Section 4
Health effects

Test Guideline No. 488

Transgenic Rodent Somatic and Germ Cell Gene Mutation Assays

30 June 2022

OECD Guidelines for the Testing of Chemicals

OECD
BETTER POLICIES FOR BETTER LIVES

OECD/OCDE 488
Adopted:
30 June 2022

OECD GUIDELINE 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 original Test Guideline 488 was adopted in 2011. In 2013, a revised guideline was adopted that updated: the age range of animals at the start of the treatment; the sections on reproductive tracts to be harvested for sperm collection; and, the correct time for rodent spermatogonial stem cells to become mature sperm and reach the cauda epididymis. In 2020, a revised guideline was adopted that updated the recommended regimen for the analysis of mutations in male germ cells. This present revision of the Test Guideline (TG) focuses on the integration of analysis of mutations in somatic tissues and germ cells, and harmonization with recently revised OECD Test Guidelines (TGs) for genotoxicity testing.

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). They use transgenic rats and mice that contain multiple copies of chromosomally integrated plasmid or 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); *lacZ* plasmid mouse; *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. 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 or plasmid shuttle 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, or ligation and electroporation of plasmids to recover the shuttle vector), 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 or *lacZ* plasmid mouse (16) (17) (18) (19)]; 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 (20)] and deletion mutations in the lambda *red* and *gam* bacteriophage genes [Spi$^-$ selection in *gpt* delta mouse and rat (14) (15) (21)]. Mutant frequency is calculated by dividing the number of plaques/plasmids containing mutations in the transgene by the total number of plaques/plasmids 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. 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 and with the *lacZ* plasmid assays (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/plasmid 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. Rodents should be housed in small groups (usually no more than five for mice and two for rats) of the same sex if no aggressive behaviour is expected, preferably in solid floor cages with appropriate environmental enrichment. Cages should conform with animal welfare standards (e.g., Directive 2010/63/EU) and be arranged to minimize possible effects due to cage placement. Animals may be housed individually only if scientifically justified.

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 (22). 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, treated with solvent or vehicle alone, and otherwise 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) (23) 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) (24). 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 or untreated control 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., 24; 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 (25) (26) (27) (28)], 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 [(24); 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 expert judgement determines that it differs from the previous distribution (see paragraph 29) (25) (26) (27) (28). 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 (29).

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 8 weeks 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 should be the dose that will be tolerated without evidence of study-limiting toxicity, relative to the duration of the study period, i.e., inducing toxic effects but not death or evidence of pain, suffering or distress necessitating humane killing (30). With most test chemicals, the dose levels used should cover a range from the maximum to little or no toxicity.

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).

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 (33), 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) (34). 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) (35) (36) (37) (38), in which the timing and kinetics of spermatogenesis have been well-defined (39) (40) (41). 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 (29). The available germ cell mutagenicity data obtained with TGR assays have been recently reviewed (5) together with modelling of mouse and rat spermatogenesis (41) 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 (42), 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 (43). 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 (40) 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 (38) (40) and ~70 days for the rat (40) (41). 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 is 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 (44). Sampling of caudal sperm should be conducted only 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) (41).

46. Germ cells extruded from seminiferous tubules comprise a mixed population of spermatogonia, spermatocytes and spermatids (35) (36) (41). 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 taking into account the known kinetics of spermatogenesis (41). 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 have received continuous treatment for the full 28 day administration period during the proliferative phase of spermatogenesis (5) (41).

47. Based primarily on extensive modelling of spermatogenesis (41) 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. According to the modelling, 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 (41). The spermatogenesis model is based on the assumption 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 this 28+28d regimen are considered conclusive.

48. Based on extensive modelling of spermatogenesis (41) 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 (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 versus only 21.6% with the 28+3d regimen (41). While theoretically not optimal, the 28+28d design is considered adequate for the evaluation of germ cell mutagenesis; it permits the assessment of mutations in 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 the 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 (41), 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. Overall, 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 euthanatised prior to completion of the test period (30).

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 (45).

