EUROPEAN COMMISSION
JOINT RESEARCH CENTRE

Institute for Health and Consumer Protection
European Union Reference Laboratory for Alternatives to Animal Testing (EURL ECVAM)

ECVAM
SCIENTIFIC
ADVISORY
COMMITTEE
(ESAC)

ESAC Working Group Peer Review Consensus Report

on

a Hadano-coordinated validation study of a Cell Transformation Assay (CTA) for carcinogenicity testing based on Bhas 42 cell line

TABLE OF CONTENTS

ESAC Working Group

This report was prepared by the "ESAC Working Group for CTA" (ESAC WG), charged with conducting a detailed scientific peer review of the JaCVAM coordinated study concerning a cell transformation assay (CTA) for carcinogenicity testing based on the Bhas 42 cell line.

The ESAC WG had been set up by the ESAC during its meeting on March 21st, 2012. Basis for the scientific review was the ECVAM request to ESAC concerning the scientific review (ESAC request Nr. 2012-02).

The ESAC WG conducted the peer review from in October 2012. This report was endorsed by the ESAC WG on October 26th, 2012, and represents the consensus view of the ESAC WG.

This ESAC WG peer review consensus report was endorsed by the ESAC on 17.12.2012.

The ESAC WG had the following members:

- Rodger Curren, Chair
- Erwin L Roggen, Rapporteur
- David Lovell. External expert
- Edgar Rivedal. External expert
- Takeki Tsutsui. External expert

ESAC Secretariat:

- Raffaella Corvi
- Patric Amcoff

NOTE ON THIS REPORTING TEMPLATE

The template follows the ECVAM modular approach and allows at the same time for the description of the analysis and conclusions concerning more specific questions. The template was approved by the ESAC through written procedure on 29 October 2010.

The template can be used for various types of validation studies (*e.g.* prospective full studies, retrospective studies, performance-based studies and prevalidation studies).

Depending on the study type and the objective of the study, not all sections may be applicable. However, for reasons of consistency and to clearly identify which information requirements have not been sufficiently addressed by a specific study, this template is uniformly used for the evaluation of validation studies.

- **Explanatory notes to the paragraph titles (in green)** have been added on 17 November 2010. These notes provide guidance on the type of information / analysis expected under each section. Depending on the purpose and scope of the study to be reviewed, some of the aspects mentioned in the explanatory notes may not be applicable or only be applicable to some extent. Moreover, the explanatory notes are not intended to represent an exhaustive list of possible issues to be addressed under the respective heading, but are thought to provide some guidance with respect to the considerations typically expected.
- For all of the template's numbered sections **the summary view of ESAC WG is given in bold** followed by more detailed comments ("general observations" and "specific observations").

ABBREVIATIONS USED IN THE DOCUMENT

Formatting Examples below

- **BLR** Between-laboratory reproducibility
- **ECVAM** European Centre for the Validation of Alternative Methods
- **ESAC** ECVAM Scientific Advisory Committee
- **ESAC WG** ESAC Working Group
- **GCCP** Good Cell Culture Practice
- **GLP** Good Laboratory Practice
- **PC** Positive Control
- **SOP** Standard Operating Procedure (used here as equivalent to 'protocol')
- **VC** Vehicle Control
- **VMT** Validation Management Team
- **WLR** Within-laboratory reproducibility

Executive summary

Following a request from ECVAM to ESAC for peer review of and scientific advice on an Hadano-coordinated study concerning a cell transformation assay (CTA) for carcinogenicity testing based on the Bhas 42 cell line, an ESAC Working Group (ESAC WG) was set up by ESAC. The ESAC WG was charged with conducting a detailed scientific peer review of this study which had assessed the reliability and transferability of the Bhas 42 CTA.

The ESAC WG met in person at ECVAM on October 1-2, and communicated further by email and teleconferences on October 10 and 24.

The ESAC WG reviewed the Bhas 42 CTA study reports and supplementary information requested by the WG (Annex 1). The ESAC WG considered that the scientific work presented was of good quality. The WG identified issues that need to be addressed, but these issues were not expected to change the overall picture of the tests with respect to WLR, transferability, BLR and predictive capacity (as defined by the number of foci as the read-out).

The objective was clearly defined as an attempt to demonstrate the transferability, reliability (reproducibility within and between laboratories) and relevance (predictive capacity) of this test system using a set of coded chemicals whose carcinogenicities are known *in vivo*.

The number of chemicals (6-well assay: 8 carcinogens (including tumor promoters (TP)), 4 non-carcinogens; 96-well assay: 12 carcinogens (including TP), 9 non-carcinogens) was acceptable for assessing the transferability and reliability of the test.

For the 6-well assay acceptance criteria were not described in the protocol. Instead, ''low transformation in the positive control and high transformation in the negative control'' were set as criteria for repetition of the test. The acceptance criteria for the 96-well assay (including the cell growth assay) were clear and detailed.

In general, the acceptance criteria were followed. However, two incomplete results were produced by the same laboratory for the 6-well assay. With respect to the 96-well assay, two incomplete results were observed. These were due to circumstances affecting compound dosing but beyond the control of the laboratories.

The statistical analysis is based upon a conservative approach that minimizes the number of false positive results. This approach is acceptable for identification of clear positive and negative compounds, but may be problematic for less unambiguous compounds.

Clarity and completeness of the protocol(s).
The protocol for the 6-well assay lacks sufficient technical detail, while the protocol for the 96-well assay was more detailed. It is the ESAC WGs opinion that a lack of sufficient detail provided in the reports may have contributed to the observed differences in the foci scored between experiments and laboratories.

Within laboratory reproducibility (WLR)
Overall the results were concordant. A different statistical method was used depending on the protocol (6-well or 96-well) which may affect the identification of the weakly positive chemicals.

In general the acceptance criteria were followed. However, incomplete results were produced by one laboratory for the 6-well assay. Two incomplete results observed for the 96-well assay were due to circumstances affecting compound dosing but beyond the control of the laboratories

Transferability

The report did not include a detailed training program, the data produced during training or the quality of these data. Post-training data were only presented for the 96-well assay, and these data were satisfactory.

The transferability of the test to a laboratory not previously experienced in CTA was not demonstrated. The WG considers the test to be transferable to laboratories with experience in cell culturing techniques, but it requires more than one day of training. Considerable training in scoring of foci (and the use of the catalogue) might be required, especially for the 6-well assay. The value of the catalogue could be enhanced with description phrases.

Between-laboratory reproducibility (BLR)

The BLR of the 6-well assay was assessed on 12 chemicals (coded) by 3 laboratories. In 9/12 (75%), concordant results (number of foci) were obtained. Caffeine gave incomplete data in one laboratory due to a lack of compliance with the protocol. The two remaining laboratories produced concordant results. It is unclear why results for anthracene and o-toluidine were not reproducible. The values for MCA and TPA range from 10 to 50 foci per well. It is not possible to determine from the information provided the reason(s) for this spread. The BLR of the 96-well assay was assessed by comparing laboratory results within the validation study. Results for both initiation and promotion were acceptable (>90% for both initiation and transformation). Twenty-two chemicals were tested by 4-6 labs.

A high level of between-laboratory reproducibility was demonstrated for both assays. However, significant quantitative differences (number of foci counted) between laboratories were observed for the positive control of the 6-well assay. It is unclear how these differences would affect the detection of less potent chemicals.

Predictive capacity

The predictive capacity of the assays was assessed by comparing the "transformation" results for each of the tested chemicals and comparing that to the reported carcinogenicity *in vivo*. In the vast majority of the cases, all the laboratories agreed on the "transformation" call, but in a few instances, noted below, the judgement of the majority of the laboratory calls was accepted for the comparison. For the 6-well assay (N=12 chemicals) this comparison resulted in 100% correlation. In ten of the twelve cases all three labs agreed on the call. However, for anthracene, only 2 of the 3 labs characterized it as negative (the "correct" call)., and for o-toluidine only 2 of the 3 labs characterized it as positive (the "correct" call). The 96-well assay (N=21 chemicals) revealed a 86% concordance, 83% sensitivity and 89% specificity. In twenty of the cases all labs (n = 2 – 8) agreed on the call. However, for phenanthrene, only 3 of the 4 labs characterized it as negative (the "correct" call).For common chemicals between the 96-well assay and the 6-well assay, sodium arsenite and o-toluidine were negative in the 96-well assay, with transformed foci on the side of the wells not scored as prescribed by the scoring criteria. Thus the 96-well method seems to need more investigation of scoring before the protocol is finalized.

While this study produced promising results supporting Sakai et al. (2010) (98 non-coded chemicals), the number of chemicals was considered too low to allow for strong statements with respect to its predictive capacity.

In spite of the differences between the applied scoring strategies, both CTA protocols were considered sufficiently standardized for assessing the cell transforming capacity of chemicals, and these Bhas 42 protocols are suggested to be developed as the basis for an OECD Test Guideline on in vitro carcinogenicity testing.

1. Study objective and design

1.1 Analysis of the clarity of the definition of the study objective

(a) ESAC WG summary of the study objective as outlined in the VSR

General observations:

This study is to demonstrate the transferability, reliability and relevance of the Bhas 42 CTA using a set of coded chemicals of known *in vivo* carcinogenicity.

The ultimate objective is to demonstrate the utility of the test for adoption as an OECD TG that

1. can detect genotoxic and non-genotoxic carcinogens;
2. is sensitive to tumor promoters (TP);
3. can discriminate between tumor-initiating and tumor-promoting activity of carcinogens.

(b) Appraisal of clarity of study objective as outlined in the VSR

The study objective is clearly formulated.

