


# Development of a neural rosette formation assay (RoFA) to identify neurodevelopmental toxicants and to characterize their transcriptome disturbances

Nadine Dreser<sup>1</sup>  · Katrin Madjar<sup>2</sup> · Anna-Katharina Holzer<sup>1</sup> · Marion Kapitza<sup>1</sup> · Christopher Scholz<sup>1</sup> · Petra Kranaster<sup>1,7</sup> · Simon Gutbier<sup>1,8</sup> · Stefanie Klima<sup>1</sup> · David Kolb<sup>3,9</sup> · Christian Dietz<sup>3,9</sup> · Timo Trefzer<sup>1,10</sup> · Johannes Meisig<sup>4</sup> · Christoph van Thriel<sup>5</sup> · Margit Henry<sup>6</sup> · Michael R. Berthold<sup>3</sup> · Nils Blüthgen<sup>4</sup> · Agapios Sachinidis<sup>6</sup> · Jörg Rahnenführer<sup>2</sup> · Jan G. Hengstler<sup>5</sup> · Tanja Waldmann<sup>1</sup> · Marcel Leist<sup>1</sup>

## Abstract

The first in vitro tests for developmental toxicity made use of rodent cells. Newer teratology tests, e.g. developed during the ESNATS project, use human cells and measure mechanistic endpoints (such as transcriptome changes). However, the toxicological implications of mechanistic parameters are hard to judge, without functional/morphological endpoints. To address this issue, we developed a new version of the human stem cell-based test STOP-tox<sub>(UKN)</sub>. For this purpose, the capacity of the cells to self-organize to neural rosettes was assessed as functional endpoint: pluripotent stem cells were allowed to differentiate into neuroepithelial cells for 6 days in the presence or absence of toxicants. Then, both transcriptome changes were measured (standard STOP-tox<sub>(UKN)</sub>) and cells were allowed to form rosettes. After optimization of staining methods, an imaging algorithm for rosette quantification was implemented and used for an automated rosette formation assay (RoFA). Neural tube toxicants (like valproic acid), which are known to disturb human development at stages when rosette-forming cells are present, were used as positive controls. Established toxicants led to distinctly different tissue organization and differentiation stages. RoFA outcome and transcript changes largely correlated concerning (1) the concentration-dependence, (2) the time dependence, and (3) the set of positive hits identified amongst 24 potential toxicants. Using such comparative data, a prediction model for the RoFA was developed. The comparative analysis was also used to identify gene dysregulations that are particularly predictive for disturbed rosette formation. This 'RoFA predictor gene set' may be used for a simplified and less costly setup of the STOP-tox<sub>(UKN)</sub> assay.

**Keywords** Neural rosettes · Cytotoxicity · Developmental toxicity · Human stem cells · Differentiation · Neural precursor cells · Phenotypic anchoring · Gene expression

## Abbreviations

AOP	Adverse outcome pathway
BMC25	Benchmark concentration leading to 25% decrease compared to control
BPA	Bisphenol A

BR	Borderline range
CNS	Central nervous system
CsA	Cyclosporin A
DMSO	Dimethylsulfoxide
DNT	Developmental neurotoxicity
FGF	Fibroblast growth factor
HDACi	Histone deacetylase inhibitor
IFNbeta	Interferon beta
KE	Key events
KNDP	Key neurodevelopmental processes
mEST	Mouse embryonic stem cell test
NAM	New approach methods
NCAM	Neural cell adhesion molecule
NDD	Neurodevelopmental distance
NEP	Neuroepithelial precursor

Tanja Waldmann and Marcel Leist shared senior authorship.

✉ Marcel Leist  
Marcel.leist@uni-konstanz.de

Extended author information available on the last page of the article

OECD	Organization for economic co-operation and development
PCMB	P-chloromercuribenzoic acid
PMA	Phorbol 12-myristate 13-acetate
PSC	Pluripotent stem cells
RA	Retinoic acid
REACH	Registration, evaluation, authorisation and restriction of chemicals (EC 1907/2006)
RoFA	Rosette formation assay
SD	Standard deviation
STOP-tox <sub>(UKN)</sub>	Stem cell-based teratogenic omics prediction-UKN toxicity assay (Shinde et al. 2016a), previously named:
UKN1	University of Konstanz (1) assay (Krug et al. 2013b)
T	Threshold
TSA	Trichostatin A
U(T)	Uncertainty of threshold
VPA	Valproic acid
Wnta	Wnt activators

## Introduction

The assessment of compounds that potentially trigger developmental neurotoxicity (DNT) is a major challenge in the field of toxicology. The standard test, according to the OECD test guideline TG426, has only been applied to few chemicals as DNT testing is not usually mandatory (Fritsche et al. 2018b; Smirnova et al. 2014; van Thriel et al. 2012). For instance in the context of REACH, there is only a requirement for DNT data, if there is a neurotoxic alert from standard guideline studies (Terron and Bennekou 2018). However, more and more evidence arises that chemical and other stressors acting during prenatal periods or early childhood are correlated with later neurodevelopmental abnormalities. These include learning deficits, susceptibility to neurodegenerative diseases and to neuropsychiatric conditions, as well as congenital malformations such as neural tube defects (e.g. spina bifida, exencephaly) (Fritsche et al. 2018b; Grandjean and Landrigan 2006, 2014; London et al. 2012; Smirnova et al. 2014).

Current DNT risk assessment is based on animal data only. This strategy makes it hard to measure human-specific health deficits, such as language impairments or effects on complex cognitive and social behavior. Even more importantly, it is associated with problems of species extrapolation.

The use of batteries of human cell-based new approach methods (NAM) has been suggested as an alternative testing approach for different regulatory areas concerned with pesticides, drugs and industrial chemicals (Bal-Price and

Fritsche 2018; Fritsche et al. 2017; Harrill et al. 2018; Terron and Bennekou 2018). The concept behind these NAM is that brain development requires the activation of so-called key neurodevelopmental processes (KNDP) (Aschner et al. 2017; Bal-Price et al. 2015), which can be studied one by one with appropriate in vitro systems. These KDNPs comprise the generation and migration of neural stem/precursor cells, the differentiation of neuroepithelial cells into specific cell types, and neuronal network building (Aschner et al. 2017; Bal-Price et al. 2018a; van Thriel et al. 2012). It is thought that DNT can occur when a compound interferes with at least one of these processes. On this basis, test methods have been established that model such KNDP (Fritsche et al. 2018a; Harrill et al. 2018; Schmidt et al. 2017). They detect toxicant effects on, e.g. cell migration (Nyffeler et al. 2017a, b), neurite growth (Krug et al. 2013a; Radio et al. 2008; Stiegler et al. 2011) or formation of electrical networks (Frank et al. 2017, 2018). An added value of the use of the NAM strategy is that a better mechanistic information about the tested chemicals' actions can be obtained, and the concept of adverse outcome pathways (AOP) (Leist et al. 2017) can be used on the basis of such data (Baker et al. 2018).

