Mutat Res Gen Tox En 839 (2019) 21–35

Contents lists available at ScienceDirect

Mutat Res Gen Tox En

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

A comparison of transgenic rodent mutation and *in vivo* comet assay responses for 91 chemicals

David Kirkland[a,*], Dan D. Levy[b], Matthew J. LeBaron[c], Marilyn J. Aardema[d], Carol Beevers[e], Javed Bhalli[f], George R. Douglas[g], Patricia A. Escobar[h], Christopher S. Farabaugh[i], Melanie Guerard[j], George E. Johnson[k], Rohan Kulkarni[f], Frank Le Curieux[l], Alexandra S. Long[g], Jasmin Lott[m], David P. Lovell[n], Mirjam Luijten[o], Francesco Marchetti[g], John J. Nicolette[p], Stefan Pfuhler[q], Daniel J. Roberts[i], Leon F. Stankowski Jr.[i], Veronique Thybaud[r], Sandy K. Weiner[s], Andrew Williams[g], Kristine L. Witt[t], Robert Young[f]

[a] *Kirkland Consulting, PO Box 79, Tadcaster LS24 0AS, UK*
[b] *US Food and Drug Administration Center for Food Safety and Applied Nutrition, College Park, MD, USA*
[c] *The Dow Chemical Company, Toxicology & Environmental Research & Consulting, Midland, MI, USA*
[d] *Marilyn Aardema Consulting LLC, 5315 Oakbrook Dr., Fairfield, OH 45014, USA*
[e] *Exponent International Ltd, Harrogate, UK*
[f] *MilliporeSigma, BioReliance Toxicology Testing Services, Rockville, MD, USA*
[g] *Environmental Health Science and Research Bureau, Health Canada, Ottawa, K1A 0K9, Canada*
[h] *Merck & Co. Inc., West Point, PA 19486, USA*
[i] *Charles River Laboratories, Skokie, IL, USA*
[j] *Roche Innovation Center Basel, pRed, F. Hoffmann-La Roche Ltd., Basel, Switzerland*
[k] *Swansea University Medical School, Swansea, UK*
[l] *European Chemicals Agency, Helsinki, Finland*
[m] *Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riß, Germany*
[n] *St George's Medical School, University of London, London, UK*
[o] *Centre for Health Protection, National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands*
[p] *AbbVie, Inc., Pre-clinical Safety, North Chicago, Illinois, USA*
[q] *Procter & Gamble, Global Product Stewardship, Mason, OH 45040, USA*
[r] *Sanofi, Vitry-sur-Seine, France*
[s] *Janssen Research & Development, Spring House, PA 19477, USA*
[t] *National Institute of Environmental Health Sciences/Division of the National Toxicology Program, Research Triangle Park, NC, USA*

ARTICLE INFO

Keywords:
Genotoxicity *in vivo*
Mutagens
DNA damage
Transgenic rodents
Risk assessment

ABSTRACT

A database of 91 chemicals with published data from both transgenic rodent mutation (TGR) and rodent comet assays has been compiled. The objective was to compare the sensitivity of the two assays for detecting genotoxicity. Critical aspects of study design and results were tabulated for each dataset. There were fewer datasets from rats than mice, particularly for the TGR assay, and therefore, results from both species were combined for further analysis. TGR and comet responses were compared in liver and bone marrow (the most commonly studied tissues), and in stomach and colon evaluated either separately or in combination with other GI tract segments. Overall positive, negative, or equivocal test results were assessed for each chemical across the tissues examined in the TGR and comet assays using two approaches: 1) overall calls based on weight of evidence (WoE) and expert judgement, and 2) curation of the data based on *a priori* acceptability criteria prior to deriving final tissue specific calls. Since the database contains a high prevalence of positive results, overall agreement between the assays was determined using statistics adjusted for prevalence (using AC1 and PABAK). These coefficients showed fair or moderate to good agreement for liver and the GI tract (predominantly stomach and colon data) using WoE, reduced agreement for stomach and colon evaluated separately using data curation, and poor or no agreement for bone marrow using both the WoE and data curation approaches. Confidence in these results is higher for liver than for the other tissues, for which there were less data. Our analysis finds that comet and TGR

Abbreviations: CA, chromosomal aberration; MN, micronucleus or micronuclei; UDS, unscheduled DNA synthesis; Carc, carcinogenicity; TGR, transgenic rodent mutation

* Corresponding author.

https://doi.org/10.1016/j.mrgentox.2019.01.007
Received 11 November 2018; Received in revised form 14 January 2019; Accepted 16 January 2019
Available online 18 January 2019

generally identify the same compounds (mainly potent mutagens) as genotoxic in liver, stomach and colon, but not in bone marrow. However, the current database content precluded drawing assay concordance conclusions for weak mutagens and non-DNA reactive chemicals.

1. Introduction

In the early decades of regulatory genotoxicity testing, the most commonly used *in vivo* somatic cell tests supported by the Organisation for Economic Co-operation and Development (OECD) guidelines only examined effects in bone marrow, namely induction of chromosomal aberrations (CA; OECD Test Guideline [TG] 475, [1]) or micronuclei (MN; OECD TG 474 [2]). The mouse spot test (OECD TG 484 [3]) was available to measure mutations in different tissues, but was performed in a limited number of laboratories and used large numbers of animals; TG484 was deleted in 2014. When the unscheduled DNA synthesis (UDS) test was developed [4], it allowed for the detection of genotoxic effects in rodent liver. This test was also developed into an OECD guideline (OECD TG 486 [5]) and was widely used to supplement the bone marrow MN and CA assays, being particularly useful at detecting hepatocarcinogens. However, tests to measure genotoxic effects in other tissues were not widely available.

In recent years, two test systems that allow determination of genotoxic effects in multiple tissues have been developed into OECD guidelines – the transgenic rodent mutation (TGR) assay, originally adopted in 2011 but updated in 2013 (OECD TG 488, [6]) and the *in vivo* alkaline comet assay (OECD TG 489, [7]). The development of the TGR OECD guideline was supported by the publication of Lambert et al. [8] and the OECD Detailed Review Paper [9], and for the comet assay by a collaborative multi-laboratory validation trial (see [10,11]). Both assays have their strengths and limitations:

- The comet assay is an “indicator” test, as not all strand breaks will necessarily result in mutation, whereas the TGR assay measures an “apical” endpoint, i.e. a mutation.
- The comet assay can be conducted in species and strains routinely used for regulatory testing, and can therefore be integrated in other *in vivo* toxicity tests, which has the advantage of reducing animal usage. However, by its very nature, the TGR assay uses specific, genetically modified rodent strains, which are not always readily available and can be costly.
- In the TGR assay, virtually any tissue can be taken and frozen for subsequent analysis whereas, it is preferable to analyse fresh tissue in the comet assay [12,13]. Currently there are no recommended protocols for the comet assay in frozen tissues, although laboratories are working on improving the reliability of data from frozen tissues, including germ cells (see [14,15]).

Mutations in germ cells can be readily detected in the TGR assay with certain protocol modifications [16], whereas the standard alkaline comet assay is not recommended for measurement of genotoxic effects in mature germ cells. As indicated in TG 489, the protocol for conducting the alkaline comet assay in sperm requires further standardization [12]. Furthermore, the recommended exposure regimen and sample collection in the guideline is not adequate for assessing genotoxicity in mature sperm.

Whilst the TGR and comet assays are available for measuring genotoxic effects in multiple rodent tissues, there has been no systematic comparison of their performance in detecting genotoxicity *per se*. A working group of the Genetic Toxicology Technical Committee (GTTC), part of the Health and Environmental Sciences Institute (HESI), was therefore established to contribute to resolving this deficiency. This manuscript describes the establishment of a database containing 91 chemicals with published TGR and *in vivo* comet assay data in at least one common tissue, and presents the outcome of various analyses addressing the above question. The results of the analysis may impact regulatory toxicity testing, although this database is relatively small and additional studies are needed to refine the comparison and further clarify the sensitivity and specificity of TGR and comet assay results. A key question the working group wished to address was whether it is necessary to follow up a chemical that has been shown to be mutagenic *in vitro* with an *in vivo* gene mutation assay (TGR test), or whether it might be equally acceptable to conduct an *in vivo* comet assay. Creation of this database is a first step towards answering that and many other outstanding questions about how best to use these assays.

2. Construction of the database

As a starting point, the TRAiD (Transgenic Rodent Assay Information Database [8]; this can be accessed by sending a request to Dr Paul A. White, Health Canada, at paul.white@canada.ca), a comprehensive database of TGR dose-response data, was examined. The TRAiD contains thousands of experimental records of transgenic mutation results from different experimental treatments (chemicals, radiation, pharmaceuticals, infectious agents, and mixtures), the majority of which were conducted in the 1990s and early 2000s before the standardization of study design when OECD TG 488 was first published in 2011 [6]. The TRAiD database that was used for this exercise (v. 7.0, updated September 2015) was originally compiled by Health Canada, has been updated with papers published since 2011, and has been carefully reviewed by members of the GTTC. Chemicals in the updated TRAiD database that had been tested in the *in vivo* comet assay were selected for comparative analysis, but alternative sources were also used to construct the database. Additional *in vivo* comet assay data available from previous database publications [17–19], along with information from the following websites, were used:

- NTP (https://ntp.niehs.nih.gov)
- PubMed (https://www.ncbi.nlm.nih.gov/pubmed)
- Toxline (https://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?TOXLINE)
- CCRIS (https://toxnet.nlm.nih.gov/newtoxnet/ccris.htm)
- European Chemicals Agency (https://echa.europa.eu)
- European Food Safety Authority (EFSA, www.efsa.europa.euwww.efsa.europa.eu).