1.2 Quality of the background provided concerning the purpose of the test method

(a) Analysis of the scientific rationale provided in the VSR

General observations:

Genotoxic and non-genotoxic carcinogens have not been defined in the VSR. This definition is controversial, but the lack of it makes it also difficult in this context to distinguish between non-genotoxic carcinogens and tumor promoters. The lack of definitions interferes with the assessment of the results from the validation study, since the basis for the assignment of a test compound to a certain test group is unclear.

It is also disputable if "tumor promoter" is a sufficiently defined property to be used as a criterion outside of a test system where the end point is a tumor. This should have been discussed more in detail in the VSR. The organ specificity previously observed for tumor promoters should also have been discussed, and explanation given for how this is considered in relation to this assay, and in the validation study. The scientific rationale behind the use of a cell line considered as "initiated" in a test for "initiating" properties should also have been discussed in more detail.

Specific observations:

The Bhas 42 cells are transformed by known tumor-promoters, including TPA, okadaic acid and lithocholic acid, without initiating treatment with a known-initiator such as MCA [Omori *et al.*, 2004], and are presumed to be initiated toward transformation by the introduced *ras* sequence [Sasaki *et al.*, 1990]. The VMT needs to describe the mechanism of initiation induction by v-Ha-ras in Bhas 42 cells. There is no reference describing the mechanism of the increased sensitivity of Bhas 42 cells to

transformation by introduction of v-Ha-ras. However, in subsequent communications with the Test Submitters (see Section 19. Annexes) an hypothesis for the mechanism was presented.

(b) Analysis of the regulatory rationale provided in the VSR

General observations:

The ESAC WG does not consider if the discrimination between "initiating" and "promoting" activity is a sufficiently clear and important property to be suggested for inclusion in a Guideline for this assay. It should have been considered whether this assay instead should focus on transforming capability. Similarly its use in discriminating genotoxic from non-genotoxic carcinogens is questionable. Assays specifically for genotoxicity should be used to discriminate between genotoxic and non-genotoxic compound.

The VSR should have discussed the advantage of the promotion part of the assay in a regulatory context.

1.3 Appraisal of the appropriateness of the study design

General observations:

Overall, the study design was appropriate in terms of number of laboratories involved, management team set-up, statistical analysis and experimental set-up.

Specific observations:

The number of laboratories (N=6) involved in the 6-well validation project exceeded the number of laboratories (N=3) involved in prevalidation studies. The validation was designed so that not all chemicals were tested in each laboratory. This allowed a good estimate of reproducibility while at the same time allowing an increased number of chemicals to be tested. Transferability, reliability and relevance of the tests were at all times assessed by data from at least 3 laboratories.

The number of laboratories (N=4, Phase I; N=3, Phase II) involved in the 96-well validation project was adequate. In Phase I all the chemicals (7) were tested in each laboratory. Phase II was designed so that only one laboratory tested all chemicals, while the other laboratories tested a subset resulting in each chemical being tested in two laboratories. As in the 6-well validation, this allowed a good estimate of reproducibility while at the same time allowing an increased number of chemicals to be tested. Transferability, reliability and relevance of the tests were at all times assessed by data from at least 2 laboratories. In the beginning of the validation study on 6-well Bhas 42 CTA an advisory committee was active. From 29.10.2008, the study was supervised by a Validation Management Team (VMT) established by the Japanese Centre for the Validation of Alternative Methods (JaCVAM). The VMT comprised representatives from JaCVAM, ECVAM and ICCVAM, as well as experts including an external statistician.

The statistical analysis was deemed appropriate for the identification of clear positive and negative responses.

The study design for assessing transferability, reliability and relevance was deemed appropriate, but certain shortcomings where identified, specifically the fact that transferability was difficult to assess

since the transfer was conducted to laboratories which were experienced in conducting the standard BALB/c 3T3 cell transformation assay.

1.4 Appropriateness of the statistical evaluation

General observations:

The criteria for declaring a compound positive or negative is described in the document as being based upon a specific set of comparisons being statistically significant. For the 6-well method two successive doses being significantly different when compared with the negative control using a one-sided Dunnett's test using a $P<0.05$ cut-off. For the 96 well plate method two successive doses being significantly different when compared with the negative control using a one-sided Chi-square test at $p<0.05$ and applying a Bonferroni correction.

The criteria seem to provide appropriate +/- decisions in the case of the test chemicals which seem to have been chosen because they are clear positive or negative compounds with good background data. It is unclear how the method would perform with weaker unknown compounds.

Specific observations:

Both Dunnett's test and the Bonferroni tests are conservative in that they reduce the number of comparisons declared statistically significant and so lowering the power of the study which may reduce the potential to detect weaker positives.

Both statistical methods depend upon the number of comparisons (or dose levels) included in the design. In this study, the +/- criteria will thus vary from chemical to chemical and from laboratory to laboratory because of the different numbers of doses in the experiments (5 – 10). In this context it is important to have it clarified as to whether or not dose levels that are toxic are included in the choice of test statistic.

The fact that that doses are assigned on a plate basis, rather than a well basis, is an important point because the statistical analysis should be applied to the experimental unit (in this case the plate). Plate to plate variability (due to position in the incubator, top or bottom of a stack, etc.) can result in apparent dose effects. The more wells the more likely it becomes to detect such artefactual variation because of pseudo replication. The chi-square test in the presence of pseudo replication is very susceptible to raised Type 1 errors. Any differences between plates will be magnified.

Based upon looking at the results tables the size of effect detected as significant using the 96-well plate is about 2.5-4.0 fold increase depending upon the negative control level.

The analysis will also be susceptible to variability between the negative control plates. It is often recommended to increase the sample size of the negative control group to get a more accurate estimate of the negative control incidence for use in making comparisons with a range of treatments.

No attempt is made to use the most powerful statistical test relevant to the design of a test for dose-related trends in the data.

While within and between laboratory reproducibility as measured in concordance is relatively high for between and within laboratory comparison, quantitative variability seems to be appreciably greater when the figures are looked at (for example Fig. 6, cadmium chloride promotion and Fig. 12

o-toluidine hydrochloride promotion). This would have implications for the assessment of chemicals with less clear-cut results.

A clearer explanation of the practical precautions that were taken to ensure that biases are not introduced in to the experiment is needed. Was there any randomization in the design, what precautions were taken to avoid biases?

The variability in counts seems appreciably less that would be expected, Much count data follow a Poisson distribution (where the mean = the variance). Although the variability associated with dose – related effects is often greater (and may lead to a loss of power and more Type II errors) the variability is less than would be expected which implies a lack of independence in the data (e.g.. Table 9b - Laboratories 1 and 4: SD^2 is smaller than the mean for most doses.).

On first review, there appeared to be several anomalies in the results reported, e.g. smaller difference in an experiment declared more significant than some larger differences (p33, table 4b). However, later communications with the Test Submitters (see Section 19. Annexes) revealed that this was a typographical error. The rejection of a clear positive result based upon the criteria stated in the protocol by using compound specific extra information (MW) (p125, table 43), was a post hoc determination by the VMT. None the less, this decision (not testing above 10 mM) followed current genetic toxicology testing standards, and this criterion has now been included in the latest protocol.

Diagnostic test (Cooper test) statistics should not be used here to show that one method is better than another. This is because the Confidence interval (CI) on sensitivity and specificity are very wide when based upon smallish sample sizes (e.g. <100), also set of samples may differ appreciably between some studies with some comparisons based upon very well-characterised clear positives or negatives and others on a more-representative general sample.

2. Collection of existing data

2.1 Existing data used as reference data

General observations:

The chemicals were selected based upon information provided by a selected number of peer-reviewed publications. The criteria applied by the submitters to define tumor promoting, carcinogenic and non-carcinogenic compounds were not provided.

Specific observations:

The chemicals list for the 6-well assay comprised compounds that were classified as tumor-promoting compounds (N=3), carcinogens (N=7) and non-carcinogens (N=4) by the OECD DRP31 (2007), Kirkland et al (2005), Lijinski et al (1992) and Reddy et al (1975). The chemicals list for 96-well assay comprised compounds that were classified as tumor-promoting compounds (N=3), carcinogens (N=7), non-carcinogens (N=7) and one non-classified compound by OECD DRP31 (2007), Kirkland et al (2005), Lijinski et al (1992), Slaga et al (1976, 1980), Reddy et al (1975) and Baird and Boutwell (1971).

The purpose of including a non-classified compound was not explained, although it is useful for reproducibility determinations.

For both tests, two of the carcinogens were selected as positive controls for the initiation assay (MCA) and the promotion assay (TPA), respectively. TPA was classified by CCRIS as a carcinogen and tumor promoter.

For both tests the validity of the TPs remains to be assessed. Tumor promotion is an operational description from animal experiments, which is challenging to transfer to *in vitro* conditions. There is a need for better understanding of the mechanisms involved, or at least more test compounds.
A number of important issues were not discussed:

- Organ specificity of tumor promoters.
- Complete carcinogens may be considered as both initiators and promoters, dependent on dose.
- The existence of pure initiators and promoters to be used in a validation situation.

2.2 Existing data used as testing data

General observations:

Source for both tests is Sakai et al (2010) describing the results for 98 (non-coded) chemicals using the 6-well method. This peer-reviewed publication was referred to in the discussion of the predictive capacity of the assays.

2.3 Search strategy for retrieving existing data

General observations:

The test submitters relied entirely on the judgement of the sources of the reference data. No strategy for retrieving the data was described. Existing in vitro data for this assay were not used in an overall analysis, so a search strategy is not applicable.

2.4 Selection criteria applied to existing data

General observations:

The test submitters did not describe nor apply any criteria for the selection of the test chemicals. The available *in vitro* and *in vivo* information was provided in the report (e.g. table 3, p24).