STOP-tox<sub>(UKN)</sub> (also referred to as UKN1) is a recently developed (Krug et al. 2013a) NAM that models the very early steps of neurodevelopment, such as neural induction and differentiation. The test method is based on the differentiation of pluripotent stem cells (PSC) into neuroepithelial precursor (NEP) cells that correspond roughly to the cells that build up the neural plate/neural groove (Balmer et al. 2012, 2014). STOP-tox<sub>(UKN)</sub> was developed to follow the KNDP "early neural differentiation", and gene expression analysis was used as the primary endpoint (Balmer et al. 2014; Krug et al. 2013a; Shinde et al. 2015; Waldmann et al. 2014, 2017). Transcriptomics has proven to be a powerful tool for assessing disturbances of differentiation, also on a fully quantitative level (Shinde et al. 2016a; Waldmann et al. 2014, 2017). Extensive characterization has been performed on how to use transcriptomics data in toxicological hit definition (Dreser et al. 2015; Pallocca et al. 2016; Shinde et al. 2016b; Weng et al. 2012). However, chemicals may also trigger transcriptome changes that are not related to altered differentiation patterns (Balmer et al. 2014; Grinberg et al. 2014; Waldmann et al. 2017). To conclude on the toxicological significance of a disturbed gene expression pattern, there is a need for a functional phenotypic anchoring of the gene expression changes to an unambiguously adverse outcome in the culture dish. This can be used to answer the question whether the gene expression changes have a (toxicologically relevant) functional consequence during subsequent development or in later life.

Such a "phenotypic anchor" could be a structural feature or a specific protein expression that can be measured in the

cell culture dish. For instance, neurite outgrowth measures of peripheral (Hoelting et al. 2016) and central neurons (Krug et al. 2013a) have been used for toxicant screening purposes (Delp et al. 2018). Alternatively, quantification of oligodendrocyte surface markers (Baumann et al. 2014, 2016) or synaptic marker proteins (Mundy et al. 2008) has also been used for toxicological approaches to connect gene expression changes to an adverse outcome. Another strategy for a phenotypic anchoring focuses on cellular functions, rather than on static features. A prime example is the mouse embryonic stem cell test (mEST) that measures the beating capacity of cardiomyocytes in vitro (Seiler and Spielmann 2011). Another functional endpoint for phenotypic anchoring is represented by the electrical network properties of mature neurons (Frank et al. 2017, 2018). One functional feature particularly useful for DNT studies could be self-organization of cells to form tissues and organoids (Beccari et al. 2018; Lancaster et al. 2017). The self-organization process relevant for neural tube formation is reflected in cell culture dishes by neural rosette formation (Chambers et al. 2009; Colleoni et al. 2011; Conti and Cattaneo 2010). Neural rosettes are nestin-positive structures formed by neuroepithelial cells in a way that ZO1 and cadherin-N are expressed only at the inside, while proliferation only occurs on the outside. They are characterized by high expression of rosette signature genes, such as *PLAGL1*, *DACH1*, and *PLZF* (Elkabets et al. 2008).

We studied here whether rosette formation could be used as a robust and quantifiable endpoint for the early neurodevelopmental toxicity test STOP-tox<sub>(UKN)</sub>. We set out to develop improved staining and quantification procedures. The main questions then were whether rosette formation correlates with (1) the transcriptome changes for different toxicant concentrations, (2) with incubation times, and (3) with effects of diverse toxicants. Using this information, statistical classifiers were developed to identify DNT compounds. Moreover, transcript changes that were particularly predictive for adverse outcomes were identified.

## Materials and methods

### Materials

Gelatine, putrescine, sodium selenite, progesterone, apotransferrin, glucose, insulin and ascorbic acid were obtained from Sigma (Steinheim, Germany). Accutase was from PAA (Pasching, Austria). FGF-2 (basic fibroblast growth factor), FGF-8b, sonic hedgehog and noggin were obtained from R&D Systems (Minneapolis, MN, USA). Y-27632, SB-43154 and dorsomorphine dihydrochloride were from Tocris Bioscience (Bristol, UK). Matrigel<sup>TM</sup> was from BD Biosciences (Massachusetts, USA). All cell

culture reagents were from Gibco/Invitrogen (Darmstadt, Germany) unless otherwise specified. The compounds used for treatment including the suppliers are listed in the supplement (Fig. S1).

### Neuroepithelial differentiation and rosette formation

The differentiation protocol was followed as described before (Chambers et al. 2009), with minor changes (noggin concentration was only 35 ng/ml, dorsomorphin was added additionally), as operated in Balmer et al. (2012) and Chambers et al. (2011). Single cells (hESC) were seeded on matrigel-coated plates (in a density of 18,000 cells/cm<sup>2</sup>) and cultured until they reached 75% confluency. Cells were supplied with fresh medium every day (previously conditioned for 24 h on mitomycin C-inactivated mouse embryonic fibroblasts; supplemented with 10 ng/ml FGF2 and 10  $\mu$ M ROCK inhibitor Y-27632). On DoD0, differentiation was initiated by adding knockout serum replacement medium (KSR: Knockout DMEM with 15% knockout serum replacement, 2 mM Glutamax, 0.1 mM MEM non-essential amino acids and 50  $\mu$ M beta-mercaptoethanol) containing 35 ng/ml noggin, 600 nM dorsomorphin and 10  $\mu$ M SB-431642. On DoD4, KSR medium was started to be replaced gradually by N2S medium (DMEM/F12 medium, 1% Glutamax, 0.1 mg/ml apotransferrin, 1.55 mg/ml glucose, 25  $\mu$ g/ml insulin, 100  $\mu$ M putrescine, 30 nM selenium and 20 nM progesterone), also supplemented with equal amounts of noggin, dorsomorphin and SB-431642 as used for KSR medium. On DoD11, cells were detached by incubation with accutase for 20 min. Cells were washed with 10 ml DMEM/F12, counted and reseeded in a density of 150,000 cells (cm<sup>2</sup>) in N2S supplemented with 20 ng/ml FGF2, 100 ng/ml FGF8, 20 ng/ml sonic hedgehog, 20  $\mu$ M ascorbic acid and 10  $\mu$ M ROCK inhibitor Y-27632. For rosette formation assay, matrigel coated 96 well plates for staining 24 well plates and for mRNA 12 well plates were used. On DoD13, medium was changed to supplemented N2S without ROCK inhibitor. On DoD15, rosettes were formed and analysed.