During the search additional chemicals were identified that had both TGR and *in vivo* comet data, and these were added to the database. Where data were available on different forms of the same chemical (e.g. free base and salt) the results were combined into a single entry, since in aqueous biological systems it would be expected that the different forms would behave in a similar way. This was done for:

- Diepoxybutane/1,2,3,4-diepoxybutane/1,2,3,4-DL-diepoxybutane
- Hydrazine, hydrazine monohydrate, hydrazine sulfate, and hydrazine 2HCl
- 2,6-Diaminotoluene (free base and 2HCl salt)
- Sodium saccharin/saccharin
- 2-Amino-3-methylimidazo[4,5-f]quinoline (IQ - free base and HCl salt)
- 2-Amino-1-methyl-6-phenylimidazo(4,5-b)pyridine (PhiP - free base and HCl salt)
- o-Anisidine (free base and HCl salt)
- Aristolochic acids I and II (and sodium salt)
- Bleomycin/bleomycin HCl/bleomycin sulphate
- Cyclophosphamide (free base and monohydrate)

- 2,4-Diaminotoluene (free base and 2HCl salt)
- Procarbazine/procarbazine HCl
- 4-Aminobiphenyl (free base + HCl salt)
- Phenobarbital/phenobarbital sodium.

Initially the group identified 92 chemicals for which there were data from both comet and TGR assays in at least one tissue. The chemicals were then divided up amongst 21 workgroup members for more in-depth searching and analysis. Through this process additional publications were found and added to the data entry spreadsheets. Detailed evaluations of each publication were made and tabulated in one of approximately 20 columns in the data entry spreadsheets. Each column prompted entry of data or comments on key aspects of methodology such as numbers of animals, dose levels, dose duration, sampling times, quality of the data, and interpretation of the results. After the initial data entry, each sheet was reviewed by alternate co-authors to minimise bias and ensure consistency across chemicals. The TGR and comet data entry spreadsheets are shown respectively in Supplementary Tables 1 and 2. When sorting through the compiled results, only 91 chemicals, shown in Supplementary Table 3, had data for both endpoints, in at least one common tissue. 2,3-epoxypropyl neodecanoate had dropped out of the comparison as there were only alkaline elution (and not comet) data available.

3. Methods of analysis of the database

Following the quality control steps described above, two different approaches were considered to evaluate the data collected in the database. In the first approach, one experienced individual reviewed all of the collected data and allocated overall calls based on weight of evidence (WoE) and expert judgement. These overall calls were applied to each set of comet or TGR results in each tissue to derive tissue specific calls, which were subsequently reviewed by the entire workgroup. We labelled this the" WoE approach". In the second approach, teams of experienced individuals developed *a priori* acceptability criteria and then reviewed each set of comet and TGR results, and only derived the tissue specific calls on data that met the acceptability criteria. We labelled the second approach the "data curation approach". An example of the differences in the two approaches is that some data which were used in the WoE approach but given low weight were excluded in the data curation approach. The two approaches are described in more detail below.

3.1. Approach based on all collected data using weight of evidence (WoE) and expert judgement

The results for all 91 chemicals in each of the two assays were summarised as a brief narrative shown in Supplementary Table 3 in the columns headed "Summary of results" (columns G and V). There were fewer results in rats than in mice, particularly for the TGR assay where the vast majority of results were in mice:

- For TGR, 13 chemicals were studied in rats only, 60 were in mice only, and 18 were in both rats and mice;
- For comet, 16 chemicals were studied in rats only, 23 in mice only, and 52 in both rats and mice.

Therefore, in order to provide sufficient data for comparison, and since either species is acceptable for *in vivo* genotoxicity testing, data from rats and mice were combined, and the same rationale was applied to data obtained in males and females. The most widely studied tissues in both assays were liver and bone marrow. However, there were several reasons (given below) to also look at responses in various segments of the GI tract:

- OECD guideline 488 [6] says "*The rationale for tissue collection should be defined clearly. Since it is possible to study mutation induction in virtually any tissue, the selection of tissues to be collected should be based upon the reason for conducting the study and any existing mutagenicity, carcinogenicity or toxicity data for the chemical under investigation. Important factors for consideration should include the route of administration (based on likely human exposure route(s)), the predicted tissue distribution, and the possible mechanism of action. In the absence of any background information, several somatic tissues as may be of interest should be collected. These should represent rapidly proliferating, slowly proliferating and site of contact tissues*".
- OECD guideline 489 [7] says "*In some cases examination of a site of direct contact (for example, for orally-administered substances the glandular stomach or duodenum/jejunum, or for inhaled substances the lungs) may be most relevant."*
- In ECHA's 2015 REACH progress report [20] it is stated that the default requirement for oral comet assays was to "*….analyse two site-of-contact tissues (glandular stomach and duodenum/jejunum), in addition to liver*".
- EFSA's Minimum Criteria for the acceptance of *in vivo* alkaline Comet Assay Reports [21] requires one site of contact tissue (e.g. stomach or duodenum) in addition to liver to be examined by default in the assay, due to the primary route of human exposure for chemicals in food.

Most GI tract results were in stomach and colon but there were also a few results in oesophagus, forestomach, ileum, jejunum, duodenum, caecum, and small intestine. Therefore, since the objective of the OECD, ECHA and EFSA statements above seemed to be to assess site-of-contact effects, it was decided to look at responses in any section of the GI tract in this WoE comparison of TGR and comet responses. In many cases, the same chemical was reported to give different results in different publications in the same tissue, or had been tested by different routes of administration. We are as yet unable to determine the reasons for divergent results within the same tissue for an assay. Differences in length of dosing, dose level, sampling time, or route of administration could have been responsible but we could not precisely associate these factors with individual cases of divergent results. Therefore, a WoE evaluation using expert judgement was employed to derive overall calls for the TGR and comet assay responses in liver, bone marrow and GI tract, and is described below. These overall calls were made initially by a single individual and the overall analysis was then reviewed by the entire workgroup. These overall calls fell into four categories, positive (+), negative (-), equivocal (E) and inconclusive (I), and are shown in the columns identified as "WoE" in Supplementary Table 3.

The following principles were applied in arriving at these overall WoE calls. Although the methods described in OECD test guidelines are considered sufficiently robust that negative results can be accepted with confidence, many of the published TGR and comet assay studies in the database were conducted before the respective OECD guidelines were adopted. Even studies published since the guidelines were adopted often deviate from the recommended experimental designs. Moreover, many of the published studies were designed to answer questions other than those for which the regulatory Test Guidelines were developed. Thus, data from studies which were fit for their intended purpose were not always easy to compare to the other studies in our database for the purpose of our analyses. Therefore, scientific judgement had to be applied to the data in order to decide whether the results could be accepted as valid. For example, a clear positive result from a study that used a shorter dosing period, fewer sampling times, or smaller group sizes could be more easily accepted as positive than a negative result from similar non-guideline study designs. Also, negative results from a recent comet or TGR study conducted to current OECD guidelines were considered more conclusive than a negative result from a study that did not comply with current test guidelines or previously recommended best practices (e.g. [22,23]), or from studies that contained little information on the test methods used. Negative results from such studies

where the robustness of the protocol was questionable were therefore allocated an "inconclusive" call. Where conflicting results were reported in the different publications, the numbers of + and – calls were not considered as important as the quality and robustness of the individual tests. Consideration was also given to whether the results had been obtained in different studies or from different publications. In those cases where the same study results were reported in different (e.g. original and review) publications, the results were considered as a single entry.

For the overall calls by the WoE approach the following criteria were adopted:

- An overall **positive (+)** call was given if: (1) there was only one study available and it provided clear evidence of a positive response regardless of whether it was conducted in mice or rats, males or females.; (2) there was clear weight of evidence from more than one study (i.e. multiple positive results outweighing a small number of negative results) or (3) if a substance was positive in one species or sex and negative in the other, and it was clear that systemic exposures (e.g. higher dose levels, evidence of target organ toxicity in the positive study) would have been greater in the positive than in the negative study. Where authors had reported a "weak positive" response, these were considered positive.
- An overall **negative (-)** call was given when the experimental design met or was close to the requirements of the current OECD guidelines, or recommended best practices (e.g. [22,23]) were fulfilled, and there was no evidence of a positive or equivocal response in any study. A negative call was made only if there was evidence that the test substance would have reached the target tissue, otherwise it was considered inconclusive (see below).
- An overall **equivocal (E)** call was given if results were ambiguous, doubtful, inconsistent (e.g. positive and negative results) within a study, or unclear (e.g. a dose-related increase in effect but the magnitude of response did not achieve statistical significance or exceed normal control ranges, and no independent repeat experiment was done to verify the response and produce a clear conclusion). An "**E**" call was also used where there were both positive and negative findings across different studies of apparently acceptable quality and validity, and where the weight of evidence did not allow a clear positive or negative overall outcome to be concluded.
- An overall **inconclusive (I)** call was given in the case of negative or unclear results, where no firm conclusion could be made in terms of the requirements of the current OECD guidelines or recommended best practices. This applied most frequently to negative TGR studies where animals had only been dosed for a few days (instead of the recommended 28 days), or where negative results were obtained but there was no proof of target cell exposure *in vivo* (for example, no analysis of test substance in blood or tissue and no target organ toxicity). Most comet studies were performed to designs that were close to the current OECD TG, and so there were fewer inconclusive calls. It was concluded that an "**I**" call should be considered as an inadequate test and treated as "no valid data".

