Section 4
Health effects

Test Guideline No. 488

Transgenic Rodent Somatic and Germ Cell Gene Mutation Assays

25 June 2025

OECD Guidelines for the Testing of Chemicals

OECD
BETTER POLICIES FOR BETTER LIVES

OECD/OCDE 488

Adopted:
25 June 2025

OECD GUIDELINES FOR THE TESTING OF CHEMICALS

Transgenic Rodent Somatic and Germ Cell Gene Mutation Assays

1. INTRODUCTION

1. The OECD Test Guidelines for the testing of chemicals are periodically reviewed in light of scientific progress, changing regulatory needs and animal welfare considerations. The Test Guideline 488 was originally adopted in 2011 and subsequently revised in 2013, 2020, and 2022. It has evolved to integrate the analysis of mutations in both somatic tissues and germ cells. To ensure harmonisation with other OECD Test Guidelines for genotoxicity testing, the 2022 revision included three equal considerations when assessing whether a response is positive or negative. First, the test chemical response was assessed as to whether there was a statistically significant increase from the concurrent negative control. Second, the response was evaluated for a concentration/dose related effect. Third, historical negative control distributions were used to assess biological relevance. As noted above, the three considerations were given equal weight. However, recent analyses suggest more nuance and flexibility should be applied to the third consideration (*i.e.,* comparison to historical controls). In particular, distributions can only serve as a proxy for normal biological variation when they are of sufficient quality for this purpose. The changes found herein address these issues by revising language in the "Evaluation and Interpretation of Results," and by updating an associated Annex.

2. A document that provides an overview of both genetic toxicity testing and the recent changes that were made to the TGs for genotoxicity testing has been developed (1). Additional information on the main changes introduced to these TGs was also published (2).

3. OECD TGs are available for a wide range of in vitro mutation assays that are able to detect chromosomal and/or gene mutations. There are TGs for several *in vivo* genotoxic endpoints (*i.e.*, chromosomal aberrations, micronuclei, unscheduled DNA synthesis, and DNA strand breaks); however, these do not measure gene mutations. While the comet and the unscheduled DNA synthesis assays are indicator tests that detect pre-mutagenic lesions, and the *Pig-a* assay is limited to the haematopoietic system, the Transgenic Rodent (TGR) gene mutation assays fulfil the need for practical and widely available *in vivo* tests for measuring gene mutations in any tissue.

4. Data from the TGR gene mutation assays have been reviewed extensively, *e.g.*, (3) (4) (5). The models covered by this Test Guideline use transgenic rats and mice that contain multiple copies of chromosomally integrated phage shuttle vectors. The transgenes contain reporter genes for the detection of various types of mutations induced in vivo by test chemicals. The purpose of the TGR gene mutation assay is to identify substances that result in mutations due to DNA damage in the tissue that is being analysed.

5. Mutations arising in a rodent are detected by recovering the transgene and analysing the phenotype of the reporter gene in a bacterial host deficient for the reporter gene. TGR gene mutation assays measure mutations induced in genetically neutral genes recovered from virtually any tissue of the rodent. These assays, therefore, circumvent many of the existing limitations associated with the study of in vivo gene mutation in endogenous genes (*e.g.*, limited tissues suitable for analysis, negative/positive selection against mutations).

6. The weight of evidence suggests that transgenes respond to mutagens in an approximately similar manner to endogenous genes, especially with regard to the detection of base pair substitutions, frameshift mutations, and deletions and insertions (3).

7. The International Workshops on Genotoxicity Testing (IWGT) have endorsed the use of TGR gene mutation assays for *in vivo* detection of gene mutations, and have recommended a protocol for their implementation (6) (7). Further analysis supporting the use of this protocol can be found in (8). The present TG is based on these recommendations for the evaluation of gene mutations in somatic tissues and includes updated recommendations for the evaluation of gene mutations in male germ cells (5).

8. The TGR gene mutation assay uses the same treatment regimen as the repeat dose toxicity study (TG 407), *i.e.*, 28-day administration, providing the option of combining the two assays into one study with the condition that performing the necropsy the day after the end of the treatment for both studies does not adversely affect the recovery of mutations. Data are also required to indicate that the performance of the repeat dose assay is not adversely affected by using a transgenic rodent strain rather than traditional rodent strains. Furthermore, it is possible to integrate additional genotoxicity endpoints into the TGR assay, such as assessment of micronuclei and *Pig-a* mutations (9). Combining studies should be based on the need to investigate specific endpoints based on existing information or specific regulatory requirements.

9. Definitions of key terms are set out in Annex 1.

2. INITIAL CONSIDERATIONS

10. TGR gene mutation assays for which sufficient data are available to support their use in this TG are: *lacZ* bacteriophage mouse (MutaMouse); *gpt* delta (*gpt* and Spi^{-}*)* mouse and rat; *lacI bacteriophage* mouse and rat (Big Blue®), as performed under non-selective conditions. In these assays, the mutations are measured in bacterial genes (*lacI*, *lacZ* and *gpt*) inserted into a lambda vector. In addition, mutations can be measured in the *cII* gene of the bacteriophage in the Big Blue® and MutaMouse models and the *red/gam* genes in the *gpt* delta model under Spi^{-} selection. The *lacZ* plasmid model has been deleted from this Test Guideline because it is no longer used routinely. Methods for the identification of mutants under selective conditions are available (see paragraph 17) and should be used preferentially. Mutagenesis in the TGR models is normally assessed as mutant frequency; if required, however, molecular analysis of the mutations can provide additional information (see Paragraphs 57-58).

11. These TGR gene mutation tests (3) are especially relevant to assessing mutagenic hazard in that the assay responses are dependent upon *in vivo* metabolism, pharmacokinetics, DNA repair processes, and translesion DNA synthesis, although these may vary among species, among tissues and among the types of DNA damage. An *in vivo* assay for gene mutations is useful for further investigation of a mutagenic effect detected by an *in vitro* system, and for investigating the underlying mode of action of tests using

other *in vivo* studies, such as a positive tumour result from carcinogenicity studies. In addition to being causally associated with the induction of cancer, gene mutation is a relevant endpoint for the prediction of mutation-based non-cancer diseases in somatic tissues (10) (11) as well as diseases transmitted through the germline (12).

12. If there is evidence that the test chemical, or relevant metabolite, will not reach any of the tissues of interest, it is not appropriate to perform a TGR gene mutation assay.

13. Before use of the Test Guideline on a mixture for generating data for an intended regulatory purpose, it should be considered whether, and if so why, it may provide adequate results for that purpose. Such considerations are not needed, when there is a regulatory requirement for testing of the mixture.

3. PRINCIPLE OF THE TEST METHOD

14. In the assays described in paragraph 10, the target gene is bacterial or bacteriophage in origin, and the means of recovery from the rodent genomic DNA is by incorporation of the transgene into a lambda bacteriophage vector. The procedure involves the extraction of genomic DNA from the rodent tissue of interest, *in vitro* processing of the genomic DNA (*i.e.*, packaging of lambda vectors), and subsequent detection of mutations in bacterial hosts under suitable conditions. The assays employ neutral non-transcribed transgenes that are readily recoverable from most tissues.

15. The basic TGR gene mutation experiment involves treatment of the rodent with a chemical over a period of time. Test chemicals may be administered by any appropriate route, including implantation (*e.g.*, medical device testing). The total period during which an animal is dosed is referred to as the administration period. Administration is usually followed by a period of time, prior to humane killing, during which the test chemical is not administered and during which unrepaired DNA lesions are fixed into stable mutations. In the literature, this period has been variously referred to as the manifestation time, fixation time or expression time; the end of this period is the sampling time (6) (8). After the animal is humanely killed, tissues are rapidly collected and frozen, after which they can be stored at or below -70°C +/- 5°C until genomic DNA is isolated from the tissue(s) of interest and purified. Tissues may be collected from moribund animals humanely killed during the last week of dosing, stored at or below -70°C +/- 5°C and analysis conducted on a case-by-case basis, if needed.

16. Data for a single tissue per animal from multiple packaging/ligations are usually aggregated, and mutant frequency is generally evaluated using a total of between 10^5 and 10^6 plaque-forming or colony- forming units per animal. When using positive selection methods, total plaque- or colony-forming units are determined with a separate set of non-selective plates.

17. Positive selection methods have been developed to facilitate the detection of mutations in both the *gpt* gene [*gpt* delta mouse and rat, *gpt*$^-$ phenotype (13) (14) (15)] and the *lacZ* gene [MutaMouse (16) (17) (18)]; whereas, no positive selection methods are available for *lacI* gene mutations in Big Blue® animals and mutations are detected through a non-selective method that identifies mutants through the generation of coloured (blue) plaques. Positive selection methodology is also in place to detect mutations arising in the *cII* gene of the lambda bacteriophage shuttle vector [Big Blue® mouse or rat, and MutaMouse (19)] and deletion mutations in the lambda *red* and *gam* bacteriophage genes [Spi$^-$ selection in *gpt* delta mouse and rat (14) (15) (20)]. Mutant frequency is calculated

by dividing the number of plaques containing mutations in the transgene by the total number of plaques recovered from the same DNA sample. In TGR gene mutation studies, the mutant frequency is the reported parameter. In addition, a mutation frequency can be determined as the fraction of cells carrying independent mutations; this calculation requires correction for clonal expansion by sequencing the recovered mutants (Paragraph 57-58).

18. The mutations scored in the lacI, lacZ, cII and gpt mutation assays consist primarily of base pair substitution mutations, frameshift mutations and small insertions/deletions which cannot be distinguished in the phenotypic assay. The relative proportion of these mutation types among spontaneous mutations is similar to that seen in the endogenous *Hprt* gene. Large deletions are detected only with the Spi^- selection in the *gpt* delta (3). Mutations of interest are *in vivo* mutations that arise in the mouse or rat. *In vitro* and *ex vivo* mutations, which may arise during phage recovery, replication or repair, are relatively rare, and in some systems can be specifically identified, or excluded by the bacterial host/positive selection system (3) (4).

4. DESCRIPTION OF THE METHOD

4.1. Preparations

4.1.1. Selection of animal species

19. Transgenic mouse and rat gene mutation detection models are currently available. Both mouse and rat models are considered equally acceptable. Justification of the model used in the TGR assay should include a consideration of: (i) laboratory proficiency with the model; (ii) availability of historical data in the tissues under investigation; (iii) known toxicity differences between the species for the substance under investigation (*e.g.*, when investigating the mechanism of carcinogenesis for a tumour seen only in one rodent species, to correlate with a toxicity study in a specific species, or if metabolism in one rodent species is known to be more representative of human metabolism); and (iv) the preferred species used in other toxicity studies in case combination with the TGR assay is foreseen.

4.1.2. Housing and feeding conditions

20. All procedures should conform to local standards of laboratory animal care. For rodents, the temperature in the experimental animal room ideally should be 22ºC (± 3ºC). Although the relative humidity should be at least 30% and preferably not exceed 70% other than during room cleaning, the aim should be 50-60%. Lighting should be artificial, with a daily sequence of 12 hours light, followed by 12 hours dark. For feeding, conventional laboratory diets may be used with an unlimited supply of drinking water. The choice of diet may be influenced by the need to ensure a suitable admixture of a test chemical when administered by this route. Animals should be group housed in small groups of the same sex; animals may be housed separately if scientifically justified. For group caging, no more than five animals should be housed per cage. Furthermore, cages should conform with animal welfare standards [*e.g.*, Directive 2010/63/EU)]. Based on Animal Care and Use Committee (or equivalent) recommendations, solid floors should be used wherever possible and appropriate environmental enrichment should be provided.

4.1.3. Preparation of the animals

21. Healthy young sexually mature adult animals (8-12 weeks old at start of treatment) should be used when germ cell data are required (see paragraph 33). For somatic tissue studies, younger animals (*e.g.*, 4-6 weeks of age at the start of treatment) are acceptable with substantial scientific or animal welfare justification, such as, for example, to avoid killing animals that have been bred but not used in a procedure. Also, provision should be made for any alteration to the historical control database such deviation in age may cause. Animals are randomly assigned to the control and treatment groups. The animals are identified uniquely using a humane, minimally invasive method (*e.g.*, by ringing, tagging, micro-chipping or biometric identification, but not ear or toe clipping). The animals are acclimated to the laboratory conditions for at least five days. At the commencement of the study, the weight variation of animals should be minimal and not exceed ± 20% of the mean weight of each sex. The selection of the sex to use is dependent on whether germ cell data is required (paragraph 33) and/or human exposure to the test chemical is sex-specific (paragraph 34).

