DRAFT SUMMARY REPORT OF THE PEER REVIEW PANEL ASSESSING THE JACVAM INITIATIVE INTERNATIONAL VALIDATION STUDIES OF THE IN VIVO RODENT ALKALINE COMET ASSAY FOR THE DETECTION OF GENOTOXIC CARCINOGENS

FORWORD

This document presents the draft peer review report for the JaCVAM initiative international validation studies of the *in vivo* rodent alkaline comet assay for the detection of genotoxic carcinogens. It also includes the draft WNT statement on the follow-up to the peer review.

In April 2013, the Working Group of the National Coordinators for the Test Guidelines Programme (WNT) endorsed the peer review report and WNT statement, as well as the pre-validation and validation reports.

The present document will be declassified and published in the Series on Testing and Assessment, as document N° 195 together with the pre-validation report (N° 193) and the validation report (N° 194) when an addendum to the validation report has been finalized. This addendum will include further analyses and surveys conducted by the expert group to address the peer review comments and support expert decisions taken during the development of the draft Test Guideline.

Agreement of the Working Group of the National Coordinators of the Test Guidelines Program on the Follow-up to the Validation Peer Review Report

The Validation Peer Review Report of the JACVAM initiative international validation studies of the *in vivo* rodent alkaline comet assay for the detection of genotoxic carcinogens was submitted for endorsement to the Working Group of National Coordinators of the Test Guidelines Program (WNT) at its April 2013 meeting.

Considering the major recommendations of the Peer Review Panel (summarized below), i.e.:

• Use of additional data including data from the literature in order to:

- address inter-laboratory reproducibility (paragraphs 26-30),

- broaden the applicability domain of the assay to classes of chemicals not included or not sufficiently represented in the validation exercise (paragraphs 31-32),

- broaden the scope of the TG to other species, gender, tissues (paragraphs 22-23).

- better describe the mechanism underpinning the assay, in particular the link between DNA migration observed in the assay and DNA damage, (paragraph 18)

- assess the advantage of the comet assay over the UDS assay (paragraph 15),

• Analyse further some of the data from the validation study, to see if using the mean or the median for all the data can reduce the variations observed between laboratories and thus improve quantitative inter-laboratory reproducibility (paragraph 25).

• Provide clearer guidance on methods of measurement of toxicity in the tissue being examined and on interpretation of test results when toxicity is found (paragraph 35).

• Revise the regulatory purpose of the test, as it should not be used as the sole predictor for carcinogenicity (paragraphs 14 and 36-37)

• Recognise the limitations of the assay:

- some organs (such as the stomach) may lead to relatively high background of DNA fragmentation and high variability (paragraphs 20-21)

- the assay has a low capability to reveal some type of damage without protocol adaptations (paragraph 20)

• Consider the need to develop some recommendations on how laboratory proficiency for tissues other than those tested in the validation study should be demonstrated.

the WNT agreed that, before finalising the development of the draft Test Guideline for the in vivo Comet assay the expert group on the comet assay should address the above recommendations as appropriate.

SUMMARY REPORT OF THE PEER REVIEW PANEL
ASSESSING THE
JACVAM INITIATIVE INTERNATIONAL VALIDATION STUDIES OF THE IN VIVO RODENT ALKALINE COMET ASSAY FOR THE DETECTION OF GENOTOXIC CARCINOGENS

PREAMBLE

This document presents the summary report of the assessment of the JaCVAM initiative international pre-validation studies of the *in vivo* rodent alkaline comet assay for the detection of genotoxic carcinogens, performed by a Panel of experts from the OECD expert group for the development of the in vivo Comet assay. The OECD expert group for the development of the in vivo Comet assay was established in September 2012 and made of experts nominated by the Working Group of the National Coordinators for the Test Guidelines Programme (WNT). Only experts who had not participated in the JaCVAM validation exercise were included in the peer review process.

The alkaline single cell gel electrophoresis or alkaline Comet assay is a simple method for measuring DNA strand breaks in eukaryotic cells. Cells embedded in agarose on a microscope slide are lysed with detergent and high salt to form nucleoids containing supercoiled loops of DNA linked to the nuclear matrix. Electrophoresis at high pH results in structures resembling comets, observed by fluorescence microscopy; the intensity of the comet tail relative to the head reflects the number of DNA breaks and the size of the resulting fragments or loops.

OECD Test Guidelines (TGs) are available for a wide range of *in vitro* genotoxicity assays that are able to detect DNA damage, gene mutations and/or chromosomal aberrations. There are TGs for *in vivo* endpoints (i.e. chromosomal aberrations, gene mutations and DNA repair as unscheduled DNA synthesis); however, these do not directly measure DNA damage. The alkaline single cell gel electrophoresis or alkaline Comet assay is presented as a practical and widely available *in vivo* test for measurement of DNA strand-breaks induction in tissues.

Origin of DNA strand breaks and alkali-labile sites in the alkaline single cell gel electrophoresis or alkaline Comet assay

Modifications of DNA structures at the origin of Comet/Tail formation/appearance can occur at different levels. Depending on the electrophoretic conditions, different mechanisms are observed.

For instance, to show the importance of pH during migration, Miyamae et al (1997) evaluated the effect of 13 products with different modes of action in the Comet assay under two different alkaline conditions (pH 12.1 and 12.6). Their results show that bleomycine induced positive results at both pH 12.1 and 12.6, demonstrating that this compound induces DNA strand breaks and not alkali-labile sites. In return, alkylating agents induce a dose-dependent response only at pH 12.6. These results confirm that at pH 12.1, only the single-stranded breaks in DNA are detected while at pH 12.6, alkali-labile sites are also highlighted.

In alkaline conditions, Collins et al. (1997) outlined the concept of relaxation of supercoiled DNA, rather than alkaline unwinding, as the primary reason for comet tail formation. Indeed, the disruption of DNA loops during the unwinding step lead to a higher mobility of this DNA making it able to partially migrate towards the anode during the electrophoresis (Collins et al. 1997; McKelvey-Martin et al. 1993).

Based on the treatment of microgels to remove proteins, Singh and Stephens (1997) showed that DNA observed in the Comet assay is associated to proteins even after electrophoresis and that the presence of DNA single-strand breaks in alkaline conditions may depend on these associations. The authors described DNA behaviour during alkaline and neutral microgel electrophoresis based on observations of the stained DNA and its migration patterns. During microgel electrophoresis, individual DNA molecules behave as if anchored at one end while the other end is free to migrate in response to the electric field.

Klaude et al. (1996) studied the behaviour of DNA under different electrophoresis conditions in mammalian cells exposed to gamma radiation. The comet tails obtained after neutral electrophoresis consist in DNA loops which are still strongly linked to the nucleus, i.e. the head of the comet. In return,

under alkaline electrophoresis conditions, 2 types of DNA fragments migrate, i.e. fragments from DNA single strand breaks that migrate relatively rapidly and double-stranded fragments of DNA whose migration from the head of the comet seems more difficult (Klaude et al. 1996).

Thus, it appears that the alkaline comet tails consist of free DNA fragments.

Summary Report of the Peer Review Panel assessing the JaCVAM initiative international validation studies of the in vivo rodent alkaline comet assay for the detection of genotoxic carcinogens

The peer review process

1. The OECD expert group for the development of the *in vivo* Comet assay was established in September 2012 and made of experts nominated by the Working Group of the National Coordinators for the Test Guidelines Programme (WNT). Only experts who had not participated in the JaCVAM validation exercise were included in the peer review process. The members of the Panel are listed in Annex 1. The work of the Panel was coordinated by the OECD Secretariat, with the support of an expert of the group (co-manager).

2. The Panel was asked to evaluate the data collected on the test method, and to assess to what extent the eight OECD validation criteria set out in the OECD Guidance Document on the Validation and International Acceptance of New or Updated Methods for Hazard Assessment (GD 34) had been met. The general questions to the Panel are included in Annex 2. Panel members were asked to base their review on two documents:
- Report of the JaCVAM initiative international pre-validation studies of the *in vivo* rodent alkaline Comet assay for the detection of genotoxic carcinogens, ver.1.4, January 14, 2013, and
- Report of the JaCVAM initiative international validation studies of the *in vivo* rodent alkaline Comet assay for the detection of genotoxic carcinogens, ver.1.4, January 14, 2013
These documents were available on clearspace (https://community.oecd.org/community/tgeg) and on the public website:
(http://www.oecd.org/env/ehs/testing/peerreviewsofecotoxicityandhumanhealthtestmethods.htm).

3. As background information, they were also provided with a first draft of the Test Guideline on the *in vivo* Comet assay, a supplementary document on additional comments on intra- and inter-laboratory reproducibility of the in vivo comet assay, developed while the validation report was being finalised by a subgroup of experts who participated in the validation process, and the following references from the literature: Bowena et al., 2011; Rothfuss at al., 2010. The link to the OECD GD 34 in the Series on Testing and Assessment was also provided (last access on February 14, 2013):
(http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono(2005)14&doclanguage=en)http://www.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono(2005)14&doclanguage=en).

4. In addition to the general charge questions (corresponding to the 8 criteria of GD 34), the experts were asked to respond to specific charge questions that were developed by a consultant for the Secretariat in September 2012. The specific charge questions to the Panel are included in Annex 2. The review work to address these specific charge questions was shared among experts. Five subgroups of 3 to 5 experts were made, each of them led by an expert of the subgroup (see Annex 1). Experts were allowed to choose their area of review and submit individual responses to general charge question and/or contribute to one or several subgroups.

5. A summary of the Panel's responses to the individual questions, provided by the peer-review co-managers, is presented in paragraphs 7 to 44. For transparency, the individual comments from the Panel members are provided anonymously and in an edited form in Annex 3a. Responses from the subgroups to the specific charge questions are presented in Annex 3b.

6. During the review process, the Panel held three teleconferences (January 17, 2013, February 21 and February 28, 2013) which were organised and coordinated by the Secretariat. The Panel members provided written responses to the charge questions to the Secretariat by February 8, 2013. Based on these responses,

a draft report taking into account all individual comments was compiled by the peer review co-managers and provided to the Panel for review and comments (February 18, 2013). The Panel discussed the draft report on February 21. Accounting for this feedback and resolving remaining open issues, a revised report was prepared by the Secretariat, sent to the PRP on February 25 and discussed on February 28. The final report was approved on March 1. This report presents the resulting approved responses of the Panel to each of the charge questions, referring, as necessary, to the responses to the specific charge questions.

General Panel responses

7. The Panel agreed that the data in the validation trial is sufficient to conclude that the in vivo Comet protocol was improved throughout the validation exercise and that its development towards a Test Guideline should continue. However, some elements require additional discussion and/or collection of more data, before moving forward.

8. The main concerns raised by the Panel regarding the quality of the validation exercise were the following:

- The lack of demonstration of intra-laboratory and inter-laboratory reproducibility,
- The need to clarify several study parameters, including criteria for dose selection, use of median or mean, use of estimate or effect for data interpretation and statistical parameters; all those criteria having a potentially high impact on the results, interpretation and/ or on the conclusion; some of these parameters could also have an impact on how to build and use the historical data,
- The limited number of tissues used,
- The limitation of the validation to male rats,
- In addition it was noted that the added value of the comet assay over the UDS assay needs to be confirmed.

9. The Panel agreed that the validation work has demonstrated that:
- the use of the stomach should not be specifically encouraged in a TG, as it is subject to high variations and
- the design used in the validation does not allow for detection of cross linking agents.
These limitations need to be mentioned in the TG.

10. In addition the Panel considered that this assay detects, in a given tissue, many but not all types of in vivo DNA damage that could potentially result in stable mutations and ultimately cancer or other diseases. Thus, it should not be used as the sole predictor for carcinogenicity, but instead is valuable as part of a battery of tests. The Panel recommends that the regulatory purpose be revised accordingly.

11. The scope of the validation study was limited in terms of tissues analysed, species and gender. As the validation exercise cannot support by itself the broadening of the scope of the Test Guideline, the Panel agreed that there is a need to go to the data from the literature and to check if the published data can support recommendations in the TG to use rodent species other than rats and only one gender (males).

12. Overall the Panel agreed that the validation criteria have been met or partially met and that the information that is missing could be requested from the VMT, collected from the literature, or gained from laboratories that have a long history of using this assay. This additional information may help:
- addressing inter/intra laboratory reproducibility including control levels,
- checking if using the mean or the median for all the data would be helpful to reduce the variations observed between laboratories and thus improve quantitative inter laboratory reproducibility,
- broadening the applicability domain of the assay to classes of chemicals not included or not sufficiently represented in the validation exercise,
- broadening the scope of the TG to other tissues as well as to mice and female animals,
- describing the mechanism underpinning the assay, in particular the link between DNA migration observed in the assay and DNA damage,
- assessing specifically the advantage of the comet assay over the UDS assay.

Panel responses to the charge questions: The eight OECD principles and criteria for test method validation

The Panel reached consensus on all eight charge questions. It was acknowledged that due to some overlap between questions, the report might include redundancies.

Charge question 1: A rationale for the test method should be available, including a clear statement of scientific basis and the regulatory purpose and need for the test method

The Panel agreed that this criterion had been partially met.

13. The validation report describes the regulatory purpose of the test as follows:
- evaluate the ability of the *in vivo* Comet assay to identify genotoxic chemicals as a potential predictor of rodent carcinogenicity, and
- consider the value of the *in vivo* Comet assay as an alternative follow-up assay to the more commonly used *in vivo* rodent Unscheduled DNA Synthesis (UDS) assay.

14. The Panel agreed that the regulatory purpose of the test was clearly described. However the Panel had a different point of view on what this assay detects. The panel considered that this assay detects, in a given tissue, many but not all types of in vivo DNA damage that could potentially result in stable mutations and ultimately cancer or other diseases and that it should not be used as the sole predictor for carcinogenicity, but instead is valuable as part of a battery of tests. The Panel recommends that the regulatory purpose be revised accordingly.

15. Regarding the capacity of the assay to be used as an alternative to the UDS assay, some reviewers noted that the additional value of the comet assay over the UDS test still needs to be confirmed. It was acknowledged that not many compounds included in the validation exercise have been tested in the UDS assay. It was noted however that a more thorough comparison could be performed by collecting additional data from the literature. As an example, it was noted that it can be referred to the paper from Kirkland and Speit (2008) comparing several in vivo assays.

16. The Panel also agreed that the scientific rationale for the test was not fully described in the validation report and that the mechanistic part could have been better described. In particular the report does not explain the mechanisms that would lead to the DNA migration observed in the assay (i.e., types of DNA damage) and impact the results (e.g., DNA repair, cell death).

Charge question 2: The relationship between the test method endpoint(s) and the biological effect and to the toxicity of interest should be described, addressing also limitations of the test methods

The Panel agreed that this criterion had been partially met.

17. The test method is described as an assay that measures %tail DNA as an indication of DNA strand breaks. This measure is also considered to be appropriate for the detection of rodent genotoxic carcinogens.

18. While the relationship between the %tail DNA and the DNA stand breaks is clearly explained in the validation report, the Panel agreed that the report does not detail either the mechanisms causing this effect, nor its consequences, i.e. there is no information about the mode(s) of action of comet formation or of reparability of the damage and there is no information either on how or how often DNA damage revealed

by the Comet assay may lead to genetic damages and to diseases including cancer, chronic cardiovascular, neurologic and immune diseases.

