Published 2008 Wiley-Liss, Inc.[†] Birth Defects Research (Part B) 83:446–456 (2008)

Review Article

State of the Art in Developmental Toxicity Screening Methods and a Way Forward: A Meeting Report Addressing Embryonic Stem Cells, Whole Embryo Culture, and Zebrafish

Robert Chapin,[1]* Karen Augustine-Rauch,[2] Bruce Beyer,[3] George Daston,[4] Richard Finnell,[5] Thomas Flynn,[6] Sidney Hunter,[7] Phillip Mirkes,[8] K. Sue O'Shea,[9] Aldert Piersma,[10] David Sandler,[11] Philippe Vanparys,[12] and Geneviève Van Maele-Fabry[13]

[1]Pfizer Global R&D, Developmental and Reproductive Toxicology Group, Groton, Connecticut
[2]Reproductive Toxicology, Bristol-Myers Squibb, Hopewell, New Jersey
[3]Sanofi-Aventis US, Drug Safety Evaluation, Malvern, Pennsylvania
[4]Central Product Safety, Procter and Gamble, Cincinnati, Ohio
[5]Center for Environmental and Genetic Medicine, Texas A&M University HSC, Houston, Texas
[6]U.S. FDA, CFSAN, Office of Applied Research and Safety Assessment, Laurel, Maryland
[7]U.S. EPA, Office of Research and Development, NHEERL, Reproductive Toxicology Division, Research Triangle Park, North Carolina
[8]Center for Environmental and Rural Health, Texas A&M University, College Station, Texas
[9]Department of Cell and Developmental Biology, University of Michigan Medical School, Ann Arbor, Michigan
[10]National Inst. For Public Health and the Environment, Bilthoven, The Netherlands
[11]Health and Environmental Sciences Institute, Washington, DC
[12]Johnson & Johnson Pharmaceutical Research & Development, Mechanistic Toxicology, Beerse, Belgium
[13]Université Catholique de Louvain, Industrial Toxicology and Occupational Medicine Unit, Brussels, Belgium

A meeting was convened so that users of three models for in vitro developmental toxicity (embryonic stem cells, whole embryo culture, and zebrafish) could share their experiences with each model, and explore the areas for improvement. We present a summary of this meeting and the recommendations of the group. *Birth Defects Res (Part B)* 83:446–456, 2008. Published 2008 Wiley-Liss, Inc.[†]

Key words: ***in vitro screens; environmental issues; pharmaceuticals***

INTRODUCTION

For more than 30 years, scientists have recognized the value of an in vitro assay that would accurately predict developmental toxicity in vivo. The reasons given are always the same and include: pharmaceutical companies could advance for further development only those compounds with a low likelihood of toxicity after having made only small amounts of possible candidates; chemical companies could compare broadly across many structures and go forward with only those that fit a certain activity profile; and regulatory agencies could quickly compare across many dozens or hundreds of compounds found in the environment and select those with the greatest probability of causing developmental toxicity for more definitive testing, etc. All of these rationales are still valid.

In addition to those considerations of speed and minimizing test compound amounts is the issue of animal use. The British Fund for the Replacement of Animals in Medical Experiments was founded in 1969, followed by the Johns Hopkins Center for Alternatives to Animal Testing in 1981 and then in 1989 by the establishment of the German Center for the Documentation and Evaluation of Alternatives to Animal Experiments within the German Federal Institute for Risk Assessment, and the European Centre for the Evaluation of Alternative Methods (ECVAM) in 1991. All of these organizations work to reduce the numbers of animals used in health research, and they advocate and carry out or sponsor research to support the development of alternative methods and models.

Many possible models have been explored, and a small sampling would include hydra regeneration (Johnson et al., 1988), chick embryo neural retina cells (Moscona,

*Correspondence to: Robert Chapin, Pfizer Global R&D, Eastern Point Road, MS 8274-1336, Groton, CT 06340.
E-mail: robert.e.chapin@pfizer.com
David Sandler's current address: Food Safety and Inspection Service, U.S. Department of Agriculture.
Received 21 April 2008; Accepted 21 April 2008
Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/bdrb.20158

1961; Daston et al., 1991), embryonic palatal mesenchyme cells (Pratt et al., 1982), mouse ovarian tumor cell attachment (Braun et al., 1979), chick embryos (Tickle, 1983), whole rat embryos in vitro (Steel et al., 1983), whole mouse embryos (Sadler et al., 1982; Van Maele-Fabry et al., 1990), mouse palatal cultures (Abbott et al., 1989), mouse limb bud reaggregates (Kistler, 1987), embryonic stem cells (Doetschman et al., 1985), and rabbit whole embryo culture (Pitt and Carney, 1999). There have been efforts to compare across multiple assays (e.g., Steele et al., 1988). In the late 1990 s, ECVAM sponsored work to ''validate'' the performance of three of these alternative systems: limb bud micro mass, rat whole embryo culture (WEC), and the embryonic stem cell assay (ESC). In this context, 'validate' means to correlate the results from animal studies in vivo and results from the in vitro studies using a strictly defined protocol, and to characterize the performance and reproducibility of the assay in several independent labs.

The innate value of this approach (using embryos or parts thereof in vitro), and the empirical case for why these models should work, was made about a decade ago in an excellent review (Daston, 1996), which reported that many of these models seem to predict animals' developmental toxicity correctly 70–80% of the time. More recently, this was again shown to be the case for the three assays that ECVAM evaluated (Genschow et al., 2002).

It is against this background that the authors of this study assembled. We represent a steering committee for a project sponsored by the Developmental and Reproductive Toxicology Technical committee of the Health and Environmental Sciences Institute, which is part of the International Life Sciences Institute. We thought that some additional progress might be made in advancing the predictivity of ''alternative screens'' if some of the users of these methods assembled and shared their data, the circumstances in which the data are used, and the occasions where the assay has failed. The hope was that themes might emerge from a larger data set that would not be apparent to individual labs, or that two or more labs would identify an area where, by working together, they would be able to accomplish more than either one alone, and would significantly advance the science.

Because the intent was to foster discussions among people who are using each of the assays currently, as well as for practical logistical reasons, the resulting workshop was limited to the consideration of just three models: whole embryo culture, mouse embryonic stem cells, and zebrafish. Whole embryos and stem cells performed best in the ECVAM validation efforts mentioned above, and the zebrafish represents an emerging model with tremendous promise.

The workshop began with overview presentations on each model system, providing some historic background and some basic biology. We also heard a presentation about the general requirements for creating predictive models.

Each model was then considered in its own breakout group, where the speakers were asked to share how they actually carried out the assay, how they interpreted their data, and how they used their data, i.e., what sort of decisions did they make based on these results. They were also asked to look forward and opine about changes they would like to see made. Are there obvious best practices that should be more broadly disseminated? How can this test be improved? Are different approaches better for pharmaceutical versus environmental compounds? Could we identify common endpoints/standards/dependent variables for these tests?

