Available online at www.sciencedirect.com
ScienceDirect
Mutation Research 627 (2007) 41–58
Genetic Toxicology and Environmental Mutagenesis
www.elsevier.com/locate/gentox
Community address: www.elsevier.com/locate/mutres

Strategy for genotoxicity testing: Hazard identification and risk assessment in relation to *in vitro* testing

V. Thybaud [a,*], M. Aardema [b], J. Clements [c], K. Dearfield [d], S. Galloway [e], M. Hayashi [f], D. Jacobson-Kram [g], D. Kirkland [c], J.T. MacGregor [h], D. Marzin [i], W. Ohyama [j], M. Schuler [k], H. Suzuki [l], E. Zeiger [m]

[a] *Drug Safety Evaluation, Sanofi-aventis, 94400 Vitry sur Seine, France*
[b] *The Procter & Gamble Co., P.O. Box 538707, Cincinnati Ohio 45253, USA*
[c] *Covance Laboratories Ltd., Otley Road, Harrogate HG3 1PY, UK*
[d] *Food Safety and Inspection Service, Office of Public Health Science, U.S. Department of Agriculture, Washington, DC 20250, USA*
[e] *Merck Research Laboratories, West Point, PA 19486, USA*
[f] *Division of Genetics and Mutagenesis, National Institute of Health Sciences, Tokyo, Japan*
[g] *Office of New Drugs, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD 20993, USA*
[h] *Toxicology Consulting Services, Arnold, MD 21012, USA*
[i] *Institut Pasteur de Lille, BP 245, 59019 Lille Cedex, France*
[j] *Yakult Central Institute for Microbiological Research, 1796 Yaho, Kunitach-shi, Tokyo 186-8650, Japan*
[k] *Pfizer Global Research & Development, Groton, CT 06340, USA*
[l] *Ina Research Inc., 2148-188 Nishiminowa, Ina-shi, Nagano 399-4501, Japan*
[m] *Errol Zeiger Consulting, Chapel Hill, NC 27514, USA*

Received 28 February 2006; received in revised form 31 July 2006; accepted 14 August 2006
Available online 27 November 2006

Abstract

This report summarizes the proceedings of the September 9–10, 2005 meeting of the Expert Working Group on Hazard Identification and Risk Assessment in Relation to *In Vitro* Testing, part of an initiative on genetic toxicology. The objective of the Working Group was to develop recommendations for interpretation of results from tests commonly included in regulatory genetic toxicology test batteries, and to propose an appropriate strategy for follow-up testing when positive *in vitro* results were obtained in these assays. The Group noted the high frequency of positive *in vitro* findings in the genotoxicity test batteries with agents found not to be carcinogenic and thought not to pose a carcinogenic health hazard to humans. The Group agreed that a set of consensus principles for appropriate interpretation and follow-up testing when initial *in vitro* tests are positive was needed. Current differences in emphasis and policy among different regulatory agencies were recognized as a basis of this need. Using a consensus process among a balanced group of recognized international authorities from industry, government, and academia, it was agreed that a strategy based on these principles should include guidance on: (1) interpretation of initial results in the "core" test battery; (2) criteria for determining when follow-up testing is needed; (3) criteria for selecting appropriate follow-up tests; (4) definition of when the evidence is sufficient to define the mode of action and the relevance to human exposure; and (5) definition of approaches to evaluate the degree of health risk under conditions of exposure of the species of concern (generally the human).

A framework for addressing these issues was discussed, and a general "decision tree" was developed that included criteria for assessing the need for further testing, selecting appropriate follow-up tests, and determining a sufficient weight of evidence to attribute a level of risk and stop testing. The discussion included case studies based on actual test results that illustrated common

* Corresponding author. Tel.: +33 1 58 93 35 38; fax: +33 1 58 93 38 44.
E-mail address: Veronique.Thybaud@sanofi-aventis.com (V. Thybaud).

1383-5718/$ – see front matter
doi:10.1016/j.mrgentox.2006.10.003

situations encountered, and consensus opinions were developed based on group analysis of these cases. The Working Group defined circumstances in which the pattern and magnitude of positive results was such that there was very low or no concern (*e.g.*, non-reproducible or marginal responses), and no further testing would be needed. This included a discussion of the importance of the use of historical control data. The criteria for determining when follow-up testing is needed included factors, such as evidence of reproducibility, level of cytotoxicity at which an increased DNA damage or mutation frequency is observed, relationship of results to the historical control range of values, and total weight of evidence across assays. When the initial battery is negative, further testing might be required based on information from the published literature, structure activity considerations, or the potential for significant human metabolites not generated in the test systems. Additional testing might also be needed retrospectively when increase in tumors or evidence of pre-neoplastic change is seen.

When follow-up testing is needed, it should be based on knowledge about the mode of action, based on reports in the literature or learned from the nature of the responses observed in the initial tests. The initial findings, and available information about the biochemical and pharmacological nature of the agent, are generally sufficient to conclude that the responses observed are consistent with certain molecular mechanisms and inconsistent with others. Follow-up tests should be sensitive to the types of genetic damage known to be capable of inducing the response observed initially. It was recognized that genotoxic events might arise from processes other than direct reactivity with DNA, that these mechanisms may have a non-linear, or threshold, dose-response relationship, and that in such cases it may be possible to determine an exposure level below which there is negligible concern about an effect due to human exposures. When a test result is clearly positive, consideration of relevance to human health includes whether other assays for the same endpoint support the results observed, whether the mode or mechanism of action is relevant to the human, and – most importantly – whether the effect observed is likely to occur *in vivo* at concentrations expected as a result of human exposure. Although general principles were agreed upon, time did not permit the development of recommendations for the selection of specific tests beyond those commonly employed in initial test batteries.

Keywords: Genotoxicity; Hazard identification; Risk assessment; Testing strategy

1. Introduction

Genetic toxicity testing is routinely performed to identify potential genotoxic carcinogens and germ cell mutagens. With regard to the identification of genotoxic carcinogens, all the minimal batteries of genetic toxicology tests recommended by regulatory agencies include at least two or three test procedures, generally an Ames test, a mammalian cell chromosome damage test, and in some cases a mammalian cell mutation assay [1–15]. Depending on the responses in the tests, the types of substances tested, and on their intended uses (*e.g.*, pharmaceuticals, pesticides, chemicals, cosmetics, *etc.*), one or more *in vivo* rodent tests (*e.g.*, bone marrow micronucleus; liver UDS) also have to be conducted [1–15].

The standard batteries of tests are selected to address two types of genetic damage of concern, *i.e.*, gene mutations and chromosome damage. Some more recent guidelines [1,15] have suggested inclusion of the *in vitro* micronucleus test to detect chromosome loss. Additional tests may be needed to clarify the substances' activity, or to determine if the activity seen in the initial testing is relevant. Such additional testing may include investigation of aneuploidy, chromosome non-disjunction, DNA interaction, and/or primary DNA damage [1,7].

The International Workshop on Genotoxicity Testing (IWGT) was formed in 1993 to bring together internationally recognized experts to examine genetic testing methods and strategies through meetings and workshops. Three workshops have been held so far, in 1993, 1999 and 2002. At the 2002 workshop, the IWGT Steering Committee initiated a discussion of testing strategies, but only limited topics were discussed and agreed upon at that time [16]. One topic that was not finalized was the selection of follow-up testing approaches following *in vitro* positive or equivocal results in the test battery. To address this issue, the IWGT Working Group reconvened during a fourth IWGT workshop in San Francisco, California, on September 9–10, 2005. The primary objective of this Working Group was to define the most appropriate follow-up testing strategy in case of positive results, and not to recommend revisions to the current batteries of tests.

During the past 30 years, genetic toxicology testing has been mainly used for hazard identification. Nevertheless, it is recognized that the discipline of regulatory genetic toxicology testing should consider moving from hazard identification towards an integrated risk assessment. At this IWGT meeting, the Working Group decided to focus on strategies for assessing the risk of cancer, although the importance of other health consequences of genetic damage to somatic and germ cells was recognized.

This publication describes those areas where a consensus was achieved among the members of the Working Group, and identifies areas that were discussed but not

resolved, or were not addressed because of time constraints or lack of available data. A general approach to determining the need for follow-up testing was defined. Case study examples are given that illustrate the extent of information (weight of evidence) needed to reach decisions about the extent of risk (risk assessment) based on the available data about the mode of action from testing results. Further, the steps needed to make recommendations about specific tests to be used in follow-up testing strategies were discussed.

2. Weight of evidence and mode of action considerations

Current regulatory practice often involves decision-making based on the results of batteries of tests designated by applicable regulatory agencies. These test batteries rely heavily on *in vitro* tests. Although there is general agreement that *in vitro* tests are useful for identifying potential genotoxic carcinogens and mutagens, the high incidence of positive findings in these *in vitro* assays [10] with agents that appear not to pose a carcinogenic health risk under certain conditions of exposure implies that reliable cancer health risk determinations cannot be made on the basis of *in vitro* findings alone. Recognition of the high rate of positive findings in *in vitro* assays has created a need for consensus agreement about how these results should be interpreted, and how appropriate follow-up testing should be structured in order to define the risks to humans.