The vast majority of GI tract results were from stomach and colon, and these were analysed separately with the data curation methodology (see section 3.2). However, for this WoE analysis it was assumed that the objective of investigating responses in the GI tract was to determine site-of-contact effects. Thus, although the majority of published results were in stomach and colon, results from other segments of the GI tract (oesophagus, forestomach, duodenum, jejunum, ileum, caecum, small intestine, and large intestine) were also considered valid for site-of-contact effects. Therefore, in this WoE approach, results from all of the different segments of GI tract were combined for the purposes of comparison. The overall calls described above therefore had to be modified slightly for GI tract as follows:

- Overall **positive (+)** calls were given where one or more segments of the GI tract gave a positive result in more than one publication, even if some other segments of the GI tract that were studied less frequently were negative.
- An overall **positive (+)** call was also given if GI tract tissues were positive using oral administration but negative using intraperitoneal (i.p.) administration (since oral administration would generally result in greater and more direct exposure to the cells that are scraped from the inner lining of the GI tract to obtain samples for the comet assay; moreover, oral gavage is preferred over i.p. administration in the OECD test guideline). However, if the only data were from i.p. administration, the positive GI tract results were accepted.
- **Equivocal (E)** calls were given as described above for bone marrow and liver data if results from a single study were ambiguous, doubtful, inconsistent or unclear. Equivocal calls were also given where there were approximately equal numbers of positive and negative results between studies of the same GI tract segment conducted in different laboratories or in different segments of the GI tract (e.g. equal numbers of positive results in stomach and negative results in colon) whether from the same or from different laboratories. For example, 2-acetylaminofluorene was positive in colon of mice but negative in stomach of mice by oral dosing, positive in stomach and colon of mice by i.p. dosing, but negative in stomach of rats by both oral and i.p. dosing, and so an overall call of E was allocated. This pattern of mixed results was found more frequently with the comet assay, but less so for the TGR assay since there were few chemicals with TGR data in both stomach and colon.
- An overall **negative (-)** call was given when all GI tract tissues evaluated were negative, even if only the i.p. route of administration was used, since i.p. administration would usually be expected to result in some GI tract exposure either by absorption into the hepatic portal system followed by enterohepatic recirculation, or by systemic bioavailability as a result of by-passing first-pass metabolism (see [24]).
- As described above, **inconclusive (I)** calls were given, for example, where treatment duration was too short in the TGR but gave a negative result, or, in one case, where a negative result was reported for a comet assay that used only a single dose level.

As in any evaluation, equivocal (E) responses form a unique category. It can be argued that, because the response is not clearly negative, "E" calls should be included in the positive response category. However, it can also be argued that they should be excluded from the positive category because they did not meet the established criteria for a positive response. For the purpose of the WoE approach, "E" calls were considered as "no valid data" and therefore excluded from the analysis. Since inconclusive calls would not contribute to a comparison of TGR and comet responses they were also excluded. Thus, WoE analyses were made on overall "+" and "- "calls only.

3.2. Step-wise approach using detailed data curation methodology

3.2.1. Step 1: verifying quality of the database

As described above, the data used for this exercise were collected from all available sources in the literature, and it was evident that the quality of some of the studies was questionable. Hence, a detailed evaluation of the quality of the study protocol and the data in each paper across all 91 chemicals was conducted prior to accepting the data into the curated database. This was achieved by objectively assigning a numerical "quality" score to each study, using the *a priori* quality criteria listed in Table 1. These criteria were developed independently for each assay although consistency across assays was intended. These quality scores were applied equally across all data in the database using evaluation criteria developed prior to considering the impact of omitting a particular study from the overall analysis. This ensured that only selected high-quality data in the database would be used for the

allocation of overall calls and subsequent analyses.

The assignment of quality scores to the TGR and comet studies was carried out separately by two independent expert groups. For the TGR expert working group, two teams of experts reviewed and scored all studies. Any scoring discrepancies were discussed by all members of the TGR expert group and a consensus call was reached for all studies. For the comet expert working group, two individuals went through the database and independently assigned quality scores. When there was a discrepancy, the consensus call was reached by the entire expert working group.

criteria developed by the workgroup or in compliance with the applicable OECD test guideline.

3.2.2. Step 2: determining tissue-specific results from discordant data

Of the 91 chemicals that had both comet and TGR data, in 28 instances, the comet assay produced discordant results in specific tissues. Differences in rodent species, dose level, sex, dose duration, and/or route of administration were among the potential causes of dichotomous results. With the exception of dicyclanil[1], these differences were not weighted in our evaluation when making final calls on tissue specific results. This was justified by the understanding that a laboratory that evaluates a previously untested chemical would be unlikely to know this information while planning an *in vivo* genotoxicity study. Therefore, when discordant results were observed, the following criteria were employed to determine an overall tissue-specific result:

Table 1
Assignment of quality scores for comet and TGR papers using pre-determined criteria.

Quality Score	*A Priori* Maximum Scoring Criteria
0 Unusable	Papers that did not present data. Human data from occupational/epidemiologic sources.[C] Alkaline elution methods were used.[C] Data generated from *in vitro* or *ex vivo* dosing.[C]
1 Problematic (not used in the curated analysis)	Methods were omitted. Summary data, abstract, or duplicated data already reported. Group size was omitted or < 3. Lack of concurrent vehicle controls. Frequency and/or dose duration omitted. Unspecified route of administration. Potential confounding cytotoxicity.[C] Most of the fields on the data checking sheets were blank. Number of nuclei scored omitted or < 50 per animal.[C] Data obtained at an atypical necropsy timepoint (e.g., 8 hours).[C]
2 Trust positive Results (negative results inconclusive)	Highest dose(s) inadequate to demonstrate an effect. Only a single dose used.[TGR] Too few repeat doses (< 28 days[TGR] or no short or long exposure[C]). Unverifiable data cited in the review of Sasaki et al. (2000).[C] Sample time after last dose not stated or not close to 3 days.[TGR]
3 Trust Results	Methods compliant with the OECD TG, or recommendations from the JaCVAM (pre)validation[C]or IWGT (Tice et al[C]).

[C] Specific to comet papers.
[TGR] Specific to TGR papers.

To assign quality scores, the criteria in Table 1 must have been met for the study being evaluated. Studies given higher quality scores met the criteria for all lower categories as well. For example, route of administration must be stated for studies scored both 2 and 3. A list of the final quality scores used in the databases is presented in Fig. 1. The final quality score for each paper was then assigned to the responses in each reported tissue. For example, for a quality 2 paper that had positive liver and negative bone marrow findings, the liver would be considered positive. However, bone marrow would be considered inconclusive.

It should be noted that the authors were aware that tissue toxicity may be a confounding factor. For chemicals that induced positive responses across different dose levels in the same study, or where the results across different papers in the same tissue agreed with one another, we did not believe cytotoxicity was likely to be a confounding factor. However, where the publications did contain histological or other evidence of cytotoxic effects, and this could have confounded a positive response, particularly if this was only at the top dose level, the findings were downgraded in that tissue.

The assignment of the *a priori* quality criteria listed in Table 1, and the application of these scores as described above, led to the following actions:

- Score of 0: Unusable; no data, abstract only, or incorrect assay.
- Score of 1: Problematic; correct assay but methodology limitations prevented study acceptance.
- Score of 2: Sub optimal; positive test results are considered valid, but limitations in study design diminished the reliability of negative test results.
- Score of 3: Valid; all test results are reliable. The studies were conducted using a protocol which conformed to standardised

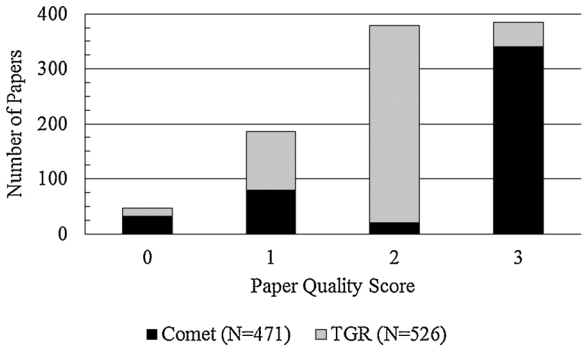

Fig. 1. Literature quality of all of the available comet and transgenic rodent assay data in the literature. A total of 91 chemicals with data in the same tissue (s) across both assays were included in the final database.

[1] Dicyclanil was a special case because of the unusually rich dataset which consistently shows a sex difference. Not only do cancer bioassays demonstrate this [25], but our database captured comet and TGR data in both sexes that were consistent with the sensitivity of females and not males.

- Papers with a quality score of 0 and 1 were not considered when determining overall tissue calls.
- Papers with a quality score of 2 were considered in the case of positive results only. Negative results from those papers were called inconclusive and not considered.
- For a clear positive or negative response, the majority (at least 70%) of papers of acceptable quality had the same result in that tissue.
- For remaining chemicals, on a tissue by tissue basis, equivocal results were assigned if the above criteria were not met. This decreased the number of chemicals with definitive results in both assays from 35 to 33 for liver, 16 to 10 for bone marrow, and 19 to 13 for GI tract (specifically, 5 chemicals had stomach data in both assays, and 8 chemicals had colon data in both assays).

For the chemicals with equivocal calls, it is unclear whether the assays were able to detect genotoxicity in that particular tissue. As mentioned, additional testing using current scientific protocols or alternate dose levels could clarify the genotoxic potential of these chemicals in the tissues studied. Hence, the analysis was conducted 3 separate ways. First, only *bona fide* positive and negative results were compared, mimicking the WoE approach. Then equivocal data were considered either positive or negative, to "bracket" the outcome of such further investigations.

Although the GI tract was evaluated as a single site of contact tissue using the WoE approach, it was decided to attempt to analyse stomach and colon separately, after data curation, for the following reasons:

- For TGR, only 3 chemicals were studied in both stomach and colon, which is too few to draw conclusions on concordance of response across tissues.
- For comet, 43 chemicals were studied in both stomach and colon and only 29 (67%) were concordant (both tissues positive or negative).

The results mentioned above for stomach and colon exemplify the diverse anatomical and physiological properties of these tissues, i.e. the parenchymal cells are different between the two tissues, chemicals may be modified by the time they reach distal GI tissue impacting actual exposure levels, and duration of exposure due to enterohepatic circulation may be different in the two compartments. Hence, the comet and TGR subgroups analysed stomach and colon separately using the curated database and did not try to combine stomach and colon results into a single category such as "GI tract".

3.3. Statistical approaches

Since the primary question that stimulated the compilation and analysis of this database was "do the TGR and comet assays detect positive responses for the same chemicals (i.e. do they have the same sensitivity to genotoxicants)", then sensitivity is the analogous descriptive statistic to evaluate. The TGR results were therefore taken as the reference point (as it measures an apical endpoint), and for each of the selected tissues (liver, bone marrow, GI tract or separately stomach and colon) the proportion of TGR-positive chemicals that were also positive in the comet assay (i.e. sensitivity) was determined; 95% confidence intervals were calculated using the modified Wald method [26].