4.1.4. Preparation of doses

22. Solid test chemicals should be dissolved or suspended in appropriate solvents or vehicles (see paragraph 23) or admixed in diet or drinking water prior to dosing of the animals. Liquid test chemicals may be dosed directly or diluted prior to dosing. For inhalation exposures, test chemicals can be administered as gas, vapour, or a solid/liquid aerosol, depending on their physicochemical properties. Other routes of exposure should be justified scientifically. Fresh preparations of the test chemical should be employed unless stability data demonstrate the acceptability of storage.

4.2. Test Conditions

4.2.1. Solvent/vehicle

23. The solvent/vehicle should not produce toxic effects at the dose volumes used, and should not be suspected of chemical reaction with the test chemical. It is recommended that wherever possible, the use of an aqueous solvent/vehicle should be considered first. Examples of commonly used compatible solvents/vehicles include water, physiological saline, methylcellulose solution, carboxymethyl cellulose sodium salt solution, olive oil and corn oil (21). If other than well-known solvents/vehicles are used, their inclusion should be supported with reference data indicating their compatibility. In the absence of historical or published control data showing that no mutations and other deleterious effects are induced by a chosen atypical solvent/vehicle, an initial study should be conducted in order to establish the acceptability of the solvent/vehicle control.

4.2.2. Positive Controls

24. Concurrent positive control animals should normally be used. This may be waived when the testing laboratory has demonstrated proficiency verification in the conduct of the test (see Paragraph 27) and has established a historical control range for the tissue under investigation (see Paragraphs 28-32). In this situation, to assure continued proficiency in detecting increases in mutant frequency, laboratories should occasionally (at least once per year) perform additional tests using tissues from mutagen-treated animals as described in paragraph 27. When a concurrent positive control group is not used, tissues from previous positive control treated animals should be included with each study to confirm the

reliability of the method. These samples should be from the same species with similar age and tissues of interest, properly stored (see Paragraph 53) and generate mutant frequencies that are consistent with previous experiments.

25. When concurrent positive controls are used, it is not necessary to administer them by the same route or duration as the test chemical; however, the positive controls should be known to induce mutations in one or more tissues of interest for the test chemical. Positive substances should reliably produce a detectable increase in mutant frequency over the spontaneous level. The doses of the positive control chemicals should be selected so as to produce weak or moderate effects that critically assess the performance and sensitivity of the assay. Examples of positive control substances and some of their target tissues are included in Table 1. Substances other than those given in Table 1 can be selected if scientifically justified.

Table 1. Examples of positive control substances and some of their target tissues

Chemical and CAS No.	Characteristics	Mutation Target Tissues/cell types	
		Rat	Mouse
N-Ethyl-N-nitrosourea [CAS no. 759-73-9]	Direct acting mutagen	Liver, glandular stomach, duodenum, jejunum, bone marrow, spleen, lung, nasal epithelium, kidney, bladder, testicular germ cells	Liver, forestomach, glandular stomach, duodenum, colon, bone marrow, spleen, lung, nasal epithelium, kidney, follicular granulosa cells, testicular germ cells, sperm
Ethyl carbamate (urethane) [CAS no. 51-79-6]	Mutagen, requires metabolism but produces only weak effects		Liver, bone marrow, spleen, forestomach, small intestine, lung
2,4-Diaminotoluene [CAS no. 95-80-7]	Mutagen, requires metabolism, also positive in the Spi$^-$ assay	Liver	Liver
Benzo[a]pyrene [CAS no. 50-32-8]	Mutagen, requires metabolism	Liver, glandular stomach, duodenum, jejunum, bone marrow, spleen, lung, nasal epithelium, kidney, bladder, omenta	Liver, forestomach, glandular stomach, duodenum, jejunum, colon, bone marrow, breast, heart, lung, kidney, bladder, testicular germ cells, sperm

4.2.3. Negative controls

26. Negative controls (see paragraph 28) treated in the same way as the treatment groups, should be included for every tissue and sampling time (however, see paragraph 23 regarding atypical solvents or vehicles).

5. VERIFICATION OF LABORATORY PROFICIENCY

5.1. Proficiency verifications

27. In order to establish sufficient experience with the conduct of the assay prior to using it for routine testing, the laboratory should have demonstrated the ability to reproduce expected results from published data (3) (22) for both mutant frequencies and transgene recovery from genomic DNA (*e.g.*, packaging efficiency). A minimum of two positive control substances (including weak response induced by low doses of positive controls) such as those listed in Table 1 (see paragraph 25) and with compatible vehicle/solvent controls (see paragraph 23) should be used. Initially, proficiency should be demonstrated in at least two tissues, preferably one for slowly dividing tissues such as liver, and one rapidly dividing tissue such as bone marrow, glandular stomach or duodenum (7) (23). If germ cell assessments are to be conducted, these should also be included in the laboratory's proficiency investigations. These experiments should use doses that give reproducible dose related increases and demonstrate the sensitivity and dynamic range of the test system in the tissue of interest. This requirement is not applicable to laboratories that have experience, *i.e.*, that have a historical database available as defined in paragraphs 28-32. Prior to conducting a study in a tissue not previously examined, a laboratory (even those that are experienced) will need to establish proficiency in the DNA extraction and transgene recovery techniques specific to that tissue, in order to establish likely mutant frequencies and packaging efficiencies. In addition, the laboratory will need to demonstrate that an acceptable positive response with a known mutagen (see Table 1) can be obtained in that tissue.

5.2. Historical control data

28. During the course of the proficiency investigation the laboratory should establish for each tissue to be investigated:

- A historical positive control range and distribution, and
- A historical negative control (*i.e.*, solvent/vehicle-treated animals or untreated animals) range and distribution.

29. When first acquiring data for a historical negative control distribution, concurrent negative controls should be consistent with published data where they exist. As more experimental data are added to the historical control distribution, concurrent negative controls should ideally be within the lower and upper bound limits of the distribution (see paragraph 32). The laboratory's historical negative control database should be compiled, analysed and regularly updated according to literature recommendations (*e.g.*, (23), (58); also see Annex 2). This should include: consideration of the minimum number of data sets required to establish a robust distribution (a minimum of 30 animals is desirable); frequency of update and methods to ensure the most recent and/or relevant data are used for assay acceptance and data evaluation (see paragraph 61). Significant deviations from these recommendations should be justified. Laboratories should use quality control methods, such as control charts that are appropriate for the distribution of the data [*e.g.*, C-charts or X-bar charts (24) (25) (26) (27)], *i.e.*, not simple distribution ranges of control data, to identify how variable the data are, and to show that the methodology is 'under control' in their laboratory.

30. Where the laboratory does not complete a sufficient number of experiments to establish a statistically robust negative control distribution (see paragraph 29) during the proficiency investigations (described in paragraph 27), it is acceptable that the distribution can be built during the first routine tests. This approach should follow the recommendations set out in the literature [(23) (58); Annex 2)] and the negative control results obtained in these experiments should remain consistent with published negative control data.

31. Any changes to the experimental protocol should be considered in terms of their impact on the resulting data remaining consistent with the laboratory's existing historical control database. Only major inconsistencies should result in the establishment of a new historical control database where sound scientific judgement determines that it differs from the previous distribution (see paragraph 29) (24) (25) (26) (27). During the reestablishment, a full negative control database may not be needed to permit the conduct of an actual test, provided that the laboratory can demonstrate that their concurrent negative control values remain either consistent with their previous database or with the corresponding published data.

32. Negative control data should consist of the mutant frequency per tissue for each animal. Concurrent negative controls should normally be within lower and upper bound limits of the distribution of the laboratory's historical negative control database. Where concurrent negative control data fall outside these control limits, they may be acceptable for inclusion in the historical control distribution as long as these data are not extreme outliers (*e.g.,* identified by an outlier test) and there is evidence that the test system is 'under control' (see paragraph 29) and there is no evidence of technical or human failure.

6. PROCEDURE

6.1. Number and Sex of Animals

33. The number of animals per group should be predetermined to be sufficient to provide the statistical power necessary to detect at least a doubling in mutant frequency. Group sizes will consist of a minimum of five animals; however, if the statistical power is insufficient, the number of animals should be increased as required. When study designs require germ cell data, male animals should be used because it is not possible to collect sufficient numbers of female germ cells to conduct the TGR assay (28).

34. When only somatic data are needed, such studies could be performed in either sex, since the mutation response is similar between male and female animals. Where human exposure to chemicals may be sex-specific, as for example with some pharmaceuticals, the test should be performed with the appropriate sex. Data demonstrating important differences between males and females (*e.g.,* differences in systemic toxicity, metabolism, bioavailability etc. including *e.g.,* in a range-finding study) would encourage the use of both sexes. If a TGR study is performed to follow up positive tumour or other toxicological findings, the selection of species and sex should be based on the species and sex of the initial study.

6.2. Administration Period

35. Based on observations that mutations accumulate with each treatment, a repeated-dose regimen is necessary, with daily treatments for a period of 28 days. This is generally

considered acceptable both for producing a sufficient accumulation of mutations by weak mutagens, and for providing an exposure time adequate for detecting mutations in slowly proliferating organs. Alternative treatment regimens may be appropriate for some evaluations and these alternative dosing schedules should be scientifically justified in the protocol. Treatments should not be shorter than the time required for the complete induction of all the relevant metabolising enzymes, and shorter treatments may necessitate the use of multiple sampling times that are suitable for organs with different proliferation rates. In any case, all available information (*e.g.*, on general toxicity or metabolism and pharmacokinetics) should be used when justifying a protocol, especially when deviating from the above standard recommendations. While it may increase sensitivity, treatment times longer than 28 days should be explained clearly and justified, since long treatment times may produce an apparent increase in mutant frequency through clonal expansion (7).

6.3. Dose Levels

36. Any existing toxicity and toxicokinetic data should be taken into consideration in setting dose levels. If a preliminary range-finding study is performed because there are insufficient suitable data already available to guide dose selection, it should be performed in the same laboratory, using the same strain (non-transgenic animals may be used), sex, and treatment route to be used in the main study.

37. In the main test, in order to obtain dose response information, a complete study should include a negative control group (see Paragraph 26) and a minimum of three, appropriately-spaced dose levels of the test chemical, except where the limit dose has been used (see Paragraph 40). Except in cases where the limit dose is applicable, the highest dose level should be chosen with the aim of inducing toxic effects but not death or severe suffering. This definition is taken from TG 407 and is used here to facilitate integration of the TGR assay with other repeat-dose studies, thereby maximising the toxicological information acquired. When limited by other factors, such as palatability for dietary or drinking water administration (and when dosing cannot be done by gavage) or explosiveness for test chemicals administered by inhalation, the highest dose level will be the maximum feasible dose. Thereafter, a descending sequence of dose levels should be selected with a view to demonstrating any dose-related response. Two- to four-fold intervals are frequently used for setting the descending dose levels when genotoxicity endpoints are evaluated. With most test chemicals, the dose levels used should cover a range from the maximum to little or no toxicity. When additional toxicity endpoints are integrated into the study, other dose spacing may be considered. The addition of a fourth test group is often preferable to using excessively large spacing between doses.

38. Test chemicals with specific biological activities at low non-toxic doses (such as hormones and mitogens), and substances that exhibit saturation of toxicokinetic properties, or induce detoxification processes that may lead to a decrease in exposure after long-term administration may be exceptions to the dose-setting criteria and should be evaluated on a case-by-case basis.

39. Care should be taken to ensure that the highest dose identified in the range finding study does not induce excessive toxicity in any tissue of interest, which may prevent the availability of sufficient cells to extract adequate quality and quantity of DNA to recover the transgene for mutation analysis. In such cases, consideration may be given to the inclusion of an additional dose group, closely spaced to the highest dose to assure the availability of the required three treatment groups for mutation analysis. For oral gavage studies, the use of alternative vehicle or split dosing (two or more treatments on the same

day separated by no more than 2-3 hours) may be considered with justification to minimize the effects leading to excessive toxicity.

