Mutat Res Gen Tox En 848 (2019) 403074

Contents lists available at ScienceDirect

Mutat Res Gen Tox En

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

Recommended criteria for the evaluation of bacterial mutagenicity data (Ames test)

Dan D. Levy[a,*], Errol Zeiger[b], Patricia A. Escobar[c], Atsushi Hakura[d], Bas-jan M. van der Leede[e], Masayuki Kato[f], Martha M. Moore[g], Kei-ichi Sugiyama[h]

[a] US Food and Drug Administration, Center for Food Safety and Applied Nutrition, College Park, MD, 20740, USA
[b] Errol Zeiger Consulting, Chapel Hill, NC, 27514, USA
[c] Merck & Co. Inc., West Point, PA, USA
[d] Tsukuba Drug Safety, Eisai Co., Ltd., Tsukuba, Ibaraki, 300-2635, Japan
[e] Non-Clinical Safety, Janssen Research & Development, a Division of Janssen Pharmaceutica N.V., Beerse, Belgium
[f] CMIC Pharma Science Co., Ltd., Hokuto, Yamanashi, Japan
[g] Ramboll US Corporation, Little Rock, AR, 72201, USA
[h] Division of Genetics and Mutagenesis, National Institute of Health Sciences, Kawasaki, Kanagawa, 210-9501, Japan

ARTICLE INFO

Keywords:
Ames
Regulatory testing
Interpretation criteria

ABSTRACT

A committee was constituted within the International Workshop on Genetic Toxicology Testing (IWGT) to evaluate the current criteria for a valid Ames test and to provide recommendations for interpretation of test results. Currently, determination of a positive vs. a negative result is made by applying various data evaluation procedures for comparing dosed plates with the concurrent solvent control plates. These evaluation procedures include a requirement for a specific fold increase (2- or 3-fold, specific to the bacterial strain), formal statistical procedures, or subjective (expert judgment) evaluation. After extensive discussion, the workgroup was not able to reach consensus recommendations in favor of any of these procedures. There was a consensus that combining additional evaluation criteria to the comparison between dosed plates and the concurrent solvent control plates improves test interpretation. The workgroup recommended using these additional criteria because the induction of mutations is a continuum of responses and there is no biological relevance to a strict dividing line between a positive (mutagenic) and not-positive (nonmutagenic) response. The most useful additional criteria identified were a concentration-response relationship and consideration of a possible increase above the concurrent control in the context of the laboratory's historical solvent control values for the particular tester strain. The workgroup also emphasized the need for additional testing to resolve weak or inconclusive responses, usually with altered experimental conditions chosen based on the initial results. Use of these multiple criteria allowed the workgroup to reach consensus on definitions of "clear positive" and "clear negative" responses which would not require a repeat test for clarification. The workgroup also reached consensus on recommendations to compare the responses of concurrent positive and negative controls to historical control distributions for assay acceptability, and the use of control charts to determine the validity of the individual test.

1. Introduction

The bacterial reverse mutation test (Ames test) using *Salmonella* and *E. coli* bacterial strains is one of the most widely used tests for the identification of mutagenic substances. The test is used among the chemical, cosmetic industry, pharmaceutical and agro-industrial fields as part of the genetic toxicity testing battery required by regulatory agencies to enable marketing of the products. Regulatory acceptance of Ames test data often requires that the test be performed according to the Organisation for Economic Cooperation and Development (OECD) test guideline (TG)471 [1] and/or ICH S2R1 [2]. OCED test guideline TG471 was adopted in 1983 [3], revised in July 1997, and has not been updated since then although the majority of the other OECD genetic toxicity test guidelines have been revised and/or updated during the last decade [4,5].

The test consists of parallel subtests in which each of the 5 OECD-recommended bacterial tester strains (i.e., *Salmonella* TA98, TA100, TA1535, TA97 or TA1537, and TA102 or *E. coli* WP2 *uvr*A or *E. coli* WP2

* Corresponding author at: 5100 Campus Drive, College Park, MD 20740, USA.
E-mail address: Dan.levy@fda.hhs.gov (D.D. Levy).

https://doi.org/10.1016/j.mrgentox.2019.07.004
Received 7 March 2019; Accepted 11 July 2019
Available online 05 August 2019

*uvr*A pKM101) are exposed to the test substance with and without a metabolic activation system derived from a rodent liver extract (S9). Each strain has a different mutation in an endogenous gene needed for synthesis of an amino acid required for growth (histidine in *Salmonella* or tryptophan in *E. coli*). If exposure to the test substance leads to a single nucleotide base change, or base insertion or deletion which allows the mutated cell to produce either the histidine or tryptophan needed to grow into a colony on an agar plate in the absence of the previously required amino acid, the substance is considered to be a mutagen. The test is considered positive if treatment with test chemical causes an increase of sufficient magnitude (discussed in detail below) in the number of revertant (mutant) colonies in any one of the subtests.

In 2017, a working committee was constituted within the International Workshop on Genotoxicity Testing (IWGT) to evaluate the currently used procedures for performing and evaluating the test, and to provide recommendations for improving test conduct and interpretation of test results. This IWGT workgroup held multiple discussions, including a face-to-face meeting during the full IWGT meeting in Tokyo, Nov. 8–10, 2017. The group developed recommendations drawing on experience from various industries, testing laboratories, experienced individuals, and regulatory agencies. This publication presents the outcome of the discussion at that meeting and subsequent discussions by this workgroup. A key facet of the new recommendations was the integration of multiple criteria, beyond simple comparisons with the concurrent control, into the overall evaluation of the test outcome. Another key facet was clearer descriptions of strategies to follow up an inconclusive result [4,5]. In the last decade, most of the Guidelines for other genetic toxicology tests were updated by the OECD. The updated guidelines contain recommendations for test data interpretation that differentiate between results which are clearly positive and negative, and results which are ambiguous or inconclusive [4,5]. Where appropriate, these workgroup recommendations for the Ames test are aligned with recommendations introduced into the recently updated OECD genetic toxicity Test Guidelines.

2. Criteria for a valid test

The workgroup developed consensus on several criteria that should be fulfilled beyond what is specified in the current OECD TG471 for a test to be regarded as acceptable. Recommendations for bacterial strain identification, maintenance and growth are addressed elsewhere [6].

2.1. Are the background (spontaneous mutation) counts in each strain consistent with literature values and the laboratory's historical solvent control values?

2.1.1. Use of published values for background (spontaneous mutation) counts

Because the solvent (negative) control is critical for the determination of a positive or negative response, the workgroup extensively discussed recommended ranges for the solvent controls based on an examination of older and more recent literature, and recent laboratory experience of the workgroup participants. Every test should be run using a concurrent "solvent control", a set of plates using the solvent or vehicle used for the test substance. An additional negative control (e.g., using saline or water) is optional unless the laboratory has not developed a historical control data set for the particular solvent. Concurrent solvent control data using different solvents are usually combined into a single historical control database if they have been shown to have similar responses, although some labs track commonly used solvents (e.g., DMSO) separately. The expected range of values for the solvent control is different for each strain; there may also be differences with and without S9 depending on the strain and the particular S9 source or concentration used. Demonstration that a particular solvent is compatible with the test is discussed elsewhere [6]. The values in Table 1, collected from available literature, represent reasonable ranges for the commonly used strains, and illustrates that there can be considerable variation in results among laboratories, and between controls with and without S9. It should be noted that variations in solvent control values in an experiment do not correlate with the response to the positive control in the same experiment. Kato et al. (see Supplementary Fig. 7 in reference [7]) plotted the mean number of revertants in the solvent control as a function of the mean for the positive control the same laboratory in each of the 5 strains tested with and without metabolic activation and reported no such correlation within each of the approximately 25 participating laboratories. Data from individual experiments also show no correlation (e.g. see Fig. 1 in Carnes et al. [8]) To further illustrate the variability among laboratories we chose data from two test strains, TA100 and TA1537 (Table 2), from recent tests conducted under Good Laboratory Practices (GLP) procedures [9–11]. For the purposes of this publication, the workgroup did not think it necessary to establish maximum or minimum values for solvent controls in each strain and recommends that each laboratory should establish and define an acceptable historical control range for each test strain with and without S9 with reference to literature findings, but using data generated by the laboratory.