### Experimental exposure

Standard treatment with toxicants was either performed in differentiating cells from DoD0-DoD6 or as indicated in the respective figures (Fig. 7; time window treatment). The compound list including the concentrations is shown in the supplemental material (Fig. S1). Cells were either harvested on DoD6 for microarray/RT-qPCR analysis or further differentiated until DoD15. To determine cytotoxicity, a resazurin assay was performed on DoD6, exactly as described previously (Krug et al. 2013b; Stiegler et al. 2011) and the EC10 concentration was determined for the

test compounds. Usually, the EC10 concentration was used for the analysis of mRNA transcript levels at DoD6 and the rosettes formation at DoD15.

## Immunostaining

For immunostaining, cells were either grown on plastic (96 well plates) or on cover slips. At DoD6 or on DoD15, cells were fixed in 4% paraformaldehyde and permeabilized in 0.3% Triton X-100 in PBS. After blocking in PBS (containing 5% bovine serum albumin and 0.1% Triton) for 1 h, primary antibodies were added (antibody concentration is indicated in Fig. S2) and incubated for 1 h at room temperature (RT). After a washing step, secondary antibodies were applied for 45 min at RT. DNA was stained with Hoechst H-33342, and cover slips were mounted in FluorSave<sup>TM</sup> reagent (Calbiochem, Merck).

## Visualizing sialic acids using metabolic glycoengineering

To visualize cellular membrane and cellular arrangement into rosette structures, we took advantage of the metabolic glycoengineering method, where cells metabolize a mannosamine variant containing an azide group and incorporate it into their membrane glycolipids/glycoproteins as sialic acids (Campbell et al. 2007). The azide-modified sialic acids can then be visualized by coupling the azide group to biotin, and its subsequent fluorescent detection by streptavidin-conjugated fluorescent dye (Sletten and Bertozzi 2009; Spate et al. 2014).

Cells were differentiated as described. On DoD11, cells were seeded in 300 µl medium/well in matrigel-coated 8-well glass bottom µ-slides (iBidi, Munich, Germany). On DoD13, medium was changed to N2S medium containing FGF2, FGF8, Shh and ascorbic acid. On DoD14, medium was replaced for N2S containing 70 µM of tetraacetylated mannosamine with an azide group (Ac<sub>4</sub>ManNAz) (Jena, Germany). 24 h later, on DoD15, the sugar precursor was washed off and sialic acids on the cellular membrane were stained by treatment with 100 µM dibenzocyclooctyne (DBCO)-biotin (Jena Bioscience, Germany) for 20 min, followed by 15 min incubation with 8 µg/ml streptavidin-AlexaFluor488 conjugate (Life Technologies, US) and 1 µg/ml Hoechst-33342. Afterwards, cells were fixed with 2% PFA containing 4% sucrose for additional 10 min and washed twice with PBS. Imaging was performed using a laser scanning confocal microscope LSM 880 (Zeiss, Germany) equipped with gallium arsenide phosphide (GaAsP) detector, using a 40x/NA 1.40

oil objective. For image processing, the Fiji program was used (<https://imagej.net/Fiji>).

## Quantification of rosette formation and viability assays

Cells were differentiated to neural rosette forming cells until DoD15 in 96 well plates. Immunostaining of tight junction protein zona occludens 1 (ZO1), Golgi matrix protein 130 (GM130) as well as nuclear staining with Hoechst H-33342 was performed as described above. For each treatment, at least three replicates and for untreated control, six replicates were stained and about 50 images of the whole well were taken by Cellomics automated microscope (Thermo Fisher Scientific) for each channel. Rosettes per well and the nuclei area per well were determined by Konstanz Information Miner software (KNIME) (Berthold et al. 2007). Data were displayed as rosettes/nuclei area relative to the untreated control. All rosette data shown in the paper were obtained at non-cytotoxic concentration [no change (of > 10%) in number of cells per well]. Cytotoxicity was also controlled at the pre-rosette stage by measuring the viability of the neuroepithelial precursor cells on DoD6 by a resazurin assay (Balmer et al. 2012, 2014; Krug et al. 2013b; Rempel et al. 2015). The cytotoxicity was determined by benchmark concentration fitting (Delp et al. 2018; Krebs et al. 2018; Krug et al. 2013a) and the BMC10 (concentration leading to a viability drop of 10% on a modelled curve) was used as the highest non-cytotoxic concentration.

## KNIME analysis

To remove noise and enhance the signal, background pre-processing was applied to all images. The total nuclear area, which is used for normalization, was determined: Hoechst H-33342 staining was used to create a segmentation of the image using the Otsu Global Thresholding Algorithm (Otsu 1979). The number of segmented pixels was counted. Since tight junctions cluster in the rosette center as a ring-like structure, ZO1 staining was used to locate the rosettes in a first step. Images that did not contain rosettes were filtered out. ZO1 spots were detected by applying Otsu thresholding and segmentation of the image. The software discriminated between rosettes and other structures (e.g. epithel, neurons) using the geometry of detected ZO1 segments as a filter. Since the cells and the organelles of a rosette orientate towards the center, the “halo like” shape of a Golgi staining (GM130) was used to identify rosettes: Voronoi Segmentation was applied to the GM130 channel using the previously discovered (ZO1) locations as seed points. This led to a segmentation of positive rosettes. The result was improved by applying a machine learning algorithm which allowed to filter out wrong segments resulting from noise or artefacts.



The classification was done using a random forest (Breiman 2001), which was learned on a training set by manual annotation. The features of the training set comprise measures of segment geometry. After classification, the positive segments were used to calculate the results. The number of rosettes (segments) per image was counted. For each well, the number of rosettes/image was summed up and normalized to the total nuclear area. Data are always displayed relative to an untreated control.