For the WoE analysis, only overall positive and overall negative results were compiled into 2 × 2 tables. Overall calls of E and I were excluded. This same approach was utilized for the curated data analysis. However, after curation many positive and negative results had moved to equivocal due to their quality scores. Hence, equivocal results were included in a sub-analysis, where they were considered either positive or negative for each tissue, except for colon, which had no equivocal data.

The analysis of the 2 × 2 tables was carried out in the same way for both WoE and curated data approaches. There is no single statistic that can capture all the aspects of a 2 × 2 table. The traditional way to assess agreement between two different measures (or scorers) has been to use Cohen's kappa statistic [27]. However, it is well known that kappa has limitations, especially when the marginal totals of the tables are not symmetrical (see [28]), as is the case for this data set, which has a high prevalence of positive results in both assays and very few negative results in either assay (see tables of results discussed below). Flight and Julious [28] note that: (1) for high values of concordance, low values of kappa can be recorded; and (2) asymmetric tables have a higher kappa than symmetric tables (the kappa "paradoxes"). Gwet [29] developed the AC1 statistic because of such limitations with the kappa. The AC1 between two or multiple raters is defined as the conditional probability that two randomly selected raters agree given that there is no agreement by chance. Flight and Julious [28] recommend using the Bias Index (BI), the Prevalence Index (PI), MAK (maximum attainable kappa) and the PABAK (the prevalence and bias adjusted kappa), and argue that the PABAK [30] is less influenced by the prevalence and the distribution of the marginal totals. They caution against relying on just the kappa value which, in certain circumstances, can be low even when there is good concordance, and instead recommend the use of all the statistics in the interpretation. The PI is the difference between the probability of each category occurring. From a standard 2 × 2 table (Table 2) it is calculated as PI = (a–d)/N. The lower the PI, the more balanced the table. Thus, larger values of PI result in a larger proportion of agreement than expected by chance and hence a smaller kappa value. The BI is a measure of the difference in proportions of times each rater classifies an object into category one. From a standard 2 × 2 table (Table 2) it is calculated as BI = (b–c)/N. Larger values of BI result in larger values of kappa. PI and BI can help explain when the paradoxes of kappa are likely to occur.

Therefore, as recommended by Flight and Julious [28], for each 2 × 2 table, sensitivity, PI, BI, kappa, PABAK and AC1 have been calculated, and the agreement/disagreement between the different measures is discussed. There are a number of different categorisations of agreements (and other measures of association) that can be applied to these metrics. One categorisation that can be used to interpret the level of agreement for kappa, which is applicable to AC1 [31] and should therefore also be applicable to PABAK, was published by Landis and Koch [32] and is shown in Table 3. Although these are useful categories they should not be over interpreted. Moreover, it has been suggested that these categorisations should be applied to the lower 95% confidence levels of each statistic because that also takes into account the size of the sample. It should be noted that for these types of statistics, sample sizes ≤ 95 are small (which is the case here, as can be seen from the tables to be discussed below), and therefore the confidence

Table 2
Standard 2 × 2 table.

	Cat. 1	Cat. 2	Total
Cat. 1	a	b	g_1
Cat. 2	c	d	g_2
Total	f_1	f_2	N

Table 3
Interpretations of the kappa statistic (from Landis & Koch, [32]).

Kappa	Agreement
< 0.20	Poor
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.8	Good
0.81-1.00	Very good

intervals, which are calculated using n, are wide.

Power simulations were also performed with AC1, which, like kappa and PABAK, ranges from -1 (complete disagreement between the 2 endpoints) through zero (no agreement) to +1 (complete agreement). The null hypothesis for these power simulations is that AC1 is zero and that there is no agreement. If there is very high agreement (e.g. near 100%) between two tests for the nature (e.g. genotoxicity) of a population of chemicals then only small samples from this population are needed to have a high (90%) probability of showing that an estimate of the level of agreement from the sample is significantly different from zero. If the 'true' level of agreement is lower, then larger sample sizes are needed to have 90% power.

The simulations showed that if the comparisons between TGR and comet were based on 11 observations, then there is 90% power that if the true value of AC1 is different from zero it will be detected as significant. As will be seen from the tables discussed below, with the exception of bone marrow (excluding equivocal calls), stomach, and colon using curated data, there were definitive comparisons for at least 11 chemicals in each tissue. This means that there is a high probability (> 90%) that if there is appreciable agreement between TGR and comet it will be detected as being statistically significant. For instance, if the association between the assays was good (e.g. AC1 = 0.6), overall calls from at least 25 chemicals in a tissue would be needed to provide 90% power that the agreement differs from zero, and if the association between the assays was only moderate (e.g. AC1 = 0.4), overall calls from at least 60 chemicals in a tissue would have been needed to provide 90% power. When the association was very poor (e.g. AC1 = 0.06) then sample sizes of over 4000 would be needed to have approximately 90% power. It is reasonable to assume that similar power would be achieved for the other statistics. It is therefore concluded that the majority of datasets analysed had sufficient power to reject the null hypothesis (that there was no agreement between the assays) when any of the statistical approaches showed moderate to good agreement.

4. Results

4.1. Analysis of overall TGR and comet responses in liver, GI tract and bone marrow based on all publications in the database using weight of evidence (WoE) and expert judgement

The 2 × 2 tables for liver, GI tract and bone marrow are shown in Table 4, with statistical analyses reported in Table 5. The following comments can be made:

- The prevalence indices reflect the high number of positive results in both TGR and comet assays, particularly in liver and GI tract. Therefore, AC1 and PABAK analyses with corrections for prevalence would be considered more informative than kappa alone.
- The bias indices are low indicating no particular bias in these data sets.
- For liver and GI tract, the AC1 and PABAK statistics showed moderate to good agreement based on the central estimates, but only poor, fair or moderate agreement based on the lower confidence limits. However, the sensitivity values for liver and GI tract were high, even when the lower confidence intervals were considered. This means that the comet assay shows very high agreement with the TGR assay at detecting positive responses in liver and GI tract.
- For bone marrow, there was poor agreement between comet and TGR for each of the statistics – there was even some evidence of disagreement based on the lower confidence limits – and the sensitivity was low.

Chemicals for which overall calls of positive (+) or negative (-) could be made for liver, GI tract and bone marrow are listed in Table 6. There were overlapping data sets in liver for 35 chemicals, in GI tract for 19 chemicals, and in bone marrow for 15 chemicals. As expected, the number of chemicals with positive test results in TGR and/or comet exceeds the number of chemicals with negative test results, particularly for liver and GI tract. Whilst many concordant results were found across different species, routes of administration, treatment times and sampling times, which might be expected from a database containing so many recognized genotoxicants with broad-ranging activity, such factors may also contribute to the observed discordant results. However, with the small numbers of discordant results, and with such varying protocols amongst the published papers, explanations of why data were discordant are not possible at this time.

4.2. Impact of data curation on the database

Detailed results of the data curation are presented within the TGR and comet databases in Supplementary Tables 1 and 2, respectively. These tables include all the parameters captured during the initial data checking exercises, the outcome of the experiment in each tissue sampled, and the classifications made during the data curation.

The comet database has 471 lines of data, with each line reporting results for one chemical in a single study of male or female test animals. During curation of the database, studies were excluded from the analysis for a variety of reasons (Table 1). In a few cases, comet data described in an abstract could not be confirmed based on reading the entire paper. More often deficiencies in assay conduct or reporting led to such low confidence in the validity of the results that they were excluded. The curated comet database included 358 lines of data for the 91 chemicals. Data for 26 of the chemicals consisted of only a single study. Removing almost a quarter of the data from the analysis had minimal impact on the number of chemicals analysed. This was because many of the studies removed for poor quality were of extremely potent alkylating agents like cyclophosphamide or similarly potent mutagens, and the remaining excluded studies were for chemicals for which there were other, higher quality studies in the database. Analysing results from multiple studies for a chemical in the database was not straightforward. For example, two lines of data could report results from two experimental groups from the same experiment (e.g. 90-day repeat dose protocol compared to 3-day exposure protocol both in the same species) or from different laboratories using the same or different species. The

Table 4
Comparison of TGR and comet assay results using the WoE approach.

A: Liver Data

Liver		TGR mutation (Mice and/or Rats)		
	Result	+	–	Total
Comet (Mice and/or Rats)	+	23	4	27
	–	3	5	8
	Total	26	9	35

B: GI Tract Data

GI tract		TGR mutation (Mice and/or Rats)		
	Result	+	–	Total
Comet (Mice and/or Rats)	+	14	3	17
	–	0	2	2
	Total	14	5	19

C: Bone Marrow Data

Bone marrow		TGR mutation (Mice and/or Rats)		
	Result	+	–	Total
Comet (Mice and/or Rats)	+	7	1	8
	–	5	2	7
	Total	12	3	15

Table 5
Statistical analyses of results for liver, GI tract and bone marrow (WoE approach).

Tissue	Sensitivity[#](95% CI)	PI	BI	Kappa (95% CI)	PABAK (95% CI)	AC1 (95% CI)
Liver	0.82(0.70-0.97)	0.51	0.03	0.46[c](0.13[e]-0.79)	0.60[c](0.31[d]-0.83)	0.68[b](0.44[c]-0.93)
GI tract	1.00(0.75-1.00)	0.63	0.16	0.50[c](0.11[e]-0.88)	0.68[b](0.37[d]-1.00)	0.77[b](0.49[c]-1.00)
Bone marrow	0.58(0.32-0.81)	0.33	−0.27	0.17[e](-0.25[g]-0.59)	0.20[e](-0.33[g]-0.73)	0.28[d](-0.29[g]-0.85)

PI = prevalence index.
BI = bias index.
CI = confidence interval.
a = very good agreement (0.81–1.00).
[#] = ability of the comet assay to detect TGR positives.
[b] = good agreement (0.61-0.80).
[c] = moderate agreement (0.41-0.60).
[d] = fair agreement (0.21-0.40).
[e] = poor agreement (< 0.20).
[g] = fair disagreement (-0.21 to -0.40).

comet data were fairly evenly divided between rats (174/358, 49%) and mice with a few studies in hamsters, guinea pigs, etc. The most common routes of administration were oral/gavage (52%) and intraperitoneal injection (38%). Less common routes of administration included drinking water/feed, dermal application, inhalation, intravenous injection, or intratracheal instillation of particles. Liver was sampled in the majority of studies (68%); other commonly sampled tissues included stomach (44%), bone marrow (40%), kidney (39%), lung (35%), bladder (30%), brain (30%), colon (27%), and blood (24%), with < 4% of studies conducted in other tissues. After curation, which led to the exclusion of certain poor quality studies, approximately 90 lines of data remained from studies conducted by Sasaki et al. published between 1996 and 2003 (see Supplementary Table 3 for references).