The vastly different state of development of these assays is reflected in the nature of the discussions. For example, because the whole embryo culture assay has been used for nearly 30 years as an investigative tool, the discussants recognized that there is a potential treasure trove of untapped data gathering dust in various laboratories. One priority ought to be the uncovering of those data and evaluating them to see if they can help create a better predictive model. At the other end of the spectrum, zebrafish have only just begun to be used to predict mammalian toxicity. Accordingly, the needs in that field are quite different, and revolve around developing optimal methods and then generating a larger public database detailing the concordance between changes in zebrafish and in the more conventional animal models (rats, mice, rabbits).

The following summarizes the main conclusions and messages from this most collegial workshop.

Whole Embryo Culture

The assay. The assay is run by explanting rodent or rabbit embryos with their yolk sacs on approximately gestation day (GD) 9.5 or 10 for rats. The conceptus is then cultured on a rotating platform in a mixture of serum and culture medium (and test article, if applicable) for 44–48 hr with increasing proportions of oxygen added in the gas overlayer (Fig. 1). At the end of the culture period, the conceptus is evaluated for the degree of maturity of various endpoints, e.g., number of somites, optic development, forelimb development, neural tube development, etc. Each of these is given a score (morphologic score based on methods developed by Brown and Fabro, 1981; Klug et al., 1985; Van Maele-Fabry et al., 1990), and the scores are summed for each concentration tested. The presence and type of any malformations are also noted. The read-out of the assay can be: 1) the concentration at which malformations begin to be evident; 2) a comparison across different compounds of the types and severity of malformations seen at a given concentration; or 3), the results from the predictive linear discriminant analytic formulae when using the ECVAM version.

Rodent WEC was included in a validation study of embryotoxicity tests conducted by the ECVAM (Piersma et al., 2004). The primary objective of the ECVAM validation study was to assess the performance of three selected in vitro tests (WEC, the micro mass test, and embryonic stem cell test) in discriminating among non-embryotoxic, weakly embryotoxic, and strongly embryotoxic compounds. Biostatistically-based prediction models were developed for each of the three tests being validated, based on the results of a preliminary study. The reproducibility of the WEC test as well as the concordance between the embryotoxic potential derived from the in vitro data and from in vivo data were good according to predefined performance criteria. The prediction model correctly classified 80% of the 20 tested compounds for all embryotoxicity classes (non-, weakly, and strongly embryotoxic). More information on the validation study, including comprehensive protocols of

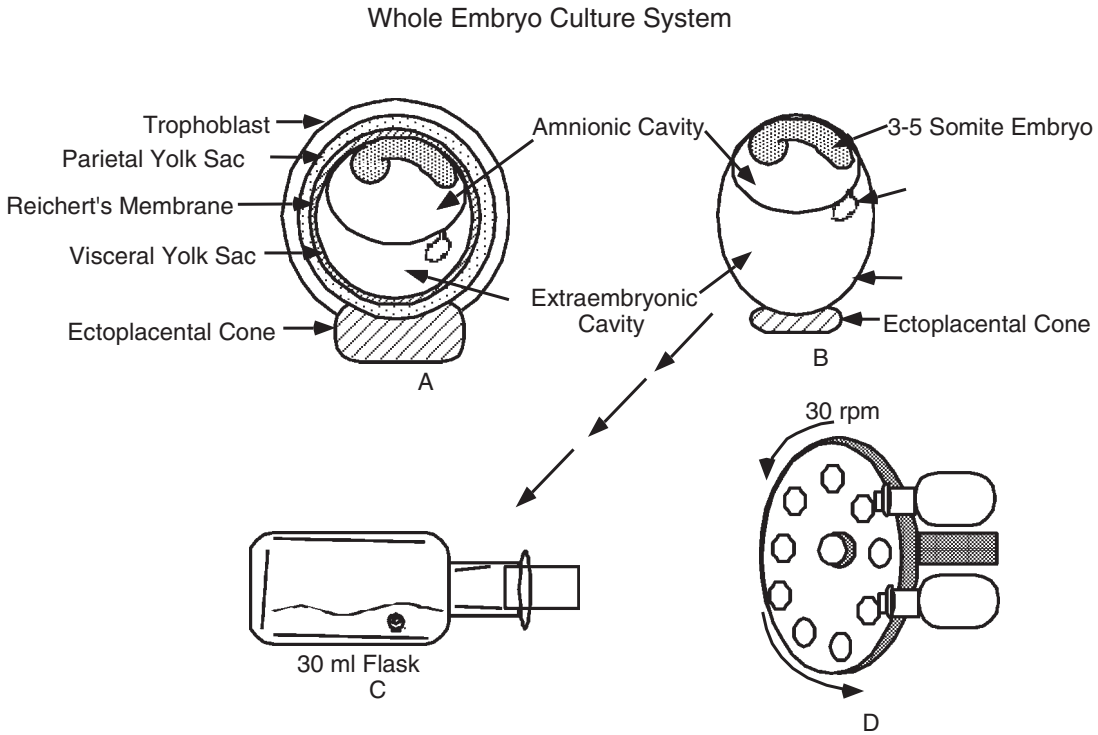

Fig. 1. Schematic representation of the technique of culturing whole rodent embryos. The conceptus is removed from the uterus (**A**), dissected free of maternal tissues, trophoblasts, the parietal yolk sac, and Reichert's membrane (**B**), and placed into a culture flask of medium (**C**) and rotated in an incubator at ca. 30 rpm and 38°C. (Figure courtesy of Dr. Sid Hunter, U.S. EPA).

the methodology, is available on the ECVAM web site (http://ecvam-dbalm.jrc.ec.europa.eu/).

Because WEC was considered to be a mature technique, the breakout session focused on current applications for WEC (pharmaceutical screening/testing, screening human populations for environmental and nutritional factors adversely affecting reproduction, and regulatory use of WEC data) and possible ways the assay could be improved or enhanced (novel endpoints and alternate species).

What does the WEC currently do well? A consensus was reached fairly quickly among the workshop participants for the need to more completely use all of the data collected as part of the WEC protocol when refining or establishing prediction models. It was believed that the ECVAM prediction model is relatively simplistic, but could be made more sophisticated by incorporating more of the endpoints already collected in WEC testing. For many labs, growth parameters (e.g., crown–rump length, protein content) are good predictors of potential embryotoxicity with the added advantage of being continuous variables (unlike morphology-based parameters). Accordingly, it is proposed to change the criteria used to evaluate embryotoxicity in WEC so as not to rely on morphologic scores alone. Therefore, the IC_{NOAEL} for a compound may be related to total morphologic score, but there is often a steep dose–response curve resulting in a rapid transition from non-embryotoxic to strongly embryotoxic effects. In these cases, the maximum inhibitory concentration (IC_{MAX}) would be equivalent to the concentration producing the highest malformation incidence. It is acknowledged that the IC_{NOAEL} and IC_{MAX} are dependent on dose selection, whereas the IC_{50} (concentration causing 50% inhibition of the response) could be derived from the dose–response curve, and would provide a more robust evaluation of the data. In addition, embryos obtained from different species provide different predictions, as is also true in vivo. Therefore, species-specific prediction models are also needed. Based on the data presented at the meeting, it was apparent that the ECVAM model does not work well as a screening tool for drug lead prioritization, although newer models developed by some pharmaceutical companies do seem to serve this function adequately. WEC should not be used for human risk assessment purposes at the present time, based in part on the absence of a maternal compartment and the very limited exposure window. Compound class-specific prediction models may improve predictivity within pharmaceutical or chemical structural classes.