### Affymetrix DNA microarray analysis and quantitative PCR

After the measurement of viability (resazurin assay), medium was removed and cells were lysed in TriFast™ (Peqlab, Germany). mRNA was isolated as described in the manufacturers protocol and reverse transcribed (iScript, Biorad). Quantitative PCR (qPCR) was performed using EVAGreen SsoFast™ mix on a BioRad Light Cycler (Biorad, Germany). For quantification, qPCR threshold cycles were normalized to reference genes [tata box binding protein (*TBP*) and ribosomal protein L13 (*RPL13A*)]. If not stated otherwise, the data of cells treated with compounds were then expressed relative to transcript levels of untreated control cells, which have been grown and differentiated for the same amount of time. For this normalization, the  $2^{-(\Delta\Delta C(T))}$  method was used (Livak and Schmittgen 2001). The primer sequences are listed in the supplement (Fig. S3). Affymetrix chip-based DNA microarray analysis (Human Genome U133 plus 2.0 arrays) was performed exactly as described earlier (Krug et al. 2013b; Rempel et al. 2015). The original data sets of HDAC inhibitors and mercurials were obtained in the context of an earlier study (GEO accession number GSE71127). These data and different forms of analyses have been published in Balmer et al. (2014) and Rempel et al. (2015). The data sets for the compounds BPA, estradiol, galnon, gleevec, cyproconazole, CsA/FK506, geldanamycin, IFN $\beta$ , LiCl, RA, BIO and CHIR have not been published before. The corresponding raw CEL files of the Affymetrix chips for these compounds will be publicly available in the Gene Expression Omnibus (GEO) database (accessible under this publication's title).

### Preprocessing of gene expression data

Three datasets (Balmer et al. 2014; Krug et al. 2013b; Rempel et al. 2015; Shinde et al. 2016a; Waldmann et al. 2014, 2017) with RMA normalized gene expression (Harbron et al. 2007) data ( $\log_2$  scale) measured on the Affymetrix HG-U133 Plus 2.0 array were combined to one gene expression dataset including 54,675 different probe sets. For each biological replicate, the control value was subtracted

from the corresponding treatment values (when multiple technical controls were available for one biological control replicate, the expression values of the technical controls were averaged to one mean control value per probe set). For the comparison with rosette data, one mean expression value was calculated for each compound and each probe set.

Correlation analysis was performed using Pearson's correlation coefficient and the corresponding correlation test for the null hypothesis of zero correlation. The resulting *p* values were multiplicity adjusted to control for the false discovery rate (FDR) by the Benjamini–Hochberg (BH) procedure (Benjamini and Hochberg 1995).

### Preprocessing of rosette formation data for the prediction models

For each biological replicate, one control value was defined (in case of multiple technical controls per biological control replicate, the rosette formation values of the technical controls were averaged to one mean control value). Next, for each biological replicate separately, the rosette formation values of the treatments were divided by the corresponding (mean) control value. Finally, for the comparison with gene expression data, one mean rosette formation value was calculated for each compound. Mean rosette inhibition values were defined by  $1 - \text{mean rosette formation values}$  (resulting negative values set to zero, so that the rosette inhibition ranges from 0 to 1) and used as response in the prediction models.

### Prediction models of rosette inhibition and associated gene sets

For 24 compounds (belinostat, BIO, BPA, CHIR, CsA/FK506, DMSO, entinostat, estradiol, galnon, geldanamycin, gleevec, HgBr $_2$ , HgCl $_2$ , IFN $\beta$ , LiCl, MeHg, panobinostat, PCMB, PMA, RA, SAHA, thimerosal, TSA, VPA) rosette formation as well as gene expression values were measured. They were used for correlation analyses and prediction models. Different gene sets were combined with different statistical algorithms to build and validate classification or regression models for the prediction of rosette inhibition based on leave-one-out cross-validation. In the case of regression models, the rosette inhibition was considered as continuous response using the original values. In the case of classification models, the rosette inhibition values were dichotomized at a threshold value of 75% (binary response: “0” = less than 75% inhibition, “1” = at least 75% inhibition). Leave-one-out cross-validation was carried out based on 24 compounds. In each cross-validation step, one compound was used as test set to predict the response and the remaining 23 compounds were used as training set to build the model. This procedure was repeated until each compound was considered once for validation. As a result of the leave-one-out

cross-validation, we obtained a predicted value of rosette inhibition for each of the 24 compounds that we compared with the true rosette inhibition values to evaluate the prediction performance of the respective model. Prediction accuracy was assessed using Pearson's correlation coefficient for continuous response, and the area under the curve (AUC) of the sensitivity–specificity plot (ROC curve) for binary response.

Based on the training set in each cross-validation step, two different gene sets were selected: (1) the top- $x$  probe sets with the largest variance in expression values (“top-variance genes”), and (2) the top- $x$  probe sets with the best univariate (positive or negative) correlation between expression values and response. For binary response, the AUC was considered as correlation measure (“top-AUC genes”) and for continuous response the Pearson's correlation (“top-corr genes”). The number of top- $x$  genes was varied with  $x = 100, 200, 1000$ . In addition, all 54,675 probe sets and the 21 neurodevelopmental distance (NDD) measured genes (Supplement Fig. S7a) (i.e. the 46 different probe sets corresponding to the 21 genes) were used to construct prediction models. In total, we compared eight different gene sets in combination with the following three statistical algorithms: regularized logistic regression with lasso penalty, regularized logistic regression with ridge penalty, and random forest. Random forest with top-200-variance genes and continuous response was chosen as final prediction model, and consequently fitted and evaluated once based on all 24 compounds. For this final model, the top-200-variance genes were determined once based on all 24 compounds (Supplement Fig. S5). The R package ‘ranger’ (version 0.10.1) was used for computing the random forest with the following default parameter settings: The random forest is an ensemble of 500 regression trees and its prediction is an average of the predicted response values returned by each tree. An individual tree is constructed based on bootstrap samples of the observations and by recursive binary splitting of the covariate space. At each node, the best splitting variable (gene) out of a randomly chosen subset of 14 genes and the best split cut point are determined. The estimated variance of the responses is used as splitting rule. A minimum node size of five observations is chosen.