19. The assay was thus described by the reviewers as an indirect biomarker of potential disease, which is nonetheless useful because the assay is relatively quick and easy to perform on specific target organ(s).

20. Based on the results obtained in the validation study, the Panel concluded to the following limitations of the assay:

- The assay has a low capability to reveal some types of damage (e.g. oxidative damage, crosslinks, maybe bulky-adducts) not all of which are identified. When the type of damage can be predicted the comet assay can be used with protocol adaptations, which make the assay much more sensitive and provide additional mechanistic information.
- The stomach model shows some organs may lead to relatively high background of DNA fragmentation and high variability and illustrate the need of refinement of experimental conditions for each tissue.

21. Most of the reviewers questioned the added value of the stomach in this assay, arguing that (i) the stomach did not help increasing the sensitivity of the assay in the validation study and gave too much variation, (ii) in the stomach only, some chemicals were reported to decrease migration relative to the negative control, which is of unknown significance (iii) neoplasms in the rodent glandular stomach are not commonly observed and in human stomach cancers an infectious agent rather than chemical exposure seems to be the most significant risk factor and (iv) stomach might only be appropriate for gavage studies but not for other routes of exposures.
One commenter however noted that the notion of local organ should be specifically mentioned in a general manner in a TG. In this context only i.e. considered as a local organ in case of oral exposure, the choice of the stomach might be relevant. Some members of the Panel considered that other tissues, including other tissues in the gastro-intestinal tract (e.g. duodenum) might be more appropriate after oral administration because they allow less variable results.

22. The panel also noted the following limitations of the validation design:

- The validation exercise has been performed on rats only, whereas the assay is described as a "rodent" assay. Literature data show however that there are differences in the outcome of the comet assay between rats and mice for some compounds depending on species specific factors.
- The validation exercise has been performed on males only, under the assumption that no significant gender differences were expected. However, there is no discussion or literature presented in the validation report to support this decision. Because of hormonal, metabolic and physiologic differences, findings in male rats might not predict findings in female rats, nor can they predict findings in male mice.
- The validation exercise has been performed on two tissues only.

23. The Panel agreed that the scope of the TG should be broadened in terms of species, gender and tissues. The Panel noted that before another tissue is used in a Test Guideline, its use should first be justified scientifically (reason why appropriate tissue) and technically (how laboratory proficiency for another tissue should be demonstrated needs to be discussed by the expert group on the comet assay). The need for modification of experimental conditions should be considered. One reviewer noted that if the scope of the TG was broadened to use other species it would be desirable to add wording to state that the use of species other than rodents needs to be justified for animal welfare reasons (greater capacity of other species to suffer, higher demands of other species concerning housing, handling, etc.). The Panel agreed however, that it might not be necessary to perform more laboratory testing but rather go first to the extensive amount of data that has been published in the literature. Some reviewers missed the existence of a Detailed Review Paper.

Charge question 3: A detailed protocol for the test method should be available

The Panel agreed that this criterion had been partially met.

24. The protocol described in phase 4-2 of the validation report appears to properly describe how DNA strand breaks are measured in the liver and stomach. However the reviewers noted that the following issues need to be more clearly defined in the Test Guideline:
- Timing of tissue collection after the last chemical exposure and how to choose a value within the given range. Toxicokinetic data if available could help.
- Method for evaluating toxicity and definition of criteria for its assessment.
- Impact of toxicity on interpretation of the results.
- Rationale for dose selection.
- How to demonstrate laboratory proficiency.
- Acceptability criteria.
- Parameters for measuring migration (mean, median) and the statistical method to be used,
- Interpretation criteria, including need for a positive control group, how to use concurrent negative and positive control data, historical data and statistical evaluation of the results.

25. It was noted that the dose selection and the statistical parameters were the most important elements that need to be clarified as they are the most susceptible to impact the results of the validation exercise. There was no general rule in the validation exercise to use one or another parameter (mean or median) and it was not discussed which of them was better. It was even not fully clear which one was used by each laboratory. The Panel agreed that clarification should be asked to the contributors of the validation exercise as this could have an impact on the results (e.g. background values, statistical analysis, inter-laboratory variability).

Charge question 4: Within- and between-laboratory reproducibility of the test method should be demonstrated

26. The intra- and inter-laboratory reproducibility, were mostly evaluated in the pre-validation report where the same few chemicals were tested several times in each laboratory and across several laboratories. However, in the pre-validation phases, not all results were shown to be reproducible and the protocol was subsequently adapted. In phase 4 of the validation exercise, using the new protocol, each laboratory tested a single chemical only, such that only the positive and negative controls can be considered for reproducibility. The Panel agreed that this was not an adequate demonstration of qualitative reproducibility. Most of the reviewers considered however, that with the use of data from the literature, there may be enough data to conclude that the *in vivo* Comet protocol results demonstrate acceptable intra- and inter-laboratory reproducibility in qualitative terms as the data shown in the validation report are comparable to those obtained from other collaborative work.

27. However, intra- and inter-laboratory quantitative reproducibility was not satisfactory for all reviewers since not only have too few chemicals been tested but also among these chemicals large variations, especially across laboratories, were observed. Variation in the quantitative response was ~8-10 fold among participating laboratories for the positive control. The concern is that a laboratory with a strong positive control response may detect damage by some chemicals as positive, while a laboratory with a weak positive response may not if the damage causes only a weak response in the assay.

28. In the validation report not all parameters that may influence the amount of DNA damage measured are described, e.g., it is not clear if the slides were manually or fully automatically scored. Other factors that

may impact the results include the camera used (difference in camera dynamics and in resolution), light source, characteristics of the electrophoresis chamber used, etc. Given the different magnitude of effects and variability of background levels, factors that may contribute to inter laboratory variability should be better described in the validation report. Furthermore, a more thorough investigation of how such factors varied among the laboratories participating in the validation may provide a unique opportunity to learn about their impact on background measured, and variability.

29. It was also mentioned that it is not clear if the 'estimates' were calculated in the same way by the different laboratories. For the validation report, the different way of calculating estimates will probably have no influence when evaluating the effect of a treatment but may be important to determine the inter-laboratory reproducibility of the negative and positive controls. Finally, reporting of the results used for statistical evaluation was considered confusing. The VMT states that "they cannot yet recommend something for statistical methods and that the performance of several approaches for statistical test will be examined through this study" whereas later in the report it is only stated that the Dunnett's test and a linear trend test were used without rationale for using this statistical test.

30. The panel agreed that inter-laboratory reproducibly was met to a certain degree by testing one positive and one negative control in many laboratories, although important variations between laboratories were observed for the positive and negative control. The panel considered however, that intra-laboratory reproducibility and qualitative inter-laboratory reproducibility were not adequately addressed by the validation exercise. The Panel considered that the data obtained with the positive and negative controls showed relatively high inter laboratory variability but not enough data was available to evaluate the impact of this variability on the capacity of the assay to detect weak genotoxic agents. More chemicals evaluated by different laboratories should have been used to better address inter laboratory reproducibility using the last protocol (Phase 4).

Charge question 5: Demonstration of the test method's performance should be based on testing of representative, preferably coded reference chemicals

The Panel agreed that this criterion had been partially met.

31. The Panel considered the number of chemicals used was sufficient, considering that the assay under validation is an in vivo assay and that the validation cannot cover the whole range of chemicals. However, it would be useful to supplement the data from the validation report with data from the literature to broaden the applicability domain of the assay. One reviewer asked how the list of chemicals was chosen to be representative of the chemical space and some reviewers indicated that the chemical applicability domain was not clearly described.

32. The Panel agreed that for some categories of chemicals, i.e. compounds that need metabolic activation, cross linking agents or chemicals inducing bulky adducts, this assay may lead to non relevant results, unless the protocol is modified (see para 21):

- Based on results with 2-AAF, subgroup 3 noted that the possibility should be considered that the assay is not sensitive to compounds that require both phase I oxidative and phase 2 reductive metabolism and possibly other classes of genotoxic carcinogens which require certain types metabolic activation.

- The Panel agreed that the validation report shows that under the conditions selected in the validation study cross-linking cannot be reliably detected, and thus chemicals for which this is the predominant form of damage are not accurately identified. Reliable detection of cross linking would require a higher background level of DNA damage that could impact the detection of other types of damage or a separate

experiment with modifications of the study conditions (e.g. increase in duration of time electrophoresis, co-exposure with an alkylating agent). The Panel considered that the Comet assay as such should not be recommended for the detection of DNA cross linking agents, and that decreases in % Tail DNA should not be considered as a signature for cross linking agents under the guideline experimental conditions. However, a decrease of DNA migration under these conditions or other information suggesting that a chemical may be a cross linker, suggests further experimentation using more appropriate test conditions.

33. Most reviewers agreed that the performance of the *in vivo* Comet assay with the substances used was in accordance with the predicted outcomes based on available genotoxicity and carcinogenicity data (and/or when unexpected data were re-visited considering all relevant information available in literature). The Panel however agreed that as discussed in para 37 , this assay detects, in a given tissue, many but not all types of DNA damage in vivo that could potentially result in stable mutations and ultimately cancer or other diseases and it should not be used as the sole predictor for carcinogenicity, but instead is valuable as part of a battery of tests. A reviewer mentioned that the VMT mainly focused on carcinogens *versus* non-carcinogens whereas genotoxic carcinogens are known to induce different types of damage. Because this assay is considered to detect both clastogens and mutagens, it would have been important to focus on the types of damage induced and detected.

Charge question 6: The performance of test methods should have been evaluated in relation to existing relevant toxicity data as well as information from the species of concern

The Panel agreed that this criterion had been partially met.

34. Data obtained in the validation study with the in vivo Comet assay were compared by subgroup 5 to those available in the literature for the in vivo UDS test and the in vivo micronucleus test. Based on the evaluation of the chemicals used in the comet assay validation study, the three assays appear to have similar sensitivity and specificity towards genotoxins or carcinogens as defined by the VMT.

35. Many reviewers underlined the importance of using historical control data for interpretation of the results. It was noted that their use would change the conclusion for some chemicals tested in the validation exercise (e.g. acrylonitrile, sodium arsenite). Several reviewers also stressed the need to clarify how cytotoxicity may affect the Comet results and how to interpret positive Comet findings in the presence of cytotoxicity findings.

36. The Panel agreed that classification of carcinogens based on the Ames test results and in vivo rodent carcinogenicity bioassay results was inappropriate as not based on the mode of action of the chemicals. The division of the test chemicals into four different categories, based on the mechanism of action (genotoxic carcinogens, genotoxic non-carcinogens, non-genotoxic carcinogens, and non-genotoxic non-carcinogens) was questioned by the reviewers of subgroup 5. The reviewers considered that assigning a compound to one of these four categories is often complicated as also illustrated by the discussion of the results in the validation report. It was noted that massive amounts of information are required to determine the mode of action for carcinogenicity and has been done thoroughly for very few chemicals (those giving unexpected results). Classification of carcinogens as “genotoxic” or “non-genotoxic” based on the proposed criteria (i.e. result of an Ames test) is inappropriate and can cause problems when interpreting the results. This is illustrated in the validation report by the discussion developed to explain the unexpected results. The rationale for discordant results is well-presented and in general convincing. The issue is how this could be applied to unknown compounds.

37. The Panel agreed that although there is a rough correlation between the outcome of the comet assay (i.e. the detection of DNA strand-breaks induced by various mechanisms) and carcinogenicity, the assay should not be used as the sole predictor for carcinogenicity. , Instead it is valuable as part of a battery of tests and its objective should stick more closely to the known mechanisms of action detected by the comet assay.

Charge question 7: All data supporting the assessment of the validity of the test method should be available for expert review

The Panel agreed that this criterion had been met.

38. The reviewers agreed that the protocols used are publically available and the validation report properly summarizes the data. It was noted however that presentation of data in the report could have been improved and that access to raw data while not impossible could have been improved. Most of the compounds were coded and it appears that some were missing or hard to identify and some files did not use standard file formats.

39. It was noted by several reviewers that the time allowed for the review of the validation exercise had been very short and prevented a thorough evaluation of the large amount of data available.

Charge question 8: Ideally, all data supporting the validity of a test method should have been obtained in accordance with the principles of Good Laboratory Practice (GLP)

The Panel agreed that this criterion had been partly met.

40. The pre-validation and validation studies were conducted in facilities that are GLP compliant and the validation tests were generated in the spirit of GLP if not fully under GLP.

Recommendations

41. The Panel agrees that this report provides a summary of their views on the status of the validation of the in vivo rodent alkaline comet assay for the detection of genotoxic carcinogens, as detailed in the responses to the questions posed to the Panel and based on the information related to the test method validation provided to the Panel.

42. The report of the Panel, along with the documents provided in the peer review package (see paragraphs 2 and 3) should form the basis for decisions on whether the validation meets the OECD principles for validation.

43. The Panel recommends that the WNT consider this report to decide any further work to finalise the validation activity which links to the development of a new OECD Test Guideline.

44. Based on the data examined, the Panel recommends the following further work and considerations:

• Use of additional data including data from the literature in order to:

- address inter-laboratory reproducibility (paragraphs 26-30),

- broaden the applicability domain of the assay to classes of chemicals not included or not sufficiently represented in the validation exercise (paragraphs 31-32),

- broaden the scope of the TG to other species, gender, tissues (paragraphs 22-23).

- better describe the mechanism underpinning the assay, in particular the link between DNA migration observed in the assay and DNA damage, (paragraph 18)

- assess the advantage of the comet assay over the UDS assay (paragraph 15),

• Analyse further some of the data from the validation study, to see if using the mean or the median for all the data can reduce the variations observed between laboratories and thus improve quantitative inter-laboratory reproducibility (paragraph 25).

• Provide clearer guidance on methods of measurement of toxicity in the tissue being examined and on interpretation of test results when toxicity is found (paragraph 35).

• Revise the regulatory purpose of the test, as it should not be used as the sole predictor for carcinogenicity (paragraphs 14 and 36-37)

• Recognise the limitations of the assay:

- some organs (such as the stomach) may lead to relatively high background of DNA fragmentation and high variability (paragraphs 20-21)

- the assay has a low capability to reveal some type of damage without protocol adaptations (paragraph 20)

• Consider the need to develop some recommendations on how laboratory proficiency for tissues other than those tested in the validation study should be demonstrated.

References:

Andreas Rothfuss, Mike O'Donovan, Marlies De Boeck, Dominique Brault, Andreas Czich, Laura Custer, Shuichi Hamada, Ulla Plappert-Helbig, Makoto Hayashi, Jonathan Howe, Andrew R. Kraynak, Bas-jan van der Leede, Madoka Nakajima, Catherine Priestley, Veronique Thybaud, Kazuhiko Saigo, Satin Sawant, Jing Shi, Richard Storer, Melanie Struwe, Esther Vock, Sheila Galloway (2010), Collaborative study on fifteen compounds in the rat-liver Comet assay integrated into 2- and 4-week repeat-dose studies, Mutation Research 702, 40–69.

Collins AR, Dobson VL, Dusinská M, Kennedy G, Stětina R. (1997), The comet assay: what can it really tell us? Mutat Res. 375(2):183-93.