The need to place findings into the context of their relevance to the health risks associated with specific exposures is, of course, not new. It has been recognized since regulatory mutagenicity testing was instituted in the mid-1970's that assessment of *in vivo* risk was an essential component of mutagenicity testing. A department-wide committee of the Department of Health Education and Welfare in the U.S. (the approximate equivalent of the current Department of Health and Human Services) was convened in the mid-1970's to recommend appropriate approaches to regulatory mutagenicity testing in the United States. The stated objective of this committee was "...to aid officials of regulatory agencies who have the responsibility for deciding: (1) advisability of promulgating test requirements for mutagenicity at the present time under any of their legislative authorities; (2) the appropriateness of mutagenicity tests for a wide range of product use and exposure categories; and (3) the reliability and interpretation of data from mutagenicity tests developed on substances of commerce within their regulatory purview in spite of the absence of formal testing requirements". This report [17], *Approaches to Determining the Mutagenic Properties of Chemicals: Risk to Future Generations*, stated:

> "It is not sufficient merely to identify substances which may pose a genetic hazard to the human population. Many such compounds will have a significant benefit factor and hence cannot reasonably be eliminated from use. Therefore, it is necessary to obtain quantitative data from relevant animal model systems from which extrapolation to humans can be made to predict virtually safe or tolerable levels of exposure".

Subsequently, the appropriate steps for conducting risk assessments and risk characterization of mutagens have been addressed [18] and the International Commission for Protection Against Environmental Carcinogens and Mutagens (ICPEMC) has delineated and published a detailed approach and recommendations [19,20]. The ICPEMC recommendations follow closely the general principles of risk assessment established by the landmark National Academy of Sciences report on risk assessment in the U.S. Federal government [21]. The evolution of strategies for assessment of mutagenic risk has been reviewed by MacGregor et al. [22].

Although the principles of risk assessment from exposures to genotoxic agents have been delineated, application of these principles varies within different regulatory agencies. Most place a major emphasis on mutagenicity data as a part of the weight of evidence for cancer risk assessment. The U.S. Environmental Protection Agency (EPA), in particular, established a procedure that incorporates information on mode of action as the focus of the risk assessment approach taken in the EPA cancer risk assessment guidelines [23]. The guidelines define the term "*mode of action*" as a sequence of key events and processes, starting with interaction of an agent with a cell, proceeding through operational and anatomical changes, and resulting in cancer formation. Increasingly, one of the key events and processes being considered is whether the carcinogenic agent is mutagenic/genotoxic or not, and if so, whether that genotoxic activity contributes to the induction of cancer (approach described in Dearfield and Moore [24]). Further, a genotoxic mode of action is coupled with more stringent regulatory control. For example, as described in a supplement to its cancer guidelines, if a mutagenic mode of action is determined for the induction of cancer, EPA will apply age-dependent adjustment factors to the cancer slope factor under certain conditions to assure further protection from early life exposures to the chemical [25]. As another example, genotoxic agents are generally restricted from use in healthy volunteers in clinical trials of new therapeutic agents [26,27].

The impact of genotoxicity data on regulatory decisions, interpretation of positive findings with respect to human risk, and the degree of quantitative risk assessment applied to genotoxicity data also varies among agencies. For example, most agencies acknowledge that mutagenic damage to germ cells constitutes a risk to subsequent generations [3,7,13,15,28], but it is quite rare to conduct a formal risk assessment of either germinal or somatic cell mutagenesis. EPA has a formal procedure for assessing germ cell risk [28], but has only conducted formal germ cell risk assessments in a few cases (ethylene oxide [29]; acrylamide [30]). For industrial chemicals [15] and for plant protection products [8] in Europe, an assessment of germ cell risk is required if a substance demonstrates genotoxic activity in somatic cells *in vivo*. The UK Committee on Mutagenicity guidance also includes assessment of germ cell risk as a separate exercise [1]. Although no formal guidance exists, experience indicates that European agencies assessing the safety of new medicines also view germ cell risk as separate from cancer risk. The U.S. Food and Drug Administration (FDA), in contrast, generally assumes that control of cancer risk also controls other health risks associated with genotoxic activity, and does not request quantitative assessments of germ cell risk. The approaches used by different agencies have recently been reviewed by Cimino [31].

In view of the differences in emphasis and policy among different regulatory agencies, and due in large part to the extent of positive findings in the genotoxicity tests with agents found not to be carcinogenic and thought not to pose a carcinogenic health hazard to humans under conditions of anticipated exposure [10,32–34], there is a general agreement among scientists and regulators in the field that a more detailed set of consensus principles for appropriate interpretation and follow-up testing when screening tests are positive would be useful. The focus of several important bodies on this issue is testimony to this general consensus; these include this IWGT effort, a recently initiated collaborative project of the International Life Sciences Institute (ILSI) on the relevance and follow-up of positive results in the genetic toxicology testing (http://www.hesiglobal.org/Committees/EmergingIssues/toxtesting/), and the issuance by the FDA of a new guidance that emphasizes a weight of evidence approach to assessing the relevance of genotoxicity test results [26,27]. In order to ensure recognition and widespread adoption of such principles, it is important that they be achieved *via* a consensus process among recognized international authorities in the field.

3. Negative results that may require follow-up testing

In most cases, when a chemical is found negative in the initial regulatory battery of tests, and appropriate conditions have been used, follow-up testing is not required. However, there are some situations in which additional testing may be necessary even when an initial regulatory battery of tests is negative. Such cases are also discussed in other reports in this volume (*e.g.*, Ku et al.; Tweats et al.). One important consideration is the relative metabolism in the laboratory model versus the human. Metabolism studies may show that humans generate a metabolite from the chemical under scrutiny that is not seen in the animal or cellular laboratory models (including rat liver S9 used in the tests). In this case, the chemical would not have been properly evaluated for human risk. Typically, if the metabolite were present at significant levels in human, additional testing with the metabolite itself (or systems that produced it) would be necessary to fully assess the potential of the chemical to induce genotoxic effects in humans [35].

In some cases, results from studies *in vivo* may suggest a need for additional genotoxicity testing. For example, positive or equivocal results in rodent carcinogenicity assays, epidemiology evidence in humans, or as suggested by some Working Group participants observation of pre-neoplastic lesions in toxicity studies, may trigger requests for additional genotoxicity testing. This situation is also discussed elsewhere in this volume (Kasper et al.). Such testing could include evaluation for the presence in the target organs of DNA adducts [36] and other DNA primary damage (*e.g.*, with assays for strand breaks), or indicators of genetic damage, such as micronucleated erythrocytes in the test animals at the end of the sub-chronic toxicity study. Newer tools, such as transgenic animals and genomics technologies may be useful in this regard.

Chemicals with structural alerts for mutagenicity but with negative results in an initial regulatory battery would usually not require additional testing, provided that the initial battery is sensitive to the type of effect indicated by the alert. The Working Group agreed that a structural alert can raise a concern, but study data are usually the final arbiter of hazard. However, if a chemical is in a structural class known to give positive results in specific genotoxicity tests or under specific experimental conditions that were not employed, then additional testing that includes these specific tests or conditions should be conducted.

4. Non-reproducible or marginal results of low concern

Occasionally, an increase in the measured genotoxicity parameter is seen that is considered significant by widely accepted criteria, such as a statistically significant increase over a negative control, or a value greater than a given fold change, *e.g.*, 2- or 3-fold, but the increase is weak or marginal. Subsequent repeats of the assay may fail to reproduce the increase. Other information, such as from other assays with a similar endpoint, or metabolism studies indicating no potentially reactive metabolite, may suggest that the singular marginal response is not a significant concern. Case Study 1 (Appendix A) presents such an example, where an initial statistically significant increase in the number of human lymphocytes with chromosome aberrations with values slightly outside the historical control range was not reproduced, and therefore was judged not to be a significant concern. The Working Group agreed that in such cases, the marginal, non-reproducible increase would be considered of no further concern for genotoxicity, and no testing beyond the standard battery of assays for that type of substance would be required. Case Study 2 (Appendix B) presents another case where there was a significant increase in the number of cells with chromosome aberrations just outside the historical control range. The Working Group concluded that it would be of low concern for human safety if repeat testing verified a weak borderline effect or showed a negative result.

There are also instances when an assay produces marginal or weak increases in response to chemical exposure, and these results are reproducible. Alternatively the chemical may produce a combination of weak and negative responses. There are several considerations that can help resolve whether these weak or equivocal results require follow-up testing. GLP study protocols generally specify in some detail conditions under which a study will be classified as positive, negative, or equivocal. While most protocols give some latitude to the professional judgment of the study director, evaluation criteria are often fairly rigid. For example, for cytogenetics studies, the criteria from one representative contract laboratory are:

> "The test article will be considered to induce a positive response when the percentage of cells with aberrations is increased in a dose-responsive manner with one or more concentrations being statistically significant and clearly outside the historical solvent control data ($p \leq 0.05$). However, values that are statistically significant but do not exceed the range of historical negative or solvent controls may be judged as not biologically significant."

Table 1
Example of a statistically but not biologically significant chromosomal aberration result due to low control values

Test article concentration	Mean (%) mitotic index	Mean (%) abnormal cells[a]
0	10	0.5
10 μg/mL	9	1
20 μg/mL	6	2
30 μg/mL	5	5.5*
Positive control	6	12

Negative historical control range 0–5%.
[a] Mean of two cell cultures: cells with structural chromosome aberrations excluding gaps.
* $p \leq 0.05$.

These criteria can lead to situations where, if one has a low value for the concomitant vehicle control, and the values for the low and mid doses of the test article appear to be dose-responsive but are well within the historical range for the vehicle control, the high dose becomes statistically different from the control but just outside the range of historical control value. An example of such data is given in Table 1.