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 and plasmid (19); *lacI* mouse (46) (47); *gpt* delta mouse (14); *gpt* delta rat (15) (48); *cII* (20). Modifications should be justified and properly documented. Data from multiple packagings 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 is 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* (49) and *lacZ* (50) 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) (51) (52) (53) (54), and in Annex 2. When evaluating the responses, all data should be taken into consideration and, in all cases, expert 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. Providing that all acceptability criteria above are fulfilled, a test chemical is considered clearly positive if all of the following criteria are met in a given tissue:

a. At least one of the treatment groups exhibits a statistically significant increase in the mutant frequency compared with the concurrent negative control;
b. When evaluated for trend (see paragraph 61), the results are dose-related (not applicable to the limit test);

c. The increase in the mutant frequency of any treatment group is outside the upper bound limit of the appropriate historical negative control distribution (see paragraphs 28-32, and Annex 2).

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

63. Providing that all acceptability criteria are fulfilled, a test chemical is considered clearly negative if all of the following criteria are met in all experimental conditions examined, in a given tissue:

a. None of the treatment groups exhibits a statistically significant increase in the mutant frequency compared with the concurrent negative control;

b. When evaluated for trend (see paragraph 61), the results are not dose-related ;

c. All results are inside the lower and upper bound limits of the appropriate historical negative control data (see paragraphs 28-32 and Annex 2);

d. Tissue exposure to the test chemical(s) or its metabolites occurred.

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. There is no requirement for further verification of a clear positive or clear negative response (see paragraphs 62 - 63).

66. In cases where the response is not clearly negative or positive, or in the case of a positive result at the only dose used in a limit test, and in order to assist in establishing the biological relevance of a result (e.g. a weak or borderline increase), further investigations of the existing experiments may be necessary. These include analyzing more plaques or mutant colonies, and analyzing more animals. If the application of expert judgement and the analysis of additional data are unable to resolve a response as either positive of negative a repeat experiment using modified experimental conditions might be needed.

67. Sequencing of mutant plaques to determine whether there is a shift in the mutation spectrum induced by the test chemical may 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 (55) (56) (57) (58).

68. In rare cases, even after further investigation, the data will preclude making a conclusion that the test chemical produces either positive or negative results, and the study will therefore be concluded as equivocal.

7.3. Test report

69. 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:

- 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 and historical negative control data with ranges, means, standard deviations, range and control limits;
- concurrent and historical positive control data;
- analytical determinations, if available (e.g. DNA concentrations used in packaging, DNA sequencing data);
- statistical analyses and methods applied;

Discussion of the results;

Results should be discussed and conclusions justified especially in case the results are not clearly positive or clearly negative

Conclusion.

8. LITERATURE

(1) OECD (2017), Overview on genetic toxicology TGs, OECD Series on Testing and Assessment, No. 238, OECD, Paris, https://doi.org/10.1787/9789264274761-en.

(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) Boerrigter, M.E., M.E. Dollé, H.-J. Martus, J.A. Gossen and J. Vijg (1995), "Plasmid-based Transgenic Mouse Model for Studying *in vivo* Mutations" *Nature*, 377(6550): 657–659.

(17) 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.

(18) 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.

(19) 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.

(20) 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.

(21) 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.

(22) 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.

(23) 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.

(24) 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.

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

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

(27) 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.

(28) 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.

(29) 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.

(30) 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.

(31) 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.

(32) 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.

(33) 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.

(34) 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

(35) 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.

(36) 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.

(37) 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.

(38) 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

(39) 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.

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

(41) 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.

(42) 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.

(43) 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

(44) 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.

(45) 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.

(46) 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.

(47) 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.

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

(49) 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.

(50) 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.

(51) 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.

(52) 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.

(53) 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.

(54) 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.

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

(56) 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.

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

(58) 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.

(59) Kluxen FM, Weber K, Strupp C, Jensen SM, Lothorn LA, Garcin J-C, Hofmann T (2021) Using historical control data in bioassays for regulatory toxicology. Regulatory Toxicology and Pharmacology 125:105024.

(60) InfinityQS International (2014) A practical guide to selecting the right control chart. Available. at:

http://www.infinityqs.com/sites/infinityqs.com/files/files/PDFs/InfinityQS_Practical_Guide_to_Selecting_the_Right_Control_Chart_Oct2013.pdf. Successfully accessed October 7, 2021.