6.4. Limit Test

40. If dose range-finding experiments, or existing data from related rodent strains, indicate that a treatment regimen of at least the limit dose (see below) produces no observable toxic effects, and if genotoxicity would not be expected based upon *in vitro* genotoxicity studies or data from structurally related substances, then a full study using three dose levels may not be considered necessary. Instead, a study with one dose level (*i.e.*, with the limit dose) is considered sufficient. Accordingly, for an administration period of 28 days (*i.e.*, 28 daily treatments), this limit dose is 1000 mg/kg body weight/day. For administration periods of 14 days or less, the limit dose is 2000 mg/kg/body weight/day (dosing schedules differing from 28 daily treatments should be scientifically justified in the protocol; see Paragraph 35). In the case of inhalation exposures, the limit test concentrations are 20 mg/L, 5 mg/L or 20,000 ppm for vapours, dusts/mists (aerosols) and gases, respectively.

6.5. Administration of Doses

41. The route should be chosen to ensure exposure to the tissue(s) of interest. The preferred route should be the anticipated route of human exposure, and other routes should be otherwise justified. Therefore, routes of exposures such as dietary, drinking water, topical, subcutaneous, intravenous, oral (by gavage), inhalation, intratracheal, or implantation may be chosen as justified. Intraperitoneal injection is generally not recommended since it is not a physiologically relevant route of human exposure, and should only be used with scientific justification. The maximum volume of liquid that can be administered at one time depends on the size of the test animal and the route of administration and should be guided by international standards related to animal welfare (31) (32). For oral gavage, the volume should not exceed 1 mL/100 g body weight for mice and rats, except in the case of aqueous solutions where a maximum of 2 mL/100 g may be used. The use of volumes greater than this should be justified. Except for irritating or corrosive test chemicals, which will normally reveal exacerbated effects at higher concentrations, variability in test volume should be minimised by adjusting the concentration to ensure a constant volume in relation to body weight at all dose levels.

6.6. Sampling Time

6.6.1. Somatic Cells

42. The sampling time is a critical variable because it is determined by the period needed for mutations to be fixed. This period is tissue-specific and appears to be related to the turnover time of the cell population (32), with bone marrow and intestine being rapid responders and the liver being much slower. The recommended protocol for the measurement of mutant frequencies in both rapidly and slowly proliferating tissues following 28 consecutive daily treatments (as indicated in Paragraph 35) is tissue collection 28 days after the final treatment (*i.e.*, 28+28d) (33). Tissue collection three days after the final treatment (*i.e.*, 28+3d), as it was recommended in previous versions of the TG, remains a valid sampling time when no germ cell data is needed.

6.6.2. Germ Cells

43. TGR assays are well-suited for the study of gene mutation induction in male germ cells (5) (34) (35) (36) (37), in which the timing and kinetics of spermatogenesis have been well-defined (38) (39) (40). Because of the low numbers of ova available for analysis, even after super-ovulation, and the fact that there is no DNA synthesis in the oocyte, female germ cells cannot be used to measure mutations using transgenic assays (28). The available germ cell mutagenicity data obtained with TGR assays have been recently reviewed (5) together with modelling of mouse and rat spermatogenesis (40) to inform on the selection of an appropriate experimental design for assessing mutagenicity in germ cells. The modelling considered that the mitotic phase of spermatogenesis (*i.e.*, stem cells, proliferating and differentiating spermatogonia) is the only spermatogenic phase where both DNA replication and cell proliferation, which are necessary to fix mutations into the transgene (41), are occurring.

44. Male germ cells can be collected as either mature sperm from the cauda epididymis or as developing germ cells from the seminiferous tubules. Developing germ cells from the seminiferous tubules can be collected by simply removing the tunica albuginea that encapsulates the testis, or by extruding them from the seminiferous tubules using either enzymatic or physical separation (42). The latter approach is preferred as it enriches the collected population for germ cells because somatic cells (*e.g.*, Leydig and Sertoli cells) present in the testis cannot be easily separated from the tubules.

45. The timing of spermatogenesis in both mouse (39) and rat (39) is well-established. The time for the progression of developing germ cells from exposed spermatogonial stem cells to mature sperm reaching the cauda epididymis is ~49 days for the mouse (37) (39) and ~70 days for the rat (39) (40). Therefore, sampling of caudal mouse and rat sperm at 28+3d does not provide meaningful mutagenicity data, because these cells represent a population of germ cells that has not undergone DNA replication during the exposure, and should, thus, not be conducted. For the mouse, there are also experimental data demonstrating that this 28+3d design does not detect the strong germ cell mutagens N-ethyl-N-nitrosourea (5) and benzo(a)pyrene (43). Sampling of caudal sperm should be conducted at a minimum of 49 days (mouse) or 70 days (rats) after the end of the 28-day administration period in those cases where it is important to assess mutations in spermatogonial stem cells (5) (40).

46. Germ cells extruded from seminiferous tubules comprise a mixed population of spermatogonia, spermatocytes and spermatids (34) (35) (40). The composition of the germ cell population collected from mouse and rat seminiferous tubules, according to the number of days of treatment received during the proliferative phase of spermatogenesis, has been described in detail for various sampling times considering the known kinetics of spermatogenesis (40). While positive results in tubule germ cells after a 28+3d regimen are informative, a negative result after a 28+3d regimen is insufficient to negate the possibility that a test chemical is a germ cell mutagen because only a limited fraction of collected germ cells from the tubules have received continuous treatment for the full 28-day administration period during the proliferative phase of spermatogenesis (5) (40).

47. *Mouse Modelling*. Based primarily on extensive modelling of spermatogenesis (40) and limited experimental data (5), collection of germ cells from the seminiferous tubules at a sampling time longer than 3 days is better for the assessment of germ cell mutagenicity. The modelling shows that the 28+28d regimen enables the evaluation of mutations in a population of mouse germ cells that has received 99.6% of the 28 days of treatment during the proliferative phase of spermatogenesis, versus only 42.2% with the 28+3d regimen

(40). The spermatogenesis model assumes that the exposure does not produce a significant induction of germ cell apoptosis or delays in the progression of spermatogenesis. However, if such effects were to occur, longer sampling times, such as provided by the 28+28d regimen, would enable recovery of spermatogenesis by allowing the testes to be repopulated with surviving stem cells and differentiating spermatogonia that have received the full 28-day administration of the test chemical during the proliferative phase of spermatogenesis. For these reasons, both positive and negative results in mouse germ cells obtained with the 28+28d regimen are considered conclusive, and the 28+28d regimen is recommended.

48. *Rat Modelling*. Based on extensive modelling of spermatogenesis (40) and the longer duration of spermatogenesis in the rat versus the mouse, the 28+28d regimen in the rat does not provide the same degree of exposure of proliferating cell stages as in the mouse using the same regimen (see Paragraph 47). The modelling of rat spermatogenesis indicates that the 28+28d regimen enables the evaluation of mutations in a population of cells that has received 80.3% of the 28-day administration period during the proliferative phase of spermatogenesis (compared to 99.6% in the mouse), while only 21.6% is obtained with the 28+3d regimen (40). While not optimal, the 28+28d design is, nevertheless, considered adequate for the evaluation of germ cell mutagenesis; furthermore, it permits the assessment of mutations in both somatic tissues and tubule germ cells from the same animals. The impact of rat proliferating germ cells receiving less than the full potential exposure should be considered when evaluating negative results obtained with this design.

49. Sampling times other than 28 days for germ cells may also be acceptable; however, the impact of using a sampling time shorter than 28 days, which reduces the degree of exposure of proliferating germ cell stages for both the mouse and the rat (40), should be considered and justified scientifically. When a sufficient number of studies become available to ascertain the benefit of any other germ cell regimen, the TG will be reviewed and, if necessary, revised in light of the experience gained.

50. In summary, when both somatic and germ cells need to be collected and/or tested, based on regulatory requirements, or toxicological information, the 28+28d regimen permits the testing of mutations in somatic tissues and tubule germ cells from the same animals.

6.7. Observations

51. General clinical observations should be made at least once a day, preferably at the same time(s) each day and considering the peak period of anticipated effects after dosing. The health condition of the animals should be recorded. At least twice daily, all animals should be observed for morbidity and mortality. All animals should be weighed at study initiation, at least once a week, and at humane killing. Measurements of food consumption should be made at least weekly. If the test chemical is administered via the drinking water, water consumption should be measured at each change of water and at least weekly. Animals exhibiting non-lethal indicators of excess toxicity should be euthanised prior to completion of the test period (29).

6.8. Tissue Collection

52. 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 genotoxicity,

carcinogenicity or toxicity data for the test chemical under investigation. Important factors for consideration should include the route of administration (based on likely human exposure route(s)), the predicted tissue exposure, and the possible target organ toxicity. In the absence of any background information, several somatic tissues as may be of interest should be collected which should represent rapidly proliferating, slowly proliferating and site of contact tissues. In addition, developing germ cells from the seminiferous tubules (as described in Paragraphs 44 and 46) should be collected and stored in case future analysis of germ cell mutagenicity is required and an appropriate sample time has been used. Relevant organ weights should be obtained, and for larger organs, the same area should be collected from all animals.

6.9. Storage of Tissues and DNA

53. Tissues (or tissue homogenates) should be quickly frozen and stored at or below −70 ºC +/-5 ºC and used as long as good high molecular weight DNA can be recovered. Isolated DNA, stored refrigerated at 4 +/-1 ºC in appropriate buffer, such as tris-EDTA, should be used optimally for mutation analysis within 1 year.

6.10. Selection of Tissues for Mutant Analysis

54. The choice of tissues should be based on considerations such as: (i) the route of administration or site of first contact (*e.g.*, glandular stomach or duodenum if administration is oral, lung or nasal epithelium if exposure is through inhalation, or skin if topical application has been used); (ii) ADME (absorption, distribution, metabolism and excretion) parameters observed in general toxicity studies, which indicate tissue distribution, retention or accumulation, or target organs for toxicity; and (iii) whether germ cell data may be required. If studies are conducted to follow up carcinogenicity studies, target tissues for carcinogenicity should be investigated. The choice of tissues for analysis should maximise the detection of chemicals that are direct-acting mutagens, rapidly metabolised, highly reactive or poorly absorbed, or those for which the target tissue is determined by route of administration (44).

55. In the absence of background information and taking into consideration the site of contact due to route of administration, the liver and at least one rapidly dividing tissue (*e.g.*, glandular stomach or duodenum, or bone marrow) should be evaluated for mutagenicity. In most cases, the above requirements can be achieved from analyses of two carefully selected tissues, but in some cases, three or more would be needed. Germ cells may also be assessed (see paragraph 52).

6.11. Methods of Measurement

56. Standard laboratory or published methods for the detection of mutants are available for the recommended transgenic models: *lacZ* lambda bacteriophage (18); *lacI* mouse (45) (46); *gpt* delta mouse (14); *gpt* delta rat (15) (47); *cII* (19). Modifications should be justified and properly documented. Data from multiple packaging can be aggregated and used to reach an adequate number of plaques or colonies. However, the need for a large number of packaging reactions to reach the appropriate number of plaques may be an indication of poor DNA quality. In such cases, data should be considered cautiously because they may be unreliable. The optimal total number of plaques or colonies per DNA sample is governed by the statistical probability of detecting sufficient numbers of mutants at a given spontaneous mutant frequency. In general, a minimum of 125,000 - 300,000

plaques are required for those TGR models with background mutant frequency in the range of 3 x10^{-5}. Models such as *gpt* delta with lower background mutant frequencies require proportionally more colony-forming units to be observed to ensure adequate statistical power. Tissues and the resulting samples should be processed and analysed using a block design, where preferably an equal number of samples from the vehicle/solvent control group, the positive control group (if used) or positive control DNA (where appropriate), and each treatment group are processed together.

6.12. Sequencing of mutants

57. Clonal amplification of early spontaneously arising mutants may occur in any animal, leading to small to large increases in mutant frequencies in individual tissues. Tissues from animals with elevated mutant frequencies outside of the historic distribution and different from other animals in the group may represent such jackpots or clonal events. Since such events are often localized within a tissue, reanalysis of a different portion of the same tissue may be one approach to assess such anomalies. Often, extra replacement animals are included in studies to accommodate animals lost due to early death or presence of jackpot mutations. Analysis of the extra animals may be appropriate in these cases.