What are the unique features of the WEC that make it advantageous? It uses an intact embryo, rather than sub-components. Thus, all components are present and able to interact and respond to exposures.

1. It employs an intact embryo, rather than sub-components. Thus all components are present and able to interact and respond to exposures.
2. The species used for WEC (mouse, rat, and rabbit), are the same species that are used most commonly in whole animal reproductive toxicity assays, thus

allowing a direct correlation between in vitro and in vivo findings, and against mouse embryonic stem cell data (vide infra).
3. Cultured rodent embryos are information rich.
4. WEC can recapitulate in vivo embryonic development for up to 48 hr.
5. Embryos can be treated in a milieu isolated from maternal effects, but some maternal effects can be introduced into the culture system if desired (e.g., obtain culture serum from treated animals, add known maternal metabolites to culture medium, hyperthermia).
6. WEC is useful for mechanistic studies, prioritization/screening of compounds, studying intrinsic differences between species, providing adjunct information for regulatory/risk assessment purposes, or to further investigate in vivo findings to increase confidence in the data while minimizing the use of animals.
7. WEC has a number of advantages relative to in vivo studies: it provides a data rich assessment of the chemical in question; it includes many targets; it recapitulates a period of in vivo morphogenesis within a teratologically important 48-hr development window; it is isolated from maternal influences (metabolism, toxicity) so that metabolites can be tested individually; and it provides the possibility of including metabolic or kinetic evaluations by using serum from different species (including humans) either treated with the drug/chemical or representing altered physiologic or medical conditions.

What are the shortcomings of the WEC? It is also recognized that WEC has some disadvantages relative to in vivo studies, including:

1. it cannot replace in vivo developmental toxicity studies at the present time because it does not recapitulate the maternal–fetal interactions or expose the conceptus for the gestational period of concern (implantation to near-term);
2. isolation from maternal influences (metabolism, toxicity) that may contribute to in vivo effects;
3. restriction to a relatively narrow developmental window that may not allow it to capture some manifestations of developmental toxicity; and
4. variation across aliquots or collections of serum.

Use of WEC to screen human populations for environmental and nutritional factors adversely affecting reproduction. WEC typically involves culture of embryos from the presomite to the early somite stages. Gestation day 9.5 or 10.5 rat embryos usually are cultured for 48 hr, then evaluated (e.g., embryonic length, structural anomalies, morphologic scoring; yolk sac morphology; protein and DNA content). One major consideration for using WEC is the composition of the culture medium. Attempts to develop a chemically defined medium have not been successful. Medium must contain from 50–90% serum depending on whether embryos are placed into culture at late- or early-somite stages, respectively. Serum is generally obtained from the same species as the embryos cultured. However, human serum obtained from individuals with diabetes (mellitus or gestational) or recurrent spontaneous abortion has been used to identify serum factors that may be related to the poor reproductive outcomes in these populations (Zusman et al., 1989; Ferrari et al., 1994).

Use of WEC by regulatory agencies. At the present time, no U.S. regulatory agencies (i.e., FDA, EPA) are using WEC for regulatory decisions because WEC has not been validated for that purpose. As a consequence, in vitro alternatives are unlikely to be accepted in the near future. However, WEC is recognized as potentially useful for the prioritization of chemicals, hazard identification, and mechanistic studies. Accordingly, a number of regulatory agencies and centers at the FDA and EPA use WEC for in-house research programs. It is recognized that WEC has some limitations: compromised maternal-fetal-placental relationships and evaluation of a limited window of developmental time. Possible refinements of the technique include use of pharmacokinetics/plasma concentration determinations, development and use of additional developmental endpoints, and following Good Laboratory Practice regulations to ensure consistency and reproducibility.

Alternative WEC species. Although rat or mouse are perhaps the species used most commonly for this method, embryos from many species can be grown in whole embryo culture [see New and Mizell (1972) for a description of the culture of opossum fetuses]. However, the use of exotic species is not a sustainable approach, and we do not know whether these species respond more like humans than do rodents, for which we have a significant historic database. Rabbit embryos can be cultured relatively easily using techniques similar to those developed for rat and mouse WEC. Typically, one rabbit embryo is allocated per test substance concentration, with continuous flow gassing using a rotating incubator. Endpoints evaluated are viability, growth, morphologic scoring [based on the rat WEC method of Brown and Fabro (1981) as modified for rabbit embryos by Carney et al. (2007)], and biochemical measures (total protein, DNA). One key difference between rat and rabbit embryos concerns the yolk sac: GD 9 rat embryos rely on histiotrophic nutrition through an inverted visceral yolk sac in which the embryo is enclosed within the yolk sac; rabbit embryos do not have an inverted yolk sac, and lie outside the yolk sac until approximately GD 13. However, yolk sac-mediated histiotrophic nutrition is specific to rodents, and may be less relevant to humans (that rely on hemotrophic nutrition). For the future, as with rat and mouse WEC, it is anticipated that functional endpoints (endocytosis, proteolysis), gene expression (yolk sac transporters), and imaging techniques (Micro CT, magnetic resonance imaging, morphometry of embryonic volume, or specific landmarks of development) will become useful endpoints for rabbit WEC analysis. Rabbit WEC can also help address species differences in developmental toxicity responses.

What are the next steps to move the assay forward?

Novel applications of rodent WEC for teratogenic assessment. An improved prediction model (PM) was discussed that requires only six embryos per tested concentration, and incorporates more of the endpoints collected in standard WEC assays. Only

10 mg compound is required at standard concentrations of 0.1, 1, and 10 μM. A preliminary PM was developed using a quick screen conducted at a single concentration of 0.1 μM. A mean morphologic score is calculated based on evaluation of the brain, somites, and spinal cord deviation because these have been found to be highly sensitive to compounds; false negative results are obtained if the compound does not affect one of these structures. The inclusion of additional endpoints and concentrations would strengthen the PM.

Novel endpoints. The classical endpoints that are typically evaluated in WEC include: mortality, morphology, developmental stage, growth (e.g., protein content), and physiologic parameters (e.g., heart rate). Other endpoints can be assessed, depending on the scientific question at hand: cell fate markers for cell death (using Lyso-tracker, caspase 3, or TUNEL), DNA integrity (Comet assay), RNA for gene expression to assess DNA damage pathways (apoptosis, oxidative stress), gene expression profiles using microarrays, and molecular function and binding. In addition, proteomics is useful to study proteins in the whole embryo or in specific anatomic areas, whereas modified proteins can be monitored using 4-HNE immunofluorescence or other immunohistochemical techniques. Finally, developmental signaling pathways can be reconstructed to integrate this information with respect to reporter transgenes (RARE-hsp lacZ) and pathway analysis to evaluate relationships (cross-talk) between different signaling pathways. Currently these are not being pursued as modifications to a predictive assay.