The available data for each of these compounds was provided in the report, but the criteria used to select or reject chemicals for this study was not described.

While the immediate objective of the study is to assess the transferability, reliability and relevance of the tests using carcinogens, the test submitters also want to demonstrate that the tests can identify non-genotoxic compounds, including tumor promoters, and that the test can discriminate tumor initiating from tumor promoting chemicals. However, the test submitters did not provide the criteria pre-defining these subsets of chemicals making the evaluation of the report in this respect problematic.

3. Quality aspects relating to data generated during the study

3.1 Quality assurance systems used when generating the data

General observations:

The present study was conducted under conditions that closely approximated Good Laboratory Practice (GLP) by all laboratories.

In general the experiments appeared to have been performed in compliance with the set criteria. In most cases (5 out of 7) tests were repeated when they did not met with the criteria. However, a few deviations were observed.

Specific observations:

1. Two incomplete results were produced by the same laboratory for the 6-well assay, due to wrongly dosing of the test chemical.
2. With respect to the 96-well assay, incomplete results were observed for each laboratory testing caprolactam and phorbol. However, these were due to circumstances affecting compound dosing but beyond the control of the laboratories.

3. Two experiments (3-methyl-cholanthrene (p90) and o-toluidine (p93)) were not repeated in spite of not complying with the test acceptance criteria.

3.2 Quality check of the generated data prior to analysis

General observations:

No detailed provisions to check the quality of the results prior to reporting were specified.

4. Quality of data used for the purpose of the study (existing and newly generated)

4.1 Overall quality of the evaluated testing data (newly generated or existing)

General observations:

Overall the quality of the testing data was good. The laboratories complied well with the acceptance criteria. However, a number of anomalies were identified.

Specific observations:

6-well assay: It was noted that the results submitted for two out of twelve test chemicals were not concordant between the naive and experienced laboratories. It was suggested that this discrepancy was due to misclassification of transformed foci by the "naive" laboratories, but no evidence was presented to support this explanation.

Seven of 12 chemicals tested were dissolved in distilled water. However, there is no indication whether the number of transformed foci shown in the figure describing controls (p.30: 3.3.2 Negative and positive controls) were produced in the negative control (water) or vehicle control (DMSO).

96 well assay: The VMT concluded that under the test conditions employed, caprolactam was to be considered negative in the initiation assay. However, all test concentrations were in excess of the 10 mM (maximum allowable concentration according to the acceptance criteria). Therefore the results of the caprolactam experiments should be considered incomplete.

One experiment (pyrene) was judged not to fulfil the assay acceptance criteria because the concentrations used by this specific laboratory (2) did not induce cytotoxicity. The data from this particular repeated experiment were however considered acceptable because three sequential concentrations induced statistically significant increases in the transformation frequency.

4.2 Quality of the reference data for evaluating reliability and relevance[1]

[1] OECD guidance document Nr. 34 on validation defines relevance as follows: "Description of relationship of the test to the effect of interest and whether it is meaningful and useful for a particular purpose. It is the extent to

General observations:

Being extracted from the literature, the quality of the reference data was assessed by several peer-reviews. On that basis, the quality of the data was deemed sufficient for evaluating the reliability and relevance of both tests as far as cell transformation is concerned.

Specific observations:

The retrospective data related to the Bhas 42 CTA were produced by researchers belonging to the same group in Japan. The 98 non-coded chemicals that were tested (89 well characterized and 9 undefined test chemicals) (Annex 16) included all the chemicals used in this study excluding pyrene and phorbol. It would have added value to this study if more chemicals had been included that had not be tested in the Bhas 42 CTA before, and that had produced clear calls in both the BALB/c 3T3 cell CTA (N=180 chemicals) and SHE cell CTA (N=500 chemicals) which both have been used by many laboratories.

4.3 Sufficiency of the evaluated data in view of the study objective

General observations:

The data and their quality were considered sufficient to assess the transferability and reliability of the tests.

5. Test definition (Module 1)

5.1 Quality and completeness of the overall test definition

General observations:

The test definition is based upon the hypothesis that some chemicals (full carcinogens) induce tumors *in vivo*, while others (tumor-promotors) do so only if a tumor-initiating event had occurred before exposure to this chemical.

Reliance on this hypothesis introduces a number of unknowns that undermine the test definition as proposed by the test submitters. For example, the mechanisms behind tumor initiation and promotion are not fully understood, but seem in some cases to be organ-specific. This makes it difficult to unambiguously define chemicals that are only tumor promoters. Consequently, the compound may not appear as a tumor promoter in a cell (e.g. Bhas 42) derived from an irrelevant tissue. This is further confounded by the fact that tumor promotion is an in vivo event, and it is not at all clear that the enhancement of transformed foci is mechanistically related.

Specific observations:

which the test correctly measures or predicts the biological effect of interest. Relevance incorporates consideration of accuracy (concordance) of a test method."

Test system description is adequate, but could be improved by a discussion of why the Ha-ras transfected cells now have a higher transformation rate with initiators (increased sensitivity as compared to other CTAs) and why they are more sensitive to tumor promoters.

The protocol for the 6-well assay is basic and could be considerably improved with more detail, especially on the criteria for the valid test procedures. Foci scoring parameters could be further aided by incorporating scoring atlas into the protocol. The atlas could be improved by describing how the individual foci of non-transformed fail to meet the definition of transformed.

The prediction model is confusing since positive results in either initiation or promotion leads to a "positive" call. Does this fit into regulatory schemes? Should laboratories stop testing when one is positive? Should this single value be used to determine predictive capacity?

5.2 Quality and completeness of the documentation concerning SOPs and prediction models

General observations:

A protocol emerged from this prevalidation study that was proposed by the test submitters to be the protocol of choice for coming studies (Annex 2). While this proposed protocol was more detailed than the protocols described in the report, there are still gaps that would make it difficult for a laboratory without experience in CTA to perform the testing.

The acceptance criteria and strategy how to deal with unqualified tests was clearly described. However, the prediction model should specify better how the results from the initiation and promotion tests are to be used.

The future usability of cells from increasing passages was not described.

Specific observations (a selection):

Preparation of cell stock (Annexes 6-8): It should be specified how long the cells can be stored at -80°C.

Fixatives and staining solutions (Annex 16): Discarding formalin, methanol, and the Giemsa solution is subjected to restrictions (see Material Safety Data Sheets). Provisions should be made to assure awareness of rules and guidelines concerning handling and management of these chemicals, or to provide the substitutes for them.

Annex 12: It should be clarified what is meant by 'the first culture' and 'the second culture'. Treatment time and temperature should be mentioned for the EDTA-PBS and trypsin incubation. It is not clear why cells were not washed with fresh medium after chemical treatment, while it was noted (p. 151, Bhas CTA validation study report) that residual chemicals may affect cells when they were not washed.
It should be specified how many "undamaged wells" are required for proper statistical analysis.

6. Test materials

6.1 Sufficiency of the number of evaluated test items in view of the study objective

General observations:

The number of chemicals (6-well assay: N=12; 96-well assay: N=21) was sufficient for assessing the transferability and reliability of the tests. While this study produced promising results supporting Sakai et al. (2010) (98 non-coded chemicals), the number of chemicals was considered too low to allow for definitive statements with respect to its predictive capacity.

Even if there had been criteria set out defining genotoxic, non-genotoxic and tumor-promoting chemicals, the number of chemicals would still have been too low for making definitive statements about the tests to discriminate between these classes of compounds.

6.2 Representativeness of the test items with respect to applicability

General observations:

The representativeness of the chemicals (6-well assay: 8 carcinogens (including TP), 4 non-carcinogens; 96-well assay: 12 carcinogens (including TP), 9 non-carcinogens) was acceptable for assessing the transferability and reliability of the test.

The list of chemicals contained a number of compounds that were Ames negative, but positive in other *in vitro* and *in vivo* test models. It seemed to be assumed that these Ames negative compounds are non-genotoxic carcinogens, but this assumption is questionable since all (except mezerein) were positive in one or more additional genotoxicity assays. Therefore they cannot clearly be defined as non-geneotoxic, and no conclusions about the detection of non-genotoxic carcinogens can be made. More chemicals clearly defined as non-genotoxins need to be tested to confirm this hypothesis. Perhaps a more useful analysis would be to asses “misleading positives” from genetox assays which are not tumorigenic *in vivo*.

It cannot be concluded from the presented data whether or not tumor-promoting chemicals can be identified because of the small number of such compounds.

7. Within-laboratory reproducibility (Module 2)

7.1 Assessment of repeatability and reproducibility in the same laboratory

General observations:

6-well assay:

Within-laboratory reproducibility (WLR) was assessed by laboratory 1 (lead laboratory). One experiment (N=12 chemicals, all coded) (experiment A) was compared with the historical data obtained by the lead laboratory (N=12, non-coded) (experiment B). Furthermore, a selection of the chemicals was tested coded (3 compounds by 2 labs, 1 compound by 1 lab) (experiment C) and compared with the same compounds tested by the same labs in a prevalidation study in which the chemicals were not coded. In addition WLR was assessed by comparing the + and – controls in each laboratory throughout the whole study.

96 well assay:

WLR was assessed during the prevalidation of the test (N=2, non-coded) and a repeat-experiment in the validation phase I using both chemicals and 1 additional compound (N=3, coded). WLR was assessed in 4 laboratories.

7.2 Conclusion on within-laboratory reproducibility as assessed by the study

General observations:

The VMT concluded that the WLR of the Bhas 43 CTA was demonstrated. This WG endorses this conclusion with respect to the 'transformation' score.