The Working Group agreed that in case of non-reproducible or marginal results, it is advisable to develop a weight of evidence approach, considering the following points:

- The dose-response relationship needs to be examined – less concern is raised if the marginal response is not part of an increasing dose-response (*i.e.*, it is not dose-related).
- If the marginal increase is seen at high cytotoxicity (*e.g.*, approaching 50% or greater cytotoxicity in a cytogenetics assay, or >80% in the mouse lymphoma assay) but no increase is seen at lower, more moderate, toxicity, then there is less concern.
- Comparison of the magnitude of the marginal increase to historical negative control observations can help assess the probability that the result occurred by chance. Marginal responses within the historical negative control range (particularly at high concentrations and/or high toxicity) are of less concern. The Working Group agreed that more weight should be given to the historical data, and that better definition of how to use historical control data was needed. Preliminary comments on the use of historical control data are included in Appendix E. However, there is a need for further discussions in the future. Alternative approaches, such as selecting the 95 or 99% upper confidence interval

to define the range, or use of data distribution including median values, instead of the absolute upper limit, may be more appropriate. Also, it should be discussed whether observations in individual cultures rather than group means are sometimes more appropriate.

- Corroborating data can be sought from other experiments. Failure to confirm the marginal increase between replicates or experiments of the same assay or between different assays examining similar endpoints indicates a lower concern.
- Structure activity data can be useful. If no structural alerts are found for the chemical and if there is some evidence that the test compound is not a DNA damaging agent (*e.g.*, Ames test negative), then the marginal increases becomes of less concern.

Weight of evidence approaches have previously been described by Brusick et al. [43,44].

Considering the points listed above, the development of such weight of evidence arguments can lead to a conclusion that, for a marginal or equivocal response, there is a low or no level of concern, and no further testing is necessary, except possibly a repeat experiment using similar experimental conditions to check the reproducibility.

5. Follow-up strategy for a clear positive assay result

In some cases a clear and reproducible positive *in vitro* result is seen, yet the other assays in the initial battery, including any required *in vivo* test, are negative. The *in vitro* result is not automatically overruled by the negative *in vivo* result, and some follow-up testing or investigation is generally necessary to determine the relevance of the *in vitro* positive result. For example, the ICH scheme [6,7] suggests follow-up testing with a second *in vivo* test in addition to the *in vivo* cytogenetics test in the initial regulatory battery. It might be assumed that the concern about the positive *in vitro* result lessens as the number and types of negative *in vivo* assay results increase. However, this assumption may not be valid since the *in vivo* assays may have different sensitivities and /or evaluate different genotoxic endpoints. It is important that relevant endpoints are examined in the most relevant tissues *in vivo*.

To understand the basis for a positive *in vitro* result in the absence of a corresponding *in vivo* result, follow-up testing may require only a few additional studies (or tests), or more extensive research. Regardless of the question(s), the testing should be based on the full knowledge of the chemical, its physico-chemical and toxicological properties, and anticipated human exposure scenarios. An understanding of the type(s) of genotoxic insult(s) induced and the nature of the response(s), with any indications of possible mechanism, is crucial. Aspects, such as formation of DNA adducts or strand breaks, involvement of reactive oxygen or nitrogen species, nucleotide pool imbalance, inhibition of DNA synthesis or topoisomerases, and disruption of mitotic spindle need to be considered in order to identify rational testing approaches to pursue. Because each situation is likely to be different, follow-up testing is not amenable to a "one size fits all" approach, and flexibility is important to determine the most appropriate follow-up strategy to pursue.

A mode of action approach is used to determine whether a chemical that has intrinsic genotoxic properties might lead to an adverse effect, such as cancer [24]. The mode of action approach takes into account all available genotoxicity information and, in combination with other available information (*e.g.*, structure activity data, pharmacokinetics, ADME (absorption, distribution, metabolism and excretion) data, other biological responses, *etc.*) helps characterize whether a chemical is likely or not to pose a risk for exposed humans. The mode of action approach allows a full examination of whether the singular positive assay result should warrant a concern for human risk or not.

Case Studies 3 (Appendix C) and 4 (Appendix D) exemplify the weight of evidence and mode of action approaches in case of clear positive results. Case Study 4 is an example of a potential aneugen which, as discussed in several publications in recent years [33,34,37,38], might be considered to act *via* an indirect mechanism with a non-linear dose-response relationship suggesting a threshold.

Further discussion and recommendations on follow-up testing in the case of clearly positive results are given below.

The workshop discussion and case studies presented in the appendices led to the development of a framework for follow-up testing (displayed in Fig. 1). As discussed above, criteria for determining when follow-up testing beyond the initial battery is needed were agreed upon. When follow-up testing is needed, it should be based on all available information on chemical structure and mode of action, and also information learned from the nature of the response observed in the initial tests and anticipated human exposure patterns. The nature of these results (endpoint, magnitude, association with toxicity, *etc.*), and available information about the biochemical and pharmacological nature of the agent, are generally sufficient to conclude that the results observed are consistent with certain mechanisms and inconsistent

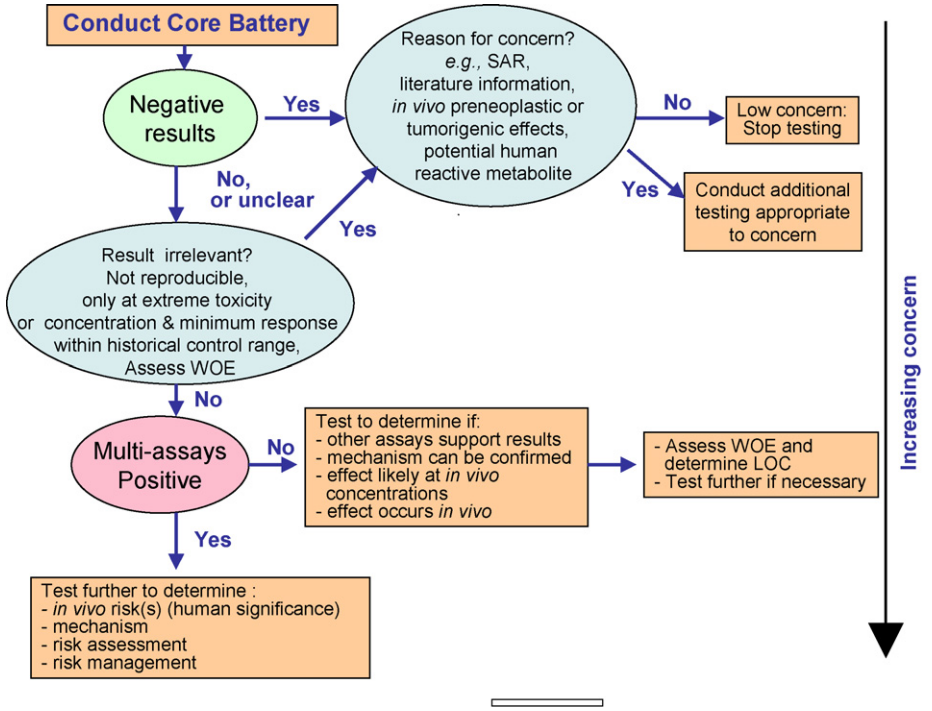

Fig. 1. Framework for interpretation of standard battery and follow-up test results. SAR = structure activity relationship, WOE = weight of evidence, LOC = level of concern.

with others. For example, it may be possible to determine whether such factors as DNA adducts or DNA strand breaks, involvement of reactive oxygen or nitrogen species, nucleotide pool imbalance and/or nucleotide mis-incorporation, interference with cell cycle kinetics, inhibition of DNA synthesis or topoisomerases, or mitotic spindle disruption are likely to be involved in the mode of action.

The different tests in the battery are selected because they measure different genotoxic effects and different genotoxic mechanisms of action, thereby providing important mechanistic information. For example, generation of chromosomal aberrations involves strand breakage and rejoining, reversion of the Ames' *Salmonella* tester strains containing specific base substitutions requires base mutations that may arise from DNA alkylation or mis-repair of bulky adducts, and reversion of the Ames' strains containing frameshift mutations requires induction of a second frameshift that is characteristic of intercalating agents and other classes, but not small molecular weight alkylating agents. These types of information should be used to guide the interpretation of the results and the selection of follow-up tests. *In vivo* endpoints should be chosen to reflect the types and the mechanisms of damage found in the initial screening battery.

It is recognized that some genotoxic agents act through disruption of cellular biochemistry that occurs only above certain treatment levels (for example, see [40]). They therefore exhibit a dose-response that is expected to be non-linear, and exhibit a threshold below which there will not be a concern. Examples of such agents include those that induce nucleotide pool imbalances (without incorporation of a nucleotide that can cause mis-pairing or strand termination at concentrations relevant to use), mitotic spindle disruption, inhibition of DNA synthesis, topoisomerase inhibition, *etc.* These agents are often referred to as "indirect gentoxins" or "non-DNA-reactive gentoxins", in contrast to those thought to act by primary reactivity with DNA. It is generally accepted that those indirect mechanisms or modes of action lead to non-linear response curves, sometimes reported as threshold effect. The Working Group recognized that when an agent can be demonstrated to act though such an "indirect" mechanism, and a dose can be established below which the effect is not observed or deemed not a concern, then appropriate margins of safety for exposure below that dose can be established. This means that human exposures sufficiently below exposures associated with the no observed effect dose should not present a significant genotoxic risk. These issues have been discussed in a special

issue of Mutation Research edited by Parry and Sarrif [45].

It will be easier to build weight of evidence or mode of action arguments when the positive result is found in a test with low specificity (*i.e.*, tests frequently positive but not confirmed in the *in vivo* studies), such as the *in vitro* chromosomal aberration test or mouse lymphoma assay. Agents that are positive in multiple tests with different endpoints, especially if they are positive *in vivo*, impart the highest level of concern and would require extensive investigations to develop a sufficient weight of evidence to establish conditions under which human exposure might be permissible.