Damian E. Bowena, James H. Whitwell, Lucinda Lillford, Debbie Henderson, Darren Kidd, Sarah Mc Garry, Gareth Pearce, Carol Beevers, David J. Kirkland (2011), Evaluation of a multi-endpoint assay in rats, combining the bone-marrow micronucleus test, the Comet assay and the flow-cytometric peripheral blood micronucleus test, Mutation Research 722, 7–19.

Kirkland, D. and G. Speit (2008), Evaluation of the ability of a battery of three *in vitro* genotoxicity tests to discriminate rodent carcinogens and non-carcinogens III. Appropriate follow-up testing *in vivo*, Mutation Research 654, 114–132.

Klaude M, Eriksson S, Nygren J, Ahnström G. (1996), The comet assay: mechanisms and technical considerations, Mutat Res. 363(2):89-96.

McKelvey-Martin VJ, Green MH, Schmezer P, Pool-Zobel BL, De Méo MP, Collins A. (1993), The single cell gel electrophoresis assay (comet assay): a European review. Mutat Res. 288(1):47-63.

Miyamae Y, Iwasaki K, Kinae N, Tsuda S, Murakami M, Tanaka M, Sasaki YF (1997), Detection of DNA lesions induced by chemical mutagens using the single-cell gel electrophoresis (comet) assay. 2. Relationship between DNA migration and alkaline condition, Mutat Res. 393(1-2):107-13.

Singh NP, Stephens RE. (1997) Microgel electrophoresis: sensitivity, mechanisms, and DNA electrostretching. (Élongation), Mutat Res. 383(2):167-75.

ANNEX 1

Members of the peer review panel who submitted individual responses to the general charge questions

Panel member	Affiliation
Eugenia Cordelli	Laboratory of Toxicology, ENEA, Rome, Italy
Abby Jacobs	Center for Food Safety and Applied Nutrition, US Food and Drug Administration.
Francesco Marchetti	Health Canada
Dan Levy	Center for Food Safety and Applied Nutrition, US Food and Drug Administration.
Birgit Mertens	Scientific Institute of Public Health, Belgium
Veronique Thybaud	Sanofi, France
Paola Villani	Laboratory of Toxicology, ENEA, Rome, Italy

Peer review co- managers: Nathalie Delrue (OECD Secretariat) and Jan van Bethem (RIVM, Netherlands)

Members of the peer review subgroups in charge of developing responses to the specific questions

Peer review Subgroup	Panel member	Affiliation
Sub group PR 1	Francesco Marchetti, Eugenia Cordelli, Maria Donner, Birgit Mertens, Paola Villani	Health Canada ENEA, Rome, Italy DuPont (BIAC) Scientific Institute of Public Health, Belgium ENEA, Rome, Italy
Sub group PR 2	Dan Levy, Abby Jacobs, Véronique Thybaud	US FDA US FDA Sanofi, France
Sub group PR 3	Stefan Pfuhler, Maria Donner, Fabrice Nesslany	Procter & Gamble (BIAC) DuPont (BIAC) Institut Pasteur de Lille, France
Sub group PR 4	Fabrice Nesslany, Maria Donner, Stefan Pfuhler, Véronique Thybaud, Anoop Kumar Sharma	Institut Pasteur de Lille, France DuPont (BIAC) Procter & Gamble (BIAC) Sanofi, France Technical University of Denmark, National Food Institute
Sub group PR 5	Birgit Mertens, Dan Levy, Fabrice Nesslany, Véronique Thybaud	Scientific Institute of Public Health, Belgium US FDA Institut Pasteur de Lille, France Sanofi, France

ANNEX 2 :

Charge questions

PRP General charge questions:

The eight OECD principles and criteria for test method validation

Charge question 1: A rationale for the test method should be available, including a clear statement of scientific basis and the regulatory purpose and need for the test method

Charge question 2: The relationship between the test method endpoint(s) and the biological effect and to the toxicity of interest should be described, addressing also limitations of the test methods

Charge question 3: A detailed protocol for the test method should be available

Charge question 4: Within- and between-laboratory reproducibility of the test method should be demonstrated

Charge question 5: Demonstration of the test method's performance should be based on testing of representative, preferably coded reference chemicals

Charge question 6: The performance of test methods should have been evaluated in relation to existing relevant toxicity data as well as information from the species of concern

Charge question 7: All data supporting the assessment of the validity of the test method should be available for expert review

Charge question 8: Ideally, all data supporting the validity of a test method should have been obtained in accordance with the principles of Good Laboratory Practice (GLP)

Recommendations of the reviewers.

PRP Specific charge questions:

- **Sub-group PR 1** - Determine whether there are sufficient data from (i) JaCVAM trial, (ii) Rothfuss et al (2010) trial and (iii) in-house/CRO sources to conclude acceptable intra- and inter-laboratory reproducibility.

- **Sub-group PR 2** - Check that the applicability domain has been appropriately defined, and that the data within the JaCVAM trial are consistent with the applicability domain. This will mean looking at expected mode of action of chemicals tested and whether expected results were obtained.

- **Sub-group PR 3** - Determine whether the use of histopathology data for determining cytotoxicity is the best or only recommended approach.

o What is the purpose of measuring cytotoxicity?

o Data on "hedgehogs" have been collected – are they useful?

o Should histopathology only be determined at time of sampling for comets, or also at later times? Do we have enough data (from all sources) to be able to provide an answer, or does more experimental work need to be done?

o Can any recommendations be made regarding supplementary measures such as Capsase 3/7 activation, Annexin V staining, TUNEL staining, Halo or neutral diffusion assay?

- **Sub-group PR 4** - Although statistical analysis was the prime criterion for determining results in the JaCVAM trial, in the TG biological relevance of results may take priority. What would describe a biologically relevant positive response? How should historical control data be used in interpretation of results?

- **Sub group PR 5** - Assess the sensitivity and specificity of the Comet assay calculated based on 40 chemicals in Phase 4-2, based on knowledge of species or tissue specificity, mode of action and other genotoxicity results.

ANNEX 3a

General charge questions: Collated initial comments from the peer review panel assessing the JaCVAM initiative international pre-validation studies of the in vivo rodent alkaline comet assay for the detection of genotoxic carcinogens

ISSUES	COMMENTS AND RECOMMENDATIONS
General comments	**Reviewer #1:** It is not clear that inter and intra-laboratory reproducibility has yet been established. Once a list of expected results for clear positive controls and challenging positive and negative controls can be established, the method described in the test protocol can be considered validated to measure some kinds of DNA damage in rat liver which has moderate sensitivity and high specificity for the identification of rat carcinogens. **Reviewer #2: -** **Reviewer #3: -** **Reviewer #4:** The data in the validation trial is sufficient to conclude that the in vivo Comet protocol has improved throughout the validation exercise and that its guideline development should continue. However, there are a few details that require additional discussion, and possibly more testing, before moving forward. In particular: - The current timing of tissue collection after the last chemical exposure of 2-6 hr may be too wide and need to be reassessed. - It may be preferable to have a more stringent criteria for an acceptable result with the positive control among laboratories. - Stronger guidance on statistical methods to be used for evaluating biological significance of the effect should be given. - It may help to have specific recommendations on the minimum set of histopathological data that would be required to be collected to exclude the influence of cytotoxicity on the comet outcome. - It remains to be established whether the variability in the magnitude of the response of the positive control among laboratories may be a limitation of the ability to detect weak genotoxins in those laboratories with a small positive

	control effect. I see this variability in the magnitude of the positive control response as a critical issue that has to be addressed before moving forward with guideline development. **Reviewer #5:** Although a standardized protocol for conduct of the comet assay has been proposed, its predictive value for anything more than DNA damage has not been demonstrated. Thus I cannot conclude that the assay is validated to predict carcinogenicity findings, because there are alternative explanations for the carcinogenicity findings and the neoplasms generally are not in the same target organs. **Reviewer #6: -**
VALIDATION CRITERIA	
1. Rationale for the test Method	**Reviewer #1:** This charge question is adequately addressed by the comments below and in the comments presented by the various groups of peer reviewers. **Reviewer #2:** Not studied in detail by the reviewer. **Reviewer #3:** The validation report states that the validation look for the "ability to the in vivo Comet assay to identify genotoxic chemicals as potential predictor of rodent carcinogens", to consider "the value of the in vivo Comet assay as an alternative follow-up to the more commonly used in vivo rodent unscheduled DNA synthesis (UDS) assay", and to ultimately "establish an OECD guideline for in vivo rodent alkaline Comet assay". Moreover it was suggested that this assay could be used as follow-up of in vitro positive results and/or as complementary/second in vivo assay. To this end the protocol was modified (i.e. 3 treatments instead of 2) during the last step of the validation in order to allow the measurement of micronuclei in peripheral blood in the same animals in order to consider the 3R's recommendations. The MNT data were not reported in the validation report when generated. It would have been interesting because they would have potentially shown the complementarities of the assays. **Reviewer #4:** The validation exercise provides reasonable support for the notion that the comet assay is a valid alternative to the UDS test.

	Reviewer #5: There is a clear scientific rationale **but how the assay will be used is not completely clear.** This is an assay for DNA damage at very high doses in rats. To say more than that (e.g., prediction of carcinogenicity) is not appropriate nor is it supported by the data. How an assay will be used affects how its predictive value will be evaluated. For use in conjunction with a repeat dose tox study per ICHS2 (R1), the use is clear. It is part of a genotoxicity screen to allow persons to receive a drug. For use of the comet as a rough screen to predict carcinogenicity of chemicals in rodents, it would seem that in conjunction with an Ames test, it is proposed to supplement an Ames test. **Reviewer #6:** Partially met. The purpose and the need of the test are discussed and clearly described. However, the scientific bases and the rationale of the test are not explicitly described **The Panel agreed that this criterion has been partly met.**
2. Relationship between the test method endpoint(s) and the biological effect and to the toxicity of interest addressing also limitations	**Reviewer #1:** 1) The description of the relationship between the test method endpoints and the biological effect as well as to the toxicity of interest could be more clearly described. The comet assay measures some kinds of DNA damage, not all of which are known. DNA damage measured by the assay sometimes results in genetic damage. Some of the genetic damage caused by the DNA damage causes cellular and tissue dysfunction in progeny of exposed cells. Some of the damaged progeny cells progress to diseases including cancer and chronic cardiac, neurologic and immune diseases. The many steps between the damage measured by the comet assay and disease makes the assay a very rough biomarker for disease which is nonetheless useful because the assay can be performed quickly whereas the diseases take 2 years to develop in rodents and decades to develop in humans. 2) The rat liver comet assay has been validated using a limited set of chemicals known to cause or to not cause cancer in rodent bioassays. This validation has demonstrated the test to have similar predictivity of the outcome of the rat bioassay as two existing tests, *in vivo* unscheduled DNA synthesis in rat liver (UDS) and *in vivo* micronucleus in rat or mouse bone marrow or peripheral blood (MN). Those tests are known to be relatively insensitive (many false negatives) but highly specific (few false positives) predictors of tumors in rat and mouse bioassays. These tests have proved useful because they permit rapid hazard assessment. The liver comet assay is likely to prove similarly useful. Many regard the *in vivo* UDS as inadequately sensitive and the *in vivo* MN as more sensitive. With the publication of a standardized validated protocol it will be possible to collect data to see whether using the liver comet assay for predicting rat carcinogens will be more like UDS or more like MN,

	although there is not enough reliable data to know than now. 3) The validation study examined data from both liver and glandular stomach. Of the compounds tested, none were uniquely positive in the stomach. Of 15 positives in liver, 4 were clearly positive in stomach and a couple more were equivocal. Bowen *et al.* tested an additional 8 genotoxic carcinogens and found only mitomycin C to be positive in stomach and not liver. Malignancies are relatively rare in the glandular stomach in rat and mouse bioassays and while stomach cancer is a significant cause of human morbidity and mortality, an infectious agent (*Helicobacter pylori*) rather than chemical exposure seems to be the most significant risk factor. While Cmax in the stomach for a gavaged test articles may be larger than in any other organ, the duration of exposure is probably shorter than in any other tissue, an effect which is compounded by the rapid turnover of the epithelial cells which are scraped from the stomach wall to be used in the comet assay. Thus both for theoretical reasons and based on the data presented in the validation report, the value added by selecting the glandular stomach for routine comet analysis is not evident. 4) The published literature contains examples of comet studies using about a dozen tissues. This is a small fraction of the approximately 4 dozen tissues collected during a standard cancer bioassay, each of which has been reported to be the site of malignancy. The proposed test guideline correctly points out that when directed by knowledge of absorption, distribution, metabolism and excretion of the test article, analysis of organs other than stomach and liver may provide insight when used to follow up a genetic toxicity hazard signal from other tests or to examine the mode of action of a known carcinogen. However, the guideline should make clear that these uses are exploratory in nature and thus yet to be validated. During the pre-validation phase four highly experienced laboratories had trouble reproducing the results for the positive control compound and for acrylamide. These difficulties were attributed to differences in interpretation of the test protocol, which was revised to be more specific. This illustrates the need for careful method development for tissue preparation for a given tissue and electrophoreses conditions, which may well need to vary for each tissue. Data must be available before the method can be said to be validated, particularly for tissues like gastrointestinal and urinary bladder epithelia which must be scraped to produce cells for the assay. The predictive value of comet in highly differentiated organs like kidney and brain remains to be demonstrated. 5) The validation was conducted only in male rats under the assumption "no significant gender differences were expected" (Section 6-1). Quantitative and qualitative differences in tumor response occur more often in rat vs. mouse bioassays than not. There is no discussion or literature presented in the validation report to support this decision. The preference for carrying out an *in vivo* test only in males seems to be unique to genetic toxicology test guidelines. A single publication which examined a very limited set of test chemicals has been cited as justification for conduct of the *in vivo* micronucleus assay in male animals only (Mutation Research 172: 151-163 (1986)). The 20 chemicals were evaluated using a quantitation method which is relatively insensitive compared to

	methods which have been since developed. This practice should be reconsidered. 6) The concordance between comet results in mice and rats is not discussed. While the report describes the assay as the "rodent" comet assay, all of the data were collected in rats. There were differences in outcome between the comet data described in the report and the comet results from literature reports of the same chemicals, many of which were evaluated in mice. Given the modest correlation between tumor formation in rats and mice, the extent to which comet data in one species does not predict comet formation in the other should discussed or at least acknowledged to be unknown. **Reviewer #2:** Not studied in detail by the reviewer. **Reviewer #3:** The validation report clearly states that the assay measures DNA strand breaks but it does not detail the mechanistic rationale (but refers to publications). Because this aspect is useful for both the selection of compounds, and the data interpretation (e.g. irrelevant positive because of apoptosis, lack of detection for cross-linking agents, and may be at least some of DNA adduct inducers) this part could have been developed in more details, as it would have been done in a detailed review paper, and knowing that there would be no DRP. **Reviewer #4:** There is strong rationale for linking the measured endpoint (%Tail DNA) with the biological effect of interest (DNA strand breaks and alkaline labile sites). More work is necessary to demonstrate that the alkaline comet assay can reliably detect other types of DNA damages such as crosslinks. Given that some of the events that have an impact on the damage detected by the Comet assay (i.e., repair of DNA breaks) are occurring rapidly after the induction of damage, the timing of tissue sampling is a critical variable and the currently proposed sampling time may not be optimal. The need for histopathology data to discriminate between a cytotoxic and a genotoxic effect is a potential limitation of the assay. **Reviewer #5:** This test has serious limitations, partially due to the limited selection of chemicals and lack of understanding about the MOA of carcinogenicity results, the mode of action of the "comet formation," and the reparability of the "damage". The comet assay predicts DNA damage that may be caused by several modes of action, most of which seem to be unknown. The comet results alone do not indicate which types of modes of action that led to the DNA damage nor does it inform on the various consequences of this DNA damage or whether the end result will be genetic damage much less genetic damage that leads to malignancy. The chemical applicability domain is not clear. For what chemicals should the assay not be used? What if exposure is by inhalation? Metals? Mixtures? This test is only applicable to orally administered chemicals, not to exposure by inhalation or by iv. These routes of