58. While for regulatory applications, DNA sequencing of mutants is not required, particularly where a clear positive or negative result is obtained (see paragraphs 62 and 63), sequencing data may be useful when high inter-individual variation is observed. In these cases, sequencing can be used to rule out the possibility of jackpots or clonal events by identifying the proportion of unique mutants from a particular tissue. Sequencing approximately 10 mutants per tissue per animal should be sufficient for simply determining if clonal mutants contribute to the mutant frequency; sequencing as many as 25 mutants may be necessary to correct mutant frequency mathematically for clonality. Sequencing of mutants also may be considered when small increases in mutant frequency (*i.e.*, just exceeding the vehicle control values) are found. Differences in the mutation spectrum between the mutant colonies from treated and untreated animals may lend support to a mutagenic effect (7). Also, mutation spectra may be useful for developing mechanistic hypotheses. When sequencing is to be included as part of the study protocol, special care should be taken in the design of such studies, in particular with respect to the number of mutants sequenced per sample, to achieve adequate power according to the statistical model used (see Paragraph 67). In this regard, Next Generation Sequencing methods are available for both *cII* (48) and *lacZ* (49) genes, which greatly facilitate the sequencing of large number of mutants.

7. DATA AND REPORTING

7.1. Presentation of results

59. Individual animal data should be presented in tabular form. The experimental unit is the animal. The report should include the total number of plaque-forming units (pfu) or colony-forming units (cfu), the number of mutants, and the mutant frequency for each tissue from each animal. The report should also include the number of packaging/rescue reactions and the number of reactions per DNA sample should be reported. While data for each individual reaction should be retained, only the total number of mutants and pfu/cfu need to be reported. Data on toxicity and clinical signs as per paragraph 51 should be

reported. Any sequencing results should be presented for each mutant analysed, and resulting mutation frequency calculations for each animal and tissue should be shown.

7.2. Statistical evaluation and interpretation of results

7.2.1. Acceptability criteria

60. The following criteria determine the acceptability of the test:

a. The concurrent negative control data are considered acceptable for addition to the laboratory historical control database (see paragraphs 28-32; Annex 2).

b. The concurrent positive controls or scoring controls should induce responses that are compatible with those generated in the historical positive control database and produce a statistically significant increase compared to the concurrent negative control (see paragraphs 24 and 25).

c. The appropriate number of doses, animal per dose, and plaque-forming units or colony-forming units have been analysed (*i.e.*, paragraphs 16, 33 and 37).

d. The criteria for the selection of the highest dose and administration route are consistent with those described in paragraphs 36-39.

7.2.2. Evaluation and interpretation of results

61. Statistical tests used should consider the animal as the experimental unit. Appropriate statistical methods can be found in the following references (6) (50) (51) (52) (53), and in Annex 2. When evaluating the responses, all data should be taken into consideration and, in all cases, sound scientific judgement applied. Where data from at least three doses plus a negative (solvent/vehicle) control are available, dose-response analysis should be conducted using an appropriate trend test.

62. For any given tissue, providing that all acceptability criteria above are fulfilled, a test chemical is considered positive if both criteria a and b, below, are met in any of the experimental conditions examined:

a. At least one of the treatment groups exhibits a statistically significant increase in the mutant frequency compared with the concurrent solvent/vehicle control; and

b. The mutant frequency responses are dose related, for example when evaluated with an appropriate trend test (Annex 2) (not applicable to the limit test).

Criteria a- and b-type analyses, both of which are based on *concurrent* negative control data, are considered the statistical analyses of primary importance for interpreting study data. *Historical* negative control data and their distribution may also be useful for contextualising the results of the current study. However, this comparison carries less weight than criterion a or b. Furthermore, the degree to which historical control data are useful for this purpose is highly dependent on the quantity and quality of the underlying data. For instance, historical control data cannot serve as a useful comparator (*e.g.*, as a proxy for normal biological variation), when inter-study variation represents the predominant source of variation. This and other concepts surrounding the appropriate use of historical control data are provided in Annex 2. Thus, only to the degree historical negative control data have been generated in sufficient quantity, and these data are deemed of sufficient quality, that in addition to criteria a and b, a third criterion, c, can be applied as follows:

c. At least one of the treatment groups exhibits a mutant frequency exceeding the upper bound limit derived from the historical negative control data distribution.

The upper bound limit derived from the historical negative control distribution should be determined using a method appropriate to the quantity and quality of the laboratory's data (e.g., control charts, tolerance intervals, or quantiles), with justification provided in the study report. As inter-laboratory variability may influence the control distribution, no fixed numerical value can be prescribed, and laboratories should define their upper bound limits in a scientifically sound and transparent manner, as described in Annex 2.

In many instances, the frequencies being considered will be the current study's mean values compared to a distribution derived from historical negative control mean values. However, there are situations that make historical negative control data and related distributions based on individual animal data useful; for example, when historical control values are low in number, or when it is of interest to estimate sources of variation. As explained in more detail in Annex 2, in these cases it is important to make appropriate comparisons. Specifically, a distribution based on individual animal data should *not* be used as a comparator for the current study's treatment group means. Rather, a distribution based on individual animal data should only be used as a comparator for the current study's individual animal data. This requirement is referred to as a "like to like" comparison in Annex 2.

Positive results indicate that, under the test conditions, the test chemical induced gene mutations in the analysed tissue.

63. For any given tissue, providing that all acceptability criteria are fulfilled, and exposure to the test chemical and/or its metabolites occurred, a test chemical is considered negative if both criteria a and b, below, are met in all experimental conditions examined:

a. No treatment group exhibits a statistically significant increase in the mutant frequency compared with the concurrent solvent/vehicle control; and

b. None of the mutant frequency responses are dose-related, for example when evaluated by an appropriate trend test (see Annex 2, section 3) (not applicable to the limit test).

Criteria a- and b-type analyses, both of which are based on *concurrent* negative control data, are considered the statistical analyses of primary importance for interpreting study data. *Historical* negative control data and their distribution may also be useful for contextualising the results of the current study. However, this comparison carries less weight than criterion a or b. Furthermore, the degree to which historical control data are useful for this purpose is highly dependent on the quantity and quality of the underlying data. For instance, historical control data cannot serve as a useful comparator (*e.g.*, as a proxy for normal biological variation), when inter-study variation represents the predominant source of variation. This and other concepts surrounding the appropriate use of historical control data are provided in Annex 2. Thus, only to the degree historical negative control data have been generated in sufficient quantity, and these data are deemed of sufficient quality, that in addition to criteria a and b, a third criterion, c, can be applied as follows:

c. None of the mutant frequencies of any of the test chemical dose groups exceed the upper bound limit derived from the historical negative control data distribution.

The upper bound limit derived from the historical negative control distribution should be determined using a method appropriate to the quantity and quality of the laboratory's data (e.g., control charts, tolerance intervals, or quantiles), with justification provided in the study report. As inter-laboratory variability may influence the control distribution, no fixed numerical value can be prescribed, and laboratories should define their upper bound limits in a scientifically sound and transparent manner, as described in Annex 2.

In many instances, the frequencies being considered will be the current study's mean values compared to a distribution derived from historical negative control mean values. However, there are situations that make historical negative control data and related distributions based on individual animal data useful; for example, when historical control values are low in number, or when it is of interest to estimate sources of variation. As explained in more detail in Annex 2, in these cases it is important to make appropriate comparisons. Specifically, a distribution based on individual animal data should *not* be used as a comparator for the current study's treatment group means. Rather, a distribution based on individual animal data should only be used as a comparator for the current study's individual animal data. This requirement is referred to as a "like to like" comparison in Annex 2.

Negative results indicate that, under the test conditions, the test chemical does not induce gene mutation in the tested tissue.

64. Evidence of exposure of the tested tissue to a test chemical or its metabolites may be gained from general toxicity (*e.g.*, reduced body/organ weight), or morphological or histopathological data obtained from determinations in the same study, or comparable toxicity studies, or other in vivo genotoxicity studies. Alternatively, ADME or TK data, or plasma analyses, obtained in the same or an independent study using the same route and same species can be used to demonstrate tissue exposure (1).

65. Regarding the statistical approaches described in paragraphs 62 and 63 above, it is important to recognise there is no single correct method of conducting a statistical analysis. A practical approach is to suggest a particular set of statistical analyses as *an example* of the sort of evaluations that can be carried out. This is the context that three types of analyses (a, b, c) described above were presented. See Annex 2 for information that testing laboratories and regulatory reviewers may find useful when considering whether a particular statistical analysis is appropriate and whether the data interpretation is scientifically sound. Whatever statistical approach is used, it is important that it is described in advance of conducting a study.

66. Especially in those instances when the criteria being considered (a and b; or a, b and c) are not all in alignment with respect to a positive or negative call, sound scientific judgment should be applied in an effort to interpret the results as either positive or negative. It should also be noted that, in general, criterion c has less weight than criteria a and b.

67. If the application of sound scientific judgement is unable to resolve whether a response is either positive or negative, further investigations of the *existing experiments* may be necessary to establish the biological relevance of a result. For example, these investigations may include analysing more plaques or mutant colonies.

68. Sequencing of mutant plaques to determine whether there is a shift in the mutation spectrum induced by the test chemical may also aid in concluding whether the response is negative or positive. As described in paragraph 58, sequencing can also help to identify

jackpot mutations. For DNA sequencing analyses, a number of statistical approaches are available to assist in interpreting the results (54) (55) (56) (57).

69. In rare cases, even after further investigation, if the data preclude concluding whether the test chemical produced a positive or negative result, the study is deemed equivocal.

70. In some cases, equivocal studies may prompt the need for a repeat experiment using modified experimental conditions, or changing the test system.

7.3. Test report

71. The test report should include the following information:

Test chemical:

- identification data and CAS n°, if known;
- source, lot number if available;
- physical nature and purity;
- physico-chemical properties relevant to the conduct of the study;
- stability of the test chemical, if known;

Solvent/vehicle:

- justification for choice of vehicle;
- solubility and stability of the test chemical in the solvent/vehicle, if known;
- preparation of dose formulations including dietary, drinking water or inhalation formulations;
- analytical determinations on formulations (*e.g.*, stability, homogeneity – for non-soluble substances, nominal concentrations);

Test animals:

- species and strain used and justification for the choice;
- number, age and sex of animals;
- source, housing conditions, diet, etc.;
- individual weight of the animals at the start of the test, including body weight range, mean and standard deviation for each group;

Test conditions:

- report whether (yes/no) blinding of the study (or parts of the study) was performed
- evidence for laboratory proficiency
- positive and negative (vehicle/solvent) control data;
- rationale for dose level selection, such as data from the range-finding study;
- details of test chemical preparation;

- details of the administration of the test chemical;
- rationale for route of administration;
- rationale for tissues/cell type analysed
- methods for measurement of animal toxicity, including, where available, histopathological or haematological analyses and the frequency with which animal observations and body weights were taken;
- methods for verifying that the test chemical reached the target tissue, or general circulation, if negative results are obtained;
- actual dose (mg/kg body weight/day) for most dose routes or for routes for diet/drinking water exposure either parts per million (ppm) or actual dose based on chemical concentration (ppm) and average food or water consumption, if applicable;
- details of food and water quality;
- detailed description of treatment and sampling schedules and justifications for the choices;
- method of euthanasia;
- procedures for isolating and preserving tissues;
- methods for isolation of rodent genomic DNA, rescuing the transgene from genomic DNA, and transferring transgenic DNA to a bacterial host;
- source and lot numbers of all cells, kits and reagents (where applicable);
- methods for enumeration of mutants;
- methods for molecular analysis of mutants and use in correcting for clonality and/or calculating mutation frequencies, if applicable;

Results:

- animal condition prior to and throughout the test period, including signs of toxicity;
- body weights and body weight changes throughout the test period;
- food and/or water consumption throughout the test period, if applicable for dosed food or drinking water studies;
- body and, if applicable, organ weights at humane killing;
- evidence of tissue exposure;
- for each tissue/animal, the number of mutants, number of plaques or colonies evaluated, number of packaging and packaging efficiency, mutant frequency;
- for each tissue/animal group, total number of mutants, mean mutant frequency, standard deviation;
- dose-response relationship, where possible;
- for each tissue/animal, the number of independent mutants and mean mutation frequency, where molecular analysis of mutations was performed;
- concurrent negative control data;
- concurrent positive control data;

- historical negative control data with number of rodents clearly specified, statistical description of the distribution including type of calculation used to generate lower and/or upper bound limits used to aid data interpretation and justification for its use (see Annex 2, section 4.2), as well as the time period covered and evidence that the assay is under control during the period covered; a representative table that captures this and other useful information is provided in Annex 2, section 6;
- historical positive control data with number of rodents clearly specified, statistical description of the distribution including type of calculation used to generate lower and/or upper bound limits used to aid data interpretation and justification for its use (see Annex 2, section 4.2), as well as the time period covered and evidence that the assay is under control during the period covered;
- analytical determinations, if available (*e.g.*, DNA concentrations used in packaging, DNA sequencing data);
- statistical analyses and methods applied;

Discussion of the results:

- Explain how scientific judgement was used in evaluating the results of the test; especially in those instances when some, but not all, of the criteria are met for a negative or a positive result;

Conclusions:

Provide a concise summary of the study findings, including:

- whether the test chemical showed evidence of gene mutation under the test conditions;
- an assessment of the biological and statistical relevance of the results;
- any limitations or uncertainties that may affect the interpretation of the findings.