Table 1
Reported means for solvent controls.

Strain	Solvent Control Range*	Reference
Salmonella		
TA 97	75–200 (−S9)	[29]
	100–200 (+S9)	[29]
	90–180	[30]
TA 98	15–60	[31
	20–50 (±S9)	[29,30]
	30–50	[7]
	5–37 (–S9)	[7]
	15–42 (+S9)	
TA 100	75–200	[31]
	75–200 (±S9)	[29]
	68–137 (−S9)	[7]
	74–156 (+S9)	[7]
TA 102	200–400	[32]
	258–376	[33]
	100–300 (−S9)	[29]
	200–400 (+S9)	[29]
	350–420 (−S9)	[34]
TA 104	200–300 (−S9)	[29]
	300–400 (+S9)	
TA 1535	3–37	[31]
	5–20 (±S9)	[29]
	4–17 (±S9)	[7]
TA 1537	4–31	[23]
	5–20 (±S9)	[29]
	3–15 (−S9)	[7]
	4–23 (+S9)	[7]
TA 1538	5–20 (±S9)	[29]
E. coli		
WP2 *uvrA*	14–33 (−S9)	[7]
	13–43 (+S9)	[7]
WP2 uvrA (pKM101)	45–151	[35]
	53–148	[33]
	90–110 (−S9)	[34]
WP2 (pKM101)	35–166	[33]
	48–53 (−S9)	[34]

* Some of the values are from individual laboratories; others are consensus ranges from multiple laboratories.

2.1.2. There was a consensus that laboratories should assemble their solvent control data in control charts and develop procedures using the negative (solvent) control charts to evaluate their tests

- Laboratories need to set "control limits" and define "rules" (defined below) describing actions to be taken when the concurrent control goes beyond their historical range (regardless of how that range is

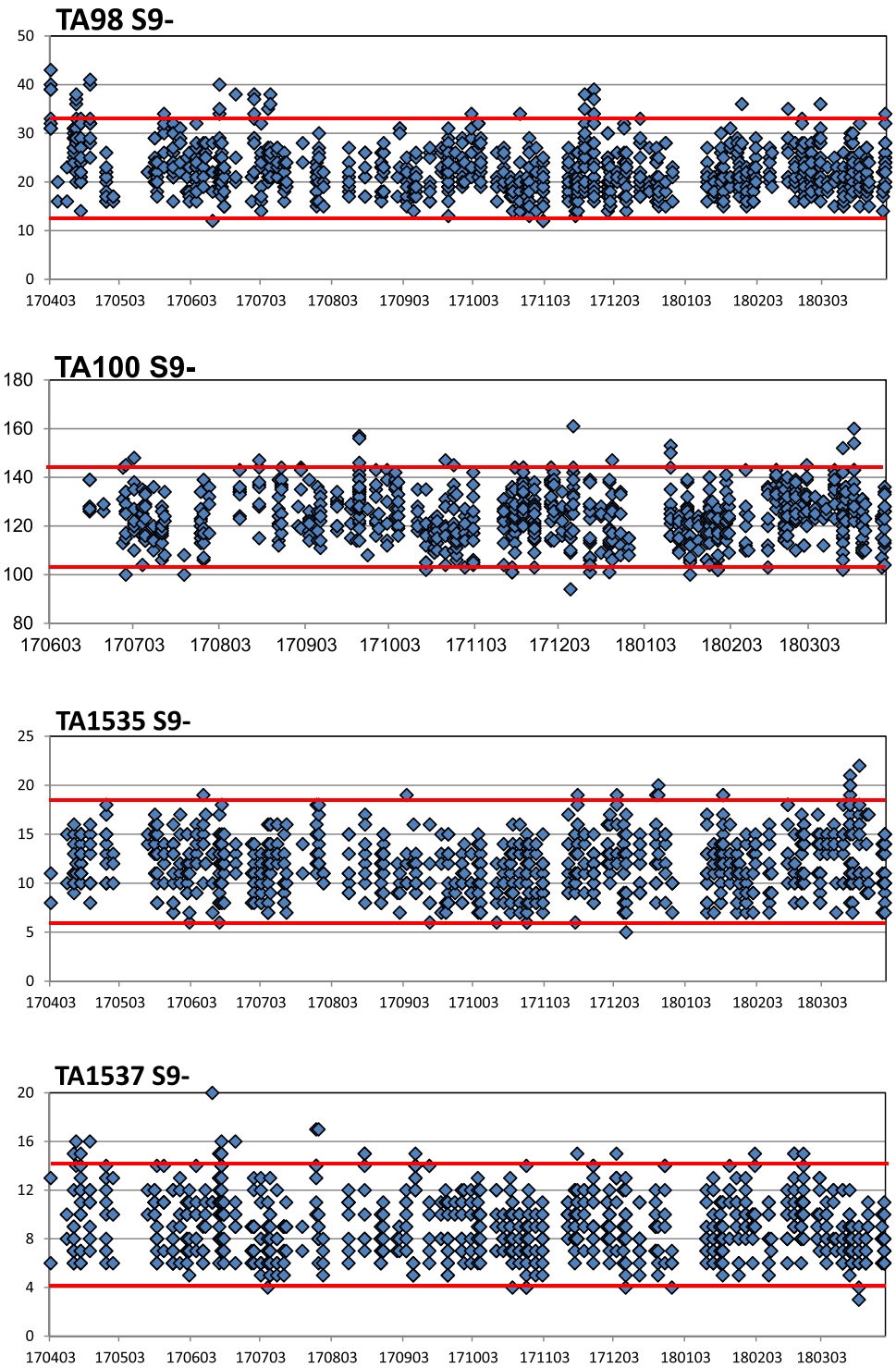

Fig. 1. Solvent control charts. Counts from each of >800 single plates (vertical axis) on the day of each test conducted between April 2017 and March 2018 (horizontal axis). Horizontal lines demarcate 2 standard deviations around the mean. Only data for 4 strains tested without metabolic activation (−S9) are shown here, TA98 (panel a), TA100 (panel b), TA1535 (panel c) and TA1537 (panel d) A complete set of charts for 5 strains with and without metabolic activation as well as a more detailed description of the methods and the statistical characterizations of the data from the laboratory of author MK are provided in Supplementary File 1.

defined) as determined by statistical analyses of their own data.

- Laboratories should create control charts and use them to inspect data for trends and establish procedures to follow when the data indicate that the experiments in one or more strains are outside of an acceptable range or may be trending out of the acceptable range (Fig. 1).
- Solvent control counts are not acceptable if they are too close to zero. If the control reversion frequency is too low the power of a test to detect a small increase is poor even though the solvent control values may be within the laboratory's acceptable confidence intervals. An experiment should not be accepted if there were no revertants on any of the solvent control plates for a strain. Based on experience with the test, most workgroup members agreed that the average for an experiment should be a minimum of 4 or 5 colonies

Table 2
Distribution of negative control counts in multiple and single laboratories.

	TA100 (no S9)					TA100 + S9				
	Multilab[a]	Single[b]	Single[c]	Single[d]	Single[e]	Multilab[a]	Single[b]	Single[c]	Single[d]	Single[e]
Number	26 labs	791 exp	20 exp	283 exp	41exp	26 labs	787 exp	20 exp	258 exp	41 exp
Mean	102	124	115	103	103	115	135	115	123	119
Std Dev.	17	10.7	14	19	15	20	9.4	15	19	14
Min	69	94	84	53	52	78	103	90	78	74
Max	132	161	140	189	155	156	175	151	179	155
Mean –2 Std Dev	68	103	87	65	73	74	116	85	85	91
Mean +2 Std Dev	137	145	143	141	133	156	154	145	161	137

	TA 1537 (no S9)				TA 1537 + S9			
	Multilab[a]	Single[b]	Single[c]	Single[d]	Multilab[a]	Single[b]	Single[c]	Single[d]
Number	25 labs	828 exp	20 exp	259 exp	25 labs	818 exp	20 exp	243 exp
Mean	9	9	7	7	13	14	10	7
Std Dev.	3.0	2.5	4	3	4.6	2.2	3	3
Min	3	3	2	3	6	3	4	2
Max	15	20	26	20	23	21	20	19
Mean –2 Std Dev	3	4	<1	1	4	10	4	1
Mean +2 Std Dev	15	14	15	13	23	18	16	13

Comparison of negative control count distributions with and without rat S9 for the strains with high (TA100) and low backgrounds (TA1537).
exp: Number of experiments contributing to laboratory values. Data are from:
[a] Multi-laboratory study in which 25 or 26 labs provided the means of duplicate plates in a single preincubation experiment [7].
[b] Summary of the single laboratory preincubation data in Fig. 1.
[c] Unpublished data from the laboratory of one of the authors. Mean of triplicate plates using preincubation.
[d] Unpublished data from the laboratory of one of the authors. Mean of triplicate plates using plate incorporation.
[e] Unpublished data from the laboratory of one of the authors. Mean of triplicate plates using plate incorporation.

per plate on two or three solvent control plates, but no specific lower limit was recommended.
- The test should be considered invalid if a concurrent solvent control for the strain is outside a predetermined control limit or if the control is contaminated or otherwise compromised.