Although these general considerations were discussed, and it was agreed that they should be the basis of selection of appropriate tests, time did not permit development of more detailed recommendations, such as identification of specific tests to follow-up on specific outcomes in standard batteries. It was agreed that consensus recommendations about specific test options would be useful, and it was suggested that the Working Group should meet again in the near future to develop such recommendations.

Considerations when choosing an *in vivo* assay as a follow-up test to an initial finding include choice of the appropriate tissue(s) in which evaluations should be performed. This depends on anticipated route(s) of exposure, tissue distribution (often known from other toxicology or pharmacokinetic studies), and metabolic degradation and/or activation in various tissues as well as target organ in long-term toxicity studies. It was noted that such considerations might dictate the use of non-standard studies (*i.e.*, not in the regulatory initial battery or standard follow-up studies but sufficiently validated) and that in those cases the use of non-standard studies would be preferable to a standard *in vivo* assay in which the endpoint or target tissue is not relevant. It was further noted that when metabolic modification of an effect is observed to occur, consideration should be given to human metabolism in relation to the laboratory models being considered. Likewise, the relationship between exposure-response information in laboratory models to blood and tissue levels from human exposures is important, and it is desirable to determine this relationship whenever possible.

6. Conclusions

A summary of the discussion and principal conclusions reached by the Working Group is as follows:

The Working Group agreed to limit the focus of the meeting to somatic cell risk in humans, with emphasis on carcinogenic risk, although the group recognized the importance of potential health risks from both somatic and germ cell mutations.

The Working Group agreed not to discuss the improvement of the standard batteries, but to focus on interpretation and appropriate follow-up testing for the tests currently in use in the standard battery. The Working Group noted that genetic toxicity test batteries vary among regulatory agencies and specific situations, but generally include a bacterial test for gene mutations and a mammalian cell test sensitive to chromosomal damage (most commonly a mammalian cell test for chromosomal aberrations and/or the mouse lymphoma $tk^{+/-}$ mutation assay), and, depending on the products and their use, an *in vivo* test in rodent bone marrow for chromosome damage.

When the initial test battery is clearly negative, there is generally no need for further testing unless there is (a) evidence that metabolites that differ from those generated in the assays may be present in the human, (b) structural alerts suggest possible activity that would not have been detected in the battery employed, or (c) evidence from the literature, previous experience, or subsequent test results suggests possible activity that would not have been identified in the battery. In these cases, follow-up testing may be indicated, and appropriate tests should be selected on the basis of the information that raised the concern.

When a non-reproducible or marginal *in vitro* positive result is obtained, and results from other assays with a similar endpoint are negative, the weight of evidence should be considered to determine if further testing is necessary or whether, based on the available data, the evidence suggests a low level of potential risk that does not require further testing. Factors that may suggest lower concern include: (a) weak effects without a strong dose relationship and values within or close to a range that could occur by chance variability (negative control historical data), (b) effects that occur only at very high levels of cytotoxicity, but not at moderate levels, in the chromosomal aberration or mouse lymphoma $tk^{+/-}$ assays (*e.g.*, approaching 50% or greater cytotoxicity in the chromosome aberration test, or >80% in the mouse lymphoma assay), (c) results that are not consistently repeatable, and (d) the absence of structural alerts or any other cause of concern. In most cases, the result is not of concern and no testing beyond the standard battery for that type of substance will be required.

When a clear positive result is obtained in *in vitro* test battery, further testing is generally indicated in order to provide a sufficient body of evidence to determine the mode of action, relevance to the human expo-

sure situation, and potential human health risk. Such testing should be based on the knowledge available about the nature and/or mode of action of the original response, *e.g.*, whether the initial result was consistent with DNA adducts versus strand breaks, involvement of reactive oxygen or nitrogen species, pool imbalance, spindle disruption, *etc*. If the evidence suggests an "indirect" mode of action (not involving direct or proximate reactivity with the DNA), such as nucleotide pool imbalance, spindle disruption, inhibition of DNA synthesis or topoisomerases, *etc.*, then tests that provide additional evidence that supports this hypothesis and that rule out direct DNA reactivity should be selected. It is often useful to determine if a positive result, found only under a specific condition in one test, is confirmed in further testing in other assays evaluating the same endpoint (*e.g.*, chromosome aberration test versus mouse lymphoma assay).

When choosing an *in vivo* assay for follow-up testing, an endpoint appropriate to the nature of the original response observed should be selected, and in addition, due consideration should be given to the route of human exposure, the expected tissues and times of highest exposure (*i.e.*, pharmacokinetic considerations), and the potential for metabolic activation and deactivation in various tissues. Although regulatory guidelines often mention *in vivo* assays that may possibly be employed for follow-up testing, non-standard studies supported by peer-reviewed published literature may, when justified, be more appropriate and informative than standard assays. The Working Group acknowledged that there are cases where mechanistic studies can be considered sufficient to support the indirect mode of action, and therefore additional *in vivo* data may not be necessary.

When addressing the relevance of findings to human health risks, the total weight of evidence should be considered. In addition to the factors noted above, when modification of an effect by a mammalian metabolizing system is observed, consideration should be given to the metabolism of the agent in the human relative to the laboratory model. Whenever possible, exposure-response information for comparison to blood and tissue levels from human exposures is desirable.

Remaining issues that were identified during the workshop, but not addressed due to time limitations, include the development of recommendations for the selection of specific assays, particularly *in vivo* assays, and the appropriate use of historical control data. It is anticipated that recommendations on these issues will be developed at a future meeting of this Working Group. Similarly, the evaluation of structural alerts was identified as a useful component of the weight of evidence approach, but not discussed in detail. This topic will have to be re-visited in the future.

Disclaimer

This manuscript represents the views of the authors and the contents do not necessarily reflect the views or policies of the authors' agencies or institutions. Mention of trade names or commercial products does not constitute an endorsement or recommendation for use.

Appendix A. Case Study 1: Compound with a weak non-reproducible increase in chromosome aberration test

Compound A is an early orally applied drug development candidate for a non-life threatening therapeutic indication and low systemic exposure. There was no evidence of increased revertant numbers in the Ames test up to the limit dose level of 5000 μg/plate, using the plate incorporation method in the presence and in the absence of metabolic activation on *Salmonella typhimurium* TA98, TA100, TA1535, TA1537 and *E. coli* WP2 uvrA.

Compound A was tested for clastogenic activity in human lymphocyte cultures. Chromosome aberrations were evaluated by metaphase analysis after 3 h treatments with and without metabolic activation at concentrations ranging from 47.2 to 84.0 μg/mL and 10.6 to 75.0 μg/mL, respectively. In addition, chromosome aberrations were evaluated after 24 h treatment without metabolic activation at concentrations ranging from 8.30 to 33.3 μg/mL. In all tests, the highest test concentration evaluated produced a 48 to 59% reduction of the mitotic index. There were no significant increases in the number of cells with chromosome aberrations at any concentration evaluated in either the 3 or 24 h tests without metabolic activation. However, Compound A produced a dose-related, statistically significant increase in the number of cells with chromosome aberrations in the initial test with metabolic activation at the two highest concentrations (58 and 68.2 μg/mL); only the highest concentration produced a value outside the range of the historical data, *i.e.*, 5.5% abnormal cells as compared to the historical control range of 0–4% (Table 2, Test 1). A confirmatory 3 h test with metabolic activation using blood of a second donor produced a statistically significant increase in chromosome damage only at the highest concentration of 84.0 μg/mL, the observed value being within the acceptable range of the historical negative control data, *i.e.*, 3.5 versus 0–4% (Table 2, Test 2). This response was not clearly reproducible between

Table 2
Lack of reproducibility of a weak increase in structural chromosomal aberrations in cultured human lymphocytes with Compound A in three independent tests with metabolic activation

Compound A (μg/mL)	Mean (%) abnormal cells[a]	Mean (%) mitotic suppression
Test 1		
1% DMSO	0.5	0
49.3	2.5	29
58.0	4.0*	52
68.2	5.5*	49
CP[b]	29.0*	45
Test 2		
1% DMSO	0.5	0
47.2	1.0	0
63.0	0.5	14
84.0	3.5*	48
CP[b]	24.3*	31
Test 3		
1% DMSO	2.5	0
52.4	1.5	21
65.5	1.0	36
81.9	2.5	56
102.0	2.0	58
128.0	1.5	65
CP[b]	58.0*	59

Negative historical control range 0–4%.
[a] Mean of two cell cultures: cells with structural chromosome aberrations excluding gaps.
[b] Cyclophosphamide at 5 μg/mL.
* Statistically significant (one-tailed Fisher's exact test $p < 0.05$).

the replicate cultures and the negative control response was low (0.5% abnormal cells as compared to the historical control range of 0–4%). A follow-up 3 h exploratory cytogenetics study with metabolic activation using blood of a third donor was performed on a larger number of concentrations in order to better characterize the cytotoxicity profile of Compound A in the lymphocyte test system. The results of this follow-up test did not reproduce the previous results using the same S9 batch but different blood donors and failed to show an increase in chromosomal aberrations at concentrations ranging from 52.4 to 128 μg/mL, which produced a 21 to 65% reduction in the mitotic index (Table 2, Test 3) and 13 to 70% reduction in cellular ATP levels (not shown). Based on the overall test results, Compound A was concluded as being equivocal for the induction of structural chromosome aberrations.

Compound A was tested for the induction of micronuclei in male and female rat bone marrow cells *in vivo*. Male and female rats were administered the vehicle only as the negative control, or Compound A at dose levels of 15, 30, and 60 mg/kg in males, and 7.5, 15, and 30 mg/kg in females by oral gavage once a day for 2 consecutive days. The numbers of micronucleated (MN) PCE in any of the test groups did not significantly increase. In conclusion, Compound A did not induce chromosome damage in the bone marrow cells of male or female rats when tested up to estimated maximum tolerated doses of 60 and 30 mg/kg, respectively. The mean maximum serum concentrations occurred 3.5 h after dosing, achieving values of 645 and 620 ng/mL in males and females, respectively, and the mean $AUC_{0–24\,h}$ exposures were 8560 and 9030 ng h/mL, respectively. The serum concentration in this study was expected to exceed the pharmacologically active concentration by approximately 60-fold.