8. LITERATURE

(1) OECD (Organisation for Economic Cooperation and Development) (2017) Overview of the set of OECD Genetic Toxicology Test Guidelines and updates performed in 2014-2015. Series on Testing and Assessment, No. 238 – 2nd edition. Available at: Overview of the set of OECD Genetic Toxicology Test Guidelines and updates performed in 2014-2015 - Second edition | OECD.

(2) Thybaud, V, E. Lorge, D.D. Levy, J. van Benthem, G.R. Douglas, F. Marchetti, M.M. Moore and R. Schoeny (2017), Main issues addressed in the 2014-2015 revisions to the OECD genetic toxicology test guidelines", *Environ. Mol. Mutagen*., 58:284-295.

(3) OECD (2009), *Detailed Review Paper on Transgenic Rodent Mutation Assays*, Series on Testing and Assessment, N° 103, ENV/JM/MONO(2009)7, OECD, Paris.

(4) OECD (2011), *Retrospective Performance Assessment of OECD Test Guideline on Transgenic Rodent Somatic and Germ Cell Gene Mutation Assays*, Series on Testing and Assessment, N° 145, OECD, Paris.

(5) Marchetti, F., M. Aardema, C. Beevers, J. van Benthem, R. Godschalk, C.L. Yauk, B. Young, A. Williams and G.R. Douglas (2018), "Identifying germ cell mutagens using OECD test guideline 488 (transgenic rodent somatic and germ cell mutation assay) and integration with somatic cell testing", Mutation Res., 832-833: 7-18. Corrigendum: Mutation Res., 2019, 844: 70-71.

(6) Heddle, J.A., S. Dean, T. Nohmi, M. Boerrigter, D. Casciano, G.R. Douglas, B.W. Glickman, N.J. Gorelick, J.C. Mirsalis, H.-J Martus, T.R. Skopek, V. Thybaud, K.R. Tindall and N. Yajima (2000), "In vivo Transgenic Mutation Assays", *Environ. Mol. Mutagen*., 35: 253-259.

(7) Thybaud, V., 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 and N. Yajima (2003), "*In vivo* Transgenic Mutation Assays", *Mutation Res*., 540: 141-151.

(8) Heddle, J.A., H.-J. Martus and G.R. Douglas (2003), "Treatment and Sampling Protocols for Transgenic Mutation Assays", *Environ. Mol. Mutagen.*, 41: 1-6.

(9) Maurice, C., D.S. Dertinger, C.L. Yauk and F. Marchetti (2019) "Integrated in vivo genotoxicity assessment of procarbazine hydrochloride demonstrates induction of Pig-a and *lacZ* mutations, and micronuclei, in MutaMouse hematopoietic cells", *Environ. Mol. Mutagen*., 60: 505-512.

(10) Erikson, R.P. (2003), "Somatic Gene Mutation and Human Disease other than Cancer", *Mutation Res.*, 543: 125-136.

(11) Erikson, R.P. (2010), "Somatic Gene Mutation and Human Disease other than Cancer: an Update", *Mutation Res.*, 705: 96-106.

(12) Jackson, M., L. Marks, G.H.W. May and J.B. Wilson (2018) "The genetic basis of disease", *Essays Biochem*., 62:643-723.

(13) Nohmi, T., M. Katoh, H. Suzuki, M. Matsui, M. Yamada, M. Watanabe, M. Suzuki, N. Horiya, O. Ueda, T. Shibuya, H. Ikeda and T. Sofuni (1996), "A new Transgenic Mouse Mutagenesis Test System using Spi⁻ and 6-thioguanine Selections", *Environ. Mol. Mutagen*., 28(4): 465–470.

(14) Nohmi, T., T. Suzuki and K.I. Masumura (2000), "Recent Advances in the Protocols of Transgenic Mouse Mutation Assays", *Mutation Res.*, 455(1–2): 191–215.

(15) Toyoda-Hokaiwado, N., T. Inoue, K. Masumura, H. Hayashi, Y. Kawamura, Y. Kurata, M. Takamune, M. Yamada, H. Sanada, T. Umemura, A. Nishikawa and T. Nohmi (2010),

"Integration of *in vivo* Genotoxicity and Short-term Carcinogenicity Assays using F344 *gpt* delta Transgenic Rats: *in vivo* Mutagenicity of 2,4-diaminotoluene and 2,6-diaminotoluene Structural Isomers", *Toxicol. Sci.*, 114(1): 71-78.

(16) Gossen, J.A., W.J. de Leeuw, C.H. Tan, E.C. Zwarthoff, F. Berends, P.H. Lohman, D.L. Knook and J. Vijg (1989), "Efficient Rescue of Integrated Shuttle Vectors from Transgenic Mice: a Model for Studying Mutations *in vivo", Proc. Natl. Acad. Sci. USA,* 86(20): 7971–7975.

(17) Gossen, J.A. and J. Vijg (1993), "A Selective System for lacZ-Phage using a Galactose-sensitive *E. coli* Host", *Biotechniques*, 14(3): 326, 330.

(18) Vijg, J. and G.R. Douglas (1996), "Bacteriophage λ and Plasmid *lacZ* Transgenic Mice for studying Mutations *in vivo*" in: G. Pfeifer (ed.), *Technologies for Detection of DNA Damage and Mutations, Part II*, Plenum Press, New York, NY, USA, pp. 391–410.

(19) Jakubczak, J.L., G. Merlino, J.E. French, W.J. Muller, B. Paul, S. Adhya and S. Garges (1996), "Analysis of Genetic Instability during Mammary Tumor Progression using a novel Selection-based Assay for *in vivo* Mutations in a Bacteriophage λ Transgene Target", *Proc. Natl. Acad. Sci. USA*, 93(17): 9073–9078.

(20) Nohmi, T., M. Suzuki, K. Masumura,M. Yamada, K. Matsui, O. Ueda, H. Suzuki, M. Katoh, H. Ikeda and T. Sofuni (1999), "Spi⁻ Selection: an Efficient Method to Detect γ-ray-induced Deletions in Transgenic Mice", *Environ. Mol. Mutagen*., 34(1): 9–15.

(21) Gad, S.C., C.D. Cassidy, N. Aubert, Spainhour B. and H. Robbe (2006) "Nonclinical vehicle use in studies by multiple routes in multiple species", Int. J. Toxicol., 25(6): 499-521.

(22) OECD (2009), Part 2: Annexes to the Detailed Review Paper on Transgenic *Rodent Mutation Assays*, Series on Testing and Assessment, N° 103, ENV/JM/MONO(2009)29, OECD, Paris.

(23) Hayashi, M., K. Dearfield, P. Kasper, D. Lovell, H.-J. Martus, V. Thybaud (2011), "Compilation and Use of Genetic Toxicity Historical Control Data", Mutation Res., 732(2): 87-90.

(24) Ryan, T.P. (2000), "Statistical Methods for Quality Improvement", 2nd ed., John Wiley and Sons, New York.

(25) Fang, Y. (2003), "C-chart, X-chart, and the Katz Family of Distributions", J. Quality Technology, 35:1-15, 2003.

(26) Lovell D.P., M. Fellows, F. Marchetti, J. Christiansen, A. Elhajouji, K. Hashimoto, S. Kasamoto, Y. Li, O. Masayasu, M.M. Moore, M. Schuler, R. Smith, L.F. Stankowski Jr, J. Tanaka, J.Y. Tanir, V. Thybaud, F. Van Goethem and J. Whitwell (2018), "Analysis of negative historical control group data from the in vitro micronucleus assay using TK6 cells", Mutat Res Genet Toxicol Environ Mutagen., 825:40-50.

(27) Dertinger S.D., J.A. Bhalli, D.J. Roberts, L.F. Stankowski Jr, B.B. Gollapudi, D.P. Lovell, L. Recio, T. Kimoto, D. Miura and R.H. Heflich (2021) "Recommendations for conducting the rodent erythrocyte Pig-a assay: A report from the HESI GTTC Pig-a Workgroup", *Environ Mol Mutagen*., 62(3):227-237.

(28) Yauk, C.L., J.D. Gingerich, L. Soper, A. MacMahon, W.G. Foster and G.R. Douglas (2005), "A lacZ Transgenic Mouse Assay for the Detection of Mutations in Follicular Granulosa Cells", Mutation Res., 578(1-2): 117-123.

(29) OECD (2000), Guidance Document on the Recognition, Assessment and Use of Clinical Signs as Humane Endpoints for Experimental Animals Used in Safety Evaluation, Series on Testing and Assessment, N°19, ENV/JM/MONO(2000)7, OECD, Paris.

(30) Diehl K.H., R. Hull, D. Morton, R. Pfister,Y. Rabemampianina, D. Smith, J.M. Vidal, C. van de Vorstenbosch and European Federation of Pharmaceutical Industries Association and European Centre for the Validation of Alternative Methods, (2001) "A good practice guide to the administration of substances and removal of blood, including routes and volumes", J. Appl. Toxicol., 21(1):15-23.

(31) Turner P.V., T. Brabb, C. Pekow and M.A. Vasbinder, (2011) "Administration of substances to laboratory animals: routes of administration and factors to consider", J. Am. Assoc. Lab. Anim. Sci., 50(5):600-613.

(32) White, P.A., G.R. Douglas, D.H. Phillips and V.M. Arlt (2017) "Quantitative relationship between lacZ mutant frequency and DNA adduct frequency in Muta™Mouse tissues and cultured cells exposed to 3-nitrobenzanthrone. *Mutagenesis*, 32(2): 299-312.

(33) Marchetti F., G. Zhou, D. LeBlanc, P.A. White, A. Williams, C.L. Yauk and G.R. Douglas (2021) "The 28 + 28 day design is an effective sampling time for analyzing mutant frequencies in rapidly proliferating tissues of MutaMouse animals", Arch. Toxicol., 95(3):1103-1116

(34) Douglas, G.R., J. Jiao, J.D. Gingerich, J.A. Gossen and L.M. Soper (1995), "Temporal and Molecular Characteristics of Mutations Induced by Ethylnitrosourea in Germ Cells Isolated from Seminiferous Tubules and in Spermatozoa of lacZ Transgenic Mice", Proc. Natl. Acad. Sci. USA, 92: 7485-7489.

(35) Douglas, G.R., J.D. Gingerich, L.M. Soper and J. Jiao (1997), "Toward an Understanding of the Use of Transgenic Mice for the Detection of Gene Mutations in Germ Cells", Mutation Res., 388(2-3): 197-212.

(36) Singer, T.M., I.B. Lambert, A. Williams, G.R. Douglas and C.L. Yauk (2006), "Detection of Induced Male Germline Mutation: Correlations and Comparisons between Traditional Germline Mutation Assays, Transgenic Rodent Assays and Expanded Simple Tandem Repeat Instability Assays", Mutation. Res., 598: 164-193.