The TGR database has 525 lines of data, each line generally reporting results for one chemical in a single group of test animals. Similar to comet data, during curation of the database, studies were removed from the analysis for a variety of reasons (see Table 1). The curated database included 404 lines of data for the 91 chemicals. As for the comet database, removal of those data from the TGR database had minimal impact on the number of chemicals analysed and for the same reasons as above: there were many studies of potent alkylating agents (various nitrosamines, ethyl- and methyl methanesulfonate) and other potent genotoxicants, or of less potent chemicals all of which were also represented in the database by higher quality studies. As mentioned, a large fraction of the studies was conducted before protocols were standardised by publication of the OECD Test Guideline, and many of these consisted of < 28 days of dosing. Negative results from such studies were interpreted as less reliable. Data were limited to only a single study for 26 of the chemicals, the same number as for the comet database. Unlike the comet data, the database consists mostly of data generated from mice, as there were relatively little rat data (74/405, 18%) and transgenic guinea pigs, hamsters, etc., are not commercially available. Routes of administration were more variable than for comet. The most common routes of administration were intraperitoneal injection (39%) with a relatively lower proportion by oral/gavage (24%), and less so in drinking water (10%), and feed (17%). Liver was sampled in the majority of studies (58%). Other tissues were sampled less frequently for TGR mutations than for comets. Listed in the same order as for comet in the preceding paragraph they include stomach (8%), bone marrow (17%), kidney (13%), lung (19%), bladder (4%), brain (4%), and colon (12%). Spleen (12%) was sampled much more often for TGR mutations than for comets, as were testes and sperm, which were not closely tracked in our analyses since they are difficult to evaluate using the comet assay.

4.3. Analysis of overall TGR and comet responses in liver, stomach, colon and bone marrow after database curation

Throughout the curation process, many calls switched either from equivocal to positive or negative, or vice versa. There were no instances in which a chemical that was judged as positive or negative using the WoE approach had the call reversed (i.e. became negative or positive, respectively) after data curation. Where there were multiple studies for the same compound, the TGR study results were frequently in agreement, and therefore there was only a single overall call of equivocal in the curated TGR data set across all tissues. Conversely, there was less agreement between different studies for the same chemical using the comet assay, which resulted in 50 overall equivocal calls in the curated comet data set across all tissues (see Supplementary Table 2 for details).

The 2 × 2 tables for liver, stomach, colon and bone marrow using positive and negative results are shown in Table 7. When equivocal results were considered as either negative or positive, separate 2 × 2 tables were created (Table 8), and the statistical analyses across both sets of data are reported in Table 9.

The following comments can be made when comparing definitive positive and negative results:

- The prevalence indices reflect the high number of positive results in both TGR and comet assays particularly in liver, stomach and colon. Therefore, AC1 and PABAK analyses with corrections for prevalence would be considered more informative than kappa alone.
- The bias indices are low, indicating no particular bias in these data sets.
- For liver, stomach and colon, the AC1 and PABAK statistics showed moderate to good agreement based on the central estimates, but only poor, fair or moderate agreement based on the lower confidence limits. However, the sensitivity values for liver, stomach and colon were high, even when the lower confidence limits were considered. This means that the comet assay shows good agreement with the TGR assay in detecting positive responses.
- For bone marrow, there was very poor or no agreement between comet and TGR for each of the statistics – there was even some evidence of disagreement – and the sensitivity of the comet assay in detecting TGR-positive responses in this tissue was low.

A list of chemicals for which overall calls of positive (+) or negative (-) could be made for liver, stomach/colon, and bone marrow are listed in Table 10. Due to attrition of lower quality data, the total number of chemical comparisons was lower post data curation (than after WoE evaluation), particularly when analysing stomach and colon as separate tissues. Thus, in liver there were overall + or - calls for 33 chemicals, in stomach for 5 chemicals, in colon for 8 chemicals, and in bone marrow for 10 chemicals. However, these numbers generally increased when E

Table 6
Lists of chemicals with overall positive/negative results in both TGR and comet assays (WoE approach).

A: Liver

TGR +/ comet + (23)	TGR +/comet – (3)	TGR -/ comet + (4)	TGR -/comet – (5)
1-Ethyl-1-nitrosourea	4-Aminobiphenyl (free base + HCL salt)	3-Chloro-4-(dichloromethyl)-5-hydroxy-2(5 H)-furanone (AKA MX)	Allura Red AC (Food Red 40)
2,4-Diaminotoluene (free base and 2HCl salt)	Chlorambucil	Acetaminophen	Chloroform
2-Acetylaminofluorene	Methyleugenol	Benzene	Hydroquinone
2-Amino-3,8-dimethylimidazo[4,5-f]quinoxaline (MeIQx)		Ethyl acrylate	Methylphenidate HCl
2-Amino-3-methylimidazo[4,5-f]quinoline (IQ - free base & HCL salt)			Nitrite, sodium
5,9-Dimethyldibenzo[c,g]carbazole			
Acrylamide			
Aflatoxin B1			
Aristolochic acids I and II (and sodium salt)			
beta-Propiolactone			
Cyclophosphamide (free base and monohydrate)			
Cyproterone acetate			
Dichlorvos			
Dicyclanil			
Ethyl methanesulphonate			
Methyl methanesulphonate			
N-Nitrosodiethylamine			
N-Nitrosodimethylamine			
N-Nitrosodipropylamine			
N-Nitroso-N-methylurea			
N-Nitrosopyrrolidine			
Urethane			
Wyeth 14,643			

B: GI tract

TGR +/ comet + (14)	TGR +/comet – (0)	TGR -/ comet + (3)	TGR -/comet – (2)
1-Ethyl-1-nitrosourea		Chromium (hexavalent)	Hydroquinone
2-Amino-1-methyl-6-phenylimidazo(4,5-b)pyridine (PhiP - free base & HCl salt)		Ethyl acrylate	Nitrite, sodium
2-Amino-3,4-dimethylimidazo[4,5-f]quinoline (MeIQ)		N-Nitrosodimethylamine	
2-Amino-3,8-dimethylimidazo[4,5-f]quinoxaline (MeIQx)			
2-Amino-3-methylimidazo[4,5-f]quinoline (IQ - free base & HCL salt)			
4-Nitroquinoline-N-oxide			
7,12-Dimethylbenz[a]anthracene			
Benzo[a]pyrene			
beta-Propiolactone			
C.I. Solvent yellow 3 (o-aminoazotoluene)			
Methyl methanesulphonate			
N-Methyl-N'-nitro-N-nitrosoguanidine			
N-Nitroso-N-methylurea			
Urethane			

C: Bone marrow

TGR +/ comet + (7)	TGR +/comet – (5)	TGR -/ comet + (1)	TGR -/comet – (2)
4-Nitroquinoline-N-oxide	1,3-Butadiene	N-Nitrosodipropylamine	2-Amino-1-methyl-6-phenylimidazo(4,5-b)pyridine (PhiP - free base & HCl salt)
7,12-Dimethylbenz[a]anthracene	2-Amino-3,4-dimethylimidazo[4,5-f] quinoline (MeIQ)		Acrylonitrile
Acrylamide	Benzo[a]pyrene		
Ethyl methanesulphonate	Chlorambucil		
Mitomycin C	Urethane		
N-Nitroso-N-methylurea			
Procarbazine HCl (natulan)/procarbazine			

calls were counted as positive or negative (Tables 7 vs 8). Similar to the WoE approach, the number of chemicals with positive test results in TGR and/or comet exceeds the number of chemicals with negative test results for liver, stomach and colon; however, there were an equal number of positives and negatives for bone marrow. It should be noted that induction of comet damage in bone marrow was detected least frequently in a compilation of comet data with over 200 rodent carcinogens, almost all of which were tested in 8 tissues including liver, stomach, colon and bone marrow [33]. This apparent insensitivity in bone marrow may be due to the nature of this tissue as a haematopoietic tissue with a high cell turnover rate. DNA damage may cause delayed cell cycle (particularly for high dose levels), whereas a stable mutation may not. Hence, the analysis of bone marrow data may be impacted by the inability of the standard comet assay to efficiently detect DNA damage in that tissue. It is possible that sampling times for the comet assay may need to be 'optimized' for bone marrow relative to other tissues.

When analysing equivocal calls as either positive or negative the following comments can be made:

- Considering the equivocal calls as positive results did not qualitatively change the statistical outcome for liver. Considering them

Table 7
Comparison of TGR and comet assay results using the data curation approach.

A: Liver Data

		TGR mutation (Mice and/or Rats)		
	Result	+	–	Total
Comet (Mice and/or Rats)	+	24	1	25
	–	4	4	8
	Total	28	5	33

B: Stomach Data

		TGR mutation (Mice and/or Rats)		
	Result	+	–	Total
Comet (Mice and/or Rats)	+	4	1	5
	–	0	0	0
	Total	4	1	5

C: Colon Data

		TGR mutation (Mice and/or Rats)		
	Result	+	–	Total
Comet (Mice and/or Rats)	+	8	0	8
	–	0	0	0
	Total	8	0	8

D: Bone Marrow Data

		TGR mutation (Mice and/or Rats)		
	Result	+	–	Total
Comet (Mice and/or Rats)	+	5	0	5
	–	5	0	5
	Total	10	0	10

negative consistently lowered all of the measures of statistical agreement (Table 9). This suggests that the equivocal calls might describe compounds which cause liver DNA damage that is challenging to detect using the comet assay (e.g. lack of reproducibility across laboratories or across experimental variables such as route of administration).