Fig. 1 shows an example of a control chart that displays the information used by a laboratory to track lab performance over time in TA98, TA100, TA1535 and TA1537 without S9. Charts for other strains and a more complete description of the methods and results are provided in Supplementary Information (File 1). There are useful discussions of the use of control charts in the literature [12–14], and how the control charts were originally developed to implement good manufacturing procedures to facilitate product quality control.

Each laboratory should set rules based on its experience with the use of the historical control values to support a decision to accept or reject a test prior to performing the test. These are referred to as “control limits” or sometimes as “action limits”. For example, if a concurrent solvent control value from a test is higher than the upper historical control limit (regardless of how that value is defined) or below the lower limit, the laboratory may set a rule automatically rejecting the test. Some laboratories set “warning limits” together with “rules” which dictate how the limit is used. For example, a laboratory might set 3 standard deviations above the mean as an upper control limit and 2 standard deviations above the mean as an upper warning limit. The warning limit is an alert which allows intervention to investigate a problem before it becomes necessary to reject valuable experimental data. A laboratory might set a rule, for example, that three consecutive experiments in a row above the warning limit (but below the control limit) nonetheless warrants some action. A laboratory might also set a rule that 10 sequential experiments with values above the mean warrants data rejection or some other action. The action could be an investigation whether there has been a change in the procedure which warrants attention. For example, it may be the first indication that control ranges need to be reset based on use of a new reagent lot (e.g., agar; nutrient broth; S9), contamination of the solvent stock solution, or deterioration or genetic drift of a bacterial stock in storage over time.

There is no standard, generally accepted, statistical parameter for evaluation of the historical solvent control data. Inspection of the data in Fig. 1, and in Kato *et al.* [7] show that using the mean ± 2 standard deviations is a good starting point for a warning level and the mean ± 3 standard deviations is a good starting point for control limits. Those figures also demonstrate that actual lab data deviate from idealized distributions on which some statistical methods rely (see for example reference [15], as well as the statistics section, below). Upper and lower limits and rules should be constructed for individual strains based on the data being generated by the laboratory, with separate determinations for each strain with and without the S9 metabolic activation system. Some laboratories prefer the use of narrower confidence intervals (e.g., the 99% confidence interval is approximately the mean ± 3 standard deviations/√N). The tolerance interval, a statistically calculated distribution range that contains 95% of the data with 95% level of confidence can be used and has the advantage of relying on the actual rather than an assumed data distribution.

There was a consensus that historical control databases and control charts should have data from at least 20 recent experiments. Some members expressed concern that control ranges might begin to get too narrow if the reference databases are too large, but there was no consensus that there should be an upper limit on the number of experiments used.

2.2. Are the mutation counts induced by positive controls in each strain consistent with the laboratory's historical positive control values?

- There was a consensus that the positive controls listed in TG471 are appropriate.
- Laboratories should assemble positive control data in control charts and develop procedures control charts to evaluate the positive control values for each strain/activation combination.
- Laboratories should set the control rules and warning limits for these acceptance criteria based on their own experience and data.
- A valid positive control response is required for a negative test in

any one strain to be considered valid.

There was no consensus on the absolute highest or lowest value or ranges for positive controls. The revertant counts induced by the positive controls are expected to be more variable than the solvent control counts among different laboratories (for example see Figs. 7–10 and Table 2 in Kato et al. [7]). This increased variability can be expected to be evident in control charts for a single laboratory. This variability depends partly on the metabolic activity of the particular S9 batch used and also on variability in the preparation of positive control solutions for the test, and variations in preparing the positive control sample for testing. The well-known positive controls listed in TG471 have the advantage that laboratories and regulatory reviewers can compare the responses to those seen in other labs, although any mutagenic substance can be used as a positive control with proper justification (e.g., literature data; structural similarity). In order for a positive control value to be valid it needs to meet the laboratory's criteria for a positive response.

2.3. Are there enough "analyzable concentrations"?

The OECD test guideline recommends data from "at least 5 analyzable concentrations" in each strain with and without metabolic activation.

- There was consensus that a solvent control and 5 concentrations using the current dose interval spacing, i.e., half-log doses, with a limit dose of 5000 μg or 5 μl/plate in the absence of toxicity or precipitation [1], is adequate for the initial experiment.
  - It sometimes becomes necessary to perform additional experiments with more closely spaced concentrations.
  - When there is evidence of cytotoxicity and no evidence of increased revertants, the concentration spacing below the lowest concentration at which cytotoxicity is observed should be sufficiently close together to assure that there is no increase in revertants that might occur between a totally non-toxic concentration and a concentration showing substantial toxicity.
  - When the initial experiment is inconclusive additional testing with a larger number of more closely spaced concentrations is often needed to differentiate between an increase induced by the test article and a high value consistent with random variation.
- When the top dose is limited by toxicity there can be disagreement about how many non-toxic doses are needed to establish a negative test.
  - The FDA Redbook [16] advises that when evaluating food additives "[i]f the doses of the test substance are limited by toxicity, then toxicity should be evident … at one or more concentrations, and no toxicity should be evident at three or more concentrations in each test, in each bacterial strain, both with and without metabolic activation …". The workgroup agreed that a preliminary toxicity screen using one or two plates per dose and one or two strains with and without metabolic activation is not an absolute requirement but usually provides enough information to set up an initial experiment with five concentrations (in addition to the solvent control) in all five strains which will produce data consistent with the Redbook recommendation
  - However, there was also consensus that rigid application of this recommendation is not helpful when there are more plates with toxicity than expected in an experiment. The appropriate number of doses and dose spacing chosen for follow-up experiments will depend on a number of factors specific to the experiment. These factors include the variability among the plate counts, reproducibility of plate counts at each concentration among repeat experiments, the ways toxicity is manifest (i.e., reduction in revertant counts and/or thinning of the background lawn) and the steepness of the toxicity response.
  - After examining several hypothetical data scenarios, it became apparent to the workgroup that there was no scientific basis for establishing a universal minimum number of non-toxic doses needed to establish the validity of a finding that a compound is non-mutagenic in the tests.
- Additional testing is not needed if a result is judged to be positive from an experiment with fewer than 5 analyzable concentrations.

3. Interpretation of test results

3.1. The overall test call vs. results for each strain

Once an experiment has been determined as meeting the above acceptance criteria the data can be interpreted as positive, negative, or inconclusive (i.e., neither positive nor negative; requiring a repeat test). The test substance is a mutagen if a positive response is seen in any one strain/activation combination, regardless of whether the other strains were tested or their responses. In contrast, a test substance is considered nonmutagenic if it is tested and shown to be negative in, at a minimum, the strains specified in TG471 with and without S9.

3.2. Distinguishing positive responses from negative responses

The current OECD TG471 Test Guideline provides minimal guidance for distinguishing positive from negative responses:

> "There are several criteria for determining a positive result, such as a concentration-related increase over the range tested and/or a reproducible increase at one or more concentrations in the number of revertant colonies per plate in at least one strain with or without metabolic activation system …. Biological relevance of the results should be considered first. Statistical methods may be used as an aid in evaluating the test results…. However, statistical significance should not be the only determining factor for a positive response. … A test substance for which the results do not meet the above criteria is considered nonmutagenic in this test." ([1]; at para. 35, 36).