Specific observations:

6-well assay: WLR was considered acceptable when 'transformation' was used as the outcome. It was noticed by the WG that discrepancy occurred primarily in the initiation test.

96-well assay: The available data on the test chemicals were limited, but sufficient to indicate WLR. In all cases positive controls were significantly different from the negative controls.

For both assays, positive controls were significantly different from the negative controls.

8. Transferability (Module 3)

8.1 Quality of design and analysis of the transfer phase

General observations:

The transferability of the test to a laboratory not previously experienced in CTA was not demonstrated because:

1. The laboratories involved in transferability assessment had extensive experience in Balbc 3T3 CTA.
2. The transfer phase did not have predefined evaluation/decision criteria.
3. There is no indication of either data generated during training or the quality of these data. Post-training data were provided for the 96-well assay.

The WG considers the tests to be transferable to laboratories with experience in cell culturing techniques provided that extensive training is offered focussing especially on scoring of foci (and use of the catalogue to do this). This is probably less important for the 96-well assay than the 6-well assay, but it should clearly be a main training point for both assays.

The value of the transformed and non-transformed foci catalogue could be improved by the use of written phrases giving specifics as to why an individual focus is considered transformed or non-transformed This would be especially helpful for the categories “negative foci” and “positive foci (sparse)”.

Specific observations:

The WG had a number of concerns related to foci assessment.

1. Were coded plates shown to members of each laboratory?
2. How was scoring expertise transferred within each laboratory? Did the Study Directors (the only ones present at the original training) do the scoring at the home laboratory or did the technicians?
3. How was proper colony assessment determined during the training?

More description on whether the difference in quantitative data between laboratories results from different scoring criteria or from experimental procedures. This could be of importance for a better understanding of the lack of concordance and reproducibility for some compounds.

6-well assay: N=7 chemicals were transferred to laboratories V and VI. In spite of the expertise of these laboratories, three discordant results (transformation) were produced, one of which (caffeine) should have been repeated.

96-well: After training, the labs checked the quality of the training by testing in house 2 compounds (non-coded). In contrast to the 6-well assay, post-training) data were provided for 2 test chemicals (Tables 23 & 24). These data seem adequate.

8.2 Conclusion on transferability to a naïve laboratory / naïve laboratories as assessed by the study

General observations:

The WG disagrees with the VMT with respect to the transferability of the Bhas 42 CTA to naïve laboratories. The data provided by the study suggest that the tests are transferable to laboratories with expertise in CTA performance, but not to what the WG considers truly "naïve" laboratories.. Indeed, none of the involved laboratories was inexperienced in this respect minimizing the challenges e.g. related to proper foci scoring.

9. Between-laboratory reproducibility (Module 4)

9.1 Assessment of reproducibility in different laboratories

General observations:

6-well assay:
Between-laboratory reproducibility (BLR) was assessed on 12 chemicals (coded) by 3 laboratories. In 9/12 (75%), concordant results (transformation) were obtained.
Caffeine gave incomplete data in laboratory VI, due to too low concentrations tested in the test not complying with the protocol. The two remaining laboratories produced concordant results. It is unclear why results for anthracene and o-toluidine were not reproducible. A comparison of plate scoring should have been conducted. It is very important to know what parameters in any assay are the most important to control.
It does appear that there are differences between positive 'initiation' control values in certain labs, and between positive 'promotion' values in other labs. The values for MCA and TPA range between 10 and 50 foci per well. It is not possible to determine from the information given the reason(s) for this variability.

96-well assay:
BLR was assessed by comparing laboratory results within the validation study. Results for both initiation and promotion were acceptable (>90% for initiation and transformation).Twenty-two chemicals were tested by 4-6 labs.

Caprolactam was rejected on the basis of lack of sufficient information in the concentration range 0-10 mM. Phorbol produced consistent but incomplete results (due to VTM decisions) and was excluded from this list. The remaining 20 chemicals performed as shown in table 1.
Table 1:

Assay	%
Initiation assay	95 (19/20)
Promotion assay	85 (17/20)
Transformation assay (Bhas 42 CTA)	95 (19/20)

9.2 Conclusion on reproducibility as assessed by the study

General observations:

With respect to the 6-well assay, the WG considered the conclusion put forward by the VMT justified for the vehicle control, but not for the positive control nor for the chemicals without further explanation for o-toluidine.

The 96-well assay performed well and the conclusion by the VMT was considered justified for vehicle control, positive control, and test chemicals.

10. Predictive capacity and overall relevance (Module 5)

10.1 Adequacy of the assessment of the predictive capacity in view of the purpose

General observations:

The predictive capacity was based upon an algorithm using the 'transformation' score.

6-well assay:

The predictive capacity was assessed by comparing transformation results (based on a majority call [see below]) from each of the 12 chemicals and comparing that to the reported carcinogenicity in vivo. This resulted in 100% correlation. These numbers are better than was found in the 89-chemical study reported by Sakai, et al. 2010.

However, there is one major inconsistency in the VM report of this study that should be corrected. o-Toluidine is reported as positive *in vitro* in two out of three labs and is therefore listed as concordant with the *in vivo* results. However, in the analysis of BLR the positive results of the two laboratories were considered by the VMT to be incorrect on the basis of 'incorrect scoring of the foci' (p.61). This result cannot be viewed both ways. Either the re-count of the foci shows that it was a negative in both laboratories which improves the reproducibility but reduces the predictive capacity, or the opposite occurs. Either way, this question must be resolved.

Lithocholic acid is classified as a non-carcinogenic compound in the OECD DRP31, but was identified as a tumor promoter by Ready et al (1975). The WG suggest ammending the 2 x 2 Contingency table for the results of the 6-well assay as follows (Table 2):

Table 2: 2 x 2 Contingency table for the results of the 6-well assay

		In vivo activity		
		Carcinogen/TP	Non-active	Total
Bhas 42 cell transformation assay	+	8	0	8
	-	0	4	4
Total		8	4	12

96-well assay:

The predictive capacity was assessed by applying the "majority rule" (based on a majority call) to the results from each of the 21 chemicals and comparing the ''call'' to the reported carcinogenicity *in vivo*. This resulted in 85% concordance, 83% sensitivity and 87% specificity. These numbers are somewhat better than was found in the 89-chemical study (6-well assay) reported by Sakai, et al. 2010, however they are not as good as the 6-well results of the current VMG report.

For common chemicals between the 96-well assay and the 6-well assay, sodium arsenite and o-toluidine were negative in the 96-well assay, with transformed colonies on the side of the wells for arsenite.

As shown in Table 49, the judgments by majority rule with respect to transformation were consistent with the reported *in vivo* carcinogenicity results except for lithocholic acid, sodium arsenite, *o*-toluidine and pyrene. The amended 2x2 contingency table for the 21 tested chemicals (Table 3) and the performance indices of the 96-well Bhas 42 CTA for the prediction of chemical carcinogenicity (Table 4) are presented here.

Table 3: 2x2 Contingency table of the results in the 96-well method Bhas 42 CTA

		In vivo activity		
		Carcinogen/TP	Non-active	Total
Bhas 42 cell transformation assay	+	10	1	11
	-	2	7	9
Total		12	8	20

Table 4: The performance of 96-well method Bhas 42 CTA for the prediction of chemical carcinogenicity

Performance index	%
Concordance	85 (17/20)
Sensitivity	83 (10/12)
Specificity	87,5 (7/8)
Positive predictive value	91 (10/11)
Negative predictive value	78 (7/9)
False negative rate	17 (2/12)
False positive rate	12,5 (1/8)

2. The potential of both tests to assess whether a chemical is likely to be a tumour initiator or tumour promoter.

Table 5 attempts to compare the results in all validation studies conducted on the 96-well Bhas 42 CTA, and is a modification Table 52 (VMT report, pp. 149-150). Of the 12 carcinogens, 6 (50%), 4 (33%), and 2 (17%) carcinogens were positive, negative, and equivocal, respectively, in the initiation assay. Although all 4 tumor promoters were positive in the promotion assay, 3 (33%), 5 (56%), and 1 (11%) of the remaining carcinogens were positive, negative, and equivocal, respectively, in the promotion assay.

The results show that it is not yet possible to judge whether or not the Bhas 42 CTA can discriminate between initiation and promotion activities of chemical carcinogens.

The set of chemicals selected does not include sufficiently described non-genotoxic carcinogens and tumor promoting chemicals for this study and the results provided in the report do therefore not support the statement that both tests can discriminate between chemicals likely to be tumour initiators or chemicals that are likely to be tumour promoters.

Table 5: Comparison of the results in all validation studies conducted on the 96-well Bhas 42 CTA

Compound	The present validation studies		Carcinogenicity *in vivo*
	Initiation assay	Promotion assay	
2-AAF	+	+	+
Benz[a]anthracene	+	+	+
B[α]P	+	–	+
Cadmium·Cl	–	+	+
Dibenz[a,h]anthracene	+	–	+
Lithocholic acid	–	+	–[a] / TP
Methapyrilene·HCl	–	+	+ / TP
3-MCA	+	–	+
Mezerein	+ / –	+	+[a] / TP
MNNG	+	–	+
$NaAsO_2$	–	–	+
o-Toluidine·HCl	+ / –	+ / –	+
TPA	–	+	+[a] / TP
Ampicillin	–	–	–
Anthracene	–	–	–
L-Ascorbic acid	–	–	–
Caffeine	–	–	–
Caprolactam		–	–
Eugenol	–	–	–
D-Mannitol	–	–	–
Phenanthrene	–	–	–
Phorbol			?
Pyrene	+	+	–

[a] *DRP 31 [Carcinogenic compounds: methapyrilene·HCl (Table 11-1, P101), mezerein (Table 11-1, P102), and TPA (Table 8, P91). Non-carcinogenic compounds: lithocholic acid, phenanthrene, and pyrene (Table 9, P94)].*

?, Positive or negative depending on mouse strains.