- Considering equivocal calls in the bone marrow as positive increased the level of agreement between comet and TGR outcomes: whereas, considering them as negative reduced the level of agreement as for the liver.
- In contrast, the opposite was observed when equivocal calls were included in the statistical analyses for the stomach. These statistical results showed better agreement when equivocal results were considered negative, which is contrary to the expectation stated above. This result could be a consequence of the very small stomach data set (Tables 7 and 8).
- "Bracketing" the results by counting all equivocal calls as either positive or negative suggests that resolving each equivocal call into a positive or negative call would not likely change the conclusions drawn about results in liver, but could have led to a conclusion of some level of agreement between the two assays in bone marrow, or to a conclusion of lack of agreement between the two assays in stomach.
- While the curation exercise resulted in a more refined data set, it had the consequence of reducing the size of data sets for analysis; when already small datasets get even smaller, confidence in their predictive power declines.
- A further consideration for strength of the curated stomach and colon data sets is that there are no negative comet data (Table 7), which increases the chance that prevalence bias reduces confidence in the result.

5. Discussion and conclusions

The key question that drove the construction of this database was whether it is necessary to follow up a chemical that has been shown to be mutagenic *in vitro* with an *in vivo* gene mutation assay (TGR test), or whether it might be equally acceptable to conduct an *in vivo* comet assay. The decision on which of these two approaches to use is related to the ultimate purpose of the *in vivo* follow-up test. Accordingly, if the intention is to confirm *in vitro* gene mutation activity in terms of genotoxicity in general, then the comet assay is an acceptable choice; however, if the intention is to confirm specifically that an *in vitro* gene mutagen induces *in vivo* gene mutations *per se*, then the TGR assay is the more appropriate test.

In general, from the analyses performed herein using either WoE evaluation of all the publications or evaluation of a subset of *a priori* curated data, the comet assay appears to yield similar results to the TGR assay in liver. The WoE analysis of aggregate GI tract data also showed statistical agreement for the comet and TGR assay outcomes. However, while analysing the curated stomach and colon data separately with the same statistical approaches showed a similar outcome, the confidence intervals were wider, thus reducing the reliability of the conclusions for those tissues. For bone marrow, neither the WoE nor data curation exercises showed sufficient agreement between the two assays.

There were a number of results considered to be equivocal (E), i.e. not clearly positive or negative. There is no general agreement as to how, or whether, such results should be included in an evaluation of test performances. The data curation results were analysed without the "E" calls, and then separately with "E" calls included with positives and

Table 8
Comparison of TGR and comet assay data for rodent tissues when equivocal calls were included in the analysis post data curation.

A: With equivocal calls counted as positive

Liver		TGR mutation (Mice and/or Rats)		
	Result	+/E	–	Total
Comet (Mice and/or Rats)	+/E	31	2	33
	–	5	4	9
	Total	36	6	42
Stomach		TGR mutation (Mice and/or Rats)		
	Result	+/E	–	Total
Comet (Mice and/or Rats)	+/E	4	3	7
	–	0	0	0
	Total	4	3	7
Bone Marrow		TGR mutation (Mice and/or Rats)		
	Result	+/E	–	Total
Comet (Mice and/or Rats)	+/E	8	0	8
	–	5	0	5
	Total	13	0	13

B: with equivocal calls counted as negative

Liver		TGR mutation (Mice and/or Rats)		
	Result	+	-/E	Total
Comet (Mice and/or Rats)	+	24	1	25
	-/E	11	6	17
	Total	35	7	42
Stomach		TGR mutation (Mice and/or Rats)		
	Result	+	-/E	Total
Comet (Mice and/or Rats)	+	4	1	5
	-/E	0	2	2
	Total	4	3	7
Bone Marrow		TGR mutation (Mice and/or Rats)		
	Result	+	-/E	Total
Comet (Mice and/or Rats)	+	5	0	5
	-/E	8	0	8
	Total	13	0	13

then with “E” calls included as negatives, to bracket a range of probable measures of agreement if results for these chemicals could be definitively resolved as either positive or negative. For liver, inclusion of “E” calls as positive did not result in a change in the level of agreement; however, including the “E” calls as negative decreased PABAK and AC1 without altering the overall conclusion (the level of agreement was reduced from “good” to “moderate”). Furthermore, including “E” calls as positive for bone marrow increased the level of AC1 agreement from “poor” to “moderate”. However, the stomach data behaved differently from the above two tissues; inclusion of “E” calls as positive decreased (rather than increasing or not changing) PABAK and AC1, resulting in poor agreement between results in the two assays. Because including the “E” calls as either positive or negative did not result in a consistent change in agreement among tissues, at this time it would not be prudent to view the equivocal results as “likely positive” or “likely negative”. Further analyses and additional experiments are required to identify the factors responsible for this variability.

This database was assembled with chemicals that had been tested in either TGR or comet assays *in vivo*. While most of the chemicals that had data across both assays in a specific tissue were positive in each of the two assays, the workgroup did not analyse combinations of results from multiple tissues, particularly tissues other than liver, stomach, colon and bone marrow, into a whole animal “overall call” for each individual assay. Thus, the only comparisons in this analysis are between comet and TGR results for the same chemical in the same tissue. There was a high prevalence of positive calls for chemicals with results in both TGR and comet assays for liver and GI tract. This pattern was less pronounced for bone marrow. This high prevalence of positive calls was absent or less obvious when reviewing all the data (not just from chemicals with both TGR and comet results) from an individual tissue in an individual assay (i.e. positive and negative results were more evenly distributed among all 73 chemicals with liver comet data as opposed to the 35 chemicals with liver data in both comet and TGR assays). As a consequence, whilst it is possible to reach meaningful conclusions regarding the agreement of positive results (i.e. the sensitivity of the comet assay relative to the TGR assay) in liver and GI tract, it is not possible to reach meaningful conclusions on the agreement between the two endpoints for negative results. To achieve this goal, more chemicals with negative results in both the TGR and comet assays would be required. It is also possible that our database is biased in favour of potent genotoxicants, which tend to be more commonly used in published studies than found during regulatory testing. Potent genotoxicants are

Table 9
A comparison of statistical analyses for liver, stomach and bone marrow after data curation.

A: Liver

Table	Sensitivity[#](95% CI)	PI	BI	Kappa (95% CI)	PABAK (95% CI)	AC1 (95% CI)
+ve and –ve	0.86 (0.72-0.99)	0.61	0.09	0.53[c](0.20[e]-0.85)	0.70[b](0.45[c]-0.86)	0.78[b](0.57[c]-0.99)
E calls + ve	0.86 (0.74-0.97)	0.64	0.07	0.44[c](0.14[e]-0.73)	0.67[b](0.43[c]-0.86)	0.76[b](0.58[c]-0.95)
E calls -ve	0.69 (0.53-0.84)	0.43	0.24	0.35[d](0.09[e]-0.60)	0.43[c](0.14[e]-.71)	0.51[c](0.24[d]-0.79)

B: Stomach

Table	Sensitivity[#](95% CI)	PI	BI	Kappa (95% CI)	PABAK (95% CI)	AC1 (95% CI)
+ve and –ve	0.80 (0.35-1.00)	0.80	0.20	NA	0.60[c](-0.20[f]-1.00)	0.75[b](0.03[e]-1.00)
E calls + ve	0.57 (0.13-1.00)	0.57	0.43	NA	0.14[e](-0.43[h]-0.71)	0.35[d](-0.59[h]-1.00)
E calls -ve	1.00 (0.42[*]-1.00)	0.29	−0.14	0.70[b](-0.01[f]- 1.00)	0.71[b](0.14[e]-1.00)	0.74[b] (0.11[e]-1.00)

C: Colon

Table	Sensitivity[#](95% CI)	PI	BI	Kappa (95% CI)	PABAK (95% CI)	AC1 (95% CI)
+ve and –ve	1.00 (0.63[*]-1.00)	1.00	0.00	NA	1.00[a](0.63[*,b]-1.00)	1.00[a](0.63[*,b]-1.00)

D: Bone Marrow

Table	Sensitivity[#](95% CI)	PI	BI	Kappa (95% CI)	PABAK (95% CI)	AC1 (95% CI)
+ve and –ve	0.50 (0.14-0.86)	0.50	0.50	NA	0.00[f](-0.60[h]-0.60)	0.20[e](-0.60[h]-1.00)
E calls +ve	0.62 (0.31-0.92)	0.62	0.38	NA	0.23[d](-0.23[g]-0.69)	0.44[c](0.58[h]-0.95)
E calls -ve	0.38 (0.08-0.69)	0.39	0.62	NA	−0.23[g](-0.69[i]-0.34)	−0.07[e](-0.80[i]-0.65)

CI = confidence interval.
PI = prevalence index.
BI = bias index.
[#] = ability of the comet assay to detect TGR positives.
[*] = Clopper-Pearson Lower Limit.
[a] = very good agreement (0.81–1.00)f = some disagreement (0 to -0.20).
[b] = good agreement (0.61-0.80) g = fair disagreement (-0.21 to -0.40).
[c] = moderate agreement (0.41-0.60) h = moderate disagreement (-0.41 to –0.60).
[d] = fair agreement (0.21-0.40)i = substantial disagreement (-0.61 to -0.80).
[e] = poor agreement (< 0.20).
[f] = some disagreement (0 to -0.20).
[g] = fair disagreement (-0.21 to -0.40).
[h] = moderate disagreement (-0.41 to –0.60).
[i] = substantial disagreement (-0.61 to -0.80).

frequently screened out early in product development and thus our results may be less predictive of the outcome of less potent genotoxicants that are more commonly submitted for *in vivo* genotoxicity testing in a regulatory environment. Unfortunately, the proprietary data for such compounds is rarely published so it is difficult to establish the extent to which this bias would affect the choice of test.