(37) Olsen, A.K., A., Andreassen, R., Singh, R., Wiger, N., Duale, P.B., Farmer, and G. Brunborg (2010), "Environmental exposure of the mouse germ line: DNA adducts in spermatozoa and formation of de novo mutations during spermatogenesis". PLoS One, 28;5(6):e11349

(38) Oakberg, E.F. (1956), "Duration of spermatogenesis in the mouse and timing of the stages of the cycle of the seminiferous epithelium", Am. J. Anat., 99: 507–516.

(39) Clermont, Y. (1972), "Kinetics of spermatogenesis in mammals: seminiferous epithelium cycle and spermatogonial renewal". Physiol. Rev. 52: 198-236.

(40) Marchetti, F., M. Aardema, C. Beevers, J. van Benthem, G.R. Douglas, R. Godschalk, C.L. Yauk, B. Young and A. Williams (2018), "Simulation of mouse and rat spermatogenesis to inform genotoxicity testing using OECD test guideline 488", Mutation Res., 832-833: 19-28. Corrigendum: Mutation Res., 2019, 844: 69.

(41) Bielas, J.H. and J.A. Heddle (2000) Proliferation is necessary for both repair and mutation in transgenic mouse cells. Proc. Natl. Acad. Sci. USA, 97: 11391-11396.

(42) O'Brien, J.M., M.A. Beal, J.D. Gingerich, L. Soper, G.R. Douglas, C.L. Yauk and F. Marchetti (2014), "Transgenic rodent assay for quantifying male germ cell mutant frequency", *J. Vis. Exp.*, 90: e51576

(43) O'Brien J.M., Beal M.A., Yauk, C.L. and F. Marchetti (2016) "Benzo(a)pyrene is mutagenic in mouse spermatogonial stem cells and dividing spermatogonia. Toxicol. Sci., 152: 363-371.

(44) Dean, S.W., T.M. Brooks, B. Burlinson, J. Mirsalis, B. Myhr, L. Recio and V. Thybaud (1999), "Transgenic Mouse Mutation Assay Systems can Play an important Role in Regulatory Mutagenicity Testing in vivo for the Detection of Site-of-contact Mutagens", Mutagenesis, 14(1): 141-151.

(45) Bielas, J.H. (2002), "A more Efficient Big Blue® Protocol Improves Transgene Rescue and Accuracy in an Adduct and Mutation Measurement", Mutation Res., 518: 107-112.

(46) Kohler, S.W., G.S. Provost, P.L. Kretz, A. Fieck, J.A. Sorge and J.M. Short (1990), "The Use of Transgenic Mice for Short-term, in vivo Mutagenicity Testing", Genet. Anal. Tech. Appl., 7(8): 212–218.

(47) Nohmi, T., K. Masumura and N. Toyoda-Hokaiwado (2017), "Transgenic rat models for mutagenesis and carcinogenesis", Genes and Environ., 39:11.

(48) Besaratinia, A., H. Li, J. Yoon, A. Zheng, H. Gao and S. Tommasi (2012) "A high-throughput next-generation sequencing-based method for detecting the mutational fingerprint of carcinogens", *Nucleic Acids Res*., 40(15):e116-6.

(49) Beal, M.A., R. Gagne, A. Williams, Marchetti F. and C.L. Yauk (2015) "Characterizing benzo(a)pyrene-induced lacZ mutation spectrum in transgenic mice using Next Generation Sequencing", BMC Genomics, 16: 812.

(50) Carr, G.J. and N.J. Gorelick (1995), "Statistical Design and Analysis of Mutation Studies in Transgenic Mice", Environ. Mol. Mutagen, 25(3): 246–255.

(51) Fung, K.Y., G.R. Douglas and D. Krewski (1998), "Statistical Analysis of lacZ Mutant Frequency Data from MutaTMMouse Mutagenicity Assays", Mutagenesis, 13(3): 249–255.

(52) Piegorsch, W.W., B.H. Margolin, M.D. Shelby, A. Johnson, J.E. French, R.W. Tennant and K.R. Tindall (1995), "Study Design and Sample Sizes for a lacI Transgenic Mouse Mutation Assay", Environ. Mol. Mutagen., 25(3): 231–245.

(53) Piegorsch, W.W., A.C. Lockhart, G.J. Carr, B.H. Margolin, T. Brooks, G.R. Douglas, U.M. Liegibel, T. Suzuki, V. Thybaud, J.H. van Delft and N.J. Gorelick (1997), "Sources of Variability in Data from a Positive Selection lacZ Transgenic Mouse Mutation Assay: an Interlaboratory Study", Mutation. Res., 388(2–3): 249–289.

(54) Adams, W.T. and T.R. Skopek (1987), "Statistical Test for the Comparison of Samples from Mutational Spectra", *J. Mol. Biol*., 194: 391-396.

(55) Carr, G.J. and N.J. Gorelick (1996), "Mutational Spectra in Transgenic Animal Research: Data Analysis and Study Design Based upon the Mutant or Mutation Frequency", Environ. Mol. Mutagen, 28: 405–413.

(56) Dunson, D.B. and K.R. Tindall (2000), "Bayesian Analysis of Mutational Spectra", Genetics, 156: 1411–1418.

(57) Lewis P.D., B. Manshian, M.N. Routledge, G.B. Scott and P.A. Burns (2008), "Comparison of Induced and Cancer-associated Mutational Spectra using Multivariate Data Analysis", Carcinogenesis, 29(4): 772-778.

(58) Dertinger SD, D Li, C Beevers, GR Douglas, RH Heflich, DP Lovell, DJ Roberts, R Smith, Y Uno, A Williams, KL Witt, A Zeller, C Zhou (2023) Assessing the quality and making appropriate use of historical negative control data: A report of the International Workshop on Genotoxicity Testing (IWGT). Environ Mol Mutagen, in press, doi: 10.1002/em.22541

Annex 1: Definitions

Administration period: the total period during which an animal is dosed.

Base pair substitution: a type of mutation that causes the replacement of a single DNA nucleotide base with another DNA nucleotide base.

Capsid: the protein shell that surrounds a virus particle.

Clonal expansion: the production of many cells from a single (mutant) cell.

Colony-forming unit (cfu): a measure of viable bacterial numbers.

Confidence interval (CI): a range of values that is likely to include a population value with a certain degree of confidence

Control limit: horizontal line(s) drawn on a statistical control chart; these are investigator-defined and use-case-dependent values; in the field of Quality Control, these values are typically the sample mean ± 3 standard deviations, but some other multiples of the standard deviation can be useful *e.g.*, 2x or 1.96x

Cos site: a 12-nucleotide segment of single-stranded DNA that exists at both ends of the bacteriophage lambda's double-stranded genome.

Deletion: a mutation in which one or more (sequential) nucleotides is lost by the genome.

Electroporation: the application of electric pulses to increase the permeability of cell membranes.

Endogenous gene: a gene native to the genome.

Extrabinomial variation: greater variability in repeat estimates of a population proportion than would be expected if the population had a binomial distribution.

Frameshift mutation: a genetic mutation caused by insertions or deletions of a number of nucleotides that is not evenly divisible by three within a DNA sequence that codes for a protein/peptide.

Insertion: the addition of one or more nucleotide base pairs into a DNA sequence.

Jackpot: a large number of mutants that arose through clonal expansion from a single mutation.

Large deletions: deletions in DNA of more than several kilobases (which are effectively detected with the Spi$^-$ selection assay).

Ligation: the covalent linking of two ends of DNA molecules using DNA ligase.

Mitogen: a chemical that stimulates a cell to commence cell division, triggering mitosis (*i.e.*, cell division).

Neutral gene: a gene that is not affected by positive or negative selective pressures.

Packaging: the synthesis of infective phage particles from a preparation of phage capsid and tail proteins and a concatemer of phage DNA molecules. Commonly used to package DNA cloned onto a lambda vector (separated by cos sites) into infectious lambda particles.

Packaging efficiency: the efficiency with which packaged bacteriophages are recovered in host bacteria.

Plaque forming unit (pfu): a measure of viable bacteriophage numbers.

Point mutation: a general term for a mutation affecting only a small sequence of DNA including small insertions, deletions, and base pair substitutions.

Positive selection: a method that permits only mutants to survive.

Reporter gene: a gene whose mutant gene product is easily detected.

Sampling time: the end of the period of time, prior to humane killing, during which the test chemical is not administered and during which unprocessed DNA lesions are fixed into stable mutations.

Shuttle vector: a vector constructed so that it can propagate in two different host species; accordingly, DNA inserted into a shuttle vector can be tested or manipulated in two different cell types or two different organisms.

Test chemical: The term test chemical is used to refer to the substance being tested.

Transgenic: of, relating to, or being an organism, whose genome has been altered by the transfer of a gene or genes from another species.

Annex 2 - Statistical analysis and data interpretation

1. Preface

1. This Annex draws partly from an International Workshop on Genotoxicity Testing (IWGT) report [Dertinger et al., 2023]. For illustrative datasets and detailed method descriptions, readers are encouraged to review this publication.

2. Statistical analysis methods used in this assay can vary, and this Test Guideline provides three examples in paragraphs 62 - 63 as practical, non-prescriptive approaches. Alternative methods are acceptable if justified with sound statistical reasoning. Investigators should document their chosen methods in a written study or validation plan before starting the study and be prepared to defend their approach.

3. The statistical approaches below are representative tools for evaluating genotoxicity, but sound scientific judgement is essential, particularly when statistical analyses exhibit conflicting results. The application of sound scientific judgement is consistent with recommendations made by an OECD genotoxicity working group that met in Ottawa, Canada in 2013 [OECD, 2017].

4. Considering the language in Test Guideline paragraphs 62 - 63 and the data evaluation tools described in this Annex, the following overarching principles should be kept in mind:

i. When concurrent negative control data fulfil study acceptability criteria, they represent the most important comparator for judging whether a particular test chemical induced a genotoxic effect.
ii. Historical control data can provide useful context for interpreting study results, but this requires supporting evidence including: a) relevant historical control data that have been generated in sufficient quantity, and; b) data that accurately describes inter-animal variability.
iii. Historical control data should be visualized before any study comparisons take place; graph(s) that show the degree to which the data are stable over time are particularly useful.
iv. Qualitative and semi-quantitative assessments of historical control data should be supplemented with quantitative evaluations including: a) the stability of the data over time, and b) the degree to which inter-study variation explains the total variability observed.
v. When inter-animal variation is the predominant source of variability, the relationship between study responses and a historical control-derived interval or upper bounds value (*i.e.*, criterion c) can be used with a strong degree of confidence in contextualizing the results.
vi. When inter-study variation is the major source of variability, comparisons between study data and the historical control-derived bounds are less useful, and consequentially, less emphasis should be placed on using historical control data to contextualize a particular study's results.
vii. Most statistical tests, including those in this Annex, assume random sampling or blocking to mitigate uncontrolled factors on study results.

2. Pairwise comparisons

5. One type of statistical test that may be implemented are pairwise comparisons that determine if the concurrent vehicle/solvent (negative) control group is significantly different from the test chemical

dose groups. Parametric analyses that use analysis of variance (ANOVA) with an appropriate multiple comparisons test are commonly used, but other methodologies are equally acceptable so long as assumptions underlying the analyses are met. For example, if data transformation does not result in homoscedasticity, weighted (variance-corrected) ANOVA and/or t-tests or appropriate non-parametric methods can be considered.

2.1 Factorial design

6. If more than one sex is used in a study, factorial design approaches are generally advantageous as it evaluates the effects of sex and test chemical dose simultaneously and examines how these factors interact. Standard statistical software can be used to analyse the data generated from this design.

7. The analysis first inspects the sex-by-dose interaction term in the ANOVA table. (Note that statisticians who take a modelling approach such as using General Linear Models may approach the analysis in a different but comparable way but will not necessarily derive the traditional ANOVA table). In the absence of a significant interaction term, the combined values across sexes or across dose levels provide valid statistical tests between the levels based upon the pooled within-group variability term of the ANOVA.