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

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 and the lacZ plasmid assays).

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 concatamer 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, Data interpretation and Historical negative control data distribution

1. Statistical analysis and data interpretation

1.1 Recommendations for evaluating TGR assay data

There is no single correct method of conducting the three types of statistical analysis of TGR data described in paragraph 61-63. In cases, where alternative methods are to be used, investigators should specify their methods before a study is initiated (*i.e.*, written study or validation plan), and should be prepared to justify their approach using sound statistical arguments.

Note that there is an important fundamental principle that all the statistical methods described below should take into consideration. Specifically, these and other statistical tests assume that the experimental design is based on random sampling, or at least a blocking approach. Randomization and/or blocking represent important methods that help mitigate the influence of factors that may have subtle, or not so subtle, effects on experimental results, but are not the main factors being studied.

1.2 Pairwise comparisons

One set of statistical tests is pairwise comparisons of mutant frequencies in the concurrent vehicle/solvent (negative) control group with those measured in 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. Generally, these types of parametric tests should be performed only when assumptions about normality and homogeneity of variance are valid. If normality and/or unequal variance is identified, an appropriate data transformation such as a logarithmic (log10) or rank transform can often be used to fulfil these requirements. If the transformation does not result in homoscedasticity, weighted (variance-corrected) ANOVA and/or t-tests can be applied, or the appropriateness of non-parametric methods can be considered. Finally, when more than one sex is used in a study, factorial design approaches that consider both treatment and sex are generally advantageous. This is described in greater detail below.

1.3 Factorial design

This design is equivalent to a two-way analysis of variance, for example, with sex and test chemical dose level as the main effects. The data can be analysed using many standard statistical software packages such as JMP, SPSS, SAS, STATA, Genstat as well as R and R Studio. Full details of the underlying methodology are available in many standard statistical textbooks and in the 'help' facilities provided with statistical packages.

1.4 Trend test

A second type of analysis described in this Test Guideline is a trend test to identify a dose-response relationship. When conducting trend testing, mutant frequency data should be available from the concurrent negative control group and each of the experimental dose groups. Note that this analysis is not normally applicable for Limit Tests (i.e., single test chemical dose level studies) as described in paragraph 51). In employing these analyses, care is needed in interpreting the results of some types of trend tests. For example, a simple linear trend test may fail to detect a trend when the dose-response is non-monotonic. Trend tests capable of detecting non-monotonicity such as the downturn protection test proposed by Bretz and Hothorn (2003) may be useful in such cases. A graph that shows the dose response can be helpful in determining what type of trend test is appropriate and/or whether a transformation is needed.

2. Historical negative control data distribution

Distributions of historical negative control data are important for assessing assay acceptability. A second important function is determining whether the mean mutant frequency of any test-chemical-exposed treatment group exceeds the upper bounds of the historical negative control data distribution. Given their importance, guidance for building historical negative control datasets is provided below.

2.1 Building historical negative control datasets

During the course of the proficiency investigations, each laboratory should establish historical negative control ranges and distributions of mutant frequencies. Such data can be derived from animals assayed as negative controls in conjunction with animals dosed with test chemicals (referred to here as concurrent negative controls). These animals typically will be dosed with the solvent/vehicle alone.

When first acquiring data for historical negative control mutant frequency distributions, negative controls should be consistent with published data, where they exist (3, 23). Experimental data should continue to be added to achieve statistically robust databases that facilitate assessments of later studies' validity, as well as comparisons of test chemical-exposed animals' mutant frequencies relative to historical negative control distributions.

Ideally, the laboratory should endeavour to acquire individual mutant frequency measurements from at least 30 negative concurrent (solvent/vehicle) control animals. To the extent possible, the data should be acquired from at least 3 independent experiments that each use animals generated from different breeding cycles, and the experiments should be conducted under conditions comparable to regulatory studies.

Laboratories should use quality control methods, such as control charts [e.g., I- and X-bar charts (26) (27)], to identify how variable their data are, and to show that the methodology is 'under control' in their laboratory. Further recommendations on how to build and use the historical data (i.e., criteria for inclusion and exclusion of data in historical databases and establishing acceptability criteria for a given experiment) can be found in the literature (e.g., 25) and are discussed below.