	administration are essential for testing drugs. Stomach is clearly not an appropriate organ for such administrations nor is it clear what timing should be used to capture comet damage that cannot be detected <24 hours post exposure. If a chemical is not stable in stomach acid, oral administration will give a negative response, even in the liver. I saw a recent example of this. **Lack of organ matching between comet results and the carcinogenicity results suggests a disconnect between cause and effect.** Most of the collated CPDB carc findings are in F344 rats from a long time ago in the U.S.—different diet- and animals from the current comet assay animals, primarily in Japan with SD rats. One cannot simply correlate a carcinogenicity finding in mice with a comet in rats, not to mention the tissue/organ mismatch. Even the few liver comet positives that *seem* to correlate to the rodent carcinogenicity liver positives, do not mean that there was a cause and effect. **Although the in vivo MN was accepted many years ago without demonstration of an ability to predict nonhematopoietic neoplasms, that is not a reason in 2013 to accept the comet as anything more than a very rough screen for DNA damage in a standardized protocol.** Even when there is an Ames positive, carcinogenicity findings may not be due to direct DNA effects (e.g., most neoplasms in studies of drugs have been shown to be receptor/hormonally related effects). Some of the reported carcinogenicity findings are misleading (e.g., you cannot correlate a rare finding in 2/50 animals such as for AZT with a comet assay). Diethanolamine was in ethanol, so liver carcinogenicity effects were contributed by the ethanol. **Reviewer #6:** Fully met **The Panel agreed that this criterion has been partly met.**
3. Availability of a detailed test method protocol	**Reviewer #1:** 1) The protocol is reasonably well detailed with a few exceptions. The first is the evaluation of chemicals which retard rather than speed DNA migration under the conditions of the assay. In phase 4-2 five compounds were reported to significantly decrease migration relative to the negative control. In each case this was in the stomach but not the liver. Two were genotoxic carcinogens (acrylonitrile and oxydianiline) one was a genotoxic non carcinogen and two were negative controls (ethionamide and sodium chloride). The minimum % tail DNA for negative controls is set at 1% to allow decreases to be detected and a 2-tailed statistical test is recommended for the same reason. Why were these not scored as detection of DNA damage? The rationale for this decision should be clearly explained. Until then judgment must be reserved about the state of validation of the assay for use in stomach or to detect chemicals which cause damage which retards migration. The possibility that electrophoresis

	or other assay parameters need to be optimized for each particular tissue should be considered. 2) As discussed below and elsewhere in the joint reports, the issue of cytotoxicity still seems to be a major challenge. Methods for evaluating cytotoxicity and the impact of cytotoxicity on interpretation of results both need to be spelled out more clearly. 3) In conjunction with my comments to charge questions 4 and 7, I think it will be unlikely that test results will be interpretable without a contemporaneous positive control and a comparison of the performance of that control to historical data for that compound in the same tissue generated by the laboratory. 4) In conjunction with my comments to charge questions 4 and 7 and in consideration of the experience of the well experienced laboratories which conducted phases 2 and 3 of the pre-validation study, a more than usually detailed section explaining how laboratories can demonstrate proficiency in the test needs to be provided. **Reviewer #2:** Not studied in detail by the reviewer but the protocol used in phase 4-2 of the validation study appears to be sufficient. However, as discussed later, further guidance on assay acceptability criteria and data interpretation (including statistical evaluation of the results) would be useful. **Reviewer #3:** - The validation report properly describes the material and methods, but is limited to rat liver and stomach. Modifications of experimental conditions would be needed for other species and tissues. - The validation report properly describes how DNA strand breaks are measured. - The validation report proposes an analysis and criteria for data evaluation. Based on what is described in the literature and on the discussions within the peer review working group both the analysis and the criteria would need further discussion to reach a consensus on what should be recommended in the guideline. - The validation report describes experimental conditions that have been refined during the validation. Those experimental conditions allowed the definition of clear criteria for negative (% tail intensity of 1-8% in liver and 1-20% in stomach) and positive (values at least 5% and 2-fold higher than the negative controls) controls that would be applicable for the guideline. **Reviewer #4:** The protocol for the in vivo Comet assay used during the last phase of the validation exercise seems detailed enough to allow the conduct of the test in a harmonized way. The acceptance criteria may need to be refined based on the data generated during this last phase. In particular, the criteria for an acceptable response in the positive control may need to be more stringent as the magnitude of the effect among the laboratory differed by 8-10 fold. Further clarification is needed on whether it is best to use the mean of means or the mean of medians and on what statistical methods to use.

	Reviewer #5: Selection of the high dose and the exact criteria for assessment of tissue cytotoxicity are not clear to me. Mild necrosis suggests to me cytotoxicity. Single cell necrosis- is probably not serious toxicity in the liver. What are the histologic criteria for toxicity in the glandular stomach? What about serious irritation or an ulcer in the stomach? At what point does cytotoxicity confound the assay? For stand-alone comet assays, why is an LD50 needed to select a dose? LD50 studies have not been recommended for any toxicity studies of drugs in the U.S. for more than 30 years. What are the dose selection criteria for a repeat dose study? It won't be anywhere near a single-dose LD50. At high doses there may be saturation of systemic absorption. For drugs, dose-spacing selection of in vivo studies considers systemic absorption. It would seem that ADME/PK should be used to help select doses, and extreme toxicity should not be needed. Will 3 hours after the last dose be applicable to all chemicals? What if the half-life of the chemical is very short? The added value of analyzing the glandular stomach is unclear. Rodents may have chemically related neoplasms in the forestomach, but not commonly in the glandular stomach. In addition, none of the chemicals solely caused neoplasms in the stomach. Metabolism in female rats may differ from that in males, but only male rats and not female rats have been tested. Because of hormonal and chemical metabolism differences, findings in male rats do not generally predict findings in female rats, not to mention male mice. **Reviewer #6:** Partially met: some incongruence appear among different parts of the text of the JaCVAM trial. For example, in the first step of the validation study (point 10 pag 103) it is stated that EMS is administered twice; on the contrary, in the protocol it is stated that EMS is administered once (pag.123). Furthermore, a question regarding statistics: It is stated that Dunnett's test was applied to Effect (diff) (pag 105), but if we well understood, the Effect (diff) is the difference between the mean tail DNA obtained in 5 treated animals and the mean obtained in 5 control animals. If this is correct, how is it possible to apply statistics to one single number (the difference of means)? Isn't it more correct to say that the Dunnett's test was applied to evaluate whether treatment induced a significant change in tail DNA? In other words, Dunnett's test was applied to the "ESTIMATE" and not to the "EFFECT". **The Panel agreed that this criterion has been partly met.**
4. Demonstration of within- and between-laboratory reproducibility	**Reviewer #1:** 1) The criteria used to establish reproducibility must be more clearly explained. A clear explanation is necessary. During the phase 4-1 of the validation study four labs tested two positive compounds and one negative compound. Both positive compounds were easily detected by each of the 3 to 4 labs testing the compound. It appears that the

	selected criteria were (1) the ability to detect a dose related increase in two compounds known to induce a robust response in the assay and (2) to find no significant difference for the negative control, a compound which was not expected to induce any physiological response in the assay. Is this correct? Are those criteria sufficient? 2) A list of challenging compounds is needed. Several compounds examined during the validation delivered challenging results. Acrylamide and 2-AAF gave inconsistent results across laboratories during early stages of the validation. Literature reports on 2-AAF are also quite variable. Thioacetemide, sodium arsenate, acrylamide and t-butylhydroquinone are examples of positive and negative controls which proved challenging to analyze. Compilation of a list of compounds which are challenging (but of course not too challenging) is the first step in establishing reproducibility of the test method. The second is using that list to generate data on intra- and inter laboratory reproducibility. Once generated, that dataset can be used to fashion recommendations to demonstrate laboratory proficiency. The outcome of a comet assay is unfortunately more sensitive to small changes in test procedures than more robust assays like micronucleus. This is common knowledge among laboratories which conduct the assay and is evident from the experience of the laboratories participating in the various phases of this validation exercise, in spite of the fact that most if not all of which were experienced in conduct of the assay. A test such as this would be required before one could say qualitative interlaboratory reproducibility had been addressed. 3) Dose administration in the final protocol included three doses of test article but only two of the positive control. It was deemed necessary to retain the two dose regimen used in earlier phases of the validation to "maintain consistency". The suggestion that there will be significant differences between two and three doses , particularly for a positive control that delivers a robust response, does not generate confidence in the reproducibility of the assay particularly since it appears that a wide range of dosing regimens will be recommended. 4) While the validation report states that one goal was to establish intra-laboratory reproducibility, it is not at all clear how that was demonstrated. For example, repeated testing by the same laboratory to demonstrate construction of appropriate positive and negative control ranges was not conducted. Analysis of challenging positive and negative controls by testing them multiple times was not conducted. Without further explanation and probably additional data it cannot be said that intra-laboratory reproducibility has been tested much less demonstrated. **Reviewer #2:** In general, the data from the JaCVAM prevalidation and validation trials complemented with those of the publications of Rothfuss *et al* and Bowen *et al* demonstrate good reproducibility of the comet assay. However, the reproducibility was much stronger in qualitative terms than in quantitative terms, especially for inter-laboratory reproducibility.

	Some comments can be made: - In the report, it is not clear if the 'estimates' were calculated in the same way by the different laboratories. When looking at the individual study reports, some laboratories calculated estimates by taking median values of 50 comets scored per slide and then calculating the average of the medians of the different slides while other labs used the mean values of 50 comets scored per slide to calculate the average for the different slides. In most cases, fold increases are more pronounced when using median tail intensities. For the validation report, the different way of calculating estimates will probably have no influence when evaluating the effect of a treatment. However, it may be important to determine the inter-laboratory reproducibility of the negative controls. For example, in the study report of Bayer both mean and median tail intensities were calculated. For liver, negative controls had a mean tail intensity of 6.79 % and a median tail intensity of 2.57 %. For stomach, negative controls had a mean tail intensity of 13.8 % and a median tail intensity of 7.28 %. These results indicate that when comparing estimates from different laboratories, the way of calculating these estimates should be taken into account. - In general, reporting of the results used for statistical evaluation is confusing. For example, it is stated that Dunnett's test was applied to Effect (diff) (P 113). However, Effect (diff) is defined by the VMT as the difference between the mean % tail DNA obtained in 5 treated animals and the mean % obtained in 5 control animals or the difference of the estimates. If this is correct, how is it possible to apply statistics to one single number? Isn't it more correct to say that the Dunnett's test was applied to evaluate whether treatment induced a significant change in the % tail DNA? In other words Dunnett's test was applied to the "ESTIMATE" and not to the "EFFECT" (this remark was also made by another member of subgroup 1). Furthermore, on P46, the VMT states that they cannot yet recommend something for statistical methods and that the performance of several approaches for statistical test will be examined through this study. However, later in the report it is only stated that the Dunnett's test and a linear trend test were used but I was not able to find the rationale of the VMT for using this statistical test. - Data-acceptance criteria were determined based on the 1st to 3rd phase pre-validation study results. Will these acceptance criteria be used in the TG? Since the protocol was adopted during the validation study, shouldn't these criteria be reconsidered based on the results of phase 4.1? - Inter-investigator variation in scoring of the slides is an important cause of inter-laboratory variability. In the validation report, it is not clear if all labs used fully automated scoring of slides. If different scoring methods were used, it would be interesting to know the impact of the scoring method on the variability.

	Reviewer #3: The validation report shows and compares data obtained within and between laboratories for negative and positive controls, and a few other genotoxic and non-genotoxic compounds. The variability was analyzed and data used for the selection of participants. Despite a relatively high variability at least for positive compounds the reproducibility was considered acceptable for such an in vivo assay. For some class of compounds (i.e. those requiring metabolic activation) the data obtained in the validation study were compared with data obtained in other collaborative work to show the reproducibility. **Reviewer #4:** There is enough data to conclude that the in vivo Comet protocol results in acceptable intra- and inter-laboratory variability in qualitative terms. However, inter-laboratory quantitative reproducibility was marginal at best with the positive control giving effects that sometimes differed by as much as ~8-10 fold among participating laboratories. This last point should be tested using appropriate statistics. If the analyses demonstrate that participating laboratories differed significantly in the magnitude of the positive control response, additional testing to better control this source of variation, or a more stringent criteria for an acceptable response for the positive control, should be implemented. The concern is that a laboratory with a strong positive control response may detect a weak mutagen as positive, while a laboratory with a weak positive response may not. This should be experimentally tested. **Reviewer #5:** This did not seem to be satisfactory—tremendous variation, especially across laboratories and too few chemicals tested. More chemicals that are not flaming positives would be a better test. **Reviewer #6:** Partially met: our comments are summarized in sub-group 1 report **The Panel agreed that this criterion has not been met.**
5. Demonstration of the test method's performance based on testing of representative reference chemicals	**Reviewer #1:** This topic is discussed in the group comments. **Reviewer #2:** This topic is discussed in the group 5 comments **Reviewer #3:** 40 compounds have been evaluated under coded after selection among 90 compounds for which enough data were considered available. However the relatively high numbers of unexpected results suggests that all information were not taken into account for the selection of compounds, and were only considered for data interpretation. The WMT mainly focused on carcinogens versus non-carcinogens probably most of the genotoxic carcinogens are know to induce

	different types of damage. However, because this assay is considered to be able to detect different types of DNA damage and not mutations per se, it would have been important to check this point and to better focus on the types of damage induced and detected. For example it is not quite sure if bulky adduct inducers that require complex metabolic activation are not properly detected because of the metabolism or because of the type of damage. It is an important point because they are adequately in the liver UDS assay, an assay that should ultimately be replaced by the liver Comet assay. Interestingly the validation study suggests that DNA cross linkers are not efficiently detected with the standard method. **Reviewer #4:** Performance of the test method using coded chemicals generated results that, by and large, were in accordance with the predicted outcomes based on available genotoxicity and carcinogenicity data. It would have been desirable to have tested a few more chemicals that required metabolic activation and that have cross-linking MOA. **Reviewer #5:** Most Ames clearly positive chemicals were potent in the assay. Since it is unclear what is predicted from the comet results other than DNA damage potential, sensitivity and specificity cannot be calculated. **Sensitivity and specificity would be a prediction of something external to the assay for which a cause and effect is plausible.** **Reviewer #6:** Fully met. The reference chemicals are all representative and all data were analyzed in blind fashion. **The Panel agreed that this criterion has been partly met.**
6. Test methods evaluation related to existing relevant toxicity data	**Reviewer #1:** This topic is discussed in the group comments. **Reviewer #2:** This topic is discussed in the group 5 comments **Reviewer #3:** Data obtained in other assays (genotoxicity and carcinogenicity assays) are summarized in the validation report and rationale for the selection of compounds is also presented. Highly detailed information are mostly provided for compounds leading to unexpected results to explain the discrepancy (false positive? false negative?). Similar level of information or same type of presentation of the data would have been also interesting for other compounds. **Reviewer #4:** Cytotoxicity evaluation using standard histopathological analyses seem to be necessary especially in cases where the increase in %Tail DNA values is marginally significant. There is the need for additional testing for