TG471 also states that "the biological relevance of the result should be considered first." The workgroup agreed that this statement was not very clear and that additional data and discussion would be needed in order to come up with clearer guidance on how to define an increase large enough to be "a clear increase". The workgroup recognized that the induction of mutation over the background level is a continuum of responses; there is no 'fine bright line' between a mutagenic (biologically relevant) and a nonmutagenic (not biologically relevant) response. Thus, developing absolute criteria proved to be difficult and different workgroup members had differing opinions as to the most appropriate approach. The discussion was complicated by the fact that these approaches include very different types of "procedures" or "approaches" (e.g., fold-rules or various statistical tests) historically used to judge the comparison of dosed plates to the concurrent control. Within each of these procedures, unique criteria need to be developed to judge the comparison between vehicle controls and test plates. No consensus was reached on how to define a "clear increase", however there was consensus that either a fold-rule or statistical approach appropriate for Ames-test data, described below, was acceptable in this assay, and that individual laboratories can select one of these based on its preference, recognizing that the two approaches can yield different outcomes depending on the individual test data.

3.3. Data evaluation procedures

The workgroup started by discussing procedures used to compare dosed plates with the concurrent solvent control and then developed an overall evaluation approach for the Ames test consistent with the approaches developed for the other OECD genetic toxicity test guidelines. Consensus was reached on several important aspects of how to identify a positive or negative result and how to conduct follow-up testing when

the results of the initial test are not clearly positive or negative. The workgroup could not reach consensus on all issues, as discussed below.

3.4. Procedures currently used to compare dosed plates to the concurrent solvent control

3.4.1. Fold rule

This rule, which is the most widely used, derives from the original methods publication of Ames *et al.* [17] "We would consider a chemical to have a negative response in the test … if the number of induced revertants compared to the spontaneous was less than 2-fold." This "rule" was subsequently modified when it was realized that, although one would feel confident that a 2-fold increase should be considered positive when the mutant values increased, for example, from a mean of 100 in the solvent control to 200 with the test substance, the same confidence would not apply of the solvent control was, for example, 5, in the case of Salmonella strain TA1537, which would result in an increase from a mean of 5–10 being considered mutagenic. For this reason, laboratories require a larger increase (usually 2.5- or 3-fold) when the solvent control is low (<10, 15, 20, etc.) (e.g., [18]). The major strength of a fold-rule is that it is easy to apply. The weakness of this approach is that it does not reflect actual biology. The induction of mutation following chemical exposure adds to the background number of mutants rather than multiplying the background. The range of solvent control values in Table 1 varies from 3 to 400, or over 100-fold. The use of a single 2-fold criterion for this wide range of mean reversion frequencies results in very different sensitivity to detect a small increase. Use of special rules for low background strains, like a 3-fold rule when the background is <10 or <20 helps, but there is still a wide range of backgrounds. That is, the rule makes no distinction between 2-fold responses in strains with a solvent control value of 40 (e.g., TA98) and those with a value of 240 (e.g., TA102). The use of the fold rule is likely to identify a small but reproducible increase for TA1535 and may miss an increase of similar biological value in TA100 which has the same mutagenic target site (see [19]). Fold rules also result in very different likelihoods that a random result will be interpreted as a positive result. As illustrated in Fig. 1, random variation in negative controls commonly resulted in values >15 in strain TA1535. That is 3-fold higher than the minimum value of 5 revertants per plate seen on occasion in that same dataset and thus the historical solvent control database contains values which, had they been seen in the same experiment, could have been described as "a clear increase". These data illustrate why it is necessary to consider other criteria (i.e., concentration related increase, increase outside historical control range) to ensure that random values obtained during testing are not mistakenly classified as positive results.

3.4.2. Statistics

There are a number of statistical approaches that are being used or have been proposed, including commercial packages, proprietary procedures, and off-the-shelf procedures, that are used as is or adapted for Ames test data. Some of these are based on different assumptions that may or may not be appropriate for the mutant colony count data obtained from the test; e.g., some statistical analyses assume normal distribution for all the tester strains and responses to negative/solvent and positive compounds, while others assume a Poisson distribution, which is mainly observed in tester strains with low spontaneous counts [15]. There are also unresolved questions about which p value to use (e.g., $p_{.05}$ vs. $p_{.01}$), whether the analyses should be one- or two-tailed, and whether there should be a correction for multiple comparisons since each test comprises 50 pairwise comparisons (5 concentrations in 5 tester strains in two metabolic activation conditions). Therefore, the various statistical analyses will not all agree on the significance of a particular, low-response, data set. It is not clear as to which procedures are sensitive to the data patterns produced by the OECD TG compliant protocol, e.g., hyper-Poisson distribution or normal distribution of replicate plate counts for strains with high (>100 revertants/plate) spontaneous frequencies, etc. In addition, each lab has its own background variability in each tester strain, and they will have to fit the statistical approaches to demonstrate that they are cognizant of their test protocol and the plate count variability in each tester strain. As illustrated in Fig. 1, actual laboratory data do not perfectly conform to the theoretical distributions. The frequency with which individual results are higher than the confidence limits defined by 2 and 3 standard deviations varies from strain to strain, and sometimes individual results are more common above the upper limits than below the lower limits indicating deviations from theoretical distributions. These patterns suggest deviations from the assumptions on which the statistical tests are based. Strains with lower mean revertant frequencies are usually closer to a Poisson distribution and strains with higher mean revertant frequencies closer to a normal distribution. For these reasons a laboratory could choose different rules and warning limits for each strain based on the laboratory's historical data. Control charts for the solvent controls can be used not only to track test performance over time but also as a check of the validity of statistical parameters used as decision criteria by the laboratory. Comparisons of fold-increase analysis with statistical analysis have shown that fold-analysis is more conservative than statistical analysis for strains with high background mutant counts, and less conservative for strains with low background counts given the plate-to-plate variability typically seen [20]. The need to adjust the decision criterion as the background frequency goes from lower to higher values is the most common defect identified by statistical analysis of the adequacy of the fold rule [20,21].

As with the fold rule, a few mutants among hundreds on a plate that are within the normal triplicate plate-to-plate variability can make the difference between a p value of 0.049 (positive) and 0.051 (not positive). While this workgroup made considerable progress in defining procedures for evaluating inconclusive Ames test data, it was unable to reach consensus conclusions on which, if any, data evaluation method can be recommended. A future workgroup might be able to complete the task but will need actual data and statistical support. Useful data would include historical control data for each strain from several labs from different global regions to ensure that the conclusions reached are robust and valid across laboratories and laboratories with varying levels of expertise with the test. It may also be helpful when evaluating the practical consequences of adopting various types of criteria for determining whether or not an increase is associated with the test article, to actually look at data from bacteria exposed to mutagens such as the data described above [7,22] to determine the performance of various proposed criteria in assessing weak mutagens or small amounts of potent mutagens.

3.5. Identification of clearly positive and clearly negative responses

Experience with the test has shown that the majority of test results will be clearly positive or negative. However, prior to addressing criteria for clearly positive or negative responses it is necessary to define the terms and concepts used. The most contentious issue during workgroup discussions was how to define "a clear increase" at any one concentration compared to the contemporaneous control. No consensus was reached.

The workgroup reached consensus that a test response is clearly positive if, for at least one strain either with or without metabolic activation, all of the following criteria are fulfilled:

1) there is a "concentration-related increase" in revertants, and
2) there is a "clear increase" in at least one concentration compared to the concurrent solvent control, and
3) there is at least one concentration with an increase "outside the distribution (based on control limits) of the laboratory's historical solvent control database".

In this situation, the workgroup agreed that no further experimental work is needed.

The workgroup reached a consensus that the test is clearly negative if:

1) there is no "concentration-related increase" in revertants, and
2) there is no "clear increase" at any concentration compared to the concurrent solvent control, and
3) there is no increase "outside the distribution (based on control limits) of the laboratory's historical solvent control database".

Definitions are discussed below.