The overall calls for the comet assay for two chemicals are worthy of specific comment since the overall conclusions reached in the above analyses were different from those reported in the JaCVAM trial [10]:

- 2-acetylaminofluorene (2-AAF) was judged as negative in the liver and stomach comet assays by the JaCVAM-organized international validation study [10]. However, this compound was positive in liver by the oral route as reported by some participating laboratories in the JaCVAM validation study [see 10]. Moreover, 2-AAF was positive in other publications following both oral and i.p. routes of administration (see references in Supplementary Table 3). Therefore, by both WoE and data curation approaches, we concluded that for liver, the comet assay was positive. Regarding stomach data, JaCVAM judged this compound negative [10]. However, our analysis included additional published data (see references in Supplementary Table 3) that resulted in a high level of discordance within the compiled 2-AFF stomach data, which precluded assignment of an overall positive or negative call in stomach. Hence, by both WoE and data curation approaches, 2-AAF was given an overall call of equivocal in stomach.
- 2,6-diaminotoluene was judged as positive by the JaCVAM-organized international validation study [10] in the liver comet assay. However, our survey included other published data (see references in Supplementary Table 3), using methods of adequate quality, that reported both positive and negative results in liver. Therefore, our overall call for the comet assay in liver, by both WoE and data curation approaches, was equivocal.

It is recognised that cytotoxicity (toxicity in the target tissue) could be a confounding factor in these assays and therefore warrants discussion. Firstly, in studies where positive comet/TGR responses were seen across several dose levels it is unlikely that tissue toxicity could explain all of the responses (particularly at middle and low doses where toxicity would have been much lower than at the highest dose). Secondly, when performing our expert judgments, most of the results across papers for a specific chemical agreed with one another across different dose levels, indicating a level of reproducibility that would not likely be due to extreme conditions such as tissue toxicity. Furthermore:

- For the TGR assay, while there are a few examples in the literature where toxicity in the form of induced cell proliferation allows DNA adducts, which would otherwise not be fixed as gene mutations (for the given dose and study design), to become full-fledged gene mutations, this is not regarded as a wide-spread problem with the assay. Rather, if anything, it increases the sensitivity of the assay to detect genotoxic agents (see [34,35]). It should be noted that there

Table 10
Lists of chemicals with overall positive /negative calls in both TGR and comet assays (data curation approach).

A: Liver

TGR +/ comet + (24)	TGR +/comet – (4)	TGR -/ comet + (1)	TGR -/comet – (4)
1-Ethyl-1-nitrosourea	4-Aminobiphenyl (free base + HCL salt)	Ethyl acrylate	Allura Red AC (Food Red 40)
2-Acetylaminofluorene	Chlorambucil		Chloroform
2-Amino-1-methyl-6-phenylimidazo(4,5-b)pyridine (PhiP - free base & HCl salt)	Chromium (hexavalent)		Hydroquinone
2-Amino-3,4-dimethylimidazo[4,5-f]quinoline (MeIQ)	Methyleugenol		Methylphenidate HCl
2-Amino-3,8-dimethylimidazo[4,5-f]quinoxaline (MeIQx)			
2-Amino-3-methylimidazo[4,5-f]quinoline (IQ - free base & HCL salt)			
5,9-Dimethyldibenzo[c,g]carbazole			
Acrylamide			
Aflatoxin B1			
Aristolochic acids I and II (and sodium salt)			
beta-Propiolactone			
Cisplatin			
Cyclophosphamide (free base and monohydrate)			
Cyproterone acetate			
Dichlorvos			
Dicyclanil			
N-Nitrosodiethylamine (diethylnitrosamine)			
N-Nitrosodimethylamine (dimethylnitrosamine)			
N-Nitrosodipropylamine [dipropylnitrosamine]			
N-Nitroso-N-methylurea			
N-Nitrosopyrrolidine			
Procarbazine HCl (natulan)/procarbazine			
Urethane			
Wyeth 14,643			

B: Stomach and colon

Stomach

TGR +/ comet + (4)	TGR +/comet – (0)	TGR -/ comet + (1)	TGR -/comet – (0)
4-Nitroquinoline-N-oxide		Ethyl acrylate	
Benzo[a]pyrene			
beta-Propiolactone			
N-Methyl-N'-nitro-N-nitrosoguanidine			

Colon

TGR +/ comet + (8)	TGR +/comet – (0)	TGR -/ comet + (0)	TGR -/comet – (0)
1-Ethyl-1-nitrosourea			
2-Amino-1-methyl-6-phenylimidazo(4,5-b)pyridine (PhiP - free base & HCl salt)			
2-Amino-3,4-dimethylimidazo[4,5-f]quinoline (MeIQ)			
2-Amino-3,8-dimethylimidazo[4,5-f]quinoxaline (MeIQx)			
2-Amino-3-methylimidazo[4,5-f]quinoline (IQ - free base & HCL salt)			
7,12-Dimethylbenz[a]anthracene			
Benzo[a]pyrene			
C.I. Solvent yellow 3 (o-aminoazotoluene)			

C: Bone marrow

TGR +/ comet + (5)	TGR +/comet – (5)	TGR -/ comet + (0)	TGR -/comet – (0)
4-Nitroquinoline-N-oxide	1,3-Butadiene		
Aristolochic acids I and II (and sodium salt)	2-Amino-3,4-dimethylimidazo[4,5-f]quinoline (MeIQ)		
Mitomycin C	Benzo[a]pyrene		
N-Nitroso-N-methylurea	Chlorambucil		
Procarbazine HCl (natulan)/procarbazine	Urethane		

is no requirement in the OECD test guideline to examine cytotoxicity in the tissues being examined, and as such, it is rarely conducted. A vast majority of the TGR studies in our database were conducted over the duration of 1-month, and de novo tissue proliferation could be a confounding factor. However, to the best of our knowledge, during 1-month of daily dose administration, observing tissue regeneration in liver or GI tract is a rare event, and this observation was not detailed in the discussion section of the papers included in the database.

- For the comet assay, it was rare to find reports of histopathology analysis in published papers before the last few years, yet the current OECD test guideline relied heavily on that early literature. We do not believe it would be appropriate to reject data in the literature for no other reason than there was no explicit measurement of organ toxicity. It was fairly rare to find older comet studies with more than one or two days of dosing, and it is unclear what data would be used to accurately predict toxicity in such an acute study. An acute toxicity study rarely provides much information about specific organ toxicity and it is unclear how the results from a longer-term repeat dose study would be used to predict a dose that was (in) appropriate for administration of the chemical one to three times over a day or two.

Hence, we believe that in the majority of the studies reviewed here, cytotoxicity was not a confounding factor. On the rare occasions we did see signs of toxicity, we explicitly used that as one of the criteria for

potential exclusion. Thus, in papers where histological findings or other cytotoxicity information was available (only a few papers), we downgraded the findings in that tissue. However, by and large we saw nothing that would lead us to conclude it was an important confounding factor. Therefore, overall, we do not believe the possible confounding effects of tissue toxicity invalidates the comparisons we have made or the conclusions we have reached.

Because the two OECD Test Guidelines were published within the last 10 years, and after most of the data leading to their adoption was collected, most of the available data (used in the current analysis) was collected using protocols not consistent with the OECD guidelines. For example, dosing periods tended to be too short for many TGR studies and histopathology was not used to evaluate cytotoxicity in many comet studies. The experts reviewing these studies identified those protocol deficiencies and indeed many results were removed during study collection, data curation and WoE analyses due to those factors. However, most of these data were non-GLP studies using varied protocols and those factors make the conclusions less reliable than would be drawn if sufficient results had been available from more recent, guideline compliant studies.

There were a number of results considered to be equivocal (E), i.e. not clearly positive or negative. There is no general agreement as to how, or whether, such results should be included in an evaluation of test performances. In order to simplify the analyses based on the total publications in the data entry spreadsheets, equivocal and inconclusive results were excluded from the WoE analyses and only positive and negative overall calls were analysed.

Analysis of TGR and comet results could be carried out for tissues other than those reported here by further interrogating the results in the final 91 chemical database, as well as the results tabulated in the data entry spreadsheets. Further analysis of how TGR and comet assays respond to *in vitro* gene mutagens and mutagenic (genotoxic) carcinogens is on-going, and will be reported elsewhere [36]. The outcome of those analyses may allow more specific recommendations on follow-up testing for regulatory purposes. Our database may also provide opportunities for additional analyses, such as reproducibility of comet and TGR data across laboratories, concordance of results across species, difference in results due to route of administration, and investigating the degree of agreement in responses across tissues (or combinations of tissues) other than those analysed here. Data curation, by removing studies of lower quality, will make such comparisons more reliable and less subject to distortion due to results from repeatedly testing the same potent genotoxicants. Such analyses may provide crucial insight into the choice of which assays to conduct to provide the most robust evaluation of genotoxicity in rodents.

The general conclusion that can be drawn from the analyses of this database is that the comet assay can detect potent genotoxicants that would otherwise be detected using the TGR assay in the liver, and GI tract; however, confidence in this GI tract association is reduced if separate regions are considered. In contrast, the bone marrow data suggests that one cannot expect the two assays to give the same result in every tissue. Assay concordance for genotoxicants which induce weak responses is less certain since our database contained fewer such chemicals. The conclusions contained herein will be of interest to regulatory agencies and laboratories during the process of selecting *in vivo* genotoxicity assays and tissues to test.

Declarations of interest

C.S. Farabaugh, D.J. Roberts and L.F. Stankowski Jr. are employed by Charles River Laboratories, which offers the comet assay commercially.

R. Kulkarni, J. Bhalli and R. Young are employees of MilliporeSigma, BioReliance Toxicology Testing Services, which offers TGR and comet assays commercially.

None of the other authors have interests to declare.

Disclaimer

This document represents the consensus of the authors' views expressed as individual scientists and does not necessarily represent the policies and procedures of their respective institutions.

Acknowledgments

The authors would like to thank Mugimane Manjanatha, Robert Heflich, Nan Mei, Wei Ding, Yan Li, Julia Kenny, Rosalie Elespuru, Jan van Benthem, Maik Schuler and Anne Doherty for their efforts in checking large amounts of published data. We would also like to thank HESI, and in particular Jenifer Tanir, Lauren Peel and Stan Parrish, for supporting the efforts of the working group members, and for providing the facilities for exchange of ideas and opinions.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.mrgentox.2019.01.007.