Future directions in research may make it possible to simplify evaluation of DNA integrity (high-throughput Comet assays, 3-D vertical Comet assays, and 96-well plates), RNA gene expression (EMAGE, spatial information like in situ hubridization to detect the location of transcripts, and specific genes), proteins (antibody microarrays to detect differences in protein expression), signaling pathways (ELISA assays of transcription factors, high throughput of DNA binding activity), and developmental signaling pathway cross-talk. A combination of high-throughput and modern imaging techniques should lead to further advances in the use of WEC. Similarly, computational analyses capable of handling large quantities of information, like systems biology, might be very useful. At present, these techniques have to prove their practicability in WEC.

Recommended enhancements to enable WEC to reach its full potential include the following:

1. Creation of a centralized database, with maintenance and curation to be determined, with the intent that this database should be readily accessible to, and updated by, researchers in academia, industry, and government. More robust statistical evaluation of the expanded database would lead to improvements in the prediction models (PMs).
2. Further standardize culture media, measurements, terminology, and dependent variables.
3. Create an atlas to illustrate findings and to provide a source for uniform terminology across laboratories including appropriate links to different active groups in the area.
4. Update the existing PMs through the incorporation of more endpoints already collected in standard WEC assessments (i.e., enhancement of the ECVAM PM), and create subject- or species-specific PMs.
5. Continue basic research into areas that may provide enhanced endpoints such as: biomarkers of toxicity or developmental signaling patterns; cell fate; molecular endpoints (genetic or proteins); cell identity (different cell populations, such as neural crest); markers to detect events not shown by morphology during the culture period (e.g., markers of limb pattern formation).
6. Revisit the need for rat serum as the culture medium. Characterize rat serum using modern proteomic techniques (e.g., MALDI-mass spectrometry) to identify serum factors essential to normal in vitro embryonic growth and development. Produce the necessary amounts of these factors using recombinant DNA technologies.

Mouse Embryonic Stem Cell Test

The assay. The proposal that embryonic stem cells could be used to evaluate the potential developmental effects of xenobiotics has led to a myriad of protocols for its implementation. However, the embryonic stem cell test (EST) as described by ECVAM (http://ecvam-dbalm.jrc.cec.eu.int/public_view_doc2.CFM?id = DC5abdf7ac30f1b7ef27e87d68aac7180bb0bc12cb10496cda74b54630a05a3291b895581f634) may have been used most extensively for such screening. Mouse embryonic stem cells (mESC) (Fig. 2) are maintained in culture as pluripotent cells by incubation with leukemia inhibitory factor (LIF). mESCs are passaged in the presence of LIF at regular intervals to maintain the optimal cell density and prevent differentiation. At the beginning of the EST, mESCs are removed from the dish and separated into a single cell suspension and grown in non-adherent conditions in the absence of LIF. Culture media are prepared with the appropriate concentrations of test agent and cells are suspended in the media to give the correct cell counts. If the cells will be exposed to an agent of interest, this exposure begins at the time of initial plating. Approximately 750–1000 cells are placed in 20 μl drops on the underside of a culture dish top that is then inverted over the bottom so that the drops hang from the plastic top. Using the ''hanging drop'' method, cells will aggregate in the bottom of the drop and form nascent embryoid bodies (EBs). After 3 days in ''hanging drop'' culture, the EBs are flooded into a culture dish and grown in suspension en masse (treated cells with others of the same concentration of test article) in fresh differentiation medium (without LIF). On Day 5, one EB is placed in each well of a 24-well plate, medium (with test article) added to each well, and the cells allowed to attach and undergo differentiation for 5 days (until study Day 10). One plate is run per concentration of test agent. On Day 10, the wells are examined for the presence or absence of any spontaneously contracting (i.e., beating) cells, which would indicate the differentiation of ESCs to cardiomyocytes. Cytotoxicity is also determined for 3T3 cells and for ES cells by an appropriate method, e.g., MTT reduction or ATP content. The predictive model (Genschow et al., 2002) requires the input into a predictive set of equations the concentrations that inhibit

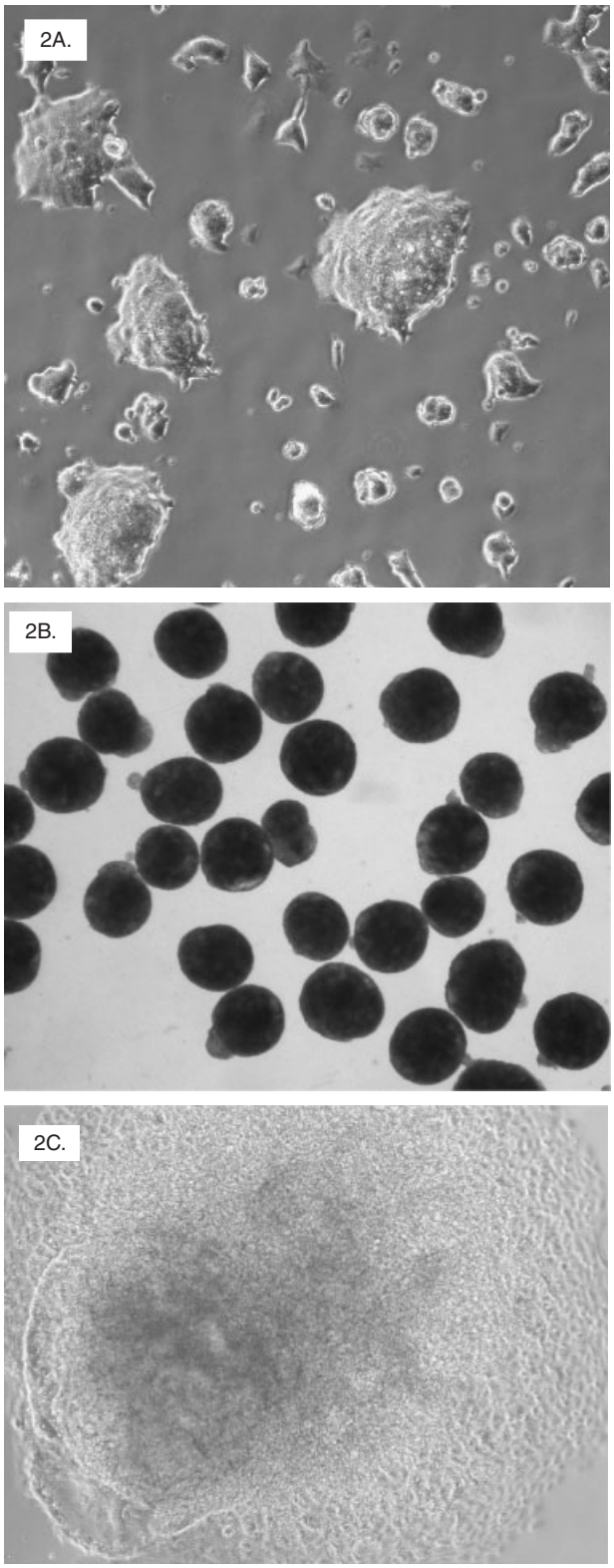

Fig. 2. Stages of mouse embryonic stem cell development. **A**: Undifferentiated ESC's. **B**: Day 5 embryoid bodies, suspended in medium and **C**: a Day 10 embryoid body attached to the culture dish substrate. The periphery of the EB is adherent and is spreading along the culture dish floor. The out-of-focus area in the left of the image is large multicellular mass of differentiating cells. It is in this area where beating cardiomyocytes are often found. (Courtesy of Don Stedman, Pfizer, Inc.).

differentiation by 50% (as measured by fewer wells with any beating cells), and that lower 3T3 and ESC viability by 50%.