10.2 Overall relevance (biological relevance and accuracy) of the test method in view of the purpose

General observations:

While this study produced promising results supporting Sakai et al. (2010) (98 non-coded chemicals), the number of chemicals was considered too low to allow for strong statements with respect to its predictive capacity.

11. Applicability domain (Module 6)

Not applicable in the context of this study.

12. Performance standards (Module 7)

Not applicable in the context of this study.

13. Readiness for standardised use

13.1 Assessment of the readiness for regulatory purposes

General observations:

Both tests produced promising results and should be considered as part of a testing strategy to provide data for risk assessment of chemicals suspected to be carcinogenic.

Motivation for this conclusion:

1. The tests seem to provide acceptable within laboratory reproducibility, transferability and between-laboratory reproducibility.
2. Available data suggest that the tests show acceptable ability to discriminate between carcinogenic and non-carcinogenic substances, of different chemical classes.
3. The tests respond to genotoxic as well as non-genotoxic carcinogens, including tumor promoters.

The WG can however not say that the tests have been acceptably proven to discriminate between tumor initiating and promoting compounds. In order to do so the WG considers that either a mechanistic understanding should be presented, or a larger number of compounds with clearly defined relevant properties as initiators or promoters should be tested.

The WG considers however the suggested end points to be of significant interest, and suggests that more studies are performed to strengthen the data and mechanistic understanding. It is of particular interest to learn more about the mechanisms underlying the ability to induce transformed foci following exposure of sparse and dense cells, and explore the possible relationship of this to tumor initiating and promoting substances.

Meanwhile it is recommended to focus on the preparation of an OECD guideline based on general ability to induce Bhas 42 cell transformation, as part of a data base to be used for cancer risk assessment.

13.2. Assessment of the readiness for other uses

General observations:

The 96-well assay, and to some extent the 6-well assay, could be useful as a tool in a screening strategy for chemicals aiming at identification of carcinogens, if the integrated judgement of the initiation and promotion assay results were accepted or justified (Yes/No output).

There was limited evidence supporting the capacity of the tests to provide specific information allowing for discrimination between genotoxic and non-genotoxic carcinogens, or tumor-initiating and tumor promoting compounds.

13.3 Critical aspects impacting on standardised use

General observations:

1. There was no clear definition provided in the report for genotoxic and non-genotoxic carcinogens.
2. Presentation of the assays for identification of genotoxic and non-genotoxic carcinogens should be reconsidered until a more systematic review of the genotoxicity and non-genotoxicity of tested compounds is conducted.
3. Compounds with TP activity were few and not well defined.
4. Scoring of foci is critical.
5. The transferability of the test to a laboratory not previously experienced in CTA is not known.
6. There was no clear rationale provided in the report for the integrated judgement of the initiation and promotion assay results.

13.4 Gap analysis

General observations:

The extent to which discordant results in both assays are due to scoring issues in the study is still not known. This should be clarified by a systematic re-evaluation of the existing plates, e.g. in a separate

scoring study on the plates. The outcome of such study will shed light on the true critical steps of the assay.

The use Bhas 42 CTA should be restricted to the assessment of the transforming capacity of chemicals, because:

1. Lack of compounds with clear 'initiator' and 'promoter' activity makes it difficult to evaluate the capacity of the tests to discriminate between both sets of chemicals.
2. Initiation versus promotion: Are the effects observed in Bhas related to different effects of compounds on sub-confluent versus confluent cells? This is interesting, and should be explored further. But is this sufficiently understood to warrant inclusion in Guideline? Or should this await better understanding of mechanisms and/or more compounds tested?
3. Because it is difficult still to unambiguously identify non-genotoxic carcinogens, a test discriminating between genotoxic and non-genotoxic carcinogens cannot yet be developed.

<u>6-well method</u>

It is still not clear that scoring has been standardized between labs – results with o-toluidine are a good example. A separate scoring study on the plates is possible and should be done so that reviewers can understand what are the true critical steps of the assay.

<u>96-well method</u>

A better definition of which foci to count needs to be created. It seems clear that sodium arsenate might be called a carcinogen if foci on the walls were counted.

14. Other considerations

None

15. Conclusions on the study

15.1 ESAC WG summary of the results and conclusions of the study

1. Study objectives

The definition of the *study objectives* were formulated clearly:

1. to demonstrate the transferability, reliability (reproducibility within and between laboratories) and relevance (predictive capacity) of this test system using a set of coded chemicals whose carcinogenicities are known *in vivo*;
2. to demonstrate the utility of the test for adoption as an OECD TG that
   - can detect genotoxic and non-genotoxic carcinogens;
   - is sensitive to TP;
   - can discriminate tumor-initiating and tumor-promoting activity of carcinogens.

2. Within-laboratory reproducibility (WLR)

WLR of the 6-well assay was assessed by laboratory 1 (lead laboratory). One experiment (N=12 chemicals, of which 8 coded) (experiment A) was compared with the historical data obtained by the lead laboratory (N=12, non-coded) (experiment B), and an experiment performed during the validation phase (N=4, selected from the 8 coded compounds in experiment A) (experiment C). Furthermore, a selection of the chemicals was tested coded (3 compounds by 2 labs, 1 compound by 1 lab) (experiment D). In addition WLR was assessed by comparing the + and – controls in each laboratory throughout the whole study.
The WLR of the 96-well assay was assessed during the prevalidation of the test (N=2, non-coded) and a repeat-experiment in the validation phase I using both chemicals and 1 additional compound (N=3, coded). WLR was assessed in 4 laboratories.
Overall the results were concordant, but some results were found difficult to understand where they came from. In study D a different statistical method was used, which may affect the score of the weakly positive chemicals.
In general the acceptance criteria were followed. However, incomplete results were produced by the same laboratory for the 6-well assay. The incomplete results observed for the 96-well assay were due to circumstances affecting compound dosing but beyond the control of the laboratories.

The VMT concluded that the WLR was shown to be satisfactory in all laboratories for the vehicle controls, the positive controls and for the test chemicals.

3. Transferability

A detailed training program, the data produce during training nor the quality of these data were provided in the report. Only for the 96-well assay post-training data were presented, and these data were satisfactory.

The conclusion of the VMT was that the tests were transferable to 'naïve' laboratories after a 1 day trainings session.

4. Between-laboratory reproducibility (BLR)

The BLR pf the 6-well assay was assessed on 12 chemicals (coded) by 3 laboratories. In 9/12 (75%), concordant results (transformation) were obtained. Caffeine gave incomplete data in laboratory VI, due to too low concentrations tested in the test not complying with the protocol. The two remaining laboratories produced concordant results. It is unclear why results for anthracene and o-toluidine were not reproducible. It does appear that there are differences between positive 'initiation' control values in certain labs, while between positive 'promotion' values in other labs. The values for MCA and TPA range between 10 and 50 foci per well. It is not possible to determine from the information given the reason(s) for this spread.
The BLR of the 96-well assay was assessed by comparing laboratory results within the validation study. Results for both initiation and promotion were acceptable (95% for initiation and 85% for promotion). Thus, 22 chemicals were tested by 4-6 labs.

The VMT concluded that of tests were reproducible also between laboratories.

5. Predictive capacity

The predictive capacity of the 6-well assay was assessed by applying the "majority rule" to the results from each of the 12 chemicals and comparing that to the reported carcinogenicity *in vivo*. This resulted in 100% correlation.
The predictive capacity of the 96-well assay was assessed by applying the "majority rule" to the results from each of the 21 chemicals and comparing that to the reported carcinogenicity *in vivo*. This resulted in 86% concordance, 83% sensitivity and 89% specificity.
For common chemicals between the 96-well assay and the 6-well assay, sodium arsenite and o-toluidine were negative in the 96-well assay, with transformed colonies on the side of the wells for arsenite.

The VMT concluded that the predictive capacity of both protocols, based on the chemicals assessed, was satisfactory.

15.2 Extent to which study conclusions are justified by the study results alone

While the advantages and potential of these tests were recognized by the WG, a number of flaws were identified that weakened the statements made by the VMT.

Some of these flaws were related to the study design that had a few shortcomings:

1. The number of chemicals (6-well assay: 8 carcinogens (including TP), 4 non-carcinogens; 96-well assay: 12 carcinogens (including TP), 9 non-carcinogens) was acceptable for assessing the transferability and reliability of the test.

2. While this study produced promising results supporting Sakai et al. (2010) (98 non-coded chemicals), the number of chemicals was considered too low to allow for strong statements with respect to its predictive capacity.
3. The transferability of the test was not convincingly proven because the laboratories which were involved in the transfer phase of the study already were experienced in Balb/c CTA performance.
4. There was no selection procedure described for genotoxic and non-genotoxic carcinogens;
5. Non of the chemicals were labelled as genotoxic or non-genotoxic before the start of the study;
6. The number of TP chemicals (N=3) was too low to demonstrate the capability of the test to discriminate tumor-initiating and tumor-promoting carcinogens.
7. For the 6-well assay no acceptance criteria were described in the protocol.

The conclusions by the WG upon evaluation of this report are:

1. The WLR was shown to be satisfactory in all laboratories for the vehicle controls, the positive controls and for the test chemicals.
2. The transferability of the test to a laboratory not previously experienced in CTA was not demonstrated since the involved laboratories had experience in Balb/c 3T3 CTA.
3. A high level of BLR was demonstrated for both assays. However, significant quantitative differences (number of foci counted) between laboratories were observed for the positive control of the 6-well assay. It is unclear how these differences would affect the detection of less potent chemicals.
4. While this study produced promising results supporting Sakai et al. (2010) (98 non-coded chemicals), the number of chemicals was considered too low to allow for strong statements with respect to its predictive capacity.