The Working Group considered that the original increase in the number of human lymphocytes with chromosome aberrations was weak, and was not clearly reproducible in follow-up testing. Based on the overall profile of this compound in the standard genetic toxicology test battery, the Working Group concluded that the genotoxicity findings of Compound A were of low concern for human safety and did not require further follow-up testing.

Appendix B. Case Study 2: Compound with marginal increase in the number of cells with chromosome aberrations at a cytotoxic concentration

Compound B is an early drug candidate in a chronic use indication with inhaled drug application and low systemic exposure. It tested negative in the Ames assay up to levels of compound insolubility and the gene mutation test in CHO cells at *hprt* locus up to cytotoxic concentrations. Compound B did not induce chromosomal aberrations in cultured human lymphocytes when tested in a 24 h exposure without metabolic activation, or a 3 h exposure with metabolic activation (Table 3). However, when it was tested in a 3 h exposure without metabolic activation, a single concentration that induced 58% mitotic suppression produced a statistically significant increase in chromosomal aberrations just outside of the historical control range (5% of cells with chromosome aberrations versus 0–4% acceptable range for negative control data). Compound B was also tested in a mouse bone marrow micronucleus assay up to a maximally tolerated dose and did not lead to an increase in the incidence of MN PCE in the bone marrow. Toxicity was seen in the bone marrow indicating compound reached this tissue.

Based on the initial results from the 3 h exposure, the Working Group came to the consensus that the response was of low concern to human safety since it occurred

Table 3
Weak increase in structural chromosomal aberrations in cultured human lymphocytes at cytotoxic concentrations of Compound B

Compound B (μg/mL)	3 h direct		24 h direct		3 h + S9	
	Mean (%) abnormal cells[a]	Mean (%) mitotic suppression	Mean (%) abnormal cells[a]	Mean (%) mitotic suppression	Mean (%) abnormal cells[a]	Mean (%) mitotic suppression
0.1% DMSO	0	0	2	0	1	0
25	–	–	1	15	–	–
28	–	–	–	–	0	20
32	–	–	3	30	–	–
35	–	–	–	–	1.5	32
41	–	–	2	55	–	–
44	–	–	–	–	1	54
101	0.5	8	–	–	–	–
120	0	28	–	–	–	–
144	5*	58	–	–	–	–

Negative historical control range 0–4% in 3 h direct test.
[a] Mean of two cell cultures: cells with structural chromosome aberrations excluding gaps.
* Statistically significant (one-tailed Fisher's exact test $p < 0.05$).

only at a high concentration that produced more than 50% cytotoxicity in human lymphocytes. Importantly, a concentration that produced 28% mitotic suppression did not generate a significant induction of aberrations. In addition, the increase was just outside the historical control range, and the compound did not produce chromosomal aberrations in the corresponding 24 h exposure. The Working Group suggested that a repeat of the positive arm of the test, to evaluate reproducibility, might contribute further to the assessment of Compound B. Taking into account the steep cytotoxicity dose-response curve, a new test should, if possible, include a concentration producing between 28 and 58% reduction in mitotic index.

If a similar pattern of results, or negative results, are obtained in the repeat test, the Working Group would consider that no further testing would be needed, and would conclude the response to be of low concern for human safety.

Appendix C. Case Study 3: Development of weight of evidence approach with a clearly positive result in cytogenetics assays

Compound C is a topically applied early drug development candidate intended for short term treatment of dermal scarring. It produced high local concentrations at the target site in the skin but was rapidly cleared systemically and showed only very low systemic exposure. There was no evidence of increased revertant numbers in the Ames test up to the limit dose level of 5000 μg/plate, using the plate incorporation method in the presence and in the absence of metabolic activation on *Salmonella typhimurium* TA98, TA100, TA1535, TA1537 and *E. coli* WP2 uvrA.

Compound C was tested for clastogenic activity in human lymphocyte cultures (Table 4). Chromosome damage was evaluated by metaphase analysis after 3 h with and without metabolic activation at concentrations ranging from 154 to 240 μg/mL and 33.4 to 52.2 μg/mL, respectively. In addition, chromosome damage was evaluated after 24 h without metabolic activation at concentrations ranging from 6.30 to 9.84 μg/mL. In all the tests, the highest concentrations produced a 52 to 57% reduction of the mitotic index. Compound C did not induce a significant increase in the number of abnormal cells at any concentration evaluated when treated in the presence of metabolic activation. In the absence of metabolic activation, Compound C produced concentration-related statistically significant increases in aberrant cells over a range of concentrations, which spanned both cytotoxic and non-cytotoxic test conditions in both 3 and 24 h exposures, producing 6.5 and 7.5% abnormal cells at 3 and 24 h, respectively.

Compound C was tested for the induction of micronuclei in Chinese hamster ovary (CHO-WBL) cells (Table 5). A slight increase (1.7- to 2.3-fold above mean background) was observed in the 24 h test without metabolic activation, whereas more substantial increases of 2.3- to 4.9-fold were observed in the test with activation. An additional 3 h test without activation was performed in response to the slight increase observed in the 24 h test and a significant increase (3.1-fold above mean background) was observed in cultures treated with concentrations of 68.5 and 107 μg/mL. Kinetochore analysis was performed to assess whether the

Table 4
Induction of structural chromosomal aberrations in cultured human lymphocytes by Compound C

Compound C (μg/mL)	3 h direct		24 h direct		3 h + S9	
	Mean (%) abnormal cells[a]	Mean (%) mitotic suppression	Mean (%) abnormal cells[a]	Mean (%) mitotic suppression	Mean (%) abnormal cells[a]	Mean (%) mitotic suppression
0.1% DMSO	1.5	0	0.5	0	1	0
6.3	–	–	2	8	–	–
7.9	–	–	5.5*	20	–	–
9.8	–	–	7.5*	52	–	–
12.3	–	–	–	–	–	–
33.4	5.5*	23	–	–	–	–
41.8	6.0*	41	–	–	–	–
52.2	6.5*	56	–	–	–	–
154	–	–	–	–	1	16
192	–	–	–	–	0.5	34
240	–	–	–	–	3	57

Negative historical control range 0–4% in 3 h direct test.
[a] Mean of two cell cultures: cells with structural chromosome aberrations excluding gaps.
* Statistically significant (one-tailed Fisher's exact test $p < 0.05$).

observed increases were the result of structural (clastogenic) or numerical (aneugenic) chromosomal events (Table 6). Over 140 micronuclei induced by a 3 h treatment with metabolic activation at soluble concentrations ranging from 107 to 168 μg/mL were evaluated using anti-kinetochore antibody staining technique. A 3- to 5-fold increase in micronucleated cells above DMSO-treated controls was again observed and these cells were predominately kinetochore negative indicating a clastogenic response.

Induction of micronuclei *in vivo* in rat bone marrow PCE was evaluated after continuous intravenous (IV) infusion at a rate of 2 mL/kg/h for approximately 24 h up to a maximum feasible dose of 655 mg/kg. Approximately 24 and 48 h following completion of the infusion, bone marrow was extracted and assessed for the induction of MN PCE. There was no bone marrow cytotoxicity and no statistically significant increases in MN PCE at any dose level or harvest time (data not shown). Mean plasma concentrations of Compound C measured after approximately 24 h of infusion were 2.93 and 3.20 μg/mL in males and females, respectively. Attempts to assess target (bone marrow and liver) tissue exposure to Compound C resulted in concentrations below assay detection limits, possibly attributed to tissue metabolism and rapid clearance. Plasma exposures

Table 5
Induction of micronuclei by Compound C in Chinese hamster ovary cells

24 h test without metabolic activation			3 h test without metabolic activation			3 h test with metabolic activation		
Compound C (μg/mL)	Cytotoxicity[a]	% MN cells[b]	Compound C (μg/mL)	Cytotoxicity[a]	% MN cells[b]	Compound C (μg/mL)	Cytotoxicity[a]	% MN cells[b]
0	0	1	0	0	1.2	0	0	0.8
1.8	17.5	1.7	68.5	33.6	3.7*	134	0	3*
3.5	25.2	1.7	107	45.2	3.7*	168	39.7	1.9*
7.0	45.9	2.3*	168	67.3	1.2	210	24.5	3.9*
14.0	81.1	INS	210	76.3	INS	262	T	T
POS[c]	15.4	9.6*	POS[d]	43.3	21.7*	POS[e]	0	5.0*

INS: insufficient number of cells for evaluation. T: toxic, less than 20% cells when compared to negative control
[a] Cytotoxicity = $100 - 100 \times \{(CBPI_T - 1/CBPI_C - 1)\}$, with CBPI = (no. of mononucleated cells + 2 × no. of binucleated cells + 3 × no. of multinucleated cells)/total number of cells.
[b] Percent micronucleated cells; evaluating a minimum of 1000 binucleated cells.
[c] Positive control mitomycin C at 0.05 μg/mL.
[d] Positive control mitomycin C at 0.4 μg/mL.
[e] Positive control cyclophosphamide at 10 μg/mL.
* Statistically significant (one-tailed Fisher's exact test $p < 0.05$).