The workgroup did not reach consensus on the need for a confirmatory test in the case of a clearly negative initial test. OECD TG471 states that a negative result may need to be confirmed by a repeat test, on a case-by-case basis [1]. Some regulatory recommendations, e.g., ICHS2(R1) have eliminated the requirement of a repeat test based on an examination of confidential industry databases [2]. The previous IWGT evaluation (Tables 10 and 11 of [23]) provided data from 7 laboratories that retested initial negative test data using the same protocol. Only 7 of the 1538 re-tests of negative results (0.45%) did not repeat. Some laboratories routinely conduct tests using both the preincubation and the plate incorporation version of the test to fulfill this requirement. The previous IWGT evaluation listed several chemical classes for which the pre-incubation assay was said to be more sensitive (Table 2 of [23]), however no data were available to address the issue of chemicals positive in the plate test but negative using preincubation. Some members of this workgroup found the available literature inadequate to support a recommendation to avoid the use of the plate incorporation assay for those chemical classes.

3.5.1. Definition of "a concentration related increase"

The workgroup reached consensus that "a concentration related increase" means increasing numbers of revertants at increasing concentrations at which there is no, or minimal, evidence of toxicity of the test article to the bacteria. The workgroup agreed that there is no expectation that the increase will be linear, and that a plateau or drop in the number of revertants with increasing concentration is an indication of toxicity of the test article whether or not there is any observable diminution or other effect on the background lawn. The maximum number of revertants per plate may be observed at the concentration immediately below the concentration at which toxicity is observed or can continue to increase at one or more of the higher concentrations (showing partial clearing of the background lawn). Unlike mammalian cell in vitro genetic toxicology tests, the most sensitive indication of cytotoxicity is often the revertant counts. Methods for measuring toxicity independent of mutagenic response have been described (e.g., see [24]) but have not been widely adopted. If statistical trend tests are used, several workgroup members felt strongly that data from high concentrations which show signs of toxicity (including a plateau or reduction in revertants with increasing concentration) should be excluded from the analysis as has long been the practice when analyzing data from this test e.g., [25]. In contrast, other workgroup members felt that some of the available statistical trend tests e.g. [26], adequately account for plateaus and other non-linearity and are valuable for evaluating Ames assay data.

3.5.2. Definition of a "clear increase"

The workgroup could not agree on a single definition of a clear increase relative to the concurrent solvent control. As such, individual laboratories and regulatory agencies must establish their own working definitions (see below). Test results which meet that definition are positive and test results which do not may still be judged positive based on the laboratory's criteria for data evaluation. Increases which are not judged to be "clear increases" are addressed in the next section.

3.5.3. Comparison of dosed plates to the historical solvent control distribution

The workgroup reached a consensus that it is appropriate to compare the response of the plates exposed to test substance to the historical control distribution (HCD) as part of the evaluation of a potentially mutagenic response. As addressed above, the procedures and acceptance criteria will necessarily vary for each laboratory and may differ among strains. The considerations discussed above in the section on control charts are applicable to determining how to undertake this comparison. Previous attempts to evaluate the use of the HCD relied on limited datasets. Carnes et al. [8] observed the weakness of the fold rule for strains with low background rates but lacked access to large enough historical control databases to properly evaluate their proposal to switch to comparison of the induced plate counts to the historical control range. Mitchell [27] compared use of historical control ranges among various genetic toxicology assays and expressed reservations about using the technique for the Ames test in contrast to the in vivo micronucleus and in vitro chromosomal aberration tests based on their observation of greater inter-test variability in the Ames test data they analyzed. Again, their historical databases were small (i.e., 16–47 experiments). However, based on experience with other assays and the use of larger historical control databases (see Example 1 in the last section of this publication), the workgroup agreed that the HCD could be helpful in determining whether an assay outcome is positive or not.

3.6. Strategies when results do not meet the criteria for a clearly positive or a clearly negative response

The workgroup reached a consensus that when one or two of the above criteria for a clearly positive response are met, but not all three, the test is neither clearly positive nor clearly negative. The initial result may still be robust enough to be considered positive or negative after a closer examination of the available data. If a test call cannot be made based on the initial result, the determination of an inconclusive (ambiguous) result is preliminary subject to further evaluation and often further experimentation.

3.6.1. Consider the reproducibility of the result and its magnitude relative to the normal variability within the assay

The appropriate interpretation of a modest increase can be the most difficult situation to resolve. Each of the decision criteria is an arbitrary point on a continuous scale. This is true whether one is using a fold rule, a statistical test like a p value, or a confidence or control interval. If two consecutive tests are near the boundary between a positive and negative response the result is more likely related to the test substance even if one value is just over the limit and the other test us just under the limit. That is, there is no biological difference between a 1.9-fold increase and a 2.1-fold increase, a statistical p value of 0.051 vs. 0.049, or a value just above or just below a 99% or 95% control limit. The presence of a reproducible concentration-related increase under the same or similar test conditions, even though the increase may not reach the level specified, i.e., fold-increase or statistical p value, strengthens a conclusion that the substance is mutagenic. Examples 2 and 3, below, illustrate this point.

Repeat testing should be designed to evaluate the reproducibility of the initial observation and may incorporate changes to the protocol in an attempt to emphasize or refute the initial response seen. Such changes can include an increased number of test substance concentrations around the dose contributing to the initial questioned response, altered metabolic activation conditions (if S9 was used), or modification of the test protocol, e.g., preincubation or vapor phase exposure instead of plate test, or a different solvent (if solubility may be in question). The workgroup agreed that any follow-up testing to clarify a weak or inconclusive response can be confined to the strain(s) and metabolic condition in which the initial result was inconclusive or was not clearly positive or negative.

The metabolic activation system can be varied in a repeat experiment when the result is inconclusive in the initial test in one of the strains using S9 mix. Based on the nature of the test article, using a variety of concentrations (e.g. 5, 10, 30%) of the same S9 mixture, use of an alternate source of S9 (e.g., mouse or hamster) or using other metabolic activation systems (e.g. hamster S9 plus FMN for azo compounds) based on literature recommendations may be appropriate. However, many laboratories do not alter metabolic activation conditions during follow up testing and instead use the strategies as described above, that is, they use a different version of the test and/or refine the concentrations selected for testing.

There are many valid reasons for a weak response that is difficult to resolve into a positive or negative call even after preforming a repeat test. The test article may be incompatible with the test system (e.g., biocides, many metals, substances which require complex metabolism not well modeled using rodent liver S9) or there may be low-level mutagenic impurities in a substance which is not itself mutagenic. It is also possible that the mutation target site (e.g., –GGG– in TA100 and TA1535) may be only weakly sensitive to the DNA-reactive moiety. Other proposed mechanisms include a response due to an extragenic suppressor so that the DNA target site is not directly mutagenized [28]. There is no broad consensus on the validity of these possibilities or a reliable way to distinguish between them and weak responses caused by the test agent or its metabolites. It was clear from the discussion within the workgroup that when the outcome of the test is neither clearly positive nor clearly negative, the most appropriate follow-up strategy for a particular test material is usually a repeat test using the same or modified test conditions.

An important consideration when evaluating the validity of small increases in a single or multiple strains is whether the response is consistent with what is known about the tester strains' responsiveness. Example 1: TA98 and TA1538 have the same *hisD*3052 mutation but differ in DNA repair capacity. The same mutagen is unlikely to cause a weak response only with metabolic activation in TA98 and a weak response only without metabolic activation in TA1538. Observation of a weak response with metabolic activation in both TA98 and TA1538 is more likely to be attributive to the test article. Example 2: TA1535 (*hisG*46) the target sequence -GGG- is reverted by a G:C → A:T base substitution mutation. TA1537 (*hisC*3076) reverted by a 1 base deletion in the -GGGG- target sequence. The same mutagen is unlikely to cause both types of mutations without also being positive in TA100 which has the same target sequence and sequence context as TA1535. These considerations have recently been discussed extensively [19]. In such situations, the test responses are not consistent with the biology of the individual strain responses, which would suggest that one or more of the responses may be artifactual, or that a mixture of reactive substances is present. Table 2 in [29] lists the reversion targets in the common strains. This is a situation which would support a repeat test using the strains and activation conditions in question and/or a chemical analysis of the test sample for possible contamination or impurity.

Consideration should be given to the presence of a low-concentration genotoxic impurity, etc. When an increase occurs only at the limit concentration in the absence of toxicity (usually 5 mg or 5 μl) and is not "a clear increase", a repeat test using higher concentrations can be considered. GLP regulations require that the test article identity, purity and composition be adequately defined and that the stability of the test item under storage and test conditions be known [9–11].