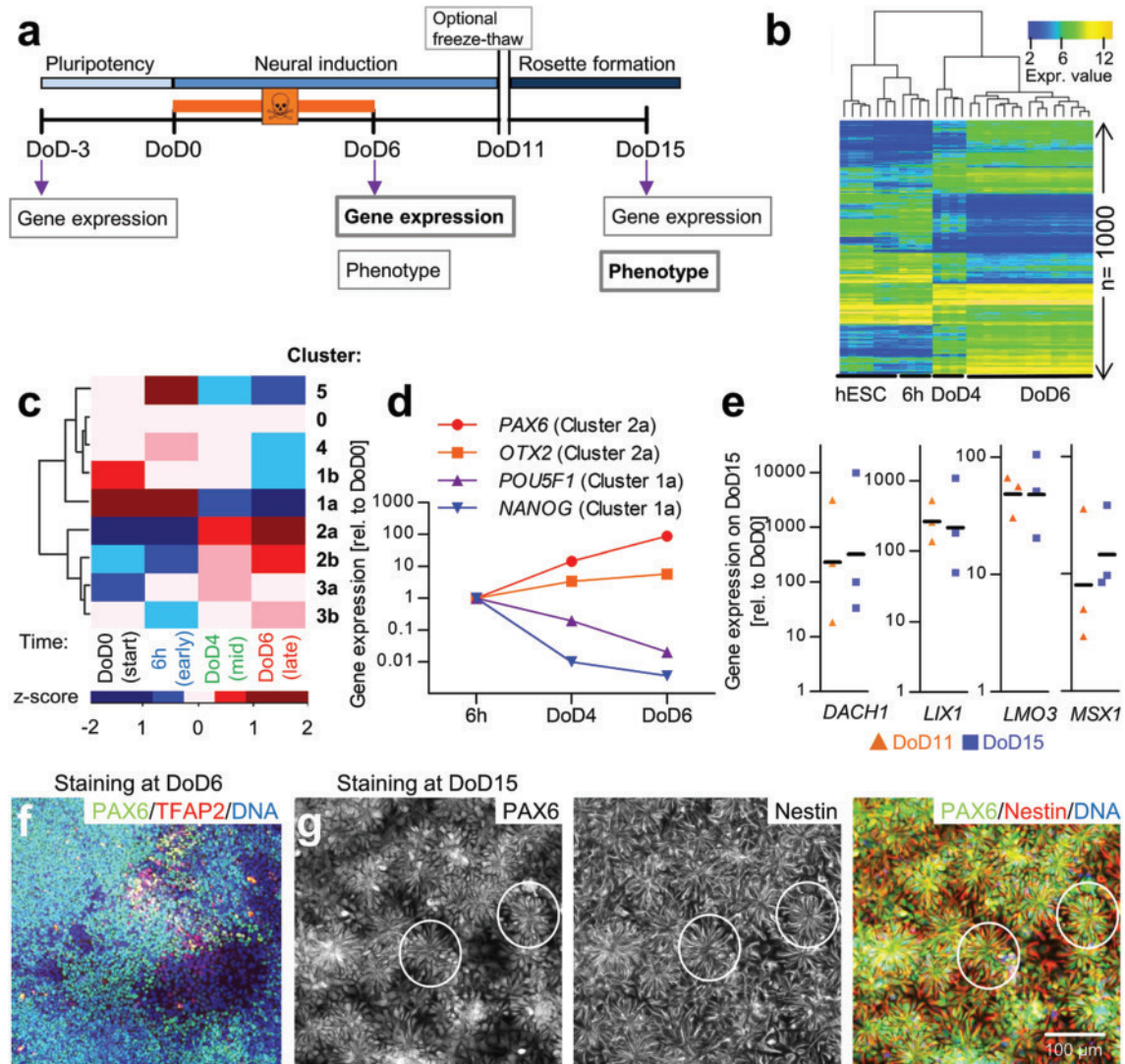
### Calculation of threshold and borderline range for the RoFA prediction model

The threshold  $T$  for the definition of positive, rosette inhibiting compounds was set as  $T = M_n - (2 \times SD_n)$  with  $M_n$ : mean of negatives and  $SD_n$ : standard deviation of negatives (Fig. 8d). The borderline range (BR) was defined as  $BR = T \pm U(T)$  ( $U(T)$ : uncertainty of  $T$ ) and  $U(T) = SD_{\text{pooled}}$  (Fig. 8e) as defined by Leontaridou et al. (2017, 2019).

## Results

### Characterization of NEP and their differentiation towards neural rosettes

Two problems are still preventing the application of the standard STOP-tox<sub>(UKN)</sub> test method for predictive toxicology. First, establishment of a general prediction model based on transcriptome changes has proven difficult; second, the functional relevance of gene expression changes alone is hard to prove. Anchoring genotypic changes induced by a toxicant to a distinct phenotype may provide a solution for these issues. Thus, the test protocol, normally ending with day of differentiation (DoD) 6 (Fig. 1a) (Krug et al. 2013b), was extended by a second phase in which cells were allowed to self-organize to form rosettes (Fig. 1a). Transcriptome data from the standard assay highlight the difficulty with building a prediction model (Fig. 1b): In such a data set, one can observe strong changes in gene expression over time, already in untreated cells. A  $k$ -means clustering ( $K = 9$ ) of the gene expression time courses (Fig. 1c) showed that the expression changes did not follow a single monotonous pattern such as upregulation/downregulation over time (cluster 1 and 2). Some clusters rather showed a wave-like pattern (e.g. cluster 4 and 5) (Fig. 1c). This makes it difficult to define marker genes that are reliable for every time point during differentiation as well as for each toxicant applied to the system. In the past, the genes *PAX6* and *OTX2* (cluster 2a) as well as *POU5F1* and *NANOG* (cluster 1a) have been used as test endpoints (Fig. 1d). Wave-like regulated genes have not been considered at all for the prediction model and it is impossible to judge the relative importance of all regulation clusters for proper development. Therefore, we sought a more integrated bifunctional endpoint. The self-organization capacity of genuine neuroepithelial cells as opposed to other cell types (Chambers et al. 2009) may provide such an endpoint (Colleoni et al. 2011). The self-organization process relevant for neural tube formation is reflected in cell culture dishes by neural rosette formation. To study rosette formation, cell culture was continued; cells were replated and allowed to form rosettes for 4 days (Fig. 1a). In a first step, we analysed the expression of relevant neural rosettes marker genes. *DACH1*, *LIX1*, *LMO3*, and *MSX1* were upregulated on gene expression level on DoD11 and DoD15 (Fig. 1e). A preliminary morphological analysis showed that the cells on DoD6 expressed the neuroepithelial marker *PAX6* (Fig. 1f) and the general neural stem cell marker nestin (data not shown) while they were mostly negative for the neural crest marker *TFAP2*. The expression of nestin and *PAX6* was maintained until DoD15 (with continued