Table 6
Classification of micronuclei induced by Compound C in Chinese hamster ovary cells using kinetochore staining

Compound C (μg/mL)	Cytotoxicity[a]	% MN[b]	K− MN[c]	K+ MN[d]	Proportion K+ MN[e]
0	0	1.6	1.3	0.3	19
107	16.8	3.6*	3.4*	0.2	6
134	31.4	6.6*	6.0*	0.6	9
168	42.1	6.2*	5.3*	0.9	14

[a] Cytotoxicity = $100 - 100 \times \{(CBPI_T - 1/CBPI_C - 1)\}$, with CBPI = (no. of mononucleated cells + 2 × no. of binucleated cells + 3 × no. of multinucleated cells)/total number of cells.
[b] Frequency of micronuclei per 100 cells.
[c] Frequency of kinetochore negative micronuclei per 100 cells indicating a chromosomal breakage.
[d] Frequency of kinetochore positive micronuclei per 100 cells indicating chromosomal loss.
[e] Percent kinetochore positive micronuclei among total micronuclei.
* Statistically significant (one-tailed Fisher's exact test $p < 0.05$).

in the rat during the *in vivo* assay were slightly (2–4-fold) higher than skin concentrations of Compound C from human biopsy specimens of 2 of 6 subjects following 12 h topical application to the intact skin of healthy volunteers.

Compound C was also tested for gene mutational activity in the *hprt* gene mutation test in CHO cells. The test in the absence of metabolic activation was conducted over a concentration range from 5 to 80 μg/mL, and the test conducted in the presence of metabolic activation spanned concentrations ranging from 75 to 250 μg/mL. Compound C did not induce a mutagenic response up to concentrations that produced significant cytotoxicity (data not shown).

There was no evidence that Compound C induced unscheduled DNA synthesis (UDS) in male rats at dose levels of 54.6 and 437 mg/kg at the 2 or 16 h time points, respectively (data not shown). Compound C was therefore evaluated as inactive in the *in vivo*/*in vitro* assay for unscheduled DNA synthesis in rat primary hepatocyte cultures at two time points when administered by continuous IV infusion up to the maximum feasible dose based on compound solubility. Mean plasma concentrations of Compound C measured in the *in vivo*/*in vitro* UDS assay after 2 and 16 h of continuous drug infusion were 0.99 and 4.29 μg/mL, respectively. Thus, slightly higher plasma levels were achieved than in the bone marrow micronucleus test, and therefore a slightly greater margin, when compared with human skin concentrations, was achieved.

Radiolabeled ^{14}C-Compound C was incubated with calf thymus DNA (1mg/mL) in triplicate at 75, 150, and 300 μg/mL for 4 h at 37 ± 2 °C (with and without metabolic activation). After the incubation, DNA was purified by treatment with RNase A, T1, and proteinase K followed by organic extraction and precipitation. Fifty microgram DNA from each sample was counted in a liquid scintillation counter. Results showed that ^{14}C-radioactivity counts obtained from the test article-treated DNA were not significantly higher than the background level counts obtained from untreated DNA. The positive control ^{14}C-benzo[a] pyrene-treated DNA with metabolic activation showed significantly higher ^{14}C-radioactivity, most likely resulting from binding of a BaP metabolite. It is concluded that, results obtained under these experimental conditions did not show evidence for tight association or binding of ^{14}C-Compound C.

Compound C was tested for the induction of micronuclei in rat skin keratinocytes (Table 7). Male and female rats were administered Compound C (2, 4, or 8 mg/animal/day) by topical application once a day for 4 consecutive days to intact skin clipped free of fur. The topical route of administration and dose levels used were selected to reflect the intended route of clinical administration and drug exposures in humans. The day following the last dose, keratinocytes were isolated from the excised skin, dropped onto glass slides, stained with acridine orange and evaluated for the presence of micronuclei. There were no significant increases in

Table 7
Induction of micronuclei in dermal keratinocytes following topical application with Compound C

Compound C (mg/animal/day)[a]	% Micronucleated keratinocytes	
	Males	Females
0[b]	0.03 ± 0.04	0.06 ± 0.05
2	0.04 ± 0.05	0.02 ± 0.03
4	0.05 ± 0	0.02 ± 0.03
8	0.05 ± 0.08	0.02 ± 0.03
CP[c]	4.19 ± 2.95	4.83 ± 2.13

[a] mg/20 cm^2 for 3 consecutive days.
[b] Ethanol, glycerol and water (70:5:25; v/v/v).
[c] 12 mg/animal/day cyclophosphamide for 4 consecutive days.

the numbers of MN keratinocytes with either sex. Concentrations of Compound C in the target tissue after 3 consecutive days of dosing at 8 mg/animal/day ranged from 72 to 175 μg/g of skin. This exceeded intended human skin concentrations by more than 20-fold.

The Working Group concluded that the initial response in the cytogenetics test and micronucleus assay when examined in isolation warranted further follow-up testing to assess the mechanism and relevance to humans exposed to the drug by dermal application. Follow-up testing yielded negative results in the *in vivo* micronucleus assay in rats, the rat liver *in vivo*/*in vitro* UDS test, and the *hprt* gene mutation test in CHO cells. This data set was considered by some members of the Working Group to be insufficient to address the concerns for Compound C. The reasons were as follows:

- The rat bone marrow micronucleus study using IV dosing produced systemic exposures that were only slightly above the human skin concentrations.
- The *in vivo*/*in vitro* UDS and *hprt* gene mutation test in CHO cells were considered to be of limited value, since they were both likely to be insensitive to compounds inducing DNA strand breakage, as clastogens would be expected to do.

The other additional studies were considered useful. The lack of DNA binding with calf thymus DNA was considered important additional information but the argument could have been strengthened by conducting the binding experiment in human lymphocyte cultures. Finally, the Working Group considered the dermal micronucleus assay as being the crucial study to address the concern for human safety since it addresses the relatively high concentration at the site of first contact and achieved safety multiples over intended human exposure.

Based on the overall weight of evidence, and in particular the DNA binding and dermal micronucleus results, the Working Group considered Compound C to be of low concern for human safety and that no further genetic toxicology testing is needed in addition to the presented package.

Appendix D. Case Study 4: Mode of action approach: a substance that does not directly react with DNA, such as a spindle poison

Compound D is intended for long-term treatment of severe autoimmune conditions, such as multiple sclerosis. In clinical trials, plasma C_{max} was 1.65 μg/mL and plasma AUC was 37.1 μg h/mL. The drug is extensively protein bound *in vivo* giving only 2.5% free drug in rats and 1.5% in humans. Approximately 60% is unbound in cell cultures.

There was no evidence of increased revertant numbers up to the limit dose level of 5000 μg/plate, when the Ames test was conducted on *Salmonella* strains TA98, TA100, TA1535, and TA1537 plus *E. coli* WP2 uvrA, using both plate incorporation and preincubation methods in the presence of S9, and plate incorporation in the absence of S9.

In the mouse lymphoma assay (MLA), cells were treated for 3 h in the absence and presence of S9 at concentrations up to 10 mM (3569 μg/mL). In the presence of S9 the treatments were very toxic, but there were no increases in mutant frequency at concentrations producing reductions in relative total growth (RTG) to <10%. In the absence of S9, at the limit concentration (10 mM, 3569 μg/mL), only 60% reduction in RTG was seen. The 24 h treatments (two separate experiments) in the absence of S9 had some concentrations where the mutant frequency increased above control by more than the global evaluation factor [39] of 126×10^{-6} for the microwell method. However, this was only at the 250 μg/mL concentrations in both experiments where relative survival (RS) was reduced to 9.88 and 8.97%, respectively (see Fig. 2a and b). Relative total growth (RTG), which is currently the recommended measure of toxicity in the MLA, did not reduce as much as RS, and was, respectively, 0.27 and 0.20 at this 250 μg/mL concentration in the two 24 h experiments in the absence of S9. The result could therefore be considered as borderline in that these responses would be excluded from consideration if judged by RS but not by RTG. At the time the study was done, both RS and RTG were considered acceptable for judgment of toxicity. No clear information on the type of damage caused at these borderline positive concentrations could be gleaned from colony sizing as frequencies of both large and small colony mutants increased. In light of the data obtained subsequently it might be argued that this unusual response in the MLA may reflect an aneugenic mode of action.

For the *in vivo* micronucleus (MN) test, rats were dosed orally on 2 consecutive days and bone marrow sampled 24 h after the second dose. Statistically significant increases in MN frequency to 3-fold concurrent control levels and slightly above the historical control range were found at the 180 and 360 mg/kg/day (Fig. 3). Although the response was relatively weak, the study was concluded as positive.

To investigate this *in vivo* finding, MN tests were conducted *in vitro* both in V79 cells and human lymphocytes.

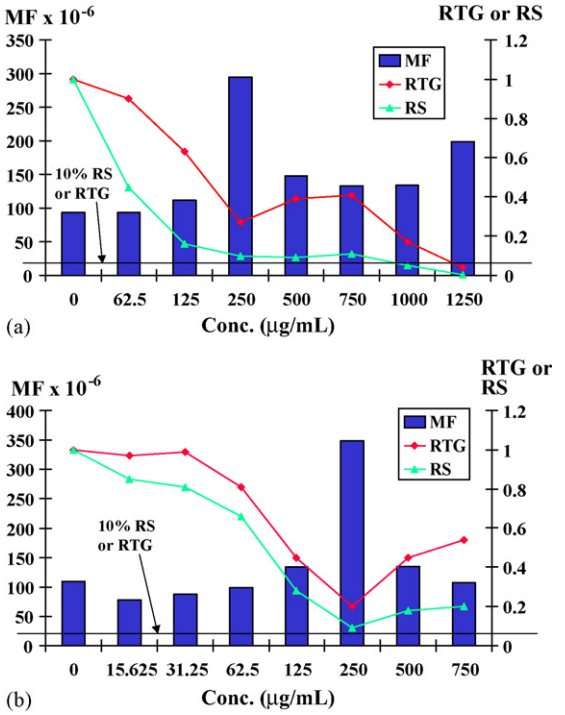

Fig. 2. Mouse lymphoma results with Compound D following 24 h treatments in the absence of S9 ((a) first independent experiment and (b) second independent experiment). MF = mutant frequency. RTG = relative total growth. RS = relative survival.