Expert judgment which, by nature, is subjective, is needed when the data show a response that cannot be judged clearly positive or negative using the formal evaluation criteria (above). In such a situation, it becomes a matter of expert judgment for which it is not possible to articulate precise objective criteria, and generally involves a weight of evidence consideration of factors specific to the particular experiment being evaluated and will vary among the "experts," and also with time, for individual researchers and regulatory reviewers as they gain more experience with the tests. In any evaluation of test results that are neither clearly positive nor clearly negative, the concepts of dose-response and reproducibility must be taken into consideration in addition to the biology of the test system (here, the tester strain sensitivities, the nature and chemistry of the test substance, and the pattern of responses among the tester strains).

3.6.2. Consider use of multiple factors when evaluating a weak increase

Determination of the significance of a small concentration-related increase or a small increase above the concurrent or historical controls often becomes a matter of expert judgment for which it is not possible to articulate precise objective criteria. There was a consensus that a single, strict criterion to determine whether a test showing a weak increase is positive or negative was not appropriate and that the use of multiple factors (i.e., requiring a concentration related response and consultation with historical controls) would lead to reliable and more consistent classification of test articles by study authors and by regulatory reviewers. There was also a consensus that a key factor in coming to a conclusion was the reproducibility of an increase when the increase is weak or inconclusive.

The workgroup reached a consensus that the relationship of the test substance response to the concurrent solvent control should generally be given more weight than comparisons against the HCD when the two comparisons result in different outcomes. (However, see Example 2, below, for a counter-example to this general rule). In addition, the greatest weight in making the overall call should be based on a determination of whether any observed increase is concentration-related. The workgroup agreed that comparing the response to the test substance to the historical control distribution for the strain, by using three rather than two criteria, provides an opportunity for a more nuanced weight-of-evidence evaluation (see Figs. 2 and 3).

Once the initial data evaluation and repeat tests are completed, the full dataset for the test substance should be evaluated. Reproducibility of the response is an important consideration when the individual responses do not reach the magnitude needed to judge the individual test positive, e.g., Chemical A in Example 3, below. Weak or inconclusive responses that are not replicated in a repeat test using the same or an alternate protocol may be considered negative.

There was a consensus that when repeat testing with appropriate protocol alterations does not resolve a weak or inconclusive result into a clear positive or clear negative result, the result can be called "equivocal" This term was developed for other OECD TGs for this situation [4] and indicates that the call is neither positive nor negative and further testing using the Ames test is unlikely to provide a more definitive call. Regulatory recommendations on how to address this situation vary and include use of computational methods and mammalian cell or in vivo tests sensitive to the assessment of gene mutation endpoints.

4. Evaluation of test results: examples using test data

4.1. Example 1: use of multiple criteria resolves inconclusive results among 74 aromatic amines

The workgroup looked at unpublished data on 74 aromatic amines in an experiment designed to evaluate the impact of using various combinations of the three criteria: concentration-related response (dose-response), comparison with the current control (fold-increase), and exceed HCD to determine test outcome [22]. All 74 compounds were tested in a single lab using 5 *Salmonella* strains with and without metabolic activation and a commercially prepared rat liver S9 mixture. Mutagenic responses were examined using each of three criteria. Comparison with the concurrent control used a criterion of 2- or 3-fold, depending on the strain.

The HCD was based on 95% tolerance intervals (defined in Section 2.1.2) from at least 200 experiments for each strain. Testing the 74 chemicals using 5 strains with 2 metabolic conditions resulted in a total

a: Concentration-related and fold-rule criteria

CR Criterion	Fold Rule Criterion	Initial Call
yes	yes	+
yes	no	inconclusive
no	yes	inconclusive
no	no	-

b: Concentration-related and historical control distribution criteria

CR Criterion	> HCD Criterion	Initial Call
yes	yes	+
yes	no	inconclusive
no	yes	inconclusive
no	no	-

c: Combination of all three criteria

CR Criterion	Fold Rule Criterion	> HCD Criterion	Initial Call
yes	yes	yes	+
yes	yes	no	likely +
yes	no	yes	likely +
yes	no	no	inconclusive
no	yes	yes	inconclusive
no	yes	no	inconclusive
no	no	yes	likely -
no	no	no	-

Fig. 2. Permutations of two or three evaluation criteria. Initial calls made for the 74 aromatic amines based on each possible combination of the criteria in the top row of the table applied to each set of plates exposed to varying concentrations in a single strain tested either with or without metabolic activation. "CR Criterion" refers to a concentration-related increase. "Fold rule Criterion" refers to the increase compared to the concurrent control using the strain-specific fold increase criterion. Another comparison metric (e.g., a statistical test) could be substituted here. "> HCD" refers to an increase above the laboratory's criterion for the historical control distribution. Fig. 2a and b illustrate use of two criteria. Fig. 2c illustrates the initial call based on a weighted combination of all three criteria (see text).

of 740 individual tests. Each test was then scored using each of the three criteria. Fig. 2a illustrates each of the 8 possible permutations of positive and negative results arising from combining a concentration related increase (CR Criterion) with an increase determined using a fold rule (Fold rule Criterion). Similarly, Fig. 2b illustrates the 8 permutations resulting from combining the CR Criterion with an increase outside of the historical control distribution (HCD Criterion). The initial call for the test was made by combining the two criteria into a positive or negative result, or inconclusive when one but not both criteria were met. Fig. 2c illustrates 24 permutations from use of all three criteria. The three criteria were weighted, giving greatest weight to whether an increase showed a concentration-related trend and least weight to the comparison with the HCD. Based on that weighting, additional "initial calls" of "likely positive" or "likely negative" were included. In some cases, uncertainty about whether the initial experimental result met one of the criteria led to a follow-up experiment which resolved an unclear initial call (likely + or likely –) into a more firm conclusion of positive, negative or truly equivocal.

Fig. 3 summarizes the result of applying the criteria in Fig. 2 to the data for the 74 aromatic amines. Results for each of the 74 aromatic amines are presented in Supplementary Table 1. Depending on how the criteria were combined, between 25 and 28 of the aromatic amines were positive in at least one strain (Supplementary Table 1). Either 9 or 14 of the 740 individual tests were inconclusive due to conflicting calls when applying two of the criteria (Fig. 3) but only 2 of 740 tests were inconclusive when all three criterial were considered. In many cases of inconclusive calls, the test was positive for the same chemical in another strain so that in the end the number of inconclusive results was limited to 3 chemicals when the CR criterion was combined with the fold rule criterion and the same number of tests were inconclusive when the CR criterion was combined with the HCD Criterion. However, when all three criteria were applied there were no chemicals with inconclusive results. As illustrated in Fig. 3, the use of all three criteria allows more flexibility in interpretation of the outcome of tests that are not clearly positive or negative. Initial calls of "positive" or "negative" using only two criteria (CR Criterion plus Fold rule Criterion or CR Criterion plus HCD Criterion) more often resulted in an unclear initial call (likely + or likely –) which would result in conducting a follow-up experiment. Evaluation of Ames test data using the combination of all three criteria will be more reliable and robust.

4.2. Example 2: a reproducible concentrated-related increase smaller than the fold-rule criterion

Fig. 4 shows data from an experiment employing TA100. In each of three experiments (a preliminary toxicity screen two confirmatory assays) a two-fold increase above the concurrent control would result in plate counts of 238–245 revertants, depending on the experiment (not shown and off the scale for this plot).

Most of the solvent controls were higher than the mean of the HCD, but not markedly higher. None of the concentration-related responses reached 2-fold over the concurrent solvent control, although in each of three experiments every plate receiving 5000 μg test substance was beyond the upper control limit defined by three standard deviations above the HCD derived from >100 previous experiments in the laboratory. Plate counts from the plates receiving 2500 μg were all above or near the same control limit, providing additional evidence of a concentration related increase. The reproducibility of the test results from the 3 experiments when taken together indicate a positive response for the test substance studied.