Any changes to the experimental protocol should be considered in terms of their impact on whether the resulting data remains consistent with the laboratory's existing historical negative control database. Only major inconsistencies should result in the establishment of a new historical negative control database. During the re-establishment, 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.

Negative control data from TGR assays should consist of the frequency of mutant frequency from each animal. Where data from individual negative control animals fall outside the historical negative control distribution, they may be acceptable for inclusion in the database provided (i)these data are not extreme (e.g., an occasional 'jackpot' mutation which has an unusually high mutant frequency is to be expected and can be excluded), (ii) there is evidence that the test system is 'under control', and (iii) there is no evidence of technical or human error.

2.2 Characterizing historical negative control distributions

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 will generally take the form of calculating an upper and/or lower bound limit that describes the boundary within which the majority of negative control animal values are expected to fall. Factors that should be considered 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. Additional information can be found in (59)(60)(61).

Inappropriate methods.

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

Confidence interval: A confidence interval is a range of estimates that is likely to include a population value (parameter) such as the mean with a defined (e.g., 95%) degree of confidence. Confidence intervals around a population mean are not useful for adequately describing the historical negative control distribution in the context of establishing useful upper and/or lower bound limits. Firstly, a confidence interval describes the current data set, not future observations; secondly, confidence intervals narrow as the sample size increases reflecting the increased precision of the estimate of the population value (parameter).

Appropriate methods.

Control limit: In the field of Quality Control, multiples of the standard deviation, usually the mean plus and minus 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 2x, 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'. Assuming a normal distribution and an 'under control' process, 99.73% of the data should fall within 3 standard deviations, and approximately 95% of the data should fall within two standard deviations. Control and/or warning limits therefore represent a useful resource for evaluating historical negative control data, and can provide useful upper and/or lower bound limits that aid in the interpretation of study data. That being said, limits that describe where approximately 95% of normally distributed data fall (± 2 standard deviations) are generally more appropriate than those based on 3 standard deviations. The former is consistent with other OECD Test Guidelines (e.g., OECD TG 474), and the latter characterizes exceptionally rare/unusual data points, thereby generating intervals that are too wide for the intended purposes. A rule

of thumb is that ≥ 25 individual data points are sufficient to derive useful control and warning limits when the process is 'under control'.

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 would be expected to fall in the range with 95% probability. (Note that, as with that of the confidence interval, this is a simplified definition, the exact definition is more complex.) Non-normal data should be transformed as necessary, with back transformation to original units for reporting and use. Alternately, some computer software programs enable a non-parametric calculation which does not assume a normal distribution. A rule of thumb is that ≥ 30 individual data points work better, and smaller sample sizes should not be used to calculate prediction intervals.

Tolerance intervals: Tolerance intervals are designed to predict numerous future observations, within user-defined 'coverage'. For example, 95% of future observations should fall within the 95% tolerance interval. (Again, this is a simplified definition, the exact definition is more complex.) Non-normal data should be transformed as necessary, with back transformation to original units for reporting and use. Alternately, some computer software programs enable a non-parametric calculation which does not assume a normal distribution. Tolerance intervals tend to be wider compared with analogous prediction intervals (since the former is designed to predict a high percentage of future values, while the latter is usually used to predict one or a few new observations). It is also important to recognize that tolerance intervals generally require much larger sample sizes compared with prediction intervals. A rule of thumb is that ≥ 100 data points work better, and smaller sample sizes should not be used to calculate tolerance intervals.

Quantiles: Quantiles are used for summarizing the rank of data points according to their 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. They establish intervals based, for instance, on percentiles, 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. A rule of thumb is that ≥ 100 individual data points work better, and smaller sample sizes should not be used to calculate quantiles.

Note that while the intervals described above will generally be calculated using individual animal mutant cell frequencies, the primary comparison described by criterion C (paragraphs 62-63) considers where treatment group mean values fall relative to the upper bound value. This is a practical recommendation, acknowledging that as an in vivo test system, sample sizes will be larger and historical control distributions more robust when based on individual animal data. Even so, this is not the only comparison that can be made. For instance, there may be times when it would be useful to consider the relationship of individual animal mutant cell frequencies to the historical negative control upper bound limit value.