15.3 Extent to which conclusions are plausible in the context of existing information

The conclusions drawn by the VMT can be considered plausible. Although the WG identified issues that need to be addressed, these issues are not expected to change the overall picture of the tests with respect to WLR, transferability, BLR and predictive capacity (as defined by the number of foci as read-out).

The ultimate objective of the Bhas42 assay, being a test that discriminates between genotoxic and non-genotoxic carcinogens, and tumor initiating and tumour promoting compounds, has not been proven sufficiently.

16. Recommendations

16.1 General recommendations

The WG considered both CTA protocols sufficiently standardized to recommend them as a basis for an OECD Test Guideline on *in vitro* cell transformation to be used as part of the tool box for cancer risk assessment. However, the WG suggests that more detail could be added.

Some attention should be given to the scoring of foci which seems to have a substantial subjective factor in the 6-well assay (e.g. how to score in case of overgrowth, overlapping foci), as well as in the 96-well assay (e.g. how to score foci that grow half on the bottom of the plate and half on the site). The addition of descriptive/explanatory lines would add to the value of the photographic catalogue.

16.2 Specific recommendations (e.g. concerning improvement of SOPs)

The SOP need more clarification, e.g. a clear description on how to do the target cell seeding in a reproducible way, scoring foci.

The introduction of exogenous metabolic activation systems into the Bhas 42 CTA would support the applicability of the assay to a broad range of chemicals.

17. References

Baird W.M. and Boutwell R.K.(1971) Tumor-promoting activity of phorbol and four 11 diesters of phorbol in mouse skin. Cancer Res., 31,1074-1079.

Kirkland D., Aardema M., Henderson L. and Müller L.(2005)Evaluation of the ability 15 of a battery of three in vitro genotoxicity tests to discriminate rodent carcinogens and 16 non-carcinogens: I. Sensitivity, specificity and relative predictivity. Mutat. Res., 584,17 1-256.18

Lijinsky W., Kovatch R.M. and Thomas, B.J.(1992)The carcinogenic effect of 35 methapyrilene combined with nitrosodiethylamine given to rats in low doses. 36 Carcinogenesis, 13,1293-1297.37

OECD(2007)Detailed review paper on cell transformation assays for detection of 3 chemical carcinogens, OECD Environment, Health and Safety Publications, Series on 4 Testing and Assessment, No. 31. 5 http://appli1.oecd.org/olis/2007doc.nsf/linkto/env-jm-mono(2007)18.6

Reddy B.S., Narisawa T., Maronpot R., Weisburger J.H., and Wynder E.L.(1975)23 Animal models for the study of dietary factors and cancer of the large bowel, Cancer Res. 24 35,3421-3426.25

Sakai A., Sasaki K., Muramatsu D. Arai S., Endou N., Kuroda S., Hayashi K., Lim Y-m., 44 Yamazaki S., Umeda M. and Tanaka N.(2010)A Bhas 42 cell transformation assay on 45 98 chemicals: The characteristics and performance for the prediction of chemical 46 carcinogenicity. Mutat. Res., 702,100-122.47

Sasaki K., Mizusawa H., Ishidate M. and Tanaka N.(1990)Establishment of a highly 9 reproducible transformation assay of a ras-transfected BALB 3T3 clone by treatment 10 with promoters, in: Y. Kuroda, D.M. Shankel, M. Waters (Eds.), Antimutagenesis and 11 Anticarcinogenesis Mechanisms II, Plenum Publishing Co., New York,pp. 411-416.12

Slaga T.J. Fischer S.M. Nelson K. and Gleason G.L.(1980)Studies on the mechanism of 18 skin tumor promotion:Evidence for several stages in promotion. Proc. Natl. Acad. Sci. 19 USA, 77 (1980) 3659-3663.20

18. ESAC Request concerning the current review

The ESAC WG is requested to deliver to the chair of the ESAC the following two documents:

1) Draft ESAC WG Report detailing its analyses and conclusions
2) Draft ESAC Opinion outlining the key findings and recommendations

The conclusions drawn in the report should be based preferably on consensus. If no consensus can be achieved, the draft Report and Opinion should clearly outline the differences in the appraisals and provide appropriate scientific justifications.

19. Annexes

Annex 1 - Supplementary information provided by the test submitters upon request by the WG.

The ESAC Working Group requested the test developers some additional clarifications regarding the validation study and report. The ECVAM secretariat forwarded the questions to the test developers and the following responses were received on 21 September, 2, 9 and 23 October.

1) *In the 96-well protocol, are the foci scored by eye or under the microscope?*
The foci are counted under a microscope (stereoscopic microscope) both in the 96-well method and in the 6-well method.

2) *Were the tumors developed in nude mice originating from single foci or from pooled foci?*
MCA-14, MCA-18, TPA-14 and TPA-15 are the name of clonal transformed cells. So, each of them was cloned from a separate single focus which had been produced in the culture treated with MCA (MCA-14 and MCA-18) or TPA (TPA-14 and TPA-15), and multiplied to obtain an enough number of cells. Each clone was inoculated at the number indicated in the figure into the shot position of nude mice.

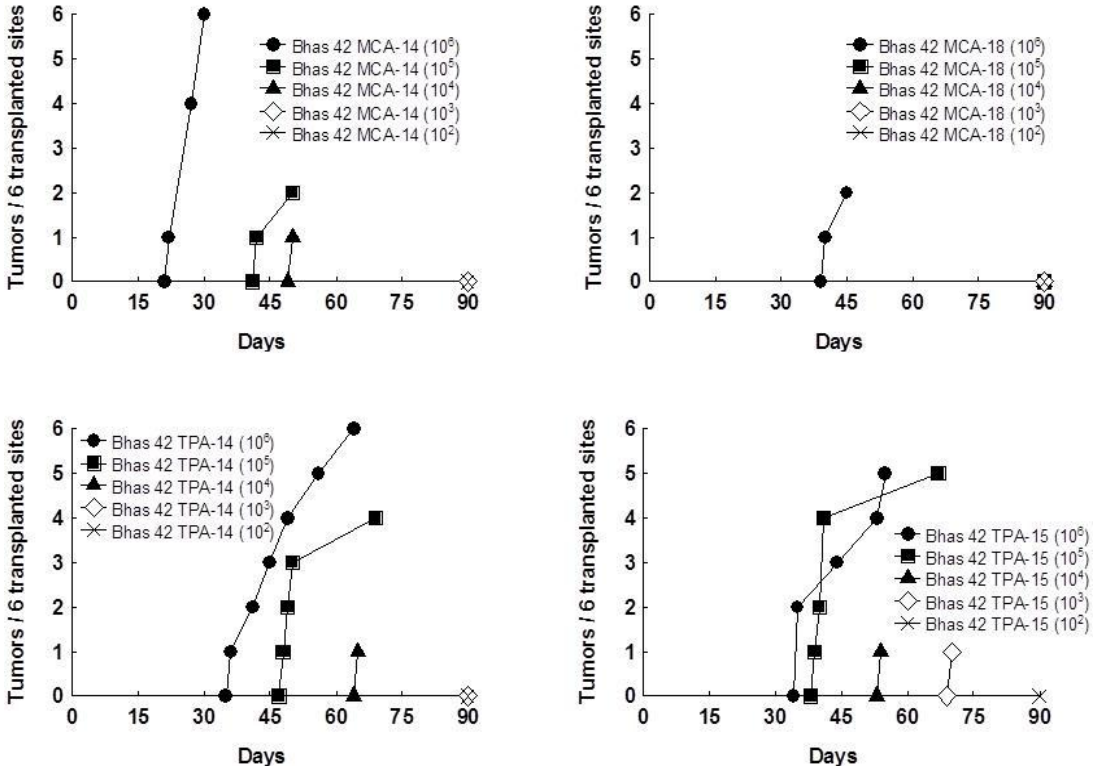

Figure 1. Transformed Bhas 42 cells are tumorigenic in nude mice (but not normal Bhas 42 and BALB/c 3T3 ells)

3) *Promotion assay: what is the cell confluency at the beginning of the promotion treatment? Is it full*

confluency? Are cells still dividing?
Sub-confluence (70-90% confluence under microscope).

4) *In the 6-well plate protocol, important differences were observed in the number of foci in the positive controls of the different labs. What could be the reason? Was any cross-laboratory evaluation of foci scoring performed to understand if the scoring was the reason for this difference? Especially in the cases of different result, e.g. o-toluidine? Could you briefly describe how the lab training was organised?*

The trainings for foci scoring which were performed for validation studies are listed in Table 1, being related to each individual validation study (for the 6-well method validation study, see 3.1.1 Study design in validation study report: page 22, lines 21-27).