What does the EST do well? In comparison with WEC, the stem cell test is relatively simple to carry out, and the main endpoint (scoring of beating) requires no in-depth knowledge of phenotype and morphologic development. The assessment of proliferation and cytotoxicity relies on standard, well established assays already in use in many laboratories. Thus, there is only limited additional training required for laboratories already adept at cell culture to carry out the EST as described in the ECVAM protocol. The assay is more amenable to relatively high-throughput modifications (robotics, scale reductions).

Advantages and limitations of EST as a predictive tool. The EST seems to work best at both ends of the activity spectrum: if a compound is classified as not a developmental toxicant, there is a very high chance (>90%) that this will be true. For strong developmental toxicants, the assay finds them all, but the assay can also classify some weak and non-toxicants as strong developmental toxicants. Embryoid bodies, by their very nature, contain many differentiating cell types. One advantage of this format is that cultured cells are easy to process to monitor the expression and levels of genes that are found in certain specific cell populations within the EB. Another advantage is that stem cells are perhaps the best hope that toxicologists have for doing work with human tissue in development. This advantage comes with the limitation that only 21 human embryonic stem cell (hESC) lines are available that can be used in federally funded research in the U.S. (see http://stemcells.nih.gov/ for current policy and stem cell registry). Inherent in selecting a hESC is the need to understand the genetics of that line (e.g., SNPs) and the predictability of any one line for a diverse population. Some of these lines have also been shown to have a differentiation bias toward different cell fates. Methods developed and optimized using mouse cells might be applied to human cells, and human cells might yield improved hazard identification (with the caveat about the very different rates of differentiation and differences in signaling pathways between mouse and human cells). The current predictive model for the EST depends largely on the absolute concentration for the various toxicities; i.e., a very low effective concentration for producing cytotoxicity will drive the model to categorize the compounds as "Strong," even if its in vivo activity is weak or less. And, like the WEC, this assay lacks a maternal component, and has limited metabolic capability (although the use of S9 can add this when necessary); this is both a strength and a weakness. The strength is that metabolites and parent compounds may be tested individually and the true active agent identified. Conversely, identifying and obtaining the metabolite(s) may be a challenge. The absence of a meaningful maternal component means that the assay is currently limited in its ability to model the direct effects of the compound on the developing system and predict dose-limiting maternal toxicity. Thus, it will be impossible to predict developmental effects produced by changes in maternal physiology (e.g., acidosis) that, in turn, alter development in vivo.

What are the shortcomings of the EST? As a reflection of the way the assay was designed, the current predictive model over-predicts some activities, such that many developmental non-toxicants are classified as toxicants by the assay. Effectively segregating non-toxic and weakly-toxic compounds is one of the greatest challenges for the assay. Another potential liability of the EST is its reliance on one differentiation outcome in the assessment. The identification of beating cells as a marker of cardiomyocyte differentiation may be confounded by an effect of the compound (test agent) on the contraction of the cardiomyocytes (such as altered energy production) or a direct cardiomyocytes toxicant. Also, there is no difference in the assessment of differentiation when one well has 10 beating cells compared with a well that contains 10,000 beating cells. Thus, this lack of discrimination in what constitutes ''differentiation'' may add to a lack of specificity in the assay. Further, the random differentiation in embryoid bodies and fact that the differentiated cells produce yet unknown growth/protective factors and cell types adds another layer of uncertainty. In general this assay also does not produce late-differentiating cell types that can be seen in, e.g., teratomas.

What are the unique features of the EST that make it advantageous? The unique features of stem cells are clearly the ability to differentiate in vitro, to all of the components of the embryo. The processes involved in establishing each embryonic layer (e.g., ectoderm, mesoderm, and endoderm) and the subsequent differentiation of these embryonic cells are recapitulated in this model. Thus, the model has the capability to assess many of the events associated with embryogenesis. One of the future advances will likely be the use of molecular markers to evaluate phenotypic differentiation. Nascent genomic modifications hint at future abilities to monitor the presence and health of numerous specific cell populations within the embryoid body, thereby providing a more subtle means of following differentiation and potentially reducing the length of time required to assess the effects of a test agent. Also, related to this is ability to derive mutant cell lines or produce ''indicator'' cell lines with reporter molecules inserted into the genome.

Another advantage of this system is that it carries out a direct comparison between differentiation and cytotoxicity/proliferation. This comparison may add to our ultimate characterization of xenobiotics as developmental toxicants in vivo. Because of the nature of the ESCs, the EST is amenable to relatively high-throughput modifications (robotics, scale reductions) for culture and a point-by-point visual evaluation of a physical structure to facilitate morphologic evaluations (such as contraction or large lipid droplets).

Also, once derived, the mESC do not require the use of animals, which is a major benefit in some contexts. In contrast, for several human ESC lines maintenance culture of the pluripotent cells does use a feeder layer of mouse embryonic fibroblasts that requires the use of additional animals. Advances are currently being made for xeno-free hESC culture that would not require animals.

What are the next steps to move the assay forward? There are several significant challenges and opportunities that face this assay.