**Fig. 1** Extension of the STOP-tox<sub>(UKN)</sub> protocol to generate rosette stage cells. **a** Exposure and differentiation scheme: Pluripotent stem cells were seeded at a low density three days before the start of differentiation (DoD-3). On the first day of differentiation (DoD0), the neural fate was induced; on DoD6, the standard STOP-tox<sub>(UKN)</sub> protocol endpoint (gene expression) was measured. For rosette formation, cells were further differentiated, with a re-plating step at DoD11. Neural rosettes were assessed at DoD15 for the rosette phenotype (new endpoint) or gene expression. Toxicant exposure occurred always between DoD0 and DoD6. **b** RNA was harvested from cells differentiated for 6 h, 4 days or 6 days, and gene expression was quantified by microarray analysis. The heat map (one lane per microarray) shows absolute expression values (log<sub>2</sub>-scaled) of the 1000 genes with the largest variance, after hierarchical 2D clustering. **c** Waves of gene regulation: The time course profiles of gene expression were clustered into nine groups (*k*-means with *k*=9), and average expres-

sion levels of the genes per groups are indicated for the four time points as row-wise *z* scores. For instance, genes in cluster-5 are first upregulated, then downregulated; genes in cluster-1b start high and are then continuously downregulated. **d** Examples of four signature genes: the gene expression levels are displayed relative to the pluripotent state (DoD0). Data are averages from four independent experiments. **e** The cells were cultured as in A, and mRNA was prepared on DoD11 (brown) and DoD15 (blue) for PCR analysis of four rosette marker genes. Their relative regulation (fold change (FC) vs DoD0) is displayed for three independent experiments; black horizontal lines indicate the average FC. **f** DoD6 cells were stained for PAX6 (neuroepithelial marker) and TFAP2 (neural crest marker). **g** The cell cultures were stained on DoD15 for the neural stem cell markers PAX6 (nuclear localization) and nestin (cytoskeletal protein), and representative images are displayed. Two typical rosettes are shown within white circles (color figure online)

absence of TFAP2 (not shown)). Morphology changes from unorganized clumps on DoD6 to organized rosettes on DoD15 (Fig. 1g).

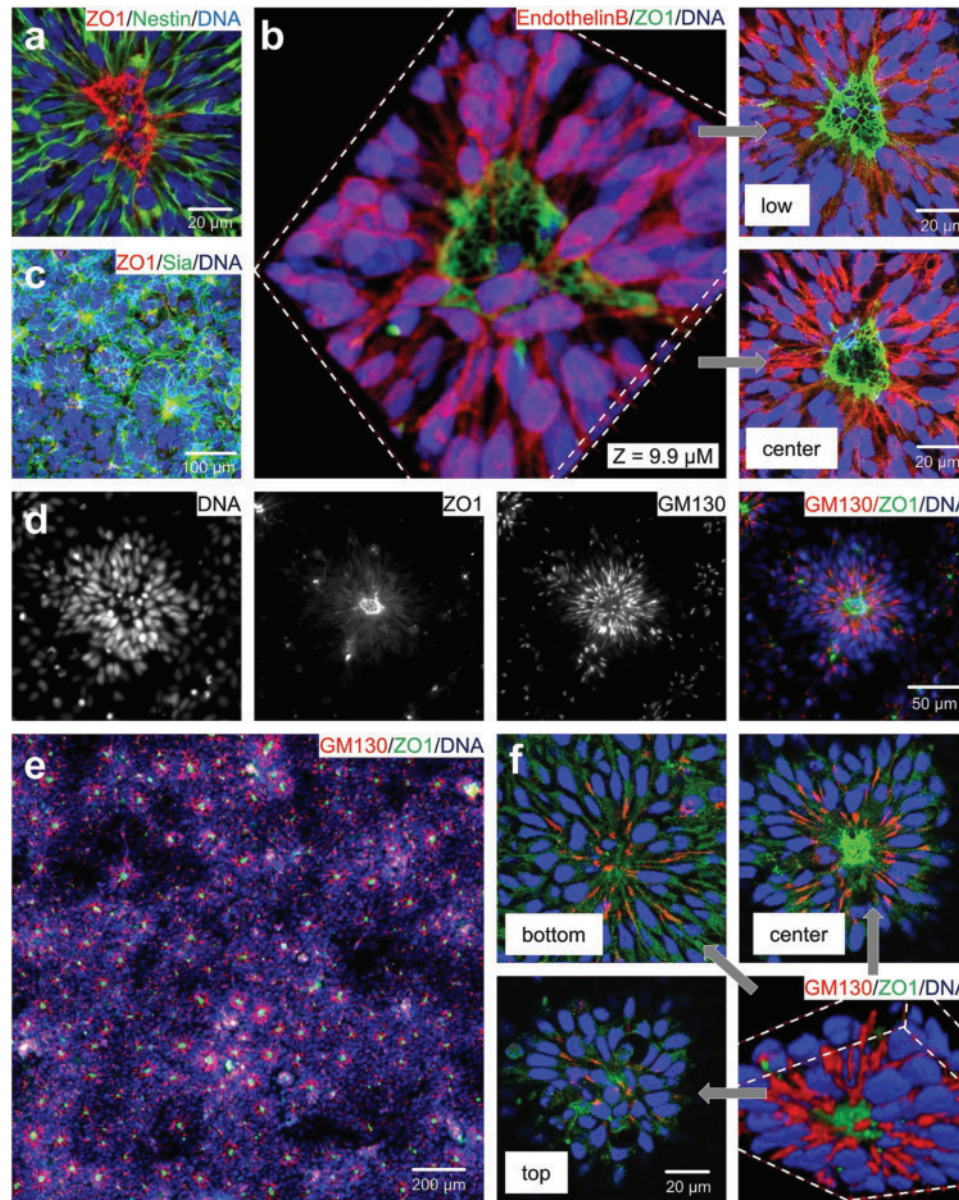
## Characterization of neural rosettes

Neural rosettes have been described as the *in vitro* counterpart to the neural tube, as they look like cross sections of the *in vivo* developing neural tube (Dhara et al. 2008; Dhara



and Stice 2008; Elkabetz et al. 2008; Zhang et al. 2010). To find an appropriate imaging technology to exploit the rosette formation as an assay endpoint, DoD15 neural rosettes were stained in different ways. ZO1 (tight junction marker) staining indicated that rosettes possessed a core structure that

looked like a net of cell membranes possibly with an open (cell-free) space in its center (Fig. 2a, b). Rosettes staining for endothelin beta indicated that this marker is distributed over the whole membrane of the rosette forming cells (Fig. 2b) (Elkabetz et al. 2008). Metabolic glycoengineering



