

Kirkland, D., Zeiger, E., Madia, F., Gooderham, N., Kasper, P., Lynch, A., Morita, T., Ouedraogo, G., Morte, J.M.P., Pfuhler, S., Rogiers, V., Schulz, M., Thybaud, V., van Benthem, J., Vanparys, P., Worth, A., Corvi, R., 2014. Can *in vitro* mammalian cell genotoxicity test results be used to complement positive results in the Ames test and help predict carcinogenic or *in vivo* genotoxic activity? I. Reports of individual databases presented at an EURL ECVAM workshop. Mutat. Res. 775–776, 55–68.

Matthews, E.J., Kruhlak, N.L., Cimino, M.C., Benz, R.D., Contrera, J.F., 2006. An analysis of genetic toxicity, reproductive and development toxicity, and carcinogenicity data: I. Identification of carcinogens using surrogate endpoints. Regul. Toxicol. Pharmacol. 44, 83–96.

Mekenyan, O.G., Petkov, P.I., Kotov, S.V., Stoeva, S., Kamenska, V.B., Dimitrov, S.D., Honma, M., Hayashi, M., Benigni, R., Donner, E.M., Patlewicz, G., 2012. Investigating the relationship between *in vitro*–*in vivo* genotoxicity: derivation of mechanistic QSAR models for *in vivo* liver genotoxicity and *in vivo* bone marrow micronucleus formation which encompass metabolism. Chem. Res. Toxicol. 25, 277–296.

Mortelmans, K., Zeiger, E., 2000. The Ames Salmonella/microsome mutagenicity assay. Mutat. Res. 455, 29–60.

Nohmi, T., Suzuki, T., Masumura, K., 2000. Recent advances in the protocols of transgenic mouse mutation assays. Mutat. Res. 455, 191–215.

Olive, P.L., Banath, J.P., 2006. The comet assay: a method to measure DNA damage in individual cells. Nat. Protoc. 1, 23–29.

Tollefsen, K.E., Scholz, S., Cronin, M.T., Edwards, S.W., de Knecht, J., Crofton, K., Garcia-Reyero, N., Hartung, T., Worth, A., Patlewicz, G., 2014. Applying Adverse Outcome Pathways (AOPs) to support Integrated Approaches to Testing and Assessment (IATA). Regul. Toxicol. Pharmacol. 70, 629–640.

United Nations (UN), 2013. Globally harmonized system of classification and labelling of chemicals (GHS) fifth revised edition. (Available at: <http://www.unece.org/trans/danger/publi/ghs/ghs_rev05/05files_e.html>).

Zeiger, E., 1998. Identification of rodent carcinogens and noncarcinogens using genetic toxicity tests: permises. Promises Perform. Regul. Toxicol. Pharmacol. 28, 85–95.