	characterizing in more details how cytotoxicity may affect the comet results and how to interpret positive comet findings in the presence of cytotoxicity findings. The validation exercise clearly demonstrated that "hedgehogs" cannot be used as an indicator of cytotoxicity. **Reviewer #5: Performance in species of concern**. This is an assay for DNA damage at high doses in male rats. More than that it is hard to say. It is not clear what the various modes of action of carcinogenicity findings were. One should only compare to SD rats of recent vintage. Hormonally related or receptor mediated neoplasms or neoplasms secondary to enzyme inhibition may be included among the Ames positive chemicals. In some cases the carcinogenicity findings were equivocal and in others no rodent carcinogenicity data were available. There are few data in SD rats and findings were not generally in liver. **A mismatch of target organs is problematic for cause and effect, since this is an in vivo assay.** If the comet assay in male SD rats cannot correlate with findings in male rats, what does it predict about carcinogenicity in any other species? Sensitivity and specificity cannot be calculated. Thus, one cannot say that the assay has been validated, only that a standardized protocol has been proposed. **It's not clear what analysis of the comet in the stomach adds to that in the liver.** **Reviewer #6:** Fully met. The existing literature on carcinogenic and mutagenic properties of chemicals relative to the species of concern was taken into account **The Panel agreed that this criterion has been partly met.**
7. Availability of all relevant data for expert review	**Reviewer #1:** 1) The validation report describes an admirable effort over the course of more than 7 years of work. While many were aware of the effort via the release of test protocols for the various phases of the study, publication of the data from some of those early phases might have allowed broader public discussion and acceptance of the results. As it is, the experts were given 3 weeks to review 7 years worth of data. The compressed timeframe prevented a thorough analysis of the data and the more refined discussion amongst experts that would help focus their evaluations. Many of the comments in this document would benefit from more discussion with the peer reviewers and with the validation management team. 2) Presentation of the data in the test report could be improved. The most important omission was the decision to report only +, - or Equivocal results for each compound in Phase 4-2 of the test report. Presentation of the quantitative results (mean and standard error for each dose and the positive and negative controls in one set of

	tables would help establish the potency of each test article in each lab and test condition. The data were presented only in tabular form in the appendix. Most of the compounds were coded and it appears that some were missing or some files did not use standard file formats. As a result, not all data could be reviewed. These can be corrected during the next phase of the discussion. 3) Evaluation of reproducibility of the assay relies, in part, in knowing the variability in the results during the validation studies. The error bars presented in figures 1-5 in the validation report should also be presented in figures 6-16. **Reviewer #2:** Not studied in detail **Reviewer #3:** - The protocols used are publically available. - The validation report properly summarizes the data. Access to raw data if not impossible is more tedious and may be less well-organized. - The protocol and experimental conditions are described in details. - Negative and positive controls data are available and target ranges are recommended. **Reviewer #4:** It would have been nice to have a few chemicals tested in multiple laboratories also during Phase 4.2. **Reviewer #5:** Modes of action of carcinogenicity for the reported carcinogenicity findings are necessary to understand if there is any cause and effect from the DNA damage. A comet assay incorporated into the repeat dose assay was not part of cross lab reproducibility or the validation studies. **Reviewer #6:** Fully met; however, we do not fully understand the choice of parameters used to create the graphics. For example, since it is often stated that the "Effect" is considered the best criterion to understand the variation of comet parameters among testing facilities, why is this parameter not used in the graphs relative to the test substances? **The Panel agreed that this criterion has been met.**
8. GLP (ideally)	**Reviewer #1:** The critical data all seems to have been obtained from studies conducted under GLP.

	Reviewer #2: Not studied in detail by the reviewer. **Reviewer #3:** The JaCVAM study and the studies used to complement the evaluation (Rothfuss and Bowen publications) were conducted according to GLP principles, except may be that the titer/concentration of treatment solution/suspensions were probably not checked. **Reviewer #4:** From the information provided it seems that the data generated during the validation test were generated in the spirit of Good Laboratory Practice. **Reviewer #5:** Why should an in vivo assay not be GLP? The FDA does not recognize accordance with principles of GLP-only in accordance with GLP or not. **Reviewer #6:** Partially met. The validation study was conducted in facilities that are GLP compliant. However the pre-validation study was conducted "under the spirit" of GLP. As far as Ruthfuss publication is concerned, it is not stated if data were obtained in accordance with the principles of GLP. **The Panel agreed that this criterion has been partly met.**

ANNEX 3b

Specific charge questions: Collated comments from the peer review subgroups assessing the JaCVAM initiative international pre-validation studies of the in vivo rodent alkaline comet assay for the detection of genotoxic carcinogens

Subgroup 1 Report

Specific charge question: Determine whether there is sufficient data from (i) JaCVAM trial, (ii) Rothfuss et al (2010) trial and (iii) in-house/CRO sources to conclude that there is acceptable intra- and inter-laboratory reproducibility

Francesco Marchetti, Eugenia Cordelli, Maria Donner, Birgit Mertens, Paola Villani

We reviewed the data presented in the prevalidation and validation JaCVAM reports and in the publications of Rothfuss et al. and Bowen et al. to assess whether the developed in vivo Comet protocol is robust and provides acceptable intra- and inter-laboratory reproducibility. In reviewing the data provided in these reports, we considered the various phases of the validation exercise independently because studies were conducted over several years and used evolving versions of the in vivo Comet protocol. Given these circumstances, and the large number of participating laboratories, we acknowledge that some discrepancies are to be expected.

The data presented strongly suggest that the in vivo Comet protocol described in the prevalidation report (Phases 1 and 2 of the validation exercise) was not fully developed and contributed to the less than acceptable reproducibility among participating laboratories observed during that phase of the validation study.

Improvements in the protocol utilized during Phase 3 of the study increased the consistency of results within and among participating laboratories, however, a few discrepancies were noted. The positive control induced significant effects in all laboratories and for both tissues, but one laboratory failed to detect the minimum increase in Effect(diff.) of 5% that was set as the criteria for acceptance of the results. Moreover, although all laboratories obtained the expected results with the coded chemicals and all positive controls produced statistically significant results, the magnitude of the effect was considerably different among laboratories. For example: the mean Effect(diff.) in liver ranged from <5% in lab 4 and >35% in lab 3 (fig 23, page 31).

In the first step of phase 4 of the validation study, good reproducibility was obtained among the participating laboratories except for 2-acetylaminofluorene. The unexpected result with this compound remains unexplained. During the second and last step of phase 4, only positive and negative control data are available to evaluate intra- and inter-laboratory variability as each laboratory tested one different coded chemical. In all participating laboratories, the Comet results for all negative and positive controls were within the data acceptable criteria established before conducting the study. Nevertheless, as for previous phases, the magnitude of the effect was considerably variable among laboratories. For example, Effect(diff.) for the positive control in liver ranged from <10% to >60% (fig 5, page 108).

In summary, our assessment is that there is sufficient data to conclude that the experimental protocol utilized during the last phase of the validation exercise for conducting the in vivo alkaline Comet assay demonstrated acceptable intra- and inter-laboratory reproducibility. We note though that the reproducibility was much stronger in qualitative terms than in quantitative terms, especially for inter-laboratory reproducibility. It would be desirable to improve the concordance of the quantitive response among laboratories.

Furthermore, Subgroup 1 makes the following comments:

- It is not clear whether the ‘estimates’ were calculated in the same way by the different laboratories. When looking at the individual study reports, some laboratories calculated

estimates by taking median values of 50 comets scored per slide and then calculating the average of the medians of the different slides, while other labs used the mean values of 50 comets scored per slide to calculate the average for the different slides. In most cases, fold increases are more pronounced when using median tail intensities. For the validation report, the different way of calculating estimates will probably have no influence when evaluating the effect of a treatment. However, it may be important to determine the inter-laboratory reproducibility of the negative controls. For example, in the study report of Bayer both mean and median tail intensities were calculated. For liver, negative controls had a mean tail intensity of 6.79 % and a median tail intensity of 2.57 %. For stomach, negative controls had a mean tail intensity of 13.8 % and a median tail intensity of 7.28 %. These results indicate that when comparing estimates from different laboratories, the way of calculating these estimates should be taken into account.

- Inter-investigator variation in scoring of the slides is an important cause of inter-laboratory variability. In the validation report, it is not clear if all labs used fully automated scoring of slides. If different scoring methods were used, it would be interesting to know the impact of the scoring method on the inter-laboratory variability.
- In Phase 3, the CV of Effect(ratio) obtained in three experiments was used as the parameter to estimate intra-laboratory variability. The limit of acceptability, set as <50%, was reached in 3 out of 4 laboratories. We suggest that the CV should have been calculated also for Effect(diff.) since this is the parameter considered in the subsequent validation phase.
- Data-acceptance criteria were determined based on the 1st to 3rd phase pre-validation study results. We suggest that these data-acceptance criteria should be reconsidered based on the results of phase 4.1.
- Despite achieving significant progress in reducing inter-laboratory variability with each subsequent version of the comet protocol, even in the last phase of the validation study, there was considerable variation in the magnitude of the positive control response among the various laboratories (Effect(diff) ranged from ~10% to ~60%). These differences should be tested statistically and, if significant, additional testing may be required to better control this source of variability. The concern is that such an inter-laboratory variability would affect the detection of weak genotoxic agents by those laboratories with small Effect(diff).

Subgroup 2 Report

Domain of Applicability

1. The validation report defines the applicability domain as: (Paragraph 97) "The high ability of the assay to deliver the expected positive and negative Comet assay outcomes, based on detailed knowledge of the compounds, their modes of action, and route-, species- and tissue specificity can be taken as convincing evidence that the applicability domain for the assay (namely to detect genotoxic and carcinogenic chemicals that induce DNA strand breakage) is appropriately defined"
2. The report identifies 6 of the tested compounds as rat liver carcinogens and one as a stomach carcinogen. The other compounds were carcinogenic in other organs but not in rat liver nor in stomach. One compound was a mouse liver carcinogen for which no rat bioassay data was available. All 6 of the rat liver carcinogens were positive in the comet assay in liver and the rat stomach carcinogen was positive in the comet assay in stomach.
3. While the report notes that it is technically possible to obtain single cell suspensions for comet assay conduct from many rat and mouse tissues in which the listed carcinogens have been shown to induce malignancy, no evidence is presented in the present report and supportive publications that damage can be or has been detected by the comet assay in a way that correlates DNA damage to malignancy in any tissue other than liver* and stomach.
4. The proposed test guideline suggests two uses for the assay: As a follow up assay to a positive result in another genetox assay (Paragraph 7) or as part of an initial screen in combination with the bone marrow micronucleus test (paragraph 10). There is no indication of whether the assay would be conducted identically under both circumstances or when or how much additional information (e.g. ADME and other data as discussed in paragraphs 8 and 11 and 13 of the proposed TG) is recommended*. The proposed guideline also suggests that "the Comet assay is a useful in vivo follow-up for chemicals inducing both gene mutations and chromosomal aberrations in vitro" (paragraph 7). While the mechanistic correlation between DNA strand breaks and chromosome aberrations is quite obvious, an in depth evaluation of the ability of the Comet assay to detect compounds that induce mainly gene mutations as a result of bulky DNA adducts or compounds which induce damage that creates alkaili-labile sites would be needed. It appears that compounds which cause bulky-adducts are not well represented in the validation, nor is it clear that the list of chemicals was chosen to be representative of the chemical space*.
5. The concordance between comet results in mice and rats is not discussed. While the report describes the assay as the "rodent" comet assay, all of the validation data were collected in a single rat strain. Other validation studies (e.g. Rothfuss *et. al*) were collected in other rat strains. There were significant differences in outcome between the comet data described in the report and the comet results from literature reports of the same chemicals, many of which were evaluated in mice. Given the modest correlation between tumor formation in rats and mice, the extent to which comet data in one species does not predict comet formation in the other is as yet unclear. It might be advisable to restrict the guideline to rat liver and stomach and to state that, while the assay has been only widely validated in rat liver and stomach, it is potentially also applicable to other species and tissues, if scientifically justified and if enough data are available to support the protocol and criteria to be used*.
6. The validation was conducted only in male rats under the assumption "no significant gender differences were expected" (Section 6-1). Quantitative and qualitative differences in tumor response occur more often in rat vs. mouse bioassays than not. There is no discussion or literature presented to support this decision*. The guideline would have to clarify that testing females or both sexes might be more appropriate in some cases (see other in vivo guidelines).

7. The report and most available comet assay literature describe only gavage as the route of administration. Safety testing of articles intended for use in food or feed is generally tested in feed or occasionally in drinking water to more closely approximate the conditions of human exposure. The suitability of these methods of administration are not discussed, particularly with regard to detection of compounds that induce damage that can be measured by comet only when animals are sacrificed <24 hours after the last administration. Moreover the Comet assay has been used for the detection of topical effect (e.g. skin) in case of unstable and reactive compounds, but this was not the purpose of this validation exercise which concentrated on oral route. A white paper exercise would have been necessary to evaluate the potential other applications of the Comet assay*.
8. It is interesting to note that only two of 8 genotoxic aryl amines were positive in this study. The aryl amines tested were 2-acetylaminofluorene , o-anisidine 2,4 diaminotolune, 4,4'oxydianiline, 9 aminoacridine, p-anisidine, 2,6-diaminotoluene and phenylenediamine. The first four of these are rodent carcinogens. While arguments are presented for the expectation of a negative response for 9-aminoacridine it has not been tested in a bioassay. Activation of 2-AAF occurs via a multi-step pathway that requires both phase I oxidative and phase 2 reductive metabolism. Other compounds in this class may require similarly complex or organ specific metabolism and many are known to be more likely to cause cancer in bladder than in liver. The possibility should be considered that the assay is not sensitive to this and possibly other classes of carcinogens which require certain types metabolic activation.
9. Two chemicals tested are expected to crosslink DNA, which is expected to retard rather than speed up DNA migration during electrophoresis. Cis-platin was clearly positive in the assay by speeding migration, consistent with its known ability to form a variety of types of damage including intra-strand adducts. Busulfan was negative in the assay. While the negative control ranges were set to be able to detect retardation by crosslinking reagents, unless they also cause other types of damage compounds which cause inter-strand crosslinks may be outside the domain of applicability unless the protocol is modified. Moreover decrease in DNA migration could be induced by other mechanisms including cytotoxicity. Therefore considering a decrease in DNA migration as the signature of DNA cross linking agents could be misleading.
10. Of the two heavy metals tested, one was clearly positive. One was judged equivocal by the VMT and negative by the testing laboratory. Large metallic ions are outside of the domain of applicability for bacterial mutagenesis assays because they cannot pass though the bacterial cell wall. Many industrial compounds including catalysts are in this class. More evaluation of the ability of comet to detect inorganic metallic test substances may be necessary to determine whether they are in the domain of applicability for the assay.
11. Comet detects some but not all types of DNA damage. Most DNA damaging agents cause multiple types of damage, as described above for cis-platin. Compounds that cause strand breaks or alkali-labile sites are visualized using comet. Techniques to augment the comet protocol with exogenous DNA repair enzymes have been reported in studies when mechanism of action is known or being determined. It is known that tissues must be collected within a few hours of the final dose administration for some damaging agents but the length of that window does not seem to be known. The validation protocol as well as validation studies by Rothfuss *et al.* and Bowen *et al.* specified harvest 3 hours after the final dosage administration. The change to 2- 6 hours in the proposed test protocol should be removed or justified. It would be useful to provide information about how the timing or type of DNA repair associated with a chemical or class of chemicals affects the chemicals which are detected by the assay and thus are within the domain of applicability*.