**Fig. 2** Immunocytochemical characterization of rosettes. Cells were differentiated for 15 days according to the scheme in Fig. 1a. Then they were fixed and stained. Nuclei were visualized with the H-33342 dye. **a** Rosette stage cultures were stained for ZO1 (tight junction protein) and nestin (neural stem cell marker). **b** 3D-image of rosettes stained for endothelin B and ZO1. On the right, two exemplary *z* planes of the rosette are depicted with focus close to the cell culture dish (low) or rather in the middle of the rosette height (center). Confocal images were obtained from *z* planes spaced 0.3  $\mu$ m apart. From these data, a 3D-display was obtained. **c** Metabolic glycoengineering was used to label surface sialic acids on rosette cells. At 8 h before

fixation, cells were fed the sialic acid precursor *N*-acetylmannosamine with a tag for immunocytochemical labelling. At 30 min before fixation, surface sialic acids were live-cell labelled for the sugar tag (Sia). Then, cells were fixed and additionally stained for ZO1 and DNA (H-33342). **d–f** Rosette stage cells were stained for ZO1 and GM130 (*cis*-Golgi marker). **e** Individual antigens of rosettes across a cell culture well are shown at a low magnification view. **f** Images were recorded for 20 *z* planes (0.35  $\mu$ m spacing) and rendered for a 3D-display. Three selected *z* planes are shown. Scale bars are included in all images for size calibration

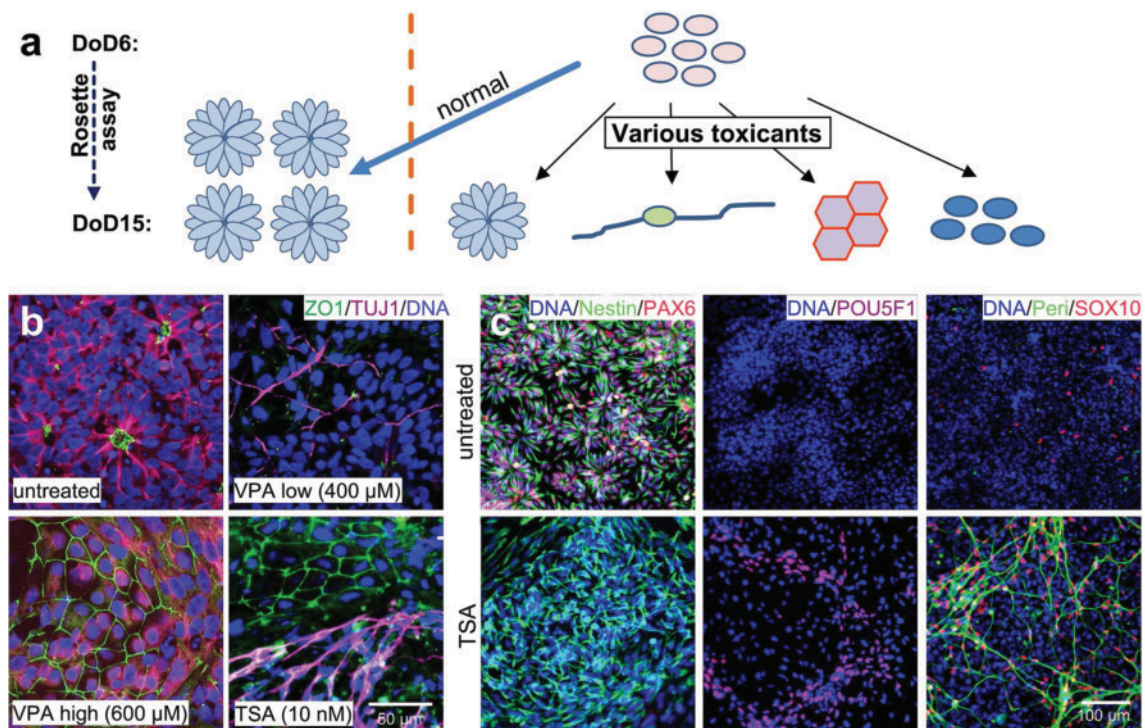


cell feeding with labelled N-acetylmannosamine was used to label surface sialic acids as typically found on poly-sialylated neural cell adhesion molecule (NCAM), a known neural stem cell marker. This revealed a cellular polarization, with the sialic acid residues rather located in the center of the rosettes (Fig. 2c). Also, all organelles showed a clear orientation within rosette cells. The Golgi apparatus was always positioned towards the middle, while cell nuclei pointed towards the outside (Fig. 2d). DoD15 rosettes had a diameter of approximately 50–100  $\mu\text{m}$  and they were equally distributed over the plastic or glass surface of the cell culture dish. This allowed to count the formed rosette structures (Fig. 2e). The Z-dimension (3-dimensionality) of the rosette was very small and they resembled flat domes with the central core appearing to be covered by a single cell layer (Fig. 2b, f).

### Disturbance of rosette formation by toxicant treatment

If rosette differentiation is to be considered as endpoint of a test method, then DNT toxicants should affect this process. We reason that exposure to chemicals may lead to a change

of the neuroepithelial differentiation. This may lead to different outcomes, such as (1) a decreased number of rosettes, (2) an accelerated differentiation towards neurons, (3) a conversion of NEP to any other neural cell type, and (4) or even a switch of differentiation track towards other cell lineages (Fig. 3a). The HDAC inhibitors, trichostatin A (TSA) and valproic acid (VPA), have been used as tool compounds in the STOP-tox<sub>(UKN)</sub> test system, and they led reproducibly to extensive changes in the gene expression of developing cells (Balmer et al. 2014; Krug et al. 2013b; Shinde et al. 2016a; Waldmann et al. 2014). Therefore, we examined the effect of these two compounds on rosette formation. Differentiating cells were treated with 400  $\mu\text{M}$  and 600  $\mu\text{M}$  VPA and 10 nM TSA from DoD0 to DoD6. On DoD6, the compounds were removed and the cells were further differentiated until DoD15 (rosette stage). Low concentrations of VPA (400  $\mu\text{M}$ ) led to a disturbance of rosette formation and the appearance of few neurons (beta III tubulin (TUJ1) positive). Treatment with higher concentrations of VPA (600  $\mu\text{M}$ ) or with TSA (10 nM) completely abolished any rosette formation. The cells formed an epithelial-like structure (Fig. 3b) and patches of neuron-like cells were detected