- One is the better use of automation to speed the throughput of test agent analysis. This could include:
  - a higher-throughput means of producing hanging drops or optimizing the production of aggregates so that the resulting EBs more closely mimic those produced by hanging drops;
  - transforming the design to a 96- or 384-well version;
  - a marker of cardiomyocytes differentiation amenable to high-throughput evaluation, such as histochemical labeling (e.g., in-cell Westerns) or promoter-reporter constructs;
  - moving from EB differentiation to monolayer, single lineage differentiation, or the derivation/availability of progenitor cell lines like partially differentiated, neuronal progenitors; and
  - developing ''chips'' containing standard targets (genes and proteins) for monitoring effects, rather like a super array.
- Another is to determine whether guided differentiation toward one germ cell layer or differentiation phenotype is a better predictive model than the undirected differentiation culture protocol used currently. Although one could rationalize the benefits of either approach, some head-to-head comparisons would be very valuable here.
- An exciting area under development is the use of molecular markers to assess multiple differentiation phenotypes in mESC after undirected differentiation. Using a quantitative approach, the relative level of mRNA for specific molecular markers (e.g., α-MHC for cardiomyocytes) can be assessed to determine differentiation to each phenotype and the relative differentiation to multiple phenotypes (e.g., ectoderm compared with mesoderm). Using lineage specific markers may also aid in the prediction of a target-tissue effect of a test agents as a developmental toxicant.
- A third goal is an improved prediction model that is not driven so heavily by cytotoxicity or, alternatively, which can take better account of cytotoxicity to predict which compounds will elicit dose-limiting maternal toxicity in vivo, and will thus never reach embryotoxic exposures. In the hands of at least one of the participants (Don Stedman, from Pfizer), many of the marketed compounds that were tested by that laboratory are reported to have no embryotoxicity in vivo, but are predicted to be at least weak embryo toxicants by the assay. It is unknown if this is inherent to the EST or represents a pharmacotoxicity difference in metabolism and disposition in vivo.
- An important feature to move the test forward would be to develop a metabolizing system capable of creating a realistic in vivo metabolite profile without that metabolizing system itself being toxic to the stem cells. Recent advances in the culture of mESC show that HepG2 cell (human liver cell line that could be used for toxicant bioactivation) conditioned medium induces differentiation of mESCs.
- Another area of work for improvement would be to compare the responses of human stem cells versus murine stem cells, and explore mechanisms of conflict

resolution when these two disagree. Dr. Rao, a participant in the workshop, reported that when gene expression is compared between mESC and hESC there is only a 50% overlap in expression patterns. It is well known that the signaling pathways required to maintain pluripotent mouse and human ESCs are different, suggesting that there may be other differences in pathways used in differentiation. Thus, a comparison between the effects of xenobiotics in mESC and hESC seems important. This type of comparison may provide for addressing situations where mouse cells predict embryotoxicity and the human cells do not. In that situation, it would be possible to confirm the prediction in mice. The reverse situation will be more difficult, and the field will have to determine how to address discrepant predictions from the two species.

Zebrafish

The assay. Much of the work with zebrafish to date has been carried out in the context of determining the hazard and risk to aquatic organisms, treating the embryos with morpholino antisense oligonucleotide sequences to reduce the expression of specific genes to understand their role in normal development, or creating small-volume models of human diseases. So far, relatively little effort has been focused on creating a predictive model of developmental toxicity. Three different efforts in this area were presented at the break-out session.

Before describing some versions of an assay, it is worth remembering what Robert Tanguay told the audience: zebrafish share most of the signaling systems found in mammalian embryos, but alterations produced by an impact on that signaling system may be species-specific. Thus, although heart development may follow the same pattern in rodents and zebrafish, alterations in the rodent limb field or digits will manifest as the relevant analogue in fish (i.e., fin ray abnormalities). A common underlying mechanism will have a species-appropriate expression that may appear superficially different in the fish (Fig. 3).

Wen Lin Seng from Phylonix, Inc. presented results of a collaborative pilot study with Bristol-Myers Squibb that evaluated a small test set of 12 blinded compounds in a zebrafish assay that compared LD_{50} values against doses that produced dysmorphology. Visual assessment of morphology was abbreviated in the sense that treated larvae were classified on "yes/no" criteria for affected morphology for a limited set of structures (body, heart, liver, and intestine) but not scored on severity or characterized extensively for types of malformations. Compounds were administered over a log-scale dose range from 0.01–1000 μM, pending compound solubility. Highly soluble compounds were additionally evaluated at a top concentration of 2000 μM. Dechorionated embryos were treated with compounds at 24 hr post-fertilization (pf), at which time the embryos were actively undergoing organogenesis. Morphology was assessed at ~120 hr pf (~4 days pf). For LD_{50} assessment, mortality was assessed daily and at the end of the assay; total mortality was used to generate the concentration–response curve using best-fit concentration–response

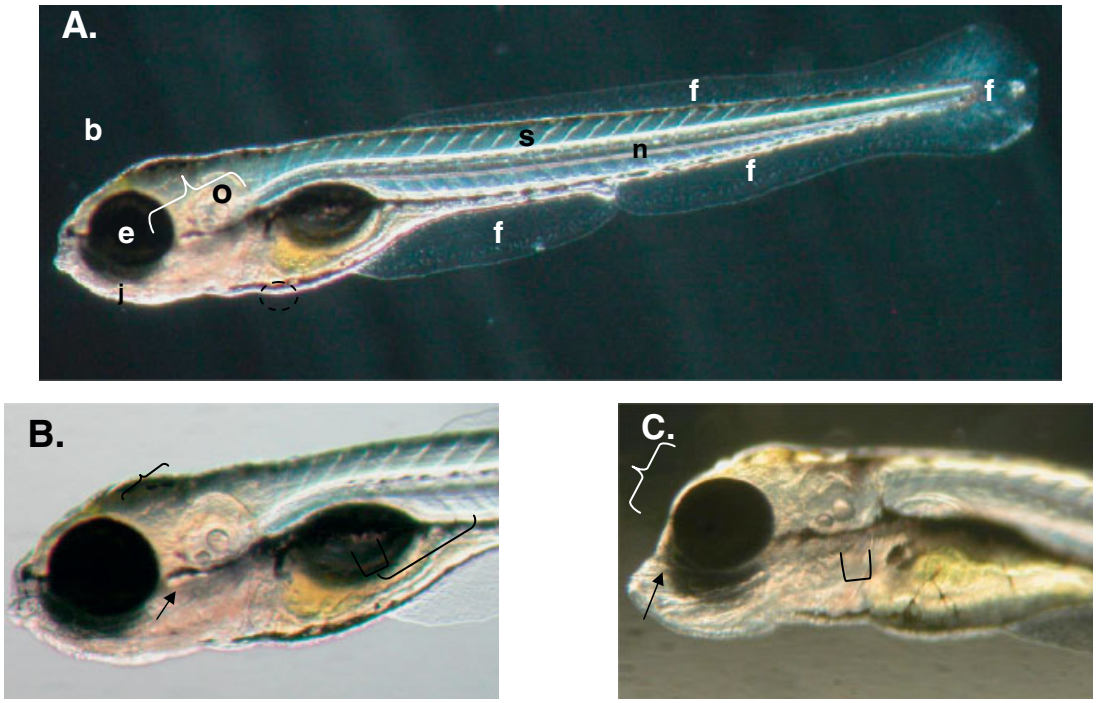

Fig. 3. Normal morphology of a Day 5 pf zebrafish larva. **A:** b, brain; s, somites; n, notochord; j, jaw; o, otic placode; f, fins; dotted circle, heart. **B:** Close-up view of cranial/upper trunk region of a normal Day 5 pf larva. Arrow, jaw; small bracket, forebrain/olfactory region; large bracket, intestine. **C:** Close-up view of cranial/upper trunk region of a Day 5 pf zebrafish larva presenting jaw abnormalities (arrow), reduced forebrain/olfactory region (small bracket) and enlarged intestine (large bracket). (Courtesy Dr. Karen Augustine, Bristol-Myers Squibb, Inc.).