8. The analysis continues by partitioning the estimate of the between-dose-level variability into contrasts, which test for linear and quadratic contrasts of the responses across the vehicle/solvent concurrent control and test chemical dose levels. When there is a significant sex x dose interaction, this term can be partitioned into linear x sex and quadratic x sex interaction contrasts. These terms test whether the dose level responses are parallel for the two sexes or there is a differential response between the two sexes.

9. The pooled within-group variability estimate allows pairwise tests between means, such as between sexes or dose levels (*e.g.*, against negative (solvent/vehicle) control levels). Where there is a significant interaction, comparisons can be made between the means of different dose levels within a sex or sexes within a dose level.

3. Trend test

10. A second type of analysis that may be used is the trend test to identify a dose-response relationship. For this test, data should be available from the negative control group and all experimental dose groups. However, trend tests are normally not applicable for Limit Tests, which involve only a single test chemical dose level as described in paragraph 40). When employing these analyses, care is needed in interpreting the results. For example, a simple linear trend test may fail to detect a trend if the dose-response is non-monotonic. In such cases, more advanced tests, like the downturn protection test proposed by Bretz and Hothorn [2003] may be useful. Creating a graph of the dose response can aid in choosing the appropriate test or determining if a transformation is needed.

4. Historical negative control data distribution

11. Distributions of historical negative control data are important for determining the minimum number of observations that should be evaluated and for assessing a particular study's acceptability and also when determining whether the results of any test chemical-exposed treatment group exceeds the upper bounds of the historical negative control data distribution ("criterion c"). Given their importance, guidance for building historical negative control datasets is provided below.

4.1 Building historical negative control datasets: Group means versus individual animal data

12. There is value in compiling historical negative control data based on both individual animal data and as study group means. First, individual animal data facilitate certain statistical analyses that cannot be accomplished with group means. One example is Variance Components Analysis aimed at estimating the extent to which inter-animal differences versus study-dependent nuisance factors explain the variation observed in an assay. This important consideration cannot be addressed when analyses are based on group mean data. Second, initially focusing on individual animal data represents a practical recommendation because, for some *in vivo* test systems, it may be difficult to achieve adequately sized databases with group means due to limitations on animal testing.

13. As a laboratory gains experience and conducts more studies, the historical control databases can expand to include both individual animal data and group mean data. Individual animal data will remain essential for evaluating the quality and stability of historical negative control data. However, the group mean historical negative control data distribution may replace the individual animal data distribution in the context of criterion c-type assessments.

14. While individual animal- and group mean-based databases are interrelated, intervals for historical negative control data based on individual animals are wider than those based on group means. Therefore, a guiding principle is *like should be compared to like* meaning individual animal results should be compared with intervals based on individual animal data, while group mean results are most appropriately compared to intervals based on group means.

4.2 Database size

15. A historical negative control database (HCD) should consist of data from at least ≥ 20 independent studies and ≥ 100 individual animals. These values are consistent with recommendations by Hayashi et al. [2011] and Igl and colleagues [2019]. A database with < 20 independent studies and < 100 individuals may still be useful but should carry less weight compared to later versions derived from ≥ 20 independent studies and ≥ 100 individual animals.

16. The goal of achieving 20 independent studies and upwards of 100 animals is complicated by the fact that some endpoints may be affected by certain experimental factors such as species, strain, age, sex, vehicle composition, etc. For the TGR mutation assay, the tissue being analysed for mutations or DNA strand breaks should also be considered. When factor(s) are found to appreciably affect the endpoint in question, separate historical control databases may become necessary, which is an unrealistically large burden that may result in many small databases that never reach the goal of 100 observations. In these instances, a practical solution may become necessary, for example conducting a relatively small number of independent experiments, each with 10 animals, and each with a different, commonly used vehicle. Once these pilot studies are complete, and so long as the data are found to be "under control" (see Section 4.4, "Evaluating the Quality of Historical Control Data"), an appropriately calculated historical control interval based on individual animal data may be used for assay acceptability and addressing criterion c. However, as explained previously, analyses conducted with such data should carry less weight relative to later databases that are more appropriately sized.

17. In most instances, the historical negative control databases will be generated with data from animals dosed with the solvent/vehicle alone. For some endpoints, historical negative control databases can benefit from the inclusion of naïve animal data. This assumes solvent/vehicle effects are not appreciable. This example stresses the importance of supplementing the historical control genotoxicity

data with the reporting of metadata that includes experimental factors such as animal model, strain, sex, solvent/vehicle, age at time of tissue harvest, and analytical equipment configuration.

18. Related to this, assuming supporting literature or other data are available, it may be appropriate to initially consider factor(s) such as sex and/or solvent/vehicle to have no effect on a genotoxicity endpoint. This helps one more efficiently construct an adequately sized database. However, as more data accumulate, these assumptions should be re-evaluated, and in this manner decisions about pooling or parsing of databases can be made on sound scientific reasoning.

4.3 Maintaining historical control databases

19. Historical control data can be from studies conducted under Good Laboratory Practices (GLP) guidelines, non-GLP studies, or from both, so long as the experimental conditions are similar, and quality control measures confirm accuracy of data not covered by a GLP-compliance statement. An established historical control database should be maintained and reissued at some consistent interval, or else by some other defined triggering event, unless no new data have been generated in the previous 12 months. While historical control databases may be updated on a predetermined schedule, assay acceptability criteria must be evaluated on a per study basis allowing for real time feedback on assay performance so as to be useful for modifying a fixed database update schedule.

20. For example, if a study does not meet assay acceptance criteria, this should trigger work aimed at identifying the underlying cause(s). If an underlying cause *is identified* (*e.g.*, an explainable technical error, reagent issue, or equipment failure), these assay control data should *not* be used in the next iteration of the historical control database. For the sake of transparency, such data should be added to the database, but omitted from the calculation of interval(s) that form the basis of assay acceptability criteria or for addressing criterion c.

21. If an assay fails assay acceptance criteria and the underlying cause *is not identified*, then a decision must be made as to whether the assay control data should be included in a future historical control database. To maximize transparency, the decision process for excluding control data from a database should most ideally be articulated *a priori*. If the unusual control data point(s) are deemed acceptable for inclusion in a future historical control database because no technical error, etc. was identified, this should ideally prompt a reissue of the database and associated analyses (*e.g.*, generation of a control chart, new interval calculation, etc.) ahead of schedule, so that the historical control database remains representative of upcoming studies.

22. Finally, a timeframe should be selected to analyse the historical control data for stability, such as 5 years as an upper bound. Accordingly, assay-specific factors will often need to be considered. For instance, if the assay is conducted infrequently, or a particular tissue is rarely studied, older data may be used to obtain approximately 20 studies and 100 independent observations, provided there have been no changes in methodology over the time period.

4.4 Evaluating the quality of historical negative control data

23. To evaluate the quality of HCD, a qualitative and semi-quantitative approach should be taken [OECD, 2017]. Initial assessments should evaluate whether (i) the genotoxicity data are consistent with published results from proficient laboratories (here, it can be useful to seek out interlaboratory validation/transferability-centric papers); (ii) the level of variation across samples within a study and across studies compare to published results from proficient laboratories; and (iii) there is an obvious drift to higher or lower frequencies with respect to time. Quantitative analyses should also be applied and

usually begin by characterizing the distribution of the data, which can be useful as residuals that do not approximate a normal distribution should not support some of the analyses described below.

24. When data transformation is deemed necessary, computational tools may be utilized to select an appropriate transformation. For instance, the Box-Cox approach available in many statistical software [Box and Cox, 1964] calculates an estimate of the lambda power parameter, which can be used to identify an optimal transformation. However, simply rounding lambda to the nearest 0.5 value helps select from common transformations. For instance, a lambda of -1 = reciprocal, 0 = logarithmic, 0.5 = square root, 1 = untransformed, and 2 = squared.

25. The next analysis should use statistical tools widely used in the fields of manufacturing and process control. Control charts are a useful for studying historical control data [Hayashi et al., Mutat. 2011] and are recommended in several genetic toxicology OECD TGs [OECD, 2017]. A control chart is a graph used to study how a process changes over time. Data are plotted in a sequential order that reflects when they were collected over time. A control chart has a central line plotted parallel to the X-axis describing the mean, an upper line for the upper control limit (UCL; mean plus 3 standard deviations, *i.e.*, sigma +3), and a lower line for the lower control limit (LCL; sigma -3). Graphs can show sigma ±2 lines, which are known as "warning limits." (Note that standard deviation values plotted on control charts are not the same standard deviation values from summary statistics of the sample; rather, control chart standard deviation values are estimated by the average range divided by a normalizing constant, d_2, see Dertinger et al., 2023). By comparing current data to these lines, conclusions can be drawn about whether a process is consistent ("under control") or variable ("out-of-control").

26. Different types of control charts with different objectives may be useful for genetic toxicology data are briefly described here. I-charts can be used with either continuous or discrete (count) data and are time-ordered sequence charts that plot individual values on the y-axis and the order of the individuals on the x-axis. C-charts also plot individual values where no attempt is made to take into consideration any grouping that may exist in the data. The primary difference between I- and C-charts is that the latter are restricted to counts (numbers of events) and assume an underlying Poisson distribution. X-bar charts are plots of the means derived from a series of experiments, each of which is made up of several individuals (*e.g.*, the mean of 5 animals in a negative control group).

27. The X-bar chart is generally more informative when there are a reasonable number of studies (*e.g.*, 20 or more), while the I-chart is useful for a smaller number of studies. However, caution is needed when using I-charts since the data may not be truly independent because, for instance, observations from the same study may have correlated responses. If a group falls outside the limits of an X-bar chart, it is probable that several individual values from the group will also fall outside the limits for the I-chart. Ultimately, these control charts should be seen as complementary tools rather than one being superior to the other.

28. Decision rules can be used with control charts to highlight when a process is out-of-control. Codified rules are used to ensure that various control charts are interpreted in a consistent manner. Western Electric rules and Nelson rules are two widely used examples that can highlight an out-of-control process based on non-random characteristics of the data [Nelson, 1984]. For instance, Nelson rule number 3 is "violated" when six (or more) data points in a row are continually increasing or decreasing.

29. Some statistical software programs supplement control charts with metrics developed for monitoring manufacturing processes, which quantify the variability of a process. One example is the Stability Index (SI) [Jensen et al., 2019] for non-grouped data that calculates by dividing the long-term standard deviation (*i.e.*, the total standard deviation of the database) by the short-term standard deviation (*i.e.*, the average of the moving range divided by d_2). SI values close to 1.0 are considered stable with

low variability. Thus SI, together with Nelson or Western Electric rules (or subsets), provide warning signals that aspect(s) of the assay have drifted to, or are in the process of drifting to, an out-of-control status.

30. Variance Components Analysis (VCA) is another category of tests that can provide insights into the quality of historical control data. There are a variety of related methods that include hierarchical or nested ANOVA, Restricted Maximum Likelihood (REML) approaches, and Bayesian modelling methods [Gelman and Hill, 2006]. Another strategy, called Evaluating the Measurement Process, has been described by Wheeler [2006]. These methods partition and estimate the proportion of total variation accounted for by various experimental factors providing an estimate of standard deviation for the source.

4.5 Calculating intervals that describe historical negative control distributions

31. There are several valid approaches for characterizing the distribution of historical negative control data, and each laboratory should use an appropriate method for describing their data. This typically involves calculating upper and/or lower bounds to define the range where most negative control animal values are expected to fall. Key considerations include sample size and whether the data are normally distributed. A brief overview of methods for calculating an upper and/or lower bound limit is provided below with further details available in the literature [Vardeman, 1992; InfinityQS International, 2014; and Kluxen et al., 2021].

4.5.1. Inappropriate methods

- Range: The range is typically based on all negative control animal values and is the difference between the minimum and maximum observed value. Range *does not adequately* describe the historical negative control distribution for the purpose of establishing useful lower and upper bound limits. This is because the range will widen as the number of samples increases and may depend on one or two extreme (unusual/outlier) values. A wide range may "reward" poorly performing laboratories by masking potentially adverse study outcomes.

- Confidence interval: A confidence interval is a range of estimates likely to include a population value (parameter), such as the mean with a defined degree of confidence (*e.g.*, 95%). While formal definitions involve considerations of repeat sampling from hypothetical populations, these are often impractical to apply in real-world contexts. Importantly, confidence intervals are *not useful* for adequately describing historical negative control distribution when establishing an upper bound limit for criterion c. This is because confidence intervals become narrower as sample size increases, reflecting the increased precision in estimating the population value (parameter), which makes them *inappropriate* for OECD criterion c data interpretation.