12. Even if not part of the validation it might be useful to mention the opportunity that the detection of oxidative damage could be improved by adding specific enzymes during cell processing such as OOG1 for ^{6}O-methyl guanine, when sufficient information about damage mechanisms is known, predicted, or being investigated. As mentioned above this issue might be the types of DNA lesions. It has always been considered that the Comet assay can detect DNA strand breaks or alkali-labile sites resulting from incomplete DNA repair process, including bulky adducts. While this has been studied *in vitro* where the large concentrations used could saturate the DNA repair capacity, it is unlikely to be relevant to the in vivo situation. This would need more evaluation. The guideline should highlight that such compounds may be outside of the domain of applicability for the current protocol and that a modified protocol, with appropriate justification, might be needed.

The applicability domain of the assay includes those chemicals which, after gavage, induce strand breakage and some types of DNA damage in the livers and stomach of male rats. A small number of rat liver carcinogens has been tested and all induced liver damage detected by comet. Detailed knowledge of the association between damage in liver detected by comet and individual cancer target tissues other than liver is not available. Detailed knowledge of the association between damage in liver detected by comet and specific mechanisms of chemical action or types of DNA repair is not available. Detailed evaluation on how the Comet assay could complement *in vitro* positive results and other *in vivo* assays is not available. With the currently available data, it is too soon to judge whether compounds which cause toxicity in an organ or tissue are within or outside the domain of applicability of the assay for that tissue. The overall association between damage detected by comet and rodent carcinogens has been assessed by the validation study and is discussed in the report of PR group 5.

*To be fair, many of these points would be more appropriate for a detailed review paper than in a report of a validation exercise. However, since no detailed review paper is available for this assay, the burden falls on this document to support the proposed test guideline.

Subgroup 3 Report

Group membership:
Maria Donner, Fabrice Nesslany, Siegfried Knasmueller, Stefan Pfuhler (coordinator)

1) Summary of comments

Specific PR3 charge question:
"Determine whether the use of histopathology data for determining cytotoxicity is the best or only recommended approach.

It is acknowledged that histopathology is a valuable tool but there is (at this time) no agreement whether this should be the 'gold standard' or even the only parameter looked at. The discussion about histopathology in the validation report seems somewhat controversial. The difficulty seems to lie in defining a cut-off at which point histopath findings may contribute to 'misleading' findings. PR3 recommends spending more time on deriving more detailed conclusions from the findings observed in the validation study, as well as from relevant literature data in order to define such a cut-off in a way that it could be applied to the Comet assay guideline. This may require the help of a pathologist.
It was also recognized that in the case when enzymatic digestion is used instead of mincing classical viability measures such as trypan blue staining can be applied.

o What is the purpose of measuring cytotoxicity?

PR Group 3 agrees that cytotoxicity can be a confounder when interpreting Comet assay results and that this parameter should be monitored in all the tissues that are included in the evaluation of an in vivo Comet assay study.

o Data on "hedgehogs" have been collected – are they useful?

From the data generated during validation it seems that "hedgehogs" are a quite insensitive parameter that, in the presence of 'high quality' histopathological data, may not contribute much to the decision whether cytotoxicity can be a confounder. It was also emphasized that there is controversy to date about what an increase in hedgehogs really means as it seems still impossible to clearly determine their origin (necrosis vs apoptosis vs strong genotoxic effect). However, data on hedgehogs are easy to collect and there will be circumstances where this will add information (e.g., when no or no 'high quality' histopathological data are available). RP3 therefore agrees that the evaluation of hedgehogs is a worthwhile effort but should not be taken as standalone criterion to determine tissue toxicity.

o Should histopathology only be determined at time of sampling for comets, or also at later times? Do we have enough data (from all sources) to be able to provide an answer, or does more experimental work need to be done?

PR Group 3 agrees that in general histopathological examination at sampling makes sense and should be sufficient. The caveat that was brought up is that in a single application situation with short sampling time (e.g, 3h) using parallel samples for histopathology could underestimate toxicity of the compound as the consequences of compound toxicity may not yet be detectable using histopathology.

o Can any recommendations be made regarding supplementary measures such as Capsase 3/7 activation, Annexin V staining, TUNEL staining, Halo or neutral diffusion assay?"

The use of new staining methods which allow to discriminate between apoptotic, necrotic and normal cells in combination with COMET assays may be useful. The caveat is that these measures are not

really specific and results generated may therefore be difficult to interpret. The question was brought up how much toxicity, according to these measures, would actually be tolerable? Also it seems like the some or all of the information that is targeted by the use of above mentioned measures will already be there from a thoroughly performed histopathological investigation. It was suggested that it may be more efficient to generate more targeted mode of action data in a follow-up study.

2) Specific comments by team members:

Specific PR3 charge question:
"Determine whether the use of histopathology data for determining cytotoxicity is the best or only recommended approach.

o What is the purpose of measuring cytotoxicity?

Maria Donner:
The purpose would be to eliminate scoring high cytotoxicity as a positive comet finding. On the other hand I believe this is a more critical question for the in vitro assay than for the in vivo assay, that also often is an integrated assay where the comet is only one endpoint. Dose levels for the actual study will often be set based on limit doses for other guidelined studies, or tolerance to the animals. This said, cytotoxicity measurements are necessary if high cytotoxicity is seen, in order to rule out interference from DNA fragmentation from direct DNA damage. Measurement of cytotoxicity might become very important when testing weak genotoxic agents which require high dose levels in order to be detected (and higher increases in cytotoxicity might therefore be seen).

FaBrice Nesslany:
I totally agree with Maria's comments regarding the rationale for measuring cytotoxicity.

It should be mentioned that the viability assessment also depends on the methods of cell/nuclei isolation used, e.g. it is not possible to used "conventional" viability assessment methods as trypan blue with isolated nuclei.

On the other hand, it is also known that the level of cytotoxicity is dependent of the method used. For instance, trypan blue is not the most relevant when apoptosis occurs and cytotoxicity may be under-estimated… and so on.

Otherwise, following previous discussions at IWGT, a consensus was obtained about histopathology which is considered as the gold standard to evaluate cytotoxicity in the *in vivo* Comet assay (Hartmann, et al., 2003).
In my opinion, I think histopathology is not always the gold standard; as any other methods, it has some limits (Sensitivity? When is the optimum time to examine tissues for histopathological changes). The Jacvam report confirmed that it is still unclear how to use histopathology for the interpretation of Comet assay results

Sigi Knasmüller:
Cytotoxicity is monitored in comet assays as it is assumed that cell death leads to DNA damage (Olive et al. 1993; Fairbairn et al. 1996). It is repeatedly stated in the literature that apoptotic and necrotic cells lead to formation of hedgehogs (HH). This assumption is based on the results of *in vitro* studies in which cells were treated with acute toxic compounds which are not DNA reactive (Hartmann and Speit 1997; Henderson et al. 1998). However, some additional points should be taken into consideration.

1) HH are not uniquely typical for apoptosis/ necrosis, they were also seen in vitro after treatment of cells with high doses of mutagens and with radiation (which does not cause apoptosis) (Burlinson et al. 2007).

2) The association between toxicity and formation of HH and normal comets depends strongly on the cell type . In rat hepatocytes and lymphoblastoid cells acute toxicity leads to comet formation, on the contrary extensive cytotoxicity does not affect comet formation in V79 cells (Hartmann et al. 2001).
3) Information about the association between acute toxicity and comet formation in inner organs is scarce. The report of Sumitomo Chemical (Study No. M4690) mentions formation of HH in the liver of rats without positive histopathology, while in the stomach, HH formation was paralleled by acute toxicity in a model study with the substance M325815.
   In an ILS report (study No. N115-001-155/156) no HH were seen in an experiment with male Sprague-Dawley rats while histopathology indicated acute toxicity.
4) It is important that an *in vitro* study with Jurkat cells (that hardly express Fas antigen) in which apoptosis was selectively induced with anti-Fas antibody, showed that not only HH but also normal comets are induced (Choucroun et al. 2001).

In conclusion, cytotoxicity should be monitored as it may lead to false positive results.

Stefan Pfuhler

I believe that the validation study does confirm that cytotoxicity can be a confounding factor (e.g. Chloroform), however, it does at the same time demonstrate that there is not easy way out of that problem. The main problem is what I would call the 'hen and egg' phenomenon – it is not so easy to find out whether cytotoxicity caused an increase in strand breaks by indirect mechanisms (necrosis, tissue inflammation), or whether cytotoxicity follows genotoxicity as is very commonly the case in genetic toxicology testing. One of the Comet specific problems is that DNA breakage is a primary effect and shows up very early in the process. From own experience I can confirm that cytotoxicity can be a confounding factor:

a) example shown in Fig 7 by Vasquez 2012 is a compound I have been working on with Marie. This I believe is a clear example of cytotoxicity triggered effect as both parameters go hand in hand. This compound was negative in all other organs tested in the Comet, and in a liver UDS test and an in vivo micronucleus test. In that case we do have confirmation that this compound triggers liver toxicity in repeated dose studies.

b) We could show that silica nanoparticles, administrated i.v, did cause an increase in a rat comet assay in the liver only (Downs et al, 2012). Looking into various additional paprameters we could demonstrate that this effect was secondary to liver inflammation demonstrated by tissue necrosis and neutrophil infiltration.

Taken together I do believe that histopathological examination of the tissues is very important and does contribute to the interpretation of the data. The difficulty seems to lie in defining a cutoff at which point histopath findings may contribute to 'misleading' findings. The draft validation report in my view does not do a good job at drawing this line. The respective section (para 101) reads: *".. it is recommended that positive findings in the Comet assay should be interpreted as relevant to in vivo genotoxicity <u>even if weakly cytotoxic changes such as single cell death/necrosis are observed by histopathology</u>, because many genotoxic carcinogens showing significant increases in % tail DNA in this validation study induced such slight cytotoxicity as observed by histopathological changes."*

More effort will be needed to define what is a 'weak' finding in histopath, and to define where the concern does start.

o Data on "hedgehogs" have been collected – are they useful?

Maria Donner:

Not really, since they are potentially apoptotic cells. I am also not aware of a clear distinction between highly damaged cells and apoptotic cells.

FaBrice Nesslany:

Up to date, it is not possible to clearly determine the origin of the hedgehogs (Ghost Cells). In that case, I think data on "hedgehogs" should be collected.

The main question is what to do with theses data.
Perhaps we should distinguish 2 possibilities:

if there is no significant increase in the % of hedgehogs, of course, no problem.

If there is a significant increase in the % of hedgehogs, (both if the Comet assay concluded negative or positive)
In that case, I think that histopathological assessment will be useful to try to determine the origin of the hedgehogs
If histopathological changes show mainly necrosis, hedgehogs are probably due to cytotoxicity and it could be an interference with the endpoint.
If histopathological changes show mainly apoptosis, hedgehogs probably due to such phenomenon. We can not exclude that apoptosis is induced by genotoxicity (via p53)
Other specific methods to display an apoptotic component could also be implemented instead of histopathology, e.g. Neutral Diffusion assay, … See below

Sigi Knasmüller:
To a certain extent yes. HH are sometimes indicative for acute toxic effects; however, as mentioned above they occur also in absence of cytotoxicty.
An important question is whether they should be evaluated and included in the comet analysis.
1) In absence of acute toxicity (i.e. when histopathology is negative), HH could be integrated in the comet analysis but it is stated in the JaCVAM paper that they should be excluded from the data analysis and recorded separately. This point deserves discussion as it may be quite difficult to define strict clear criteria for HH, and it is also unclear if the exclusion is justified.
2) In the presence of acute toxicity the outcome of a comet experiments should be regarded as inconclusive, but this does not depend on the occurrence of HH or not as a cell death may lead to normal comets (see above).

Stefan Pfuhler
To keep it very short: I believe 'hedgehogs', in an *in vivo* setting, are a quite insensitive parameter that may not contribute to the decision whether cytotoxicity can be a confounder. This seems to be confirmed by the data generated during validation (see para 102)

o Should histopathology only be determined at time of sampling for comets, or also at later times? Do we have enough data (from all sources) to be able to provide an answer, or does more experimental work need to be done?

Maria Donner:
I am not sure what later timepoint is intended. However, histopathology should be determined at any timepoint considered relevant for the interpretation of the comet results.
This might be an area where some more work needs to be done.

FaBrice Nesslany:
I totally agree with Maria's comments.
Taking into account that DNA fragmentation is assessed early after the last treatment, I think that histopathology should be performed at the same time. In other words, how explain that the photograph of DNA fragmentation is performed at T0 while the one of histopathologic changes is done later? If so, there is a possibility of measuring the consequence of the damage (DNA damage) rather than the occurrence of DNA lesions.

Furthermore, if preliminary toxicity assay is well perform, theoretically, no lethal or too toxic dose should be kept for main genotoxicity assay even at late stage

I did not find data from in vivo experiments which allow to answer this question; in a study of Morley et al. (2006) the time course of comet formation and apoptosis was investigated in vitro and it seems that both events take place at the same time and that apoptosis persists while DNA damage vanishes as a result of repair.

Stefan Pfuhler
I tend to agree with Maria and FaBrice. This question, I believe, is mostly relevant for a Comet assay protocol that uses single dosing. In such a situation I can see that using parallel samples could be underestimating the toxicity of the compound for the early sampling protocol (e.g, 3h) when the consequences of toxicity may not yet be clearly detectable using histopathology. Also it is often ignored that by using such an early sampling the animals could be dosed higher than LD50 since they may survive lethal doses for such a short time.

Sigi Knasmüller
I did not find data from *in vivo* experiments which allow to answer this question; in a study of Morley et al. (2006) the time course of comet formation and apoptosis was investigated *in vitro* and it seems that both events take place at the same time and that apoptosis persists while DNA damage vanishes as a result of repair.

o Can any recommendations be made regarding supplementary measures such as Capsase 3/7 activation, Annexin V staining, TUNEL staining, Halo or neutral diffusion assay?"

Maria Donner:
These are all different measurements for apoptosis. It would be good if consensus could be found on which measurement can delineate strong apoptotic false positives. Maybe a combination of doing a TUNEL assay and Annexin V staining? On the other hand, I think the TG should leave some flexibility to the performing laboratory as well and not tie this in too strictly.