curve calculations. The LD_{50} values generally correlated with in vivo teratogenic potential: those for which one could generate a numeric value had in vivo teratogenic potential and those that were in vivo non-teratogens could not produce a LD_{50} value up to the top dose evaluated. Visual assessment results were not as successful in accurately classifying compounds as in vivo teratogens or non-teratogens. A ratio of LD_{50} concentration/LOAEL of dysmorphology findings was conducted and, based on the ratios of the test set, indices for in vivo teratogenic classification were defined. Out of 12 tested compounds, one compound could not be classified due to the lack of LD_{50} value to ratio against the LOAEL dysmorphology value (ascorbic acid); one compound was incorrectly classified as a non-teratogen, when it was actually a weak teratogen (diphenylhydantoin); and 10 compounds were classified correctly as in vivo non-teratogens or teratogens, with accurate prediction of respective teratogenic potency. There was a 91% success rate in positive classification of in vivo teratogenic and non-teratogenic compounds. The initial study design was heavily weighted with retinoid class compounds (5 of 12 compounds were retinoids) and the study was generally limited in total number of compounds. Additional evaluation of larger/more diverse numbers of compounds would be needed to thoroughly assess the assay's predictivity.

Anita Marguerie presented a review of the in vitro developmental toxicity assay being generated at Danio-Labs Inc. A maximum tolerated concentration was determined by identifying a LD_{50} in Day 5 pf larvae as a surrogate for general adult toxicity. A dose range was then established for the respective compound, typically in half-log concentration increments, with four concentrations selected for the embryotoxicity assay. Dechorionated embryos were used in the assay. Two types of treatment regimens were evaluated. The first regimen included treating embryos at the 4 or 24 hr post-fertilization (hpf) stage and evaluating the embryos 24 hr later (at the 24 and 48 hpf stage). The combined effects on morphologic integrity were then assessed. A subsequent treatment regimen involved treating the embryos with compound at the 2-cell stage and conducting morphologic assessment at 24 and 48 hpf. A pilot study was run for ECVAM and Pfizer assessing 18 compounds representing three classes of in vivo teratogenic potency (non-, weak, or potent teratogens). The results suggested that the second treatment regimen was generally better in correct classification of in vivo outcome, with a 67%, 100%, and 50% success rate at correctly classifying non-, weak, and strong teratogens, respectively. Overall concordance in correct classification was 72%.

Kimberly Brannen (Bristol-Myers Squibb) presented an overview of the assay development and validation process executed by the Bristol-Myers Squibb Reproductive Toxicology group. A test set containing 24 compounds, which were a mix of ECVAM validation compounds and Bristol-Myers Squibb pharmaceutical compounds with characterized in vivo teratogenic potential, were evaluated. Dedechorionated embryos were used in these studies. Two measurements of general toxicity were evaluated in prediction modeling. The first measurement adapted a practice previously used by ECVAM in predictive modeling of the rodent whole embryo culture and mouse embryonic stem cell assays, where the respective compounds were evaluated in a dose range in NIH3T3 fibroblast cells. This cell culture model was used as a surrogate for adult toxicity and an IC_{50} concentration for each respective compound was determined. The second general toxicity measurement involved evaluating the compounds in a concentration range in zebrafish embryos and determining the general toxicity concentration based on a 25% lethal dose concentration (LC_{25}). In the definitive assay, dedechorionated embryos were treated with compound at approximately 4–6 hr pf, and the embryos evaluated for viability at 24 hr pf. At 5 days pf, the larvae were assessed for viability and dysmorphology. A morphologic scoring system was developed that assessed various structures and organs in the Day 5 pf larvae; this included a score for the severity of the dysmorphology. Predictive model classification of each compound involved calculating a ratio of the general toxicity concentration to the no-observed-adverse-effect concentration (NOAEC) based on gross morphology. The preliminary results of this study suggested that neither the use of an IC_{50}-value determined by the NIH3T3 assay nor a larval LC25 concentration enhanced performance of the prediction model. In this ongoing study, the results to date indicate that the cumulative concordance of the prediction model outcome with in vivo teratogenicity data was 92%, with a 94% success rate in positively identifying in vivo teratogens and a 86% success rate in positively identifying in vivo non-teratogens. In addition, there was a 87.5% success rate in positively characterizing in vivo morphologic outcome (either no adverse effect on fetal morphology or positive identification of at least one affected structure/organ system associated with in vivo exposure to the compound).

In summary, the zebrafish assay is still evolving. These predictivity rates are promising, but only after a common protocol is established and the predictive models are in place will this assay begin to be used more widely.

What do zebrafish do well? Zebrafish take genomic modification very well. Morpholino antisense oligonucleotide approaches are well established in the zebrafish as a tool to conduct loss-of-function studies of targeted genes. However, morpholino antisense molecules only induce loss-of-function transiently on the level of inhibiting a targeted gene's translation into protein, so generation of mutant strains is not achievable by this approach. Some advances in transgenic technology in the zebrafish include the use of transgenes with floxed alleles and transposon applications. High genomic incorporation efficiency is achieved with the transposon approach, enabling integration rates as high as 70%. Such technical improvements have contributed to generation of transgenic fish lines that can provide cell-specific reporting models or suitable genomic backgrounds for improved characterization of teratogenic mechanisms. As a teratogenic screening tool, zebrafish offers an entire organism and all stages of development, not a conceptus for a limited part of development or isolated cells in cell culture. This intact model allows all the cells and tissue layers of the conceptus to interact normally, and brings a completeness unavailable in other models. The small size of the embryos also means that compound requirements are minimal, which is of benefit when testing novel

compounds that must be created de novo before testing. Additionally, one can produce an allelic series of hypomorphic embryos, where progressive knock-down of gene expression produces successively more-impacted phenotypes.

Advantages and limitations of zebrafish as a predictive tool. Early indications suggest that this model has the potential to provide predictivity that is at least as good as existing models, and perhaps better. However, our collective experience remains limited, and the ability of any one version of the assay to be translated into another laboratory and still function well is unknown.

What are the shortcomings of zebrafish? The necessity to physically dechorionate each embryo, to allow the tail to straighten, and provide better visual imaging of the embryo and any malformations, adds a layer of manipulation that slows the assay. The chorion is not thought to provide a real barrier to entry for most compounds, but the examination of the embryo's structure is much improved by having the embryo extended and relatively quiescent. One small but ever-present hurdle is the necessity to educate and reassure our colleagues about the relevance of the zebrafish to our more common rodent models, and to get them to trust the predictions. Generally this can be accomplished by sharing the "validation" test data. This will also be aided by the presence of a commonly-accepted and standardized assay and set of endpoints.