In both cases, substantial and statistically significant increases in MN frequency were seen after the 24 h treatments (but not after short treatments) in the absence of S9. The positive effects were seen at low levels of toxicity (Figs. 4 and 5). The MN in the human lymphocyte preparations were probed with a pan-centromeric

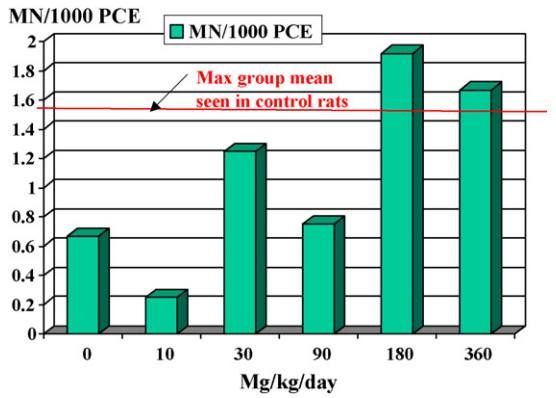

Fig. 3. Results of rat bone marrow micronucleus (MN) test with Compound D. Rats were dosed orally 2× daily and bone marrow sampled 24 h after the second dose. Data are accumulated from two separate experiments. Two thousand polychromatic erythrocytes (PCE) were scored per animal.

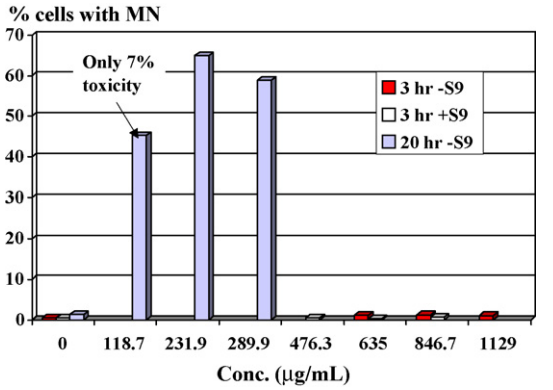

Fig. 4. Micronucleus (MN) data in V79 cells treated with Compound D.

DNA probe, using a fluorescent *in situ* hybridization method. 77% of the MN in the drug-treated cultures were centromere-positive, which is comparable to the 70% value observed with carbendazim tested as the positive control aneugen, and much higher than the 7% centromere-positive MN seen with the clastogenic positive control, cyclophosphamide (Table 8). There were too few MN from solvent controls to be probed for presence or absence of centromeres, but the historical control range for centromere-positive MN in human lymphocytes in the testing facility was 9–14.5%.

The pattern of results observed with this drug, (*i.e.*, borderline 24 h MLA positive responses in the absence of S9, the strong induction of micronuclei *in vitro* and high frequency of centromere-positive MN induced in human lymphocytes), was consistent with chromosome loss (aneuploidy). The result of the *in vivo* MN assay was very weak. At doses where no associated increase in MN was seen, the plasma exposures (C_{max} or AUC) were 75–125× the proposed human exposure levels.

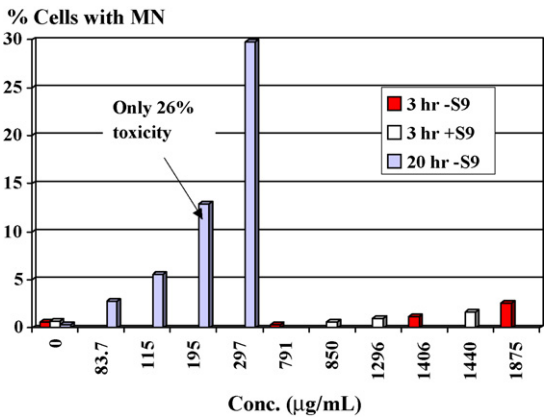

Fig. 5. Micronucleus (MN) data in human lymphocytes treated with Compound D.

Table 8
Centromere labeling with FISH probe for pan-centromeric DNA, for Compound D in micronucleus test in human lymphocytes

Treatment	No. of micronucleated binucleated cells probed	% C+	% C−
Compound D 194.5 μg/mL	100	77	23
Carbendazim 2.5 μg/mL	100	70	30
Cyclophosphamide 3.125 μg/mL	27	7	93

%C+: percent of cells with micronuclei containing a centromere. %C−: percent of cells with micronuclei containing no centromere. Pan-centromeric probing was not performed on concurrent solvent control slides. However, the historical frequency of C+ MN in human lymphocytes in the testing laboratory is 9–14.5%.

The Working Group agreed that, from the existing data, it was not clear whether the MN *in vivo* were solely due to chromosome loss. Even though 77% of the MN drug-treated cultures *in vitro* were centromere-positive, a secondary (*e.g.*, clastogenic) mechanism based on a direct DNA-mediated effect could be involved. It was suggested that it was necessary to demonstrate the absence of a direct DNA mechanism, and this could be done either *in vitro* or *in vivo*. Approaches that could be considered were:

- Demonstrate absence of DNA binding
- Demonstrate absence of induction of structural chromosomal aberrations
- Demonstrate absence of induction of DNA strand breaks.

The investigation of chromosomal aberration induction *in vivo* could be performed either in an acute study (scoring bone marrow) or in a repeat dose (*e.g.*, 28-day) study (scoring blood lymphocytes).

The Working Group discussed the evidence that is needed to convincingly describe a non-linear or threshold mechanism. It became clear that in addition to obtaining evidence in favour of a non-linear mechanism it might be necessary also to obtain evidence against a linear mechanism, *i.e.*, ruling out direct DNA reactivity. In this case study, evidence was obtained that was consistent with a chromosome loss (aneugenic) mode of action. However, this was not considered sufficient on its own. The mode of action should be more fully supported to be plausible. Further information on potential to affect the mitotic apparatus, or negative results in a rigorous test for chromosome breakage or for potential to damage DNA, would be very supportive. In this case the aneugenic mechanism of action would be accepted, no further genotoxicity testing would be required, and safety margins could be determined.

Appendix E. Use of historical control data in assessment of genotoxic response

Historical control data have been used as a convenient tool for explanation of the lack of relevance of a weak positive response, usually revealed by statistical significance when the response is compared with the concurrent negative control. For this purpose, the minimum and maximum values of the accumulated control data in the laboratory are usually used. The historical control data, however, should be more accurately and appropriately based on the distribution of the control data accumulated in the laboratory. To achieve this, certain experimental conditions, as summarized by Margolin and Risko [41] need to be achieved as follows:

- The experimental protocol must have remained fixed throughout the period covered by the historical data and the current experiment
- The method of scoring the response must be unchanged during the period
- The experimental unit must be comparable throughout the period
- The data must have been gathered by the same investigators within the same laboratory
- There must exist no known systematic differences between the various control groups, current and historical, that would produce systematic differences in response.

All control data should be included, even out-of-range data, unless there is a convincing rationale for exclusion. As long as the criteria above for obtaining historical control data have been met, these data may be combined over some different experimental conditions (*e.g.*, solvents, sampling times).

When historical control data fulfils the requirements above, they can provide a good guide for evaluation of experimental data. They can be compared with concurrent control data by using statistical methods. An example of a procedure for data analysis using historical control data has been proposed by Hayashi et al. [42].

References

[1] COM (Committee on Mutagenicity of Chemicals in Food, Consumer Products and the Environment), COM guidance on a strategy for testing of chemicals for mutagenicity, United Kingdom, December 2000. (http://www.advisorybodies.doh.gov.uk/com/).