4.3. Example 3: small variations can change test results when arbitrary cutoff values are used

Using an arbitrary cut-off on a continuous function can lead to arbitrary test calls of positive or negative. This is illustrated in Table 3 for the fold rule but applies equally to statistical tests when, for example, the cut-off is a p value of 0.05 and the data result in values slightly above or below that cutoff. Although chemical A would be judged nonmutagenic and chemical B mutagenic using a strict interpretation of the 2-fold rule, the plate-to-plate variability at 5000 μg in chemical B, and the difference between the means of 197 ± 9 and 202 ± 20 at 5000 μg/plate is within counting error, indicating that the two chemicals are providing the same response. The same concerns would also be present if the tester strain had a low spontaneous background frequency and the positive response required a 3-fold increase.

In a case like this where Chemical A did not quite meet the 2-fold criterion, the result would be considered together with the obvious dose-related increase and possibly also comparison with the historical control distribution. Even if the test call was not "clearly positive", the test could be judged positive without need for confirmatory testing.

5. Summary

- The workgroup did not recommend a preferred test evaluation procedure such as fold increase or statistical significance but, instead, recommended that a combination of approaches such as a fold-increase or statistics, be used in combination with historical control values and expert judgment.
- There was consensus on the need for each laboratory to develop and

Test condition		Individual test responses and resulting overall call																	
		Overall Call Negative				Inconclusive (2 Criteria)			Inconclusive (2 Criteria)			Inconclusive (3 Criteria)				Overall Call Positive			
Strain	Metabolic Condition	CR	Fld	HCD	N	CR	Fld	N	CR	HCD	N	CR	Fld	HCD	N	CR	Fld	HCD	N
TA98	-S9	-	-	-	69			0			0				0	+	+	+	5
		-	-	+	1														
	+S9	-	-	-	56	-	+	1	-	+	2	-	+	+	1	+	+	+	16
TA1537	-S9	-	-	-	70			0	+	-	1				0	+	+	-	1
																+	+	+	3
		-	-	+	1				-	+	1								
	+S9	-	-	-	65	+	-	1	+	-	1	+	-	-	1	+	+	+	7
TA100	-S9	-	-	-	68	+	-	1			0				0	+	-	+	1
																+	+	+	5
		-	-	+	3											+	-	+	2
	+S9	-	-	-	60	+	-	2	-	+	3				0	+	+	+	9
TA1535	-S9	-	-	-	69	+	-	2			0				0	+	-	+	2
																+	+	+	3
	+S9	-	-	-	67	+	-	1			0				0	+	-	+	1
																+	+	+	6
TA102	-S9	-	-	+	1														
		-	-	-	67			0	-	+	1				0	+	+	+	6
	+S9	-	-	+	5											+	-	+	1
		-	-	-	63	+	-	1	-	+	5				0	+	+	+	5
Total per test condition					665			9			14				2				73
A compound conclusion is based on consolidation of the overall calls for each strain																			
Compound Conclusion					46			3			3				0				28

Fig. 3. Impact of combining 2 or 3 criteria on test outcome for 74 aromatic amines. Response for each test of each of 74 aromatic amines in 5 strains with or without metabolic activation [22]. This is a summary of data presented in Supplementary Table 1. Individual column headings indicate the call based on a single criterion: either concentration-related response (CR), 2- or 3-fold increase (depending on strain) above concurrent control (Fld), or above historical control distribution (HCD). Positive (+) or negative (−) calls are indicated in the white boxes. Criteria were combined using the "rules" (initial calls) in Fig. 2.
In both tables, green indicates a negative call, red a positive call and yellow inconclusive. The numbers (N) in the colored boxes to the right of the white boxes denote the number of individual experiments with the pattern illustrated in the boxes for that particular strain and metabolic activation status. The numbers were adjusted based on follow-up experiments when the initial experiment did not present a clear result for one or more of the individual criteria. The number in the row "Total per test condition" indicates the sum of individual test results in the column above the total. The final row, "Compound conclusion," consolidates the call for each of the 74 compounds by eliminating negative, inconclusive, and multiple positive calls for compounds judged to be positive in at least one strain under one metabolic condition.

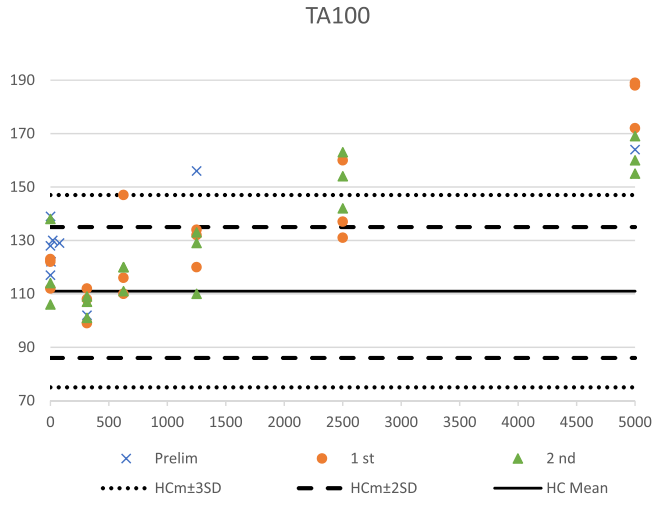

Fig. 4. Test data from the files of one of the authors. Each point represents the single plate count (vertical axis) at an applied concentration (μg/plate, horizontal axis) with strain TA100 without metabolic activation. Data from three experiments are shown: a preliminary range finding tests (blue crosses, one plate per concentration) and two confirmatory tests (orange circles and green triangles, respectively, three plates per concentration). The horizontal lines illustrate the historical control distribution from >100 tests: the solid line is the mean (HC mean), dashed lines represent values two standard deviations above and below the mean (HCm ±2SD), the dotted line represent values three standard deviations above and below the mean (HCm ±3SD).

Table 3
Comparison of two chemicals judged by a strict 2-fold rule.

Chemical A		Chemical B	
μg/plate	revertants/plate	μg/plate	revertants/plate
0	100 ± 7	0	100 ± 7
10.	122 ± 7	10.	122 ± 7
67.	149 ± 9	67.	149 ± 9
100.	178 ± 12	100.	178 ± 12
667.	190 ± 8	667.	190 ± 8
5000.	197 ± 9	5000.	202 ± 20
increase	*1.97x = nonmutagen*	*increase*	*2.02x = mutagen*

maintain historical positive and solvent control records for each tester strain with and without metabolic activation (S9). Universal criteria cannot be developed for how to use these records due to expected variations among laboratories. Each laboratory must develop rules and criteria based on its own data.

- Control charts should be monitored to determine whether intentional changes (e.g., new lot or supplier of a reagent or agar or use of a new solvent) or unintentional changes (e.g. S9 or reagent degradation, laboratory cell banks) are affecting solvent control counts or the ability of the bacteria to respond to mutagens.
- The relationship of the test substance response to the concurrent solvent control should generally be given more weight than comparisons against the HCD when the two comparisons result in different outcomes. In addition, the greatest weight in making the overall call should be based on a determination of whether any

observed increase is concentration-related (and reproducible if a repeat experiment is conducted).

- When evaluating the data from an individual test, the dynamics of the test system needs to be taken into consideration, including the plate-to-plate variability and the relationship of the response to the historical control values for the particular tester strain/S9 combination. A test can be described as negative if a modest increase is not reproducible and was consistent with variability seen in concurrent and historical negative controls.
- A positive response in any one strain/activation combination is sufficient to classify the test substance as a mutagen. Responses that are clearly positive do not have to be repeated.
- Tests yielding increased responses that are judged not clearly positive or negative generally need to be repeated. Modified protocols designed to examine the increased response in greater detail are often appropriate.
- When repeat testing with appropriate protocol alterations does not resolve an inconclusive result into a clear positive or clear negative result, the result can be called “equivocal”.
- For a classification of nonmutagenicity, the substance must be judged negative in all strains tested (i.e., all TG471- recommended tester strains) with and without metabolic activation. There was no consensus on the need for a confirmatory test when the initial test is clearly negative.

Disclaimer

The recommendations described in this work are those of the individual authors and not intended to represent the policies of the companies or agencies at which they are employed.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgement

Funding for Open Access to this publication was provided by the Bacterial Mutagenicity Study group of the Japanese Environmental Mutagen Society (JEMS/BMS).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.mrgentox.2019.07.004.