FaBrice Nesslany:
As I mentioned before, apoptosis could be determined but by a scientifically recognized method, and accepted as sufficiently specific. In that way, TUNEL is unfortunately not specific of apoptosis as indicated by its name (Terminal déoxynucléotidyl transférase dUTP Nick-End Labeling). Indeed, the 3'OH ending of DNA strand breaks are not generated only during the apoptotic process but also when genotoxicity (or necrosis) occurs.
I think that if a recommendation is done, it should be something like: "
If apoptosis is assessed, it should be carried out be by a scientifically recognized method as sufficiently specific"

Anyway, even if the DNA fragmentation is attributed to apoptosis, we can not exclude that apoptosis is induced by genotoxicity (via p53) and we could not qualify such effect as false positive. The main interest may be to determine a threshold dose if it exists...

Sigi Knasmüller:
I think that the main problem is that we don't know how strong the impact of acute toxicity is on comet formation in inner organs. Therefore we can not define the conditions which are acceptable (how much acute toxicity is tolerable???). For in vitro assays it is specified that the viability should be in general 70-80% (Anderson and Plewa 1998; Henderson et al. 1998; Tice et al. 2000) but this statement is not based on sound scientific data. I think that additional data are required to clarify this question.

According to Vasquez (2012), the inclusion of LMW diffusion assays in in vivo studies will provide information on apoptosis/necrosis; but it is not clearly explained what the advantages are in comparison to the use of conventional histopathology.
A very promising approach may be the apo/necro comet assay which allows to detect DNA migration separately in viable, apoptotic and necrotic cells by use of a combination of stains (Morley et al. 2006). This approach was used so far only in in vitro experiments but if it is possible to adapt it for in vivo experiments, it may substantially contribute to solve the problem associated with cytotoxicity.
In regard to the use of the TUNEL assay it should be kept in mind that it is rather unspecific and fails to discriminate between apoptosis, necrosis and autolytic cell death (Grasl-Kraupp et al. 1995).

Stefan Pfuhler
Well, according to the validation report 'apoptosis could indicate both cellular toxicity and DNA damage" (para 26) so this question would be irrelevant. I don't necessarily agree with that statement. I do agree with the comments of Maria, FaBrice and Sigi that the measures are not really specific. Also, in cases where we used Annexin V (Downs et al, 2012), this worked fine but did not really add much to what we already knew from histopath. In general I do, however, believe that additional mechanistic information generated certainly can add to resolving a problematic data set but this would be challenging to do in a general setup and would better be done in a targeted follow-up study

References (not yet checked for completeness)

Anderson D, Plewa MJ (1998) The International Comet Assay Workshop. Mutagenesis 13 (1):67-73.

Burlinson B, Tice RR, Speit G, Agurell E, Brendler-Schwaab SY, Collins AR, Escobar P, Honma M, Kumaravel TS, Nakajima M, Sasaki YF, Thybaud V, Uno Y, Vasquez M, Hartmann A (2007) Fourth International Workgroup on Genotoxicity testing: results of the in vivo Comet assay workgroup. Mutat Res 627 (1):31-35.

Choucroun P, Gillet D, Dorange G, Sawicki B, Dewitte JD (2001) Comet assay and early apoptosis. Mutat Res 478 (1-2):89-96.

Fairbairn DW, Walburger DK, Fairbairn JJ, O'Neill KL (1996) Key morphologic changes and DNA strand breaks in human lymphoid cells: discriminating apoptosis from necrosis. Scanning 18 (6):407-416.

Grasl-Kraupp B, Ruttkay-Nedecky B, Koudelka H, Bukowska K, Bursch W, Schulte-Hermann R (1995) In situ detection of fragmented DNA (TUNEL assay) fails to discriminate among apoptosis, necrosis, and autolytic cell death: a cautionary note. Hepatology 21 (5):1465-1468.

Hartmann A, Kiskinis E, Fjallman A, Suter W (2001) Influence of cytotoxicity and compound precipitation on test results in the alkaline comet assay. Mutat Res 497 (1-2):199-212.

Hartmann A, Speit G (1997) The contribution of cytotoxicity to DNA-effects in the single cell gel test (comet assay). Toxicol Lett 90 (2-3):183-188.

Henderson L, Wolfreys A, Fedyk J, Bourner C, Windebank S (1998) The ability of the Comet assay to discriminate between genotoxins and cytotoxins. Mutagenesis 13 (1):89-94.

Morley N, Rapp A, Dittmar H, Salter L, Gould D, Greulich KO, Curnow A (2006) UVA-induced apoptosis studied by the new apo/necro-Comet-assay which distinguishes viable, apoptotic and necrotic cells. Mutagenesis 21 (2):105-114.

Olive PL, Frazer G, Banath JP (1993) Radiation-induced apoptosis measured in TK6 human B lymphoblast cells using the comet assay. Radiat Res 136 (1):130-136.

Tice RR, Agurell E, Anderson D, Burlinson B, Hartmann A, Kobayashi H, Miyamae Y, Rojas E, Ryu JC, Sasaki YF (2000) Single cell gel/comet assay: guidelines for in vitro and in vivo genetic toxicology testing. Environ Mol Mutagen 35 (3):206-221.

Vasquez MZ (2012) Recommendations for safety testing with the in vivo comet assay. Mutat Res 747 (1):142-156.

Subgroup 4 Report

"Although statistical analysis was the prime criterion for determining results in the JaCVAM trial, in the TG biological relevance of results may take priority. What would describe a biologically relevant positive response? How should historical control data be used in interpretation of results?"

Fabrice Nesslany (Lead), Véronique Thybaud, Stefan Pfuhler, Anoop Kumar Sharma, Maria Donner

What would describe a biologically relevant positive response?

In order to put this first question specifically to the Comet assay in the broader context of OECD genotoxicity guidelines, it was deemed necessary to recall the use of statistical analysis and biological relevance including historical data in current OECD guidelines:

In the current OECD draft guideline it is recommended to consider for data interpretation the biological relevance first and the statistical analysis (only) as an aid.

Recently during the revision process, there was a proposal to consider the following points for data interpretation, at least in the *in vitro* guidelines knowing that this approach would have to be adapted for *in vivo* guidelines.

Text in the last drafts as follows:

Providing that all acceptability criteria are fulfilled, the following criteria are considered for the evaluation of results:

(1) the increase is dose-related,

(2) at least one of the test concentration exhibits a statistically significant increase compared to the concurrent negative control,

(3) the result is reproducible (e.g. between duplicates or between experiments),

(4) the result is outside the distribution of the historical negative control data (e.g. 95% confidence interval).

In conclusion in the current guidelines and revised draft guidelines, the statistical analysis is only one element of the data interpretation, and biological relevant including comparison with historical data pay an important role. Therefore in addition to statistical analysis, among the criteria that could be considered for the data interpretation are 1) the dose-effect relationship, 2) reproducibly between slides and animals, 3) any other relevant information like cytotoxicity, and 4) comparison with spontaneous background (i.e. negative historical control data) discussed afterwards in the document (specific question).

1) Dose-effect relationship:

In the validation report, dose-response effect was evaluated using a trend test in addition to pair-wise analysis. Even for the well-known genotoxic compounds that have been tested, many different types of dose-effect relationships were observed (linear and increasing, linear and decreasing, plateau, bell-shaped, decrease(s) at the highest dose(s) generally linked to toxicity). Taking into account such numerous possibilities with different genotoxic compounds, the Sub-Group PR4 agreed that dose-effect relationship should be used as one of the criteria for biological significance but would recommend a cautious evaluation. For example, a compound could be concluded as genotoxic in the Comet Assay without inducing a clear dose-effect relationship if the highest dose only shows a clear increase in DNA strand breaks. Contrarily if only the lowest dose gives an effect while there are no signs of toxicity in the other two doses, then this effect of the low dose should be cautiously interpreted. Similarly, caution should be taken when a compound induces a clear dose-effect relationship but with none of the dose showing a statistically significant difference with the concurrent

negative control data and/or if all mean % Tail DNA are within the historical data for negative response.

2) Reproducibility between slides, animals and experiments

The heterogeneity between slides and animals, and the reproducibility if the experiment is repeated should also considered as key elements in data interpretation. Increases only observed in one out of two slides or one or two animals would have less weight than increases observed in all slides and animals from the same group. Unfortunately due to the limited amount of time dedicated to the peer review the work was not able to review all validation data in details to check this point.

3) Use of cytotoxicity data and any other relevant information

The discussion of unexpected results in the validation report clearly exemplifies how toxicity and any other relevant information could be used for data interpretation and questions still remain about how the results of histopathology analysis and % hedgehogs should be considered when interpreting Comet assay results.

Anyway, a positive response in the Comet assay in the presence of histopathological changes or hedgehogs should be interpreted with caution. It may even warrant a repeat assay at lower doses or even a different assay. Members of SG PR4 recommend referring to SG PR3 conclusions.

4) How should historical control data be used in interpretation of results?

Criteria for positive results: statistics versus historical data

Interestingly data-acceptance criteria for positive control group are: 1) Effect (difference of means of % tail DNA between groups of EMS and vehicle control) is statistically significantly increased **and** 2) is at least 5% higher than negative control levels (primary criteria); and 3) Effect (ratio of means of % DNA in tail between groups of EMS and vehicle control) is 2-fold or higher, while for the coded compounds only the first part, i.e. statistically significance increase in at least one dose group was applied (complemented by a trend test).

Whether positive control criteria (sort of cut-off values or global evaluation factor) could be applied for the interpretation of the results obtained with an unknown compound was not discussed in the validation report, but the question was raised in the review group. It was noted that not all tested compounds concluded as positive showed changes in the % tail DNA (Diff) higher than 5% or more than "2-fold when compared to the concurrent negative control" (see Table 2 Summary of 4th Phase-2nd Step Validation Study Results). Moreover, even if an important dispersion in % Tail DNA in negative controls was seen from one lab to another, the normal range for negative controls was considered to be 1-8% for liver and 1-20% in stomach. This point should also be taken into consideration if such a cut-off approach or minimum increase over the background is considered.

Moreover, in the validation report, we can easily note that there is an even more important dispersion in % Tail DNA positive responses from one lab to another. For instance, for the liver % Tail DNA for EMS 200 mg/kg (X2) used as positive reference compound ranged from +8 to +62 in terms of Effect difference while Effect ratio varied from X2 to X40. For the stomach % Tail DNA ranged from +15 to +70 in terms of Effect difference while Effect ratio varied from X2 to X14.

Members of SG PR4 underline the importance of using historical control data for interpretation.

Considerations on how to build historical database were addressed and SG PR4 suggests referring to Hayasi et al (2011) recommendations, *i.e.*:

- A minimum set of data resulting from independent experiments is recommended to create the historical data set (note that the SG PR4 did not fix a precise number of independent assays),
- It is not appropriate to use the simple range (minimum and maximum value observed during the data accumulation period) of the accumulated historical, especially negative, control data for an assessment. Rather, the distribution of the data together with appropriate descriptive statistics should be considered (e.g., confidence intervals, 95-99% percentiles).

Otherwise, because of the prerequisite for the negative controls it should be highlighted that the broader (and fully acceptable) range will be 1-8 % for liver and 1-20 %for stomach.

CONCLUSION:

First, members of SG PR4 would like to point out that there was not enough time to allow definitive conclusions. Up to date, only the following preliminary conclusions can be proposed:

The description of a biologically relevant positive response should include Effect together with biological relevance including historical control ranges.

The effect should be preferably expressed as Effect (difference) rather as Effect (ratio).

The definition of a clear positive result could correspond to:

- **a dose-related increase (or clear understanding of the absence of dose-effect),**
  and
- **% tail intensity value(s) is(are) outside the historical negative control range,**
  and
- **a statistically significant increase over the concurrent negative control group in at least one dose,**

Other criteria as reproducibility between slides, animals and experiments (if repeated) but also cytotoxicity and any relevant information should also be taken into account.
Anyway, members of SG PR4 state that more discussion is needed before a final conclusion can be drawn.

Consequently, several the points were identified during our discussions that would need further discussion at the OECD meeting:

- **the use of cut-off values or absolute increases for a biological significance,**
- **how taking into account histopathological changes or hedgehogs,**
- **parameter(s) to be used for stat analysis and group comparison,**
- **the stat analysis itself**

Subgroup 4 report - Appendix 1

To illustrate the capital interest of the criteria retained for the decision of a biologically significance, a member of SubGroup PR4 has gone through the raw data from the laboratory reports of five chemicals for which the lab judgement was different from the final judgement of the Validation Management Team (see Table 2 in Report of the JaCVAM initiative international validation study of the in vivo rodent alkaline Comet assay for the detection of genotoxic carcinogens: the 5th (definitive) phase -1st step, dated November 19, 2012, draft version-3). It should be acknowledged that such an exercise was not possible for most of the data set because no historical data were available.

Both the biological relevance and the lab's reported historical control data were taken into account. With these criteria, conclusions are different from the final judgement in three out of the five chemicals (see hereafter).

a) Acrylonitrile:

Stomach: No increases in the dosed group compared to controls.

Liver: Significant increase in Dunnett's test (highest dose compared to controls) and linear trend test, however the levels were well within the 95% confidence interval of historical controls. Relatively weak increases compared to the controls (the highest dose was in absolute values only 2.8 % tail DNA higher compared to the control group). This small increase is not biological relevant.
Since the values of the dosed groups are within the historical control range, I would conclude that Acrylonitrile is negative in the liver.

b) Sodium Arsenite:

Liver: Lab O's conclusion is equivocal/negative ("equivocal in liver tissue in the Comet assay. However, the mean % tail DNA value for the dose group showing a statistically significant increase was very close to laboratory historical vehicle control data").
The historical control data were: n=35, mean±SD: 2.36±1.04.In the report the 95% Confidence interval was not reported, however by calculations of the SD it would roughly be: 0.28-4.44.The top dose gave a mean response of % tail DNA of 2.3% (statistically significantly different from the control. That is well within the 95% confidence interval. Moreover, the top dose increase in absolute values compared to the control group was only 1% tail DNA (2.3% tail DNA vs. 1.3%). This very minor increase I do not find biological relevant.

My conclusion would be that on basis of the test results sodium arsenite is negative.
It was tested positive by lab M (I cannot find the lab report anywhere). The final judgement is that it is equivocal. If one laboratory (lab M) tested it positive and lab O's result are negative, my conclusion would also be equivocal.

c) Thioacetamide

Stomach: the % tail DNA showed a significant effect at the medium and high dose compared to the control group. There was also a significant linear trend. The values of % tail DNA in the medium and high dose group were outside (higher) than the 95% confidence interval for historical controls. The increase in % tail DNA was observed to be concomitant with an increase in % "clouded" cells, and with limited histopathological changes associated with tissue toxicity in two animals in the high dose group. The laboratory's conclusion was that the increase in % tail DNA was likely to be due to toxicity related activity and not solely due to genotoxicity.

Liver: Histopathological analysis showed toxicity in the dosed groups. The labs conclusion was that increases in % tail DNA was concomitant with effects in the liver and therefore could not be attributed solely to genotoxic activity.