Table 1. Chronological order of training and validation studies for the Bhas 42 CTA

Training or Study	Date	Participating Laboratory[1)]
Workshop and training for foci scoring in the 6-well method	31/01/2007	I (3), II (4), III (1), IV (2) and two other labs
6-Well pre-validation study[2)]	31/01/2007-31/03/2007	I (3), II (4), III (1), IV (2) and two other labs
Workshop and training for foci scoring in the 6-well method	27/08/2007	V, VI
6-Well validation study	01/10/2007-13/10/2009	I (3), II (4), III (1), IV (2), V and VI
Workshop and training for foci scoring in the 96-well method	18/11/2008	1 (III), 2 (IV), 3 (I), and 4 (II)
96-well validation study—pre-validation phase	18/11/2008-13/01/2009	1 (III), 2 (IV), 3 (I), and 4 (II)
Joint meeting for evaluation of 96-well pre-validation phase study—harmonization of focus counting in the 96-well method (cross-laboratory evaluation of foci scoring)	26/01/2009	1 (III), 2 (IV), 3 (I), and 4 (II)
96-well validation study—phase I	29/01/2009-16/09/2009	1 (III), 2 (IV), 3 (I), and 4 (II)
96-well validation study—phase II	06/01/2010-14/07/2010	1 (III), 2 (IV), and 3 (I)

[1)] Labs I, II, III, IV, V and VI participated in the 6-well validation study. Labs 1, 2, 3 and 4 participated in the 96-well validation study. Lab I and Lab 3, Lab II and Lab 4, Lab III and Lab 1, and Lab IV and Lab 2 are the same laboratories. See text page 18.

[2)] Bhas 42 cell transformation assay validation study report dose not include this study.

- *What could be the reason?*

It needs training and experience to score transformed foci, i.e. to discriminate transformed foci from non-transformed foci. Therefore, it was possible that the differences in the number of transformed foci in the positive controls between laboratories were caused by insufficient training for foci scoring, although it might not be the only reason.

- *Was any cross-laboratory evaluation of foci scoring performed to understand if the scoring was the reason for this difference?*

I regret that the cross-laboratory evaluation of foci scoring was not performed in the 6-well method validation study.

- *Especially, in the cases of different result, e.g. o-toluidine?*

As I described in 5.6 Discrepancy of results between the 6-well method and 96-well method in the validation study report (page 151, lines 38-48), ***o***-toluidine hydrochloride and ***o***-toluidine were always negative in the initiation assay and equivocal (predominantly negative) in the promotion assay in both 6-well and 96-well methods in the repeated examinations by lead laboratory HRI (Lab I). Labs V and VI reported for ***o***-toluidine hydrochloride to be positive both in the initiation assay and in the promotion assay in the 6-well method validation study. Lab VI reported for anthracene to be positive in the promotion assay, whereas the other two laboratories reported for the chemical to be negative. Labs V and VI are naive in the Bhas 42 CTA. They had only one opportunity to learn foci scoring. They might have wavered in foci scoring. However, Labs V and VI participated in the validation study from foreign countries and their data submission delayed, and ,therefore, under the restricted NEDO budget there was no time and no room to discuss the differences in the results and/or foci scoring

among laboratories. The data submitted from the laboratories were simply presented in the validation study report.

In the 96-well method, comparable results for transformation frequency of negative and positive controls were obtained between laboratories. One of the reasons for this fact is considered to be based on advantageous way of quantifying transformation frequency as described in 5.3 Quantification of transformation frequency in the validation study report (page 148, lines 17-33). As an additional reason we consider that the participating laboratories had got experiences of Bhas 42 CTA for the procedures and foci scoring through 6-well method validation study. The 6-well method and the 96-well method are fundamentally the same assay, and the criteria of transformed foci are the same between two methods (see Annex 12 Recommended Protocol for the Bhas 42 Cell Transformation Assay: 6.1 Record of transformation frequency: line 523). Therefore, we consider that comparable results for the number of foci in the positive controls would have been obtained in the 6-well method validation study if the participating laboratories had been trained a little more. Nevertheless, the number of foci in each positive control was statistically significantly different from that of corresponding negative control in every assay in all the laboratories in the 6-well method validation study. This fact should be remarked.

- *Where the plates from the 6-well study stored?*

Each individual laboratory should store them, although I have not confirmed if all the laboratories still store them. HRI has the plates.

- *Would it eventually be possible to organise an exercise to cross check the foci count in the future?*

I am not sure if we correctly understand the meaning of this question, i.e. what do “exercise” and “cross check” specifically mean? The jobs that need money and/or budget may not be performed, since the NEDO Project finished.

For example, 1) and 2) will be possible but 3) will be impossible among following three assumed items:

1) It will be possible that we ask the participating laboratories to rescore foci of 10 specimens among the plates HRI stores from the 6-well validation study and examine the difference in the number of foci scored, if the laboratories undertake the volunteer work. However, there will be some problems as follows:

- It will be need a long time, since all the laboratories must score foci in the same specimens and therefore the same specimens are circulated among all the laboratories.
- Some (not a few) of the experimenters who took part in the validation study have already left their laboratories.
- It would be possible that the variation may be increased further in this rescoring than in scoring in the validation study, because long time has passed since the validation study was finished.

2) It will be possible for HRI to hold a training course of foci scoring in Bhas CTA or a workshop of Bhas CTA (including the exercise of foci scoring) for the experimenters who are going to start or have started Bhas CTA, after OECD/TG of Bhas CTA is developed. In HRI, the staff sometimes performs Bhas CTA for contract studies or their own research.

3) It will be impossible that for cross-check all the laboratories rescore the foci in all the plates, even if limited to the positive- and negative-control’s plates, produced in all the participating laboratories in the 6-well validation study, considering cost and labor.

In our view, CTAs should be considered to be sufficiently robust when on the whole the number of foci in each individual positive control is statistically significantly different from that in the corresponding negative control, even if there are differences in the number of foci in the positive controls between experimenters and/or between laboratories. It is well known as a characteristic common among CTAs that the batch of fetal bovine serum (FBS) affects transformation frequency. It is meaningless to insist on absolute numbers of foci produced in the positive controls, although too

many foci are unable to be scored and a too small number of foci does not meet statistical analysis. Therefore, it is important that 1) the FBS batch is selected before scheduling CTA as described in Annex 12, the recommended protocol (page 4, lines 155-164: 2.8 Selection of FBS) and 2) the number of transformed foci in the positive control is statistically significantly different from that in the corresponding negative control in each experiment. Meanwhile, the concentrations of positive controls used in the Bhas CTA (MCA, 1 μg/mL; TPA, 0.05 μg/mL) is set to be lower than those used in conventional CTAs (MCA in the standard BALB/c 3T3 and C3H10T1/2 CTA, 2-5 μg/mL; TPA in the two stage BALB/c 3T3 CTA, 0.1-1 μg/mL) . When these lower concentrations of positive controls result in statistically significant increases in the number of transformed foci, it is considered that Bhas CTA is carried out with sufficient sensitivity.

5) *How was the BHAS 42 clone selected? It was not transforming, but were there other criteria, e.g. level of ras expression, etc.?*

Dr. Sasaki transfected the BALB/c 3T3 cells with active v-Ha-*ras* gene, isolated clones containing *ras* (examined by dot blot analysis and Southern hybridization), and selected morphologically non-transformed but TPA- sensitive clones. One of those clones was the Bhas 42.

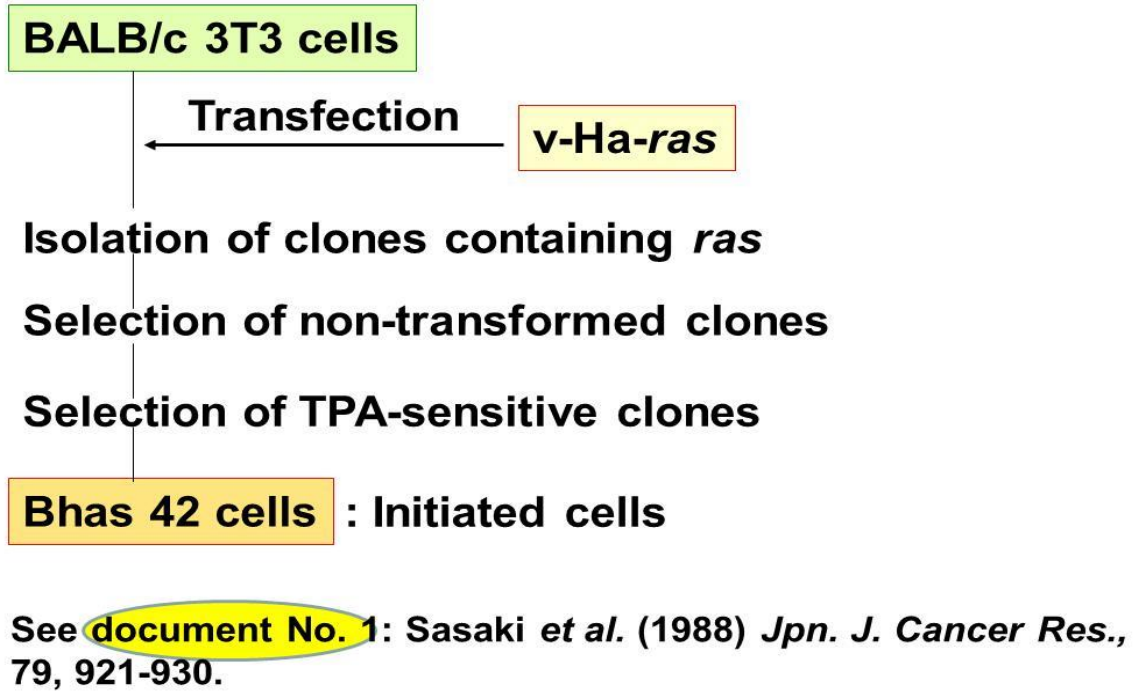

Figure 2. Development of Bhas 42 cells.

At the stage of Bhas isolation, the expression of *ras* was not examined, but the rough amount of integrated *ras* was examined by Southern hybridization. The amount of *ras* sequence integrated into a cell was not related to whether the particular clone is transformed or not, i.e. there were clones which had incorporated a large amount of *ras* sequence but morphologically non-transformed, and those which had a little *ras* sequence but morphologically transformed.