References

[1] OECD Guidelines for the Testing of Chemicals, Genetic Toxicology: Mammalian Bone Marrow Chromosomal Aberration Test, Organisation for Economic Co-operation and Development, Paris, 2016 TG475 adopted 29 July.
[2] OECD Guidelines for the Testing of Chemicals, Genetic Toxicology: Mammalian Erythrocyte Micronucleus Test, Organisation for Economic Co-operation and Development, Paris, 2016 TG474 adopted 29 July.
[3] OECD Guidelines for the Testing of Chemicals, Genetic Toxicology: Mouse Spot Test, Organisation for Economic Co-operation and Development, Paris, 1986 TG484 adopted 23 October.
[4] B.E. Butterworth, J. Ashby, E. Bermudez, D. Casciano, J. Mirsalis, G. Probst, G. Williams, A protocol and guide for the in vivo rat hepatocyte DNA-repair assay, Mutat. Res. 189 (1987) 123–133.
[5] OECD Guidelines for the Testing of Chemicals, Genetic Toxicology: Unscheduled DNA Synthesis (UDS) Test With Mammalian Liver Cells in Vivo, Organisation for Economic Co-operation and Development, Paris, 1997 TG486 adopted 21 July.
[6] OECD Guidelines for the Testing of Chemicals, Genetic Toxicology: Transgenic Rodent Somatic and Germ Cell Gene Mutation Assays, Organisation for Economic Co-operation and Development, Paris, 2013 TG488 adopted 28 July 2011, revised 26 July.
[7] OECD Guidelines for the Testing of Chemicals, Genetic Toxicology: in Vivo Mammalian Alkaline Comet Assay, Organisation for Economic Co-operation and Development, Paris, 2016 TG489 adopted 29 July.
[8] I.B. Lambert, T.M. Singer, S.E. Boucher, G.R. Douglas, Detailed review of transgenic rodent mutation assays, Mutat. Res. 590 (2005) 1–280.
[9] OECD Series on Testing and Assessment, Number 103, Detailed Review Paper on Transgenic Rodent Mutation Assays, Organisation for Economic Co-operation and Development, Paris, 2009 23 July.
[10] Y. Uno, H. Kojima, T. Omori, R. Corvi, M. Honma, L.M. Schechtman, R.R. Tice, C. Beevers, M. De Boeck, B. Burlinson, C.A. Hobbs, S. Kitamoto, A.R. Kraynak, J. McNamee, Y. Nakagawa, K. Pant, U. Plappert-Helbig, C. Priestley, H. Takasawa, K. Wada, U. Wirnitzer, N. Asano, P.A. Escobar, D. Lovell, T. Morita, M. Nakajima, Y. Ohno, M. Hayashi, JaCVAM-organized international validation study of the in vivo rodent alkaline comet assay for detection of genotoxic carcinogens: II. Summary of definitive validation study results, Mutat. Res. 786–788 (2015) 45–76.
[11] OECD Series on Testing and Assessment, Number 196, Report of the JaCVAM Initiative International Validation Studies of the in Vivo Rodent Alkaline Comet Assay for the Detection of Genotoxic Carcinogens, Organisation for Economic Co-operation and Development, Paris, 2015 15 July.
[12] G. Speit, M. Vasquez, A. Hartmann, The comet assay as an indicator test for germ cell genotoxicity, Mutat. Res. 681 (2009) 3–12.
[13] G. Speit, H. Kojima, B. Burlinson, A.R. Collins, P. Kasper, U. Plappert-Helbig, Y. Uno, M. Vasquez, C. Beevers, M. De Boeck, P.A. Escobar, S. Kitamoto, K. Pant, S. Pfuhler, J. Tanaka, D.D. Levy, Critical issues with the in vivo comet assay: a report of the comet assay working group in the 6th International Workshop on Genotoxicity Testing (IWGT), Mutat. Res. 783 (2015) 6–12.
[14] P. Jackson, K.S. Hougaard, A.M. Boisen, N.R. Jacobsen, K.A. Jensen, P. Moller, G. Brunborg, K.B. Gutzkow, O. Andersen, S. Loft, U. Vogel, H. Wallin, Pulmonary exposure to carbon black by inhalation or instillation in pregnant mice: effects on liver DNA strand breaks in dams and offspring, Nanotoxicology 6 (2012) 486–500.
[15] L. Recio, G.E. Kissling, C.A. Hobbs, K.L. Witt, Comparison of Comet assay dose–response for ethyl methanesulfonate using freshly prepared versus cryopreserved tissues, Environ. Mol. Mutagen. 53 (2012) 101–113.
[16] F. Marchetti, M.J. Aardema, C. Beevers, J. van Benthem, R. Godschalk, A. Williams, C.L. Yauk, R. Young, G.R. Douglas, Identifying germ cell mutagens using OECD test

guideline 488 (transgenic rodent somatic and germ cell gene mutation assays) and integration with somatic cell testing, Mutat. Res. 832-3 (2018) 7–18.
[17] D. Kirkland, L. Reeve, D. Gatehouse, P. Vanparys, A core in vitro genotoxicity battery comprising the Ames test plus the in vitro micronucleus test is sufficient to detect rodent carcinogens and in vivo genotoxins, Mutat. Res. 721 (2011) 27–73.
[18] D. Kirkland, E. Zeiger, F. Madia, R. Corvi, Can in vitro mammalian cell genotoxicity test results be used to complement positive results in the Ames test and help predict carcinogenic or in vivo genotoxic activity? II. Construction and analysis of a consolidated database, Mutat. Res. 775–776 (2014) 69–80.
[19] D. Kirkland, P. Kasper, H.-J. Martus, L. Müller, J. van Benthem, F. Madia, R. Corvi, Updated recommended lists of genotoxic and non-genotoxic chemicals for assessment of the performance of new or improved genotoxicity tests, Mutat. Res. 795 (2016) 7–30.
[20] ECHA, Evaluation under REACH: progress report 2015, Safer Chemicals – Focussing on What Matters Most. Reference: ECHA-15-R-20-EN, European Chemicals Agency, Helsinki, Finland, 2016 February.
[21] EFSA scientific report, minimum criteria for the acceptance of *in vivo* alkaline comet assay reports, european food safety authority Parma, italy, Efsa J. 10 (2012) 2977.
[22] R.R. Tice, E. Agurell, D. Anderson, B. Burlinson, A. Hartmann, H. Kobayashi, Y. Miyamae, E. Rojas, J.C. Ryu, Y.F. Sasaki, Single cell gel/comet assay: guidelines for in vitro and in vivo genetic toxicology testing, Environ. Mol. Mutagen. 35 (2000) 206–221.
[23] V. Thybaud, S. Dean, T. Nohmi, J. de Boer, G.R. Douglas, B.W. Glickman, N.J. Gorelick, J.A. Heddle, R.H. Heflich, I. Lambert, H.-J. Martus, J.C. Mirsalis, T. Suzuki, N. Yajima, In vivo transgenic mutation assays, Mutat. Res. 540 (2003) 141–151.
[24] K. Sekihashi, T. Sasaki, A. Yamamoto, K. Kawamura, T. Ikka, S. Tsuda, Y.F. Sasaki, A comparison of intraperitoneal and oral gavage administration in comet assay in mouse eight organs, Mutat. Res. 493 (2001) 39–54.
[25] JECFA, Toxicological Evaluation of Certain Veterinary Drug Residues in Food, WHO Food Additives Series, 2000 No. 45.
[26] A. Agresti, B.A. Coull, Approximate is better than "exact" for interval estimation of binomial proportions, Am. Stat. 52 (1998) 119–126.
[27] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1960) 37–46.
[28] L. Flight, S.A. Julious, The disagreeable behaviour of the kappa statistic, Pharm. Stat. 14 (2015) 74–78.
[29] K.L. Gwet, Handbook of Inter-Rater Reliability. The Definitive Guide to Measuring the Extent of Agreement Among Raters, 2nd edition, Advanced Analytics, LLC, Gaithersburg, MD 20886–2696, USA, 2010.
[30] T. Byrt, J. Bishop, J.B. Carlin, Bias, Prevalence and kappa, J. Clin. Epidemiol. 46 (1993) (1993) 423–429.
[31] N. Wongpakaran, T. Wongpakaran, D. Wedding, K.L. Gwet, A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples, BMC Med. Res. Methodol. 13 (61) (2013).
[32] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174.
[33] Y.F. Sasaki, K. Sekihashi, F. Izumiyama, E. Nishidate, A. Saga, K. Ishida, S. Tsuda, The comet assay with multiple mouse organs: comparison of comet assay results and carcinogenicity with 208 chemicals, selected from the IARC monographs and U.S. NTP Carcinogenicity Database, Crit. Rev. Toxicol. 30 (2000) 629–799.
[34] F. Tombolan, D. Renault, D. Brault, M. Guffroy, O. Périn-Roussel, F. Périn, V. Thybaud, Kinetics of induction of DNA adducts, cell proliferation and gene mutations in the liver of MutaMice treated with 5,9-dimethyldibenzo[c,g]carbazole, Carcinogenesis 20 (1999) 125–132.
[35] F. Tombolan, D. Renault, D. Brault, M. Guffroy, F. Périn, V. Thybaud, Effect of mitogenic or regenerative cell proliferation on lacZ mutant frequency in the liver of Muta™Mice treated with 5, 9-dimethyldibenzo[c,g]carbazole, Carcinogenesis 20 (1999) 1357–1362.
[36] D. Kirkland, Y. Uno, M. Luijten, C. Beevers, J. van Benthem, B. Burlinson, S. Dertinger, G.R. Douglas, S. Hamada, K. Horibata, D.P. Lovell, M. Manjanatha, H-J. Martus, N. Mei, T. Morita, W. Ohyama, A. Williams, 7th International Workshop on Genotoxicity Testing: Report of the in vivo strategies working group, submitted for publication.