[2] K.L. Dearfield, M.C. Cimino, N.E. McCarroll, I. Mauer, L.R. Valcovic, Genotoxicity risk assessment: a proposed classification strategy, Mutat. Res. 521 (2002) 121–135.
[3] FDA, Toxicological Principles for the Safety of Food Ingredients. IV.C.1. Short-Term Tests for Genetic Toxicity, Redbook, 2000.
[4] FDA, Guidance for Industry Recommended Approaches to Integration of Genetic Toxicology Study Results. Draft Guidance. Genotoxicity. Center for Drug Evaluation and Research, November 2004.
[5] Health Protection Board (Canada), The Assessment of Mutagenicity. Health Protection Branch Mutagenicity Guidelines, Environ. Mol. Mutagen. 21 (1992) 15–37.
[6] ICH (International Cooperation on Harmonization), ICH Harmonized Tripartite Guideline. S2A. Guidance on Specific Aspects of Regulatory Genotoxicity Tests for Pharmaceuticals, July, 1995. (http://www.ich.org).
[7] ICH (International Cooperation on Harmonization), ICH Harmonized Tripartite Guideline. S2B. Genotoxicity: A Standard Battery for Genotoxicity Testing of Pharmaceuticals, July, 1997. (http://www.ich.org).
[8] Council directive concerning the placing of plant protection products on the market (91/414/CEE), 15 July 2001.
[9] D.J. Kirkland, Genetic toxicology testing requirements: Official and unofficial views from Europe, Environ. Mol. Mutagen. 21 (1993) 8–14.
[10] D. Kirkland, M. Aardema, L. Henderson, L. Muller, Evaluation of the ability of a battery of three genotoxicity tests to discriminate rodent carcinogens and non-carcinogens. I. Sensitivity, specificity and relative predictivity, Mutat. Res. 584 (2005) 1–256.
[11] P.G.N. Kramers, A.G.A.C. Knaap, C.A. van der Heijden, M.R.D.F. Taalman, G.R. Mohn, Role of genotoxicity assays in the regulation of chemicals in The Netherlands: considerations and experiences, Mutagenesis 6 (1991) 487–493.
[12] OECD (Organisation for Economic Co-operation and Development), OECD Guidelines for Testing of Chemicals. Introduction to the OECD Guidelines on genetic toxicology testing and guidance on the selection and application of assays, OECD, Paris, France, 1986.
[13] T. Sofuni, Japanese guidelines for mutagenicity testing, Environ. Mol. Mutagen. 21 (1993) 2–7.
[14] E. Zeiger, Identification of rodent carcinogens and noncarcinogens using genetic toxicity tests: premises, promises, and performance, Regul. Toxicol. Pharmacol. 28 (1999) 85–95.
[15] TGD, Technical Guidance Document, second ed., European Chemicals Bureau, Joint Research Centre, Italy, December 2003 (http://ecb.jrc.it/new-chemicals/documents), 2003.
[16] L. Müller, D. Blakey, K.L. Dearfield, S. Galloway, P. Guzzie, M. Hayashi, P. Kasper, D. Kirkland, J.T. MacGregor, J.M. Parry, L. Schechtman, A. Schmidt, N. Tanaka, D. Tweats, H. Yamasaki, Strategy for genotoxicity testing and stratification of genotoxicity test results-report on initial activities of the IWGT Expert Group, Mutat. Res. 540 (2003) 177–181.
[17] Approaches to determining the mutagenic properties of chemicals: risk to future generations, Subcommittee On Environmental Mutagenesis, Committee to Coordinate Toxicology and Related Programs, U.S. Department of Health Education and Welfare (DHEW), Washington, D.C., 1977, pp. 1–58. Also published as: W.G. Flamm, L.R. Valcovic, F.J. deSerres, W. D'Aguanno, L. Fishbein, S. Green, H.V. Malling, V. Mayer, M. Prival, G. Wolff, E. Zeiger. Approaches to determining the mutagenic properties of chemicals: risk to future generations. J. Environ. Pathol. Toxicol. 1 (1977) 301–352.
[18] B.A. Bridges, ICPEMC Perspectives on the scientific basis of risk assessment in genetic toxicology, in: P. Oftedal, A. Brogger (Eds.), Risk and Reason: Risk Assessment in Relation to Environmental Mutagens and Carcinogens, Alan R. Liss, New York, 1986, pp. 175–182.
[19] D.J. Brusick, Methods for Genotoxic Risk Assessment, Lewis Publishers, CRC Press, Boca Raton, FL, 1994.
[20] J. Lewtas, D.M. DeMarini, J. Favor, D.W. Layton, J.T. MacGregor, J. Ashby, P.H.M. Lohman, R.H. Haynes, M.L. Mendelsohn, Risk characterization strategies for genotoxic environmental agents, in: D.J. Brusick (Ed.), Methods for Genotoxic Risk Assessment, Lewis Publishers, CRC Press, Boca Raton, FL, 1994, pp. 125–169.
[21] National Research Council (NRC), Risk assessment in the federal government, National Research Council, National Academy Press, Washington, D.C., 1983.
[22] J.T. Macgregor, D. Casciano, L. Müller, Strategies and testing methods for identifying mutagenic risks, Mutat. Res. 455 (2000) 3–21.
[23] U.S. Environmental Protection Agency (EPA), Guidelines for carcinogen risk assessment, EPA/630/P-03/001F, March 2005 (http://www.epa.gov/cancerguidelines).
[24] K.L. Dearfield, M.M. Moore, Use of genetic toxicology information for risk assessment, Environ. Mol. Mutagen. 46 (2005) 236–245.
[25] U.S. Environmental Protection Agency (EPA), Supplemental guidance for assessing cancer susceptibility from early-life exposure to carcinogens, EPA/630/R-03/003F, March 2005 (http://www.epa.gov/cancerguidelines).
[26] D. Jacobson-Kram, A. Jacobs, Use of genotoxicity data to support clinical trial or positive genetox findings on a candidate pharmaceutical or impurity ... Now what? Int. J. Toxicol. 24 (2005) 129–134.
[27] FDA (Food and Drug Administration) CDER (Center for Drug Evaluation and Research), Guidance for industry and review staff: recommended approaches to integration of genetic toxicology study results, January 2006.
[28] U.S. Environmental Protection Agency (EPA), Environmental Protection Agency. Guidelines for mutagenicity risk assessment, Washington, DC, Fed Register 51 (1986) 34006–34012.
[29] L. Rhomberg, V. Dellarco, C. Siegel-Scott, K. Dearfield, D. Jacobson-Kram, A quantitative estimation of the genetic risk associated with the induction of heritable translocations at low-dose exposure: ethylene oxide as an example, Environ. Mol. Mutagen. 16 (1990) 104–125.
[30] K.L. Dearfield, G.R. Douglas, U.H. Ehling, M.M. Moore, G.A. Sega, D.J. Brusick, Acrylamide: a review of its genotoxicity and an assessment of heritable genetic risk, Mutat. Res. 330 (1995) 71–99.
[31] M.C. Cimino, Comparative overview of current international strategies and guidelines for genetic toxicology testing for regulatory purposes, Environ. Mol. Mutagen. 47 (2006) 362–390.
[32] R.D. Snyder, J.W. Green, A review of the genotoxicity of marketed pharmaceuticals, Mutat. Res. 488 (2001) 151–169.
[33] D.J. Kirkland, L. Müller, Interpretation of the biological relevance of genotoxicity test results: the importance of thresholds, Mutat. Res. 464 (2000) 137–147.
[34] L. Müller, P. Kasper, Human biological relevance and the use of threshold-arguments in regulatory genotoxicity assessment: experience with pharmaceuticals, Mutat. Res. 464 (2000) 19–34.

[35] US FDA, Center for Drug Evaluation and Research (CDER), Guidance for Industry: Safety Testing of Drug Metabolites. Draft, June 2005.
[36] E.R. Nestmann, D.W. Bryant, C.J. Carr, Toxicological significance of DNA adducts: summary of discussions with an expert panel, Regul. Toxicol. Pharmacol. 24 (1996) 9–18.
[37] S. Madle, W. Von der Hude, L. Broschinski, G.R. Jänig, Threshold effects in genetic toxicology: perspective of chemicals regulation in Germany, Mutat. Res. 464 (2000) 117–121.
[38] M. Kirsch-Volders, A. Vanhauwaert, U. Eichenlaub-Ritter, I. Decordier, Indirect mechanism of genotoxicity, Toxicol. Lett. 140/141 (2003) 63–74.
[39] M.M. Moore, M. Honma, J. Clements, G. Bolcsfoldi, B. Burlinson, M. Cifone, J. Clarke, R. Delongchamp, R. Durward, M. Fellows, B. Gollapudi, S. Hou, P. Jenkinson, M. Lloyd, J. Majeska, B. Myhr, M. O'Donovan, T. Omori, C. Riach, R. San, L.F. Stankowski Jr., A.K. Thakur, F. Van Goethem, S. Wakuri, I. Yoshimura, Mouse lymphoma thymidine kinase gene mutation assay: mouse lymphoma thymidine kinase gene mutation assay: follow-up meeting of the international workshop on genotoxicity testing – Aberdeen, Scotland, 2003 – assay acceptance criteria, positive controls, and data evaluation, Environ. Mol. Mutagen. 47 (2006) 1–5.
[40] N.E. McCarroll, A. Protzel, Y. Ioannou, H.F. Stack, M.A. Jackson, M.D. Waters, K.L. Dearfield, A survey of EPA/OPP and open literature on selected pesticide chemicals. III. Mutagenicity and carcinogenicity of benomyl and carbendazim, Mutat. Res. 512 (2002) 1–35.
[41] B.H. Margolin, K.J. Risko, The statistical analysis of *in vivo* genotoxicity data: case studies of the rat hepatocyte UDS and mouse bone marrow micronucleus assay, in: J. Ashby, et al. (Eds.), Evaluation of Short-Term Tests for Carcinogens. Report of the International Program on Chemical Safety's Collaborative Study on *In vivo* Assays, vol. 1, Cambridge University Press, Cambridge, pp. 129–142.
[42] M. Hayashi, I. Yoshimura, T. Sofuni, M. Ishidate Jr., A procedure for data analysis of the rodent micronucleus test involving a historical control, Environ. Mol. Mutagen. 13 (1989) 347–356.
[43] D. Brusick, J. Ashby, E. de Serres, P.H.M. Lohman, T. Matsushima, B.E. Matter, M.L. Mendelsohn, M.D. Waters, Weight-of-evidence scheme for evaluation and interpretation of short-term results, in: C. Ramel, B. Lambert, J. Magnusson (Eds.), Genetic Toxicology of Environmental Chemicals, Part B: Genetic Effects and Applied Mutagenesis, Alan R. Liss, New York, 1986, pp. 121–129.
[44] D.J. Brusick, J. Ashby, E.J. de Serres, P.H.M. Lohman, T. Matsushima, B.E. Matter, M.L. Mendelsohn, D.H. Moore, S. Nesnow, M.D. Waters, A method for combining and comparing short-term genotoxicity test data, Mutat. Res. 266 (1992) 1–60.
[45] J.M. Parry, A. Sarrif (Eds.), Dose-response and threshold-mediated mechanisms in mutagenesis, Mutat. Res. Spec. Issue 465 (1999) 1–160.