6) *How stable are the BHAS cells? For how many passages can they be stable? Do you have some figures?*

The v-Ha-*ras* gene in the Bhas 42 cells at the current passage, passage 18th, was confirmed by Fluorescence in situ hybridization (FISH). All the cells have the v-Ha-*ras* gene as shown in the following figure 3.

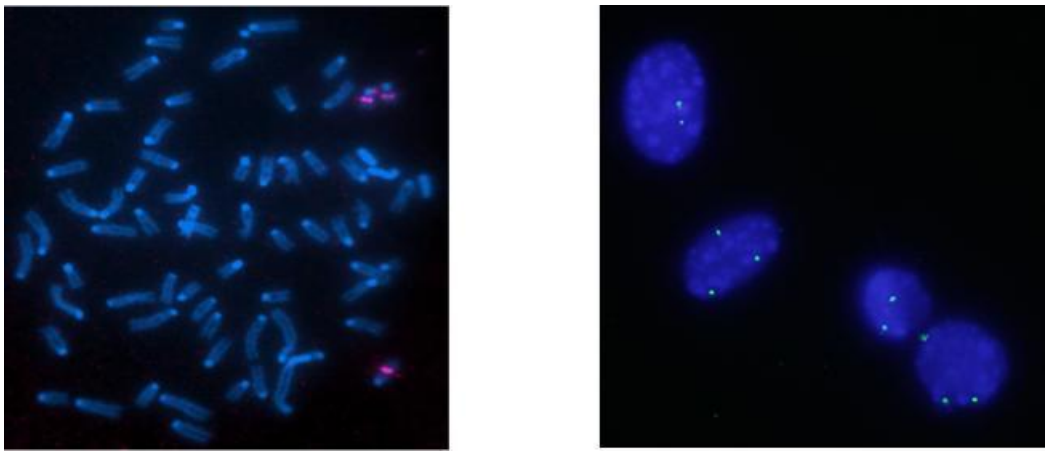

Figure 3. v-Ha-*ras* in the Bhas 42 cells (FISH)

The stability of copy number and expression of v-Ha-*ras* gene in Bhas 42 cells was examined by a quantitative PCR (Figure 4). The copy number and expression assays for the v-Ha-*ras* gene were performed in the Bhas 42 cells at the present passage, passage 18th, and in their subcultures (A and C), and transformed Bhas 42 cells (B and D).

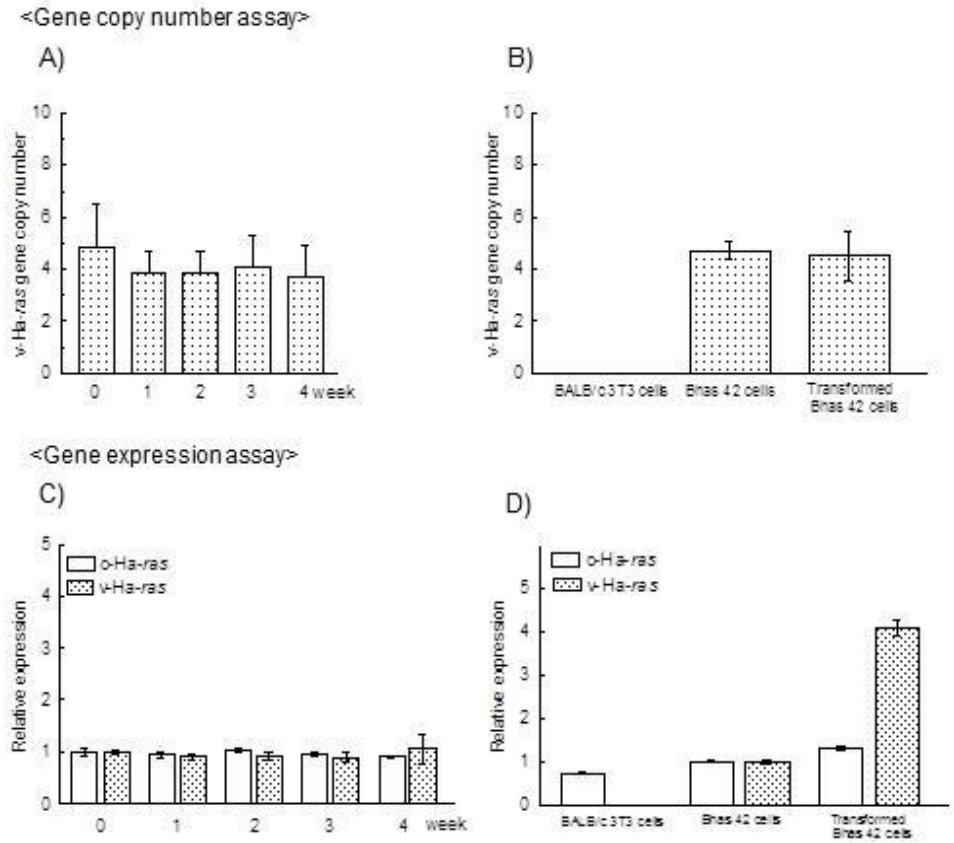

Figure 4. The copy number and expression of v-Ha-*ras* gene in Bhas 42 cells

The results showed that Bhas 42 cells had about 4 copies of v-Ha-*ras* gene per cell and the copy number was unchanged during 4 weeks including 4 passages (Figure 4, A) and between before and after transformation (B). The expressions of v-Ha-*ras* and c-Ha-*ras* gene were maintained at a

constant level in the subcultured Bhas 42 cells for 4 weeks (4 passages) (C). The increased expression of v-Ha-*ras* gene was observed in transformed Bhas 42 cells whereas the expression level of c-Ha-*ras* gene was not affected by transformation (D).

- *The ESAC Working Group would like to confirm if their interpretation of the text is correct. Cells at passage 18 are cultured for 4 passages. Is this considered passage 22?*

Yes, the first column (or columns: 0 week) is of passage 18 and the fifth column (or columns: 4 weeks) is of passage 22 in Figure 4-A (or -C) in the attachment of previous e-mail.

- *Have higher passages been evaluated for cell transformation?*

No, we have never carried out CTA using Bhas 42 cells at the passages higher than 18 in HRI. However, there is the other team in HRI that has been performing CTA for contract studies for more than 10 years, using a working stock of Bhas 42 cells at passage 12. The culture history of our cell stock at passage 17 is the same till passage 12 as that of their cell stock, and their data are comparable with those of our team. Therefore, Bhas 42 cells should also be stable for CTA at the higher passages, if experimenters hereafter prepare cell stocks keeping the conditions described in Annex 12, the recommended protocol (page 3, lines 125-138: 2.6 Stock cells).

- *Why have cells been cultivated for 18 passages after transfection?*

There is no special reason. One reason is that we want to avoid consuming Bhas 42 cells at young passages. Bhas 42 cells have been frequently used in HRI for our research and contract studies and often requested from other laboratories. Therefore, we have sequentially prepared sub-muster stocks. The passage 17 is the sub-master stock which was prepared for the NEDO project and used for in-house and collaborative studies including the validation studies. HRI and the other validation-participating laboratories prepared the working stocks at passage 18 from the passage 17 distributed for collaboration.

HRI still stores the master stock at passage 10. As above mentioned, the other team is still using the cells at passage 12 for CTA.

7) *In the report, page 33, table 4bLab VI, concentration 15 µg/mL,*
Can you confirm that the number of foci is correct (11.7 ± 4.5) and at the same time is not statistically significant?
The number of foci (11.7 ± 4.5) is correct and statistically significant. Please add "*". (cf. Annex 4, Results Submitted from Laboratories in the Validation Study of 6-Well Method, AAF, Lab VI, Promotion, where the data and the results of the statistical analysis are presented). I am very sorry. I mistyped or dropped "*" during editing.
I have checked data in the other tables by comparing them with raw data in Annexes 4, 9, 10 and 11.

8) *If the BHAS cells are initiated cells, how can we explain that they are still able to detect initiator chemicals?*
Bhas 42 cells gained the mutated ras gene, v-Ha-ras, by transfection. Ras is strongly related to carcinogenesis among cancer related genes. Therefore, in the multistage carcinogenesis, Bhas 42 cells is apt to transform more than mother cells (BALB/c 3T3), because of skipping one step to go up malignant stage. And also Bhas 42 cells still keep the character of BALB/c 3T3. Transformed foci produced in Bhas 42 cells cause tumors in nude mice (in contribution by Sasaki et al.). This cell system is clearly different from the primary SHE cell CTA system.

In the course of development of Bhas 42 cells from BALB/c 3T3 cells, Sasaki transfected v-Ha-ras into BALB/c 3T3 cells, isolated clones having v-Ha-ras, selected morphologically non-transformed clones and then selected TPA sensitive clones. One of those clones was Bhas 42.

The initiated cells are in the process toward complete transformation and presumed to have a mutated gene which render the cells predisposed for transformation and/or sensitive to tumor-promoter (In case of Bhas 42, the mutated gene is v-Ha-ras). Thus, initiated cells are sensitive to transformation by tumor-promoter.

Transformation by initiators (genotoxic carcinogens) is caused by accumulation of genetic alterations. Generally, it is considered that mutation in a single gene does not result in transformation (carcinogenesis) and transformation (carcinogenesis) requires mutations in multiple genes. Therefore, normal (BALB/c 3T3) and initiated (Bhas 42) cells are transformed by genotoxic chemicals through the accumulation of genetic alterations in their genome. In this theory, the initiated cells must be a little more sensitive to transformation by genotoxic chemicals than normal cells, since the initiated cells have already obtained at least one alteration in the genome.