4.5.2. Appropriate methods

32. Methods for calculating an interval that describe the distribution of historical negative control data are listed below. A database of ≥ 100 individual rodents/data points from ≥ 20 independent experiments is preferred. However, smaller intervals may suffice, particularly if the data are normally distributed or can be normalized, and the assay is well-control. Intervals from smaller datasets should be weighted less heavily than those from larger databases, as they are less likely to cover the desired proportion.

33. Although historical control data may fall within these intervals, this does not mean that the data are appropriate for criterion c-type assessments. Excessive variability can result in limit/interval values that are too wide. This would be indicative of poor performance by the laboratory and should not be mistaken for the "normal, biological variability" of the assay. This cautionary statement applies to all intervals described below and underscores the recommendation to use methods such as Nelsen or Western Electric rules, SI, and/or VCA to evaluate the quality of historical control data before calculating distribution intervals.

- Control limit: In the field of Quality Control, multiples of the standard deviation, usually the mean ± 3 standard deviations, are used as control limits. These values are lines plotted on a control chart and may be accompanied by other standard deviation multiples, for instance the mean ± 2 standard deviations, which are referred to as warning limits. In conjunction with control charts, control and warning limits are valuable tools for assessing the degree to which a repeated process or test is under control.

  When setting up control charts, one should consider that the standard deviation used is not the same as the standard deviation from summary sample statistics. Multiplying this standard deviation by 2 or 3 yields values different from those used to create the warning and control limits. This discrepancy arises because statistical software often estimates the standard deviation using the "average moving range," as the true population standard deviation is unknown and needs to be estimated from the data. The sample estimate of the standard deviation is adjusted accordingly to minimize biases or errors.

  Assuming a normal distribution and an "under control" process, 99.73% of the negative control data should fall within 3 standard deviations of the mean, and approximately 95% of the data should fall within two standard deviations. Control and/or warning limits represent a useful resource for evaluating historical negative control data and for providing upper and/or lower bound limits that aid in data interpretation.

  When considering study data in the context of criterion c-type assessments, warning limits that describe where approximately 95% of normally distributed data fall (± 2 standard deviations) are generally more appropriate than control limits that are based on 3 standard deviations. The former is consistent with many OECD Test Guidelines (*e.g.*, OECD TG474, 2016), while the latter characterizes exceptionally rare/unusual data points, thereby generating intervals that are too wide for the criterion c use case. However, control limits based on 3 standard deviations may be useful for data acceptability criteria.

- Prediction intervals: Prediction intervals are designed to predict one, or several, future observation(s) based on existing data. For example, with a 95% prediction interval, a new result (or specified number of future observations) from a negative test article would be expected to fall in the prediction interval with 95% probability. Non-normal data should be transformed as necessary prior to calculating the interval, which can be back transformed to original units for use and reporting. Alternately, some computer software programs enable a non-parametric calculation that does not assume a normal distribution.

  Prediction intervals typically address a specified *number* of future observations, usually one or several, rather than a percentage of future observations. This may pose challenges for using a prediction interval to judge study validity for criterion c-type purposes, unless one accompanies the calculation with rationale for choosing a particular number of future observations.

- Tolerance intervals: Tolerance intervals, like prediction intervals, are forward-looking but are designed to predict future observations with specified coverage and confidence levels. Coverage refers to the percentage of future observations defined by the calculated lower and upper bounds (*e.g.*, 95%

or 99%), modeled with a specified confidence level. While tolerance intervals are often calculated using 95% confidence, this parameter can be user defined.

Considerations for normality and optional non-parametric approaches discussed for prediction intervals also apply to tolerance intervals. In general, tolerance intervals are wider than prediction intervals since they are designed to predict a high percentage of future values, while prediction intervals typically focus on estimating one or a few new observations.

- Quantiles: Quantiles are used for summarizing the rank of data points based on size without assuming any specific probability distribution. Quantiles are widely used in many biomedical applications where non-normality because of outliers and/or skewness is common. For instance, intervals based on percentiles are often used to help interpret test results. Confidence intervals for quantiles can be calculated to provide estimates of uncertainty around the quantile measurement. These can help evaluate the quality of the underlying data set. Quantile confidence intervals will be especially wide for the tails of the distribution unless the sample size is large.

4.6 Using historical control data to interpret study data

34. Low quality historical control data are not suitable for criterion c-type assessments. With this in mind, this section describes the use of historical control data to interpret study data and is based on the assumption that the database is large enough, and of an acceptable quality, for this purpose.

35. A common approach for fulfilling criterion c involves conducting a large number of studies, then using the study mean values to calculate a historical control distribution-based interval (or else just the upper bound limit value). In this case, the aforementioned “like to like” comparison translates to a straight-forward assessment. That is, for any particular study, the treatment group mean values are compared to the upper bound limit value (derived from historical negative control study means) that separates “expected” from “elevated” values. This is a valid and useful approach, with the caveat that such a strategy requires many separate studies and consequentially large numbers of rodents, and the use of laboratory animals for this and other purposes is becoming more challenging in many regions of the world.

36. As discussed previously, it is also possible to use *individual* animal data to evaluate the quality of historical control data, and to calculate a useful interval (or else just an upper bound limit value) that can be used for criterion c purposes (at least until enough studies have been performed to justify a study mean-based interval). However, after arriving at an upper bound that separates expected from elevated values, it may be less obvious how one should proceed to evaluate study results on a per animal basis. In this situation, make use of well documented, sound scientific reasoning.

37. When sound scientific reasoning is used to assess individual animal data against an appropriately derived upper bound limit value, at least four factors should be considered. First, how many animals and numbers of studies were used to calculate the interval or upper bound limit? For example, a database and upper limit based on 10 studies and 50 animals should carry more weight relative to 5 studies and 25 animals. Second, how many test chemical treated animal responses were above the upper bound value? That is, the greater the number, the higher the concern. Third, which treatment group(s) exhibited elevated values? For example, all the elevated value(s) appearing in the low dose group may be indicative of a technical issue as opposed to a treatment-related effect. Fourth, what was the magnitude of the elevated response(s)? That is, greatly elevated values would be of higher concern than values that barely exceeded the upper bound limit.

38. This is not an exhaustive list of factors that should be considered. Rather, this subset focuses on the interpretation of elevated individual animal responses against an upper bound limit, and how they may contribute to the contextualization of statistical analyses (*i.e.*, criteria a and b). Other important considerations, such as the distribution and metabolism of a test substance, kinetics of appearance and disappearance of the endpoint, etc., are additional important elements that must be considered when assessing biological plausibility.

5. Interpretation of study results

39. The statistical approaches described in the preceding sections are regarded as representative tools for evaluating whether the test chemical induces genotoxicity. Accordingly, sound scientific judgement must take a leading role when considering any data set. This is especially obvious when statistical analyses exhibit, what on the surface, may appear to be conflicting results. This paradigm was reinforced by an expert OECD genotoxicity working group that met in Ottawa, Canada, 2013 [OECD, 2017].

40. In some cases, even after applying scientific judgement, appropriate statistical tools, and possibly evaluating additional data, it will not be possible to classify a response as positive or negative. In these cases, the response is equivocal and further testing may be required to resolve the genotoxicity of the test chemical.

6. Representative historical negative control table

41. As indicated in the Test Report section, study reports should provide information regarding the manner in which historical negative control data were collected, evaluated, and utilised. The following table is offered as a guideline.

Annex 2, Table 1. Representative historical negative control table*

	Information	Description and Comments
1. General description	Date range	Period over which the HCD were collected.
	Test system	The endpoint investigated and the model used. For an in vivo study, this may include the species, strain, sex, and age range of the animals at the time of dose initiation and tissue harvest, and the randomization or blocking procedures employed.
	Analytical system	Method of data collection.
2. Basic metrics	Number of studies and observations	Number of studies the data were collected from, and the total number of observations used in the HCD, including a statement on whether the observations are individual animals or group means.
	Median, mean** and Std. Dev.	Median and mean of the HCD, with its standard deviation. **Alternative measures of central tendency may be used.
	Observed range	Minimum and maximum values observed in the HCD.
	Group structures	Describe the rationale for combining or separating historical control databases based on experimental factors such as sex, vehicles, dose routes, age groups, amount of available data, and so forth.
3. Quality metrics	Transformation	A description of the analysis performed to select the appropriate transformation and justification for the final transformation selected. Visualization of the transformed HCD is recommended.
	Control chart	A longitudinal graphical display of the HCD, that includes the type of control chart presented and displays both warning and control limits.
	Variability over time	An assessment of the stability of HCD over time, including a brief description of the methods used.
	Sources of variation	An assessment of the predominant source of variability in the HCD, including a brief description of the methods used.
	Distribution interval	Interval that describes the distribution of the HCD, together with a brief description of the method used to calculate the interval and a justification for why the specific interval was selected.

*Given the amount and type of information requested above, one practical suggestion would be for sponsors to supply these data in an appendix or annex to a regulatory study report. It would remain consistent for some period of over time, until the historical negative control database has been updated with new data/distributions.

7. References for Annex 2, in order of appearance

Dertinger SD, D Li, C Beevers, GR Douglas, RH Heflich, DP Lovell, DJ Roberts, R Smith, Y Uno, A Williams, KL Witt, A Zeller, C Zhou (2023) Assessing the quality and making appropriate use of historical negative control data: A report of the International Workshop on Genotoxicity Testing (IWGT). Environ Mol Mutagen, in press, doi: 10.1002/em.22541.

OECD (Organisation for Economic Cooperation and Development). (2017) Overview of the set of OECD Genetic Toxicology Test Guidelines and updates performed in 2014-2015. Series on Testing and Assessment, No. 238 – 2nd edition. Available at: https://www.oecd.org/en/publications/overview-of-the-set-of-oecd-genetic-toxicology-test-guidelines-and-updates-performed-in-2014-2015-second-edition_61eca5cd-en.html.

Bretz F, LA Hothorn (2003) Statistical analysis of monotone or non-monotone dose-response data from in vitro toxicological assays. ATLA-Altern Lab Animal 31(Suppl 1):81-96.

Hayashi M, K Dearfield, P Kasper, D Lovell, H-J Martus, V Thybaud (2011) Compilation and use of genetic toxicity historical control data. Mutat Res 723:87-90.

Igl B-W, A Bitsch, F Bringezu et al. (2019) The rat bone marrow micronucleus test: Statistical considerations on historical negative control data. Regul Toxicol Pharmacol 102:13-22.

Box GEP, DR Cox (1964) An Analysis of Transformations, Journal of the Royal Statistical Society: Series B (Methodological), 26(2):211-243.

LS Nelson (1984) The Shewhart Control Chart—Tests for Special Causes. Journal of Quality Technology, 16(4):237-239.

WA Jensen, J Szarka III, J., K White (2019). Stability assessment with the stability index. Quality Engineering, 31(2):289-301.

A Gelman, J Hill (2006) Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge: Cambridge University Press.

DJ Wheeler (2006) EMP III (Evaluating the Measurement Process) Using Imperfect Data. SPC Press, Knoxville, TN.

Vardeman SB (1992) What about the other intervals? The American Statistician 46:193-197.

InfinityQS International (2014) A practical guide to selecting the right control chart. pp 1-21. Available at: http://www.infinityqs.com/sites/infinityqs.com/files/files/PDFs/InfinityQS_Practical_Guide_to_Selecting_the_Right_Control_Chart_Oct2013.pdf.

Kluxen FM, K Weber, C Strupp, SM Jensen, LA Hothorn, J-C Garcin, T Hofmann (2021) Using historical control data in bioassays for regulatory toxicology. Reg Toxicol Pharmacol 125:105024.

OECD (Organisation for Economic Cooperation and Development) (2016) OECD Guideline for the testing of chemicals: Mammalian erythrocyte micronucleus test. Test Guideline 474. Available at: https://www.oecd.org/en/publications/2016/07/test-no-474-mammalian-erythrocyte-micronucleus-test_g1g6fb16.html.