My conclusion: The increase in of % tail DNA in the dosed groups compared to the control group was minimal (in absolute values less than 1%) and there was no statistical significance compared to the control. Based on these facts alone (not considering the histopathological findings nor the historical control range) I would conclude that there was no effect in the liver.
The results from the stomach are more difficult. There was only limited hiostopathological changes in the high dose group, not in the medium group, which was also statistically significantly different (% tail DNA) from controls (and outside the historical control range). The mean % of clouds was higher in the medium group compared to the control group (14.42 vs. 7.8). However, the historical control range of mean % clouds ranges from 3.5 to 20 (minimum and maximum values of n=35), the 95% confidence interval was not reported, that is a pity, because that would give us valuable information. So even if there are more "cloud cells" in the medium dose group, I would conclude that the results from the stomach are equivocal. Because there were no histopathological changes in the medium group. Clouded cells may express toxicity but without histopathological changes, I am not fully convinced.

d) *t*-Butylhydroquinone

Liver: Significant increase in Dunnett's test and linear trend test, however the levels were within historical controls. Relatively weak increases compared to the controls (the highest dose was in absolute values only 3.3% tail DNA higher compared to the control group). And the mean value of % tail DNA in the high dose group was within the historical control range (111 animals: minimum-maximum: 0.29-8.31). It is a pity that the 95% confidence interval for historical control range was not reported, that gives much more valuable information compared to mean and minimum and maximum values.
My conclusion: Negative.

Stomach: There was statistical significance compared to the control with the low dose group and the medium dose group but not the high dose group. There was no linear trend. The low dose group gave the highest response and that was in absolute values 3.06% higher than the control group which is a small increase. The values in the low dose group and medium dose group were within the historical control range. Again, It is a pity that the 95% confidence interval for historical control range was not reported, that gives much more valuable information compared to mean and minimum and maximum values.
My conclusion: Negative.

e) 2,4-Diaminotoluene

Stomach cells: No increase in % tail DNA compared to the controls.

Liver cells: A dose related increase (linear trend test significant). All three dose groups were statistically significantly different from controls. There were slight to moderate histopathological changes in all three dose groups.

The vehicle control dose group was very low (mean value of 0.64% tail DNA). The absolute increases in % tail DNA in the dosed groups compared to controls were low (maximum 2.41%). I would say that it is not biologically relevant. Moreover the % tails DNA in the dosed groups were well within the 95% confidence interval for historical control values. I would conclude that there is no effect in the liver.

Summary table of Stat, Final and Own judgements

Chemical	Organ	Lab judgement	Stat Judgement	Final judgement	My Judgement
Acrylonitrile (107-13-1)	Liver Stomach	Negative Negative	Increase Decrease	Positive in liver	Negative in liver and stomach
Sodium Arsenite (7784-46-5)	Liver Stomach	Stomach (Lab O): negative Liver (Lab O): equivocal/negative Liver (Lab M): positive	Liver (Lab O): Equivocal Liver (Lab M): Equivocal	Equivocal	Negative in stomach and liver based on Lab O's data. (I could not find Lab M's report).
Thioacetamide (62-55-5)	Liver Stomach	Negative Negative	Equivocal (liver)	Equivocal	Liver: Negative Stomach: equivocal
t-Butylhydroquinone	Liver Stomach	Negative Negative	Increase in liver Stomach: equivocal	Positive	Liver: Negative Stomach Negative
2,4-Diaminotoluene	Liver	Negative	Increase	Positive	Liver: Negative

Subgroup 4 report - Appendix 2

Following proposition of a member of SG PR4, it was deemed necessary to recall how and why statistical analysis was used in the validation report?

Rationale for choosing statistical evaluation versus historical data or other criteria in the validation report:

In the validation report (phase IV) WMT explained that "Because many participating laboratories did not have extensive historical negative control data with the exact protocol being used, statistical analysis was considered to be the most appropriate way to determine a positive response in this validation trial." In most cases the statistical analysis are in agreement with the laboratory data interpretation. Only in rare cases the conclusion of the laboratories disagreed with the statistical analysis applied by WMT because despite the statistical significance the values remained within the distribution of the negative historical data (t-butylhydroquinone, acrylonitrile), or because cytotoxicity was observed. Nevertheless it is difficult to know which data set have been effectively compared to the laboratory historical data in the validation study, and to evaluate what would have been the added value of the use of historical data for data interpretation.

However the by default use of statistical analysis in the JaCVAM validation study (for the reasons explained in the report, see above) does not mean that negative historical control data could not be recommended for data interpretation in the guideline, when there are available. It should be emphasised that the laboratories will have to conduct studies (at least on a limited number of animals for ethical reasons) before the implementation of the assay and in order to demonstrate the laboratory proficiency (to be described in the guideline).

Selection of statistical evaluation and parameters in the validation report

The rationale for the selection of the statistical analysis used in the validation report is "Dunnett's test (two-sided, $p<0.05$ and linear trend (two-sided, $p<0.05$) were applied to effect (Diff) in the groups of coded test chemicals. Two-sided because both increases and decreases in the Comet parameter could be detected." Effect (Diff) was first selected "… as the more appropriate for comparison of variation/reproducibility … Therefore Effect (diff) is mainly used to reveal the results of data analysis, in this document." Effect is defined as a difference designated as Effect (difference) or Effect (ratio) of a mean of an Estimate between a negative control group and a treatment group.

While the use of Effect could be appropriate for intra- and inter-laboratory comparison the reason why Effect and not Estimate (Mean of DNA% for each animal) was also used for data interpretation should be clarified. Moreover it should be clarified why means and not medians are used.

Subgroup 5 Report

Specific charge question: Assess the sensitivity and specificity of the Comet assay calculated based on 40 chemicals in Phase 4-2, based on knowledge of species or tissue specificity, mode of action and other genotoxicity results

Birgit Mertens, Dan D. Levy, Fabrice Nesslany, Veronique Thybaud

We reviewed the data presented in the prevalidation and validation JaCVAM reports and in the publications of Rothfuss *et al.* and Bowen *et al.* to assess the sensitivity and specificity of the Comet assay for the identification of "genotoxic chemicals as potential predictors of rodent carcinogenicity". As the *in vivo* Comet assay is proposed as an alternative follow-up assay for the *in vivo* rodent UDS assay, data obtained in the validation study with the *in vivo* Comet assay were compared with those available in literature (data provided in the validation report) for the *in vivo* UDS assay. The comet and UDS assays both measure DNA damage, although in different ways. It is valid to ask how accurately each of them predict genetic damage which results from DNA damage. It is equally valid to ask how each of them predict the outcome of the rodent cancer bioassay. The Validation Management Team (VMT) has also chosen to combine those two endpoints to assess how comet predicts hybrid endpoints: genotoxic carcinogens, non-genotoxic carcinogens, genotoxic non-carcinogens and non-genotoxic non-carcinogens. For completeness, data obtained in the validation study with the *in vivo* Comet assay were also compared with those available in literature (data provided in the validation report) for the *in vivo* micronucleus test.

	Comet +	**Comet +*.**	**UDS +**
Genotoxic carcinogens	**12/18 (67%)**	**7/10 (70%)**	**7/10 (70%)**
Genotoxins[$]	**13/23 (57)**	**8/12 (67)**	**9/12 (75)**
Carcinogens	**12/25 (48)**	**7/12 (58)**	**7/12 (58)**
Non-genotoxins[$]	**1/15 (6.7)**	**0/3 (0)**	**0/3 (0)**
Non-genotoxic non-carcinogens	**1/8 (1.3)**	**0/1 (0)**	**0/1 (0)**

	Comet +	**Comet +*.**	**MN +**
Genotoxic carcinogens	**12/18 (67%)**	**10/17 (59%)**	**13/17 (76%)**
Genotoxins[$]	**13/23 (57)**	**11/21 (52)**	**15/21 (71)**
Carcinogens	**12/25 (48)**	**10/23 (43)**	**13/23 (57)**
Non-genotoxins[$]	**1/15 (6.7)**	**1/11 (9)**	**0/11 (0)**
Non-genotoxic non-carcinogens	**1/8 (1.3)**	**1/5 (20)**	**0/5 (0)**

*All ratios for the Comet assay are based on VMT calls of a positive or negative result. Equivocal results not counted. [$]These categories are combinations of those described by the VMT in the validation report: "Genotoxins" combines their "Genotoxic carcinogens" with their "genotoxic non-carcinogens". "Non-genotoxins" combines their "non-genotoxic carcinogens" with their "non-genotoxic non-carcinogens". *examines only chemicals for which there is data from both assays.*

Based on the evaluation of these particular chemicals the three assays appear to have similar sensitivity and specificity towards genotoxins or carcinogens as defined by the VMT.

In the JaCVAM trial, 10 carcinogens called genotoxic by the VMT and for which UDS results are available in literature were tested. Of these 10 compounds 6 or 7/10 were positive in the Comet assay, depending on whether the lab or the VMT call on acrylonitrile is accepted. Seven out of 10 of these

compounds were also positive in literature reports of the UDS assay. The two assays were discordant for both 2-AAF and 1,3 DCP but in opposite ways so the overall concordance between rodent carcinogenicity results and Comet is identical based on this data set. The 3 carcinogens described as non-genotoxic used in the validation study and for which UDS results are available displayed negative results in both the Comet assay and the UDS test.

In the validation described by Rothfuss *et al*. 5 additional compounds were tested for which UDS results were also reported from the literature. One compound was negative in both assays, one positive in both assays and 3 negative in UDS but positive in Comet: gemifloxacine, phenobarbital and acrylamide. The comet results for the first two compounds were somewhat surprising since they were not expected to be genotoxic. However, as discussed in the paper, those two compounds are considered as able to induce DNA strand-breaks via indirect mechanisms, topoisomerase inhibition and oxidative stress respectively. Furthermore, it should also be noted that comet detection of acrylamide was inconsistent between 4 laboratories during the early phases of the JaCVAM pre-validation study. When data of the two studies are combined, data from literature reports of UDS are available for 18 chemicals examined in the two validation studies which are called 'genotoxic' by the VMT. Comet identified 12/18 and UDS 8/18. This is not an overwhelming difference, especially since the accuracy of at least two of the positive comet calls is questionable and the denominators are small. Thus comet might be slightly more sensitive than UDS at picking up genotoxic compounds and the assays are of about equal sensitivity at identifying genotoxic carcinogens.

However, several important remarks can be made:

1. Tissues investigated

Based on the validation report the three assays appear to have similar sensitivity and specificity towards genotoxins or carcinogens as defined by the VMT. In contrast to a previous publication (Kirkland and Speit, 2008), the Comet assay was thus not more predictive than the *in vivo* UDS test. One explanation might be that JaCVAM trials were systematically performed on liver and stomach. Several compounds have other target organs than liver and stomach, but these were not included in the validation study. Consequently, whether the *in vivo* comet assay performed on a larger number of organs is more predictive than the *in vivo* UDS assay which only examines damage in liver was beyond the scope of these validation studies. Data from other studies should be used to select the organs including data from toxicokinetic studies, sub-acute or chronic studies or from studies with compounds that belong to the same chemical class. A special sentence on this issue should be included in the introduction of the test guideline. It should also be noted that experimental conditions (at least for the preparation of nuclei) and criteria (e.g. target tail intensity) would potentially have to be adapted for the other tissues as highlighted by the difference between liver and stomach experimental conditions and criteria. Furthermore, some compounds have specific mechanisms of genotoxic action that are more difficult to detect in the Comet assay without modifying the protocol (e.g. addition of exogenous DNA repair enzymes such as OGG1 when cells are being processed). This requires knowledge or prediction of the mechanism of DNA damage (i.e. some metals and hydroquinone which may act via oxidizidation of other entities which then damage DNA).

Consequently, the subgroup notes that use of the assay beyond the procedures described in the test guideline may be useful to obtain more complete information about chemical hazard.

2. Acceptance and data interpretation criteria

When evaluating the results of genotoxicity tests, several acceptance criteria are considered including the use of the range of historical negative control data. However, in the JaCVAM trials, the historical control range was not taken into account by the VMT. Unless the proposed guideline is changed to differ from all

other genotox test guidelines (and provides a basis for that statement), the criteria for a positive result should be applied consistently. It is particularly true for the Comet Assay where normal ranges are defined for negative controls (1-8 % for liver and 1 - 20 % for stomach). Consequently, the result of acrylonitrile should be judged negative (as was done by the laboratory) or at best equivocal since the results were all within the laboratory historical control range. The same holds true for sodium arsenite considering the fact that the mean at the highest dose is almost equal to the mean of the historical negative control of the lab. Although rodents are known to be a poor model for arsenic carcinogenesis, this would certainly not be known in advance of testing. Sensitivity to metals is particularly important since they are outside of the domain of applicability of bacterial mutagenesis assays. The impact on whether or not to use the historical control range was further illustrated by the example of t-butylhydroquinone. As the compound induced a statistically significant effect, the result was considered positive by the VMT whereas the laboratory judged it negative because the highest exposure induced an increase that was just outside of the actual range.

The way by which cytotoxicity should be taken into account when evaluating a result is not clear either. For example, the laboratory call for thioacetamide was overruled because the laboratory was concerned about the reliability of the results due to signs of cytotoxicity. Clear guidance on how to use hedgehog measurements and the results of histopathology in conjunction with assay calls is lacking.

In general, clear guidance on assay acceptability criteria and data interpretation (statistical analysis including dose relationship, comparison with the distribution of historical negative controls, homogeneity between slides and animals) would be very useful.

3. Definition of the different categories

The VTM divides the test chemicals used in phase 4-2 of the validation study into four different categories: genotoxic carcinogens, genotoxic non-carcinogens, non-genotoxic carcinogens, and non-genotoxic non-carcinogens, in which genotoxicity is defined as a positive result in the Ames test or in standard *in vivo* genotoxicity tests such as the bone-marrow micronucleus assay. However, assigning a compound to one of these four categories is often complicated as also illustrated by the discussion of the results in the validation report. Massive amounts of information are required to determine the mode of action for carcinogenicity and has been done thoroughly for very few chemicals. Classification of carcinogens as "genotoxic" or "non-genotoxic" based on the proposed criteria is inappropriate and can cause problems when interpreting the results. This is illustrated in the validation report by the discussion developed to explain the unexpected results. The rationale for discordant results is well-presented and in general convincing. The issue is how this could be applied to unknown compounds. Recommendation on information to consider for better data interpretation (negative or positive) would be useful in the guideline and introduction document.

Furthermore, the VMT states that for genotoxic non-carcinogens 'negative results' will be preferred, but clear criteria why these negative results are expected are lacking. Within this category, 9-Aminoacridine is also misclassified as a non-carcinogen. While it is true that "there is no evidence that it is carcinogenic" in the absence of a bioassay it is unfair to group it with compounds for which bioassays have been run and found to be negative. A recent study of a large number of drugs demonstrated that many intercalating agents which do not covalently react with DNA can nonetheless be positive in bioassays. It is thus not clear that this compound belongs in any of the categories being used to assess expectations of the comet result.

The subgroup considers that the three weeks allocated for review of such a large amount of previously unreleased data representing over 6 years of study makes it difficult to perform a thorough data review and an accurate evaluation of sensitivity and specificity of the in vivo comet assay for rodent carcinogens.

As a result, and considering the purpose of the validation identification of “genotoxic chemicals as potential predictors of rodent carcinogenicity” it is recommended consolidating the categories into “rodent carcinogens” and “non rodent carcinogens” noting that many carcinogens act though non-genotoxic mechanisms.