What are the unique features of zebrafish that make it advantageous? The zebrafish is an intact organism at a size that is convenient for cell culture. It has been a preferred model for geneticists carrying out manipulations, so there are many tools available for modulating gene expression. The fish are fecund and can deliver hundreds of eggs every morning, so getting large numbers for testing is inexpensive and relatively easy. The developmental trajectory is quite well-defined and widely published, making it easy to learn. The embryo is transparent, allowing visual observations of internal structure over time. The zebrafish larvae can be kept in 100-µl volumes in 96- or 384-well plates for several days. Because the larvae can be kept in 96-wells, only small amounts of compound are needed that are dissolved directly into water or in the presence of DMSO as carrier. Finally, there is a significant electronic resource of images and developmental biology (http://www.fishnet.org.au/FishNet/index.cfm), featuring images and cross-sections of the larvae and embryos at every developmental stage.

What are the next steps to move the assay forward? The use of zebrafish for predicting developmental toxicity is still in its infancy. One of the most meaningful things to move the assay forward would be the publication and sharing of as many predictive models as possible, which will allow users to pick the assay that seems best suited to their needs. If one assay seems to be the consensus choice, the approach it adopts might also spur optimization efforts for the other assays. Key features will be which endpoints are measured, when they are measured, whether a measure of cytotoxicity (e.g., 3T3 IC_{50}) is included in the predictive model, and whether it is useful to know how much compound actually gets into the embryos.

Workshop Summary

This workshop identified several tasks that, if completed, might lead to significant improvements in the conduct and performance of these three assays. There was much support for creating a meta-file of WEC data, and encouraging the exploration of those data as one way to streamline the assay and focus on the most relevant endpoints. The stem cell assay will be probed for ways of incorporating other measures of differentiation and development that might be more inclusive than looking at the appearance of a single cell type (contracting cardiomyocytes). An effort in Europe is aimed at completely revamping the EST; this will likely play out through 2008 or so. Zebrafish will be most helped by the sharing of successful assay methods, so that this new model can be more fully evaluated with many more compounds. Pursuing and completing these tasks will help to advance the field of in vitro predictive models for developmental toxicity.

This workshop convened a group of creative scientists, strongly motivated to improve the outcomes for in vitro assays that predict developmental toxicity. Ultimately, the challenges facing these assays should be solvable.

ACKNOWLEDGMENTS

The authors would like to thank Ms. Regina Graham for her excellent work in the organization and conduct of this workshop.

REFERENCES

Abbott BD, Harris MW, Birnbaum LS. 1989. Etiology of retinoic acid-induced cleft palate ovaries with the embryonic stage exposed. Teratology 40:533–554.

Braun AG, Emerson DJ, Nichinson BB. 1979. Teratogenic drugs inhibit tumor cell attachment to lectin-coated surfaces. Nature 282:507–509.

Brown NA, Fabro S. 1981. Quantitation of rat embryonic development in vitro: A morphological scoring system. Teratology 24:65–78.

Carney EW, Tornesi B, Keller C, Findlay HA, Nowland WS, Marshall VA, Ozolins TRS. 2007. Refinement of a morphological scoring system for postimplantation rabbit conceptuses. Birth Defects Res (Part B) 80:213–222.

Daston GP 1996. The theoretical and empirical case for in vitro developmental toxicity screens, and potential applications. Teratology 53:339–344.

Daston GP, Baines D, Yonker JE 1991. Chick embryo neural retina cell culture as a screen for developmental toxicity. Toxicol Appl Pharmacol 109:352–366.

Doetschman TC, Eistetter H, Katz M, Schmidt W, Kemler R. 1985. The in vitro development of blastocyst-derived embryonic stem cell lines: Formation of visceral yolk sac, blood islands, and myocardium. J Embryol Exp Morph 87:27–45.

Ferrari DA, Gilles PA, Klein NW, Nadler D, Weeks BS, Lammi-Keefe CJ, Hillman RE, Carey SW, Ying YK, Maier D, Olsen P, Wemple DW, Greenstein R, Muechler EK, Miller RK, Mariona FG. 1994. Rat embryo development on human sera is related to numbers of previous spontaneous abortions and nutritional factors. Am J Obstet Gynecol 170:228–236.

Genschow E, Spielmann H, Scholz G, Seiler A, Brown N, Piersma A, Brady M, Clemann N, Huuskonen H, Paillard F, Bremer S, Becker K. 2002. The ECVAM international validation study on in vitro embryotoxicity tests: Results of the definitive phase and evaluation of prediction models. Altern Lab Anim 30:151–176.

Johnson EM, Newman LM, Gabel BEG, Boerner TF, Dansky LA. 1988. An analysis of the Hydra assay's applicability and reliability as a developmental toxicity prescreen. J Am Coll Toxicol 7:111–126.

Kistler A. 1987. Limb bud cell cultures for estimating the teratogenic potential of compounds. Validation of the test system with retinoids. Arch Toxicol 60:403–414.

Klug S, Lewandowski C, Neubert D. 1985. Modification and standardization of the culture of early postimplantation embryos for toxicological studies. Arch Toxicol 58:84–88.

Moscona A. 1961. Rotation-mediated histogenetic aggregation of dissociated cells. Exp Cell Res 22:455–475.

New DA, Mizell M. 1972. Opossum fetuses grown in culture. Science 175:533–536.

Piersma AH, Genschow E, Verhoef A, Spanjersberg Q I, Brown N A, Brady M, Burns A, Clemann N, Seiler A, Spielmann H. 2004. Validation of the postimplantation rat whole-embryo culture test in the international ECVAM Validation Study on three in vitro embryotoxicity tests. Altern Lab Anim 32:275–307.

Pitt JA, Carney EW. 1999. Development of a morphologically-based scoring system for postimplantation New Zealand White rabbit embryos. Teratology 59:88–101.

Pratt RM, Grove RI, Willis WD. 1982. Prescreening of environmental teratogens using cultured mesenchymal cells from human embryonic palate. Teratog Carcinog Mutagen 2:313–318.

Sadler TW, Horton WE, Warner CW. 1982. Whole embryo culture: A screening technique for teratogens? Teratogen Carcinogen Mutagen 2:243–253.

Steele CE, Trasler DG, New DAT. 1983. An in vivo/in vitro evaluation of the teratogenic action of excess vitamin A. Teratology 28: 209–214.

Steele VE, Morrissey RE, Elmore EL, Gurganus-Rocha D, Wilkinson BP, Curren RD, Schmetter BS, Louie AT, Lamb 4th JC, Yang LL 1988. Evaluation of two in vitro assays to screen for potential developmental toxicants. Fundam Appl Toxicol 11:673–684.

Tickle C. 1983. Positional signaling by retinoic acid in the developing chick wing. Prog Clin Biol Res 110:89–98.

Van Maele-Fabry G, Delhaise F, Picard JJ. 1990. Morphogenesis and quantification of the development of post-implantation mouse embryos. Toxicol In Vitro 4:149–156.

Zusman I, Yaffe P, Raz I, Bar-On H, Ornoy A. 1989. Effects of human diabetic serum on the in vitro development of early somite rat embryos. Teratology 39:85–92.