References

[1] OECD, Guideline for the Testing of Chemicals: Bacterial Reverse Mutation Test No. 471 OECD Environment, Health and Safety Publications Series on Testing and Assessment Organization for Economic Cooperation and Development, Paris, 1997.
[2] ICH, Guidance on genotoxicity testing and data interpretation for pharmaceuticals intended for human use S2(R1), International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, (2011).
[3] OECD, Guideline for the Testing of Chemicals: Bacterial Reverse Mutation Test No. 471 (Archived), OECD Environment, Health and Safety Publications Series on Testing and Assessment Organization for Economic Cooperation and Development, Paris, 1983.
[4] OECD, Overview of the Set of OECD Genetic Toxicology Test Guidelines and Updates Performed in 2014–2015, Series on Testing & Assessment, Organization for Economic Cooperation and Development, Paris, 2016.
[5] V. Thybaud, E. Lorge, D.D. Levy, J. van Benthem, G.R. Douglas, F. Marchetti, M.M. Moore, R. Schoeny, Main issues addressed in the 2014–2015 revisions to the OECD Genetic Toxicology Test Guidelines, Environ. Mol. Mutagen. 58 (2017) 284–295.
[6] D.D. Levy, A. Hakura, R.K. Elespuru, P.A. Escobar, M. Kato, J. Lott, M.M. Moore, K. Sugiyama, Demonstrating laboratory proficiency in bacterial mutagenicity assays for regulatory, Mutat. Res. (2019) .
[7] M. Kato, K.I. Sugiyama, T. Fukushima, Y. Miura, T. Awogi, S. Hikosaka, K. Kawakami, M. Nakajima, M. Nakamura, H. Sui, K. Watanabe, A. Hakura, Negative and positive control ranges in the bacterial reverse mutation test: JEMS/BMS collaborative study, Genes Environ. 40 (2018) 7.
[8] B.A. Carnes, S.S. Dornfeld, M.J. Peak, A quantitative comparision of a percentile rule with a 2-fold rule for assessing mutagenicity in the Ames assay, Mutat. Res. 147 (1985) 15–21.
[9] OECD, Principles on Good Laboratory Practice, OECD Series on Principles of Good Laboratory Practice and Compliance Monitoring, Paris, 1997.
[10] Japan Industrial Safety and Health Association, Good Laboratory Practice in Industrial Safety and Health Law (in Japanese), (1991).
[11] U.S. Food and Drug Administration, Good Laboratory Practice for Nonclinical Laboratory Studies, 21 U.S. Code of Federal Regulations, Part 58, (2002).
[12] U.S. National Institute of Standards and Technology, Process or product monitoring and control, Chapter 6, Engineering Statistics Handbook, (2013).
[13] Nordic Committee on Food Analysis (NMKL), Control Charts and Control Materials in Internal Quality Control in Food Chemical Laboratories, (2016) www.nmkl.org/dokumenter/prosedyrer/en/NMKLProc3_2016Eng.pdf.
[14] L.P. van Reeuwijk, V.J.G. Houba, Internal quality control of data, Chapter 8, Guidelines for Quality Management in Soil and Plant Laboratories. (FAO Soils Bulletin - 74), United Nations Food and Agriculture Organization, 1998, http://www.fao.org/docrep/w7295e/w7295e0a.htm#8.3%20control%20charts.
[15] B.S. Kim, B.H. Margolin, Statistical methods for the Ames Salmonella assay: a review, Mutat. Res. 436 (1999) 113–122.
[16] U.S. Food and Drug Administration, Redbook 2000: IV.C.1.a Bacterial Reverse Mutation Test, Toxicological Principles for the Safety Assessment of Food Ingredients, 2000.
[17] B.N. Ames, J. McCann, E. Yamasaki, Methods for detecting carcinogens and mutagens with the Salmonella/mammalian-microsome mutagenicity test, Mutat. Res. 31 (1975) 347–364.
[18] V.C. Dunkel, E. Zeiger, D. Brusick, E. McCoy, D. McGregor, K. Mortelmans, H.S. Rosenkranz, V.F. Simmon, Reproducibility of microbial mutagenicity assays: I. Tests with Salmonella typhimurium and Escherichia coli using a standardized protocol, Environ. Mutagen. 6 (Suppl. 2) (1984) 1–251.
[19] R. Williams, D.M. DeMarini, L.F. Stankowski, P.A. Escobar, E. Zeiger, J. Howe, R. Elespuru, K.P. Cross, Are all bacterial strains required by OECD mutagenicity test guideline TG471 needed? Mutat. Res. (2019) in press.
[20] N.F. Cariello, W.W. Piegorsch, The Ames test: the two-fold rule revisited, Mutat. Res. 369 (1996) 23–31.
[21] C. Hamada, T. Wada, Y. Sakamoto, Statistical characterization of negative control data in the Ames Salmonella/microsome test, Environ. Health Perspect. 102 (Suppl. 1) (1994) 115–119.
[22] B. van der Leede, A. Schuermans, J. Van Gompel, Are we applying the correct evaluation criteria in the Ames test: an investigation with test results of aromatic amines, UKEMS Annual Conference, (2017).
[23] D. Gatehouse, S. Haworth, T. Cebula, E. Gocke, L. Kier, T. Matsushima, C. Melcion, T. Nohmi, T. Ohta, S. Venitt, et al., Recommendations for the performance of bacterial mutation assays, Mutat. Res. 312 (1994) 217–233.
[24] M.J. Prival, Anomalous mutagenicity profile of cyclohexanone oxime in bacteria: cell survival in background lawns, Mutat. Res. 497 (2001) 1–9.
[25] L. Bernstein, J. Kaldor, J. McCann, M.C. Pike, An empirical approach to the statistical analysis of mutagenesis data from the Salmonella test, Mutat. Res. 97 (1982) 267–281.
[26] T. Jaki, L.A. Hothorn, Statistical evaluation of toxicological assays: Dunnett or Williams test-take both, Arch. Toxicol. 87 (2013) 1901–1910.
[27] I.D. Mitchell, R.W. Rees, P.J. Gilbert, J.B. Carlton, The use of historical data for identifying biologically unimportant but statistically significant results in genotoxicity assays, Mutagenesis 5 (1990) 159–164.
[28] S. Albertini, E. Gocke, Phenobarbital: does the positive result in TA1535 indicate genotoxic properties? Environ. Mol. Mutagen. 19 (1992) 161–166.
[29] K. Mortelmans, E. Zeiger, The Ames Salmonella/microsome mutagenicity assay, Mutat. Res. 455 (2000) 29–60.
[30] D.M. Maron, B.N. Ames, Revised methods for the Salmonella mutagenicity test, Mutat. Res. 113 (1983) 173–215.
[31] L.D. Kier, D.J. Brusick, A.E. Auletta, E.S. Von Halle, M.M. Brown, V.F. Simmon, V. Dunkel, J. McCann, K. Mortelmans, et al., The Salmonella typhimurium/mammalian microsomal assay. A report of the U.S. Environmental Protection Agency gene-tox program, Mutat. Res. 168 (1986) 69–240.
[32] D.E. Levin, M. Hollstein, M.F. Christman, E.A. Schwiers, B.N. Ames, A new Salmonella tester strain (TA102) with A X T base pairs at the site of mutation detects oxidative mutagens, Proc. Natl. Acad. Sci. U. S. A. 79 (1982) 7445–7449.
[33] P. Wilcox, A. Naidoo, D.J. Wedd, D.G. Gatehouse, Comparison of Salmonella typhimurium TA102 with Escherichia coli WP2 tester strains, Mutagenesis 5 (1990) 285–291.
[34] K. Watanabe, K. Sakamoto, T. Sasaki, Collaborative study of interlaboratory variability in Salmonella typhimurium TA102 and TA2638 and Escherichia coli WP2/pKM101 and WP2 uvrA/pKM101. Association of Microbial Mutation Testing Laboratory C.O, Mutagenesis 10 (1995) 235–241.
[35] S. Venitt, C. Crofton-Sleigh, R. Forster, Bacterial mutation assays using reverse mutation, in: S. Venitt, J.M. Parry (Eds.), Mutagenicity Testing: A Practical Approach, IRL Press, Oxford, 1984, pp. 45–97.