**Fig. 3** Qualitative features of disturbed rosette formation. **a** Overview of potential toxicant effects during early differentiation: Under normal control conditions, cells were differentiated towards a neural fate and form rosettes on DoD15. Treatment with teratogens during differentiation may lead to an altered phenotype on DoD15: (1) reduced numbers of rosettes, (2) emergence of differentiated neurons or of epithelial-like structures, or (3) generation of other, non-neural cell populations. **b** Cells were differentiated for 15 days according to

the scheme in Fig. 1a. Treatment (400  $\mu\text{M}$  and 600  $\mu\text{M}$  VPA; 10 nM TSA) was performed from DoD0–DoD6. Cultures were fixed and stained for ZO1 and  $\beta$ III-tubulin (TUJ1 antibody, neuronal marker). Nuclei were visualized with the H-33342 dye. **c** Characterization of cells after TSA treatment: Cells were fixed on DoD15 and stained for nestin, PAX6, POU5F1, peripherin (Peri) and SOX10. Nuclei were visualized with the H-33342 dye. Exemplary images of untreated and TSA-treated cells are depicted

(Fig. 3b). Further immunostaining experiments showed that reduced formation of rosettes was associated with a variety of alternative differentiations. For instance, treatment with TSA did not affect the neural stem cell marker nestin, but it abolished the expression of the neuroepithelial marker PAX6 almost completely. It also led to a stable expression of the pluripotency marker POU5F1 in some cells, while normally all rosette stage cells are negative for POU5F1. Furthermore, SOX10 (precursors of oligodendrocytes and neural crest cells) was upregulated and peripherin-positive cells (peripheral neurons) emerged (Fig. 3c). This is consistent with findings that VPA/TSA leads to a change of the differentiation track from NEP to a more neural crest-(like) cell fate in chicken (Murko et al. 2013). Therefore, we hypothesized that one of the developmental disturbances that can be triggered in our test is a switch from CNS neuroepithelial precursors to neural crest cells and their derivatives, such as peripheral neurons (Fig. 3c). In summary, the reduction of rosette formation appeared to be the most universal feature of the developmental toxicants tested; in contrast to this, the generation of alternative cell types was highly dependent on concentration and type of toxicant, and it also appeared to depend more strongly on the time of endpoint assessment. Therefore, we decided to use overall rosette formation as endpoint.

### Quantification of neural rosettes and establishment of the rosette formation assay (RoFA)

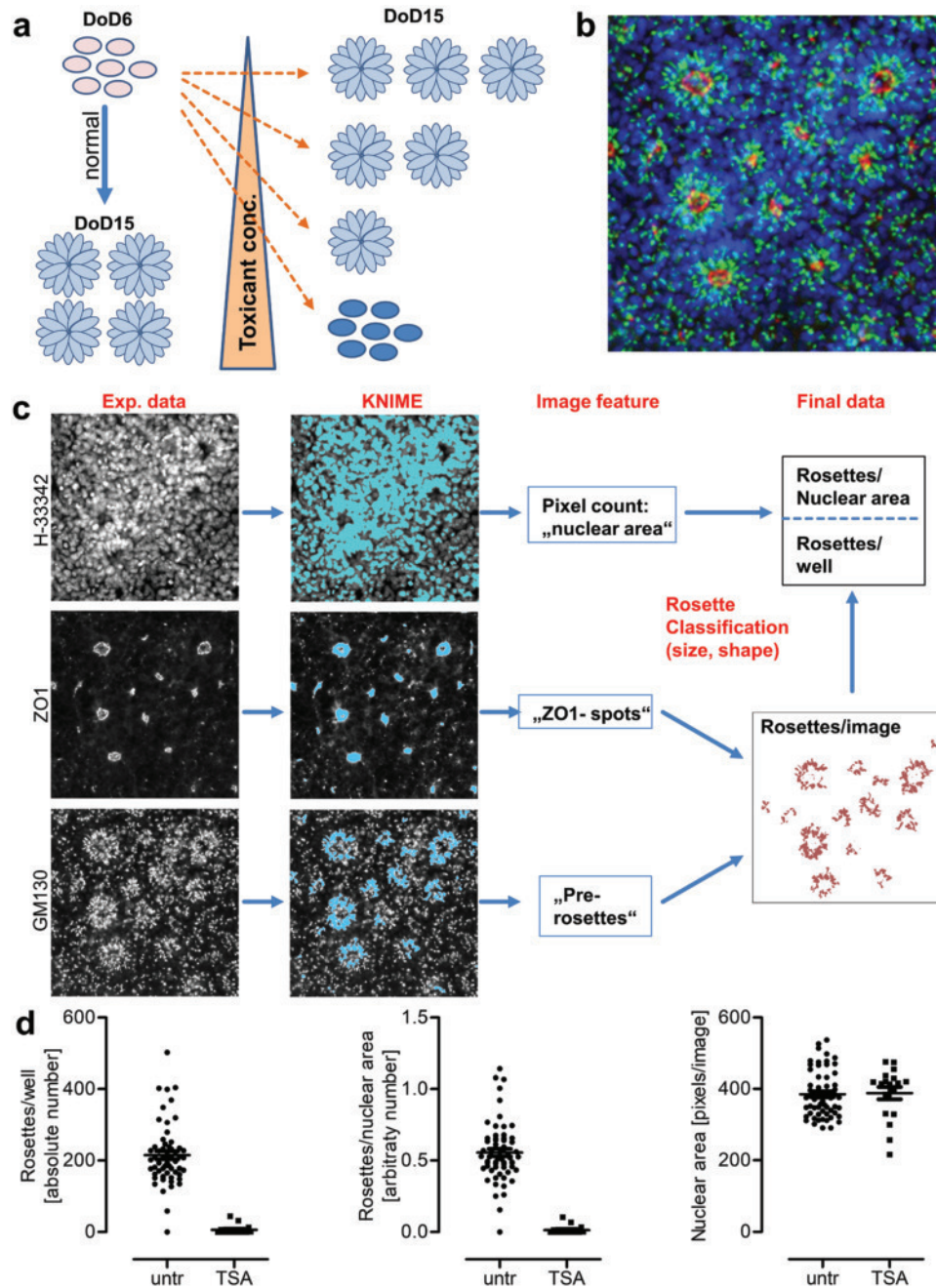
We hypothesized that a neurodevelopmental toxicant inhibits rosette formation in a concentration dependent manner, so that a fully quantitative method to evaluate this endpoint was desirable (Fig. 4a). Therefore, a counting algorithm was developed, based on the production of images from the immunostained rosettes. We decided on labelling the inner tight junction marker ZO1 in combination with staining of the Golgi apparatus by antibodies against the Golgi matrix protein GM130, as this approach yielded the clearest representation of rosettes (Fig. 4b). The program we developed uses the combined information from the ZO1 spot in the middle surrounded by a halo of GM130. This was combined with a filter for size and shape (roundness) to identify rosettes. In the next step, the number of rosettes is normalized to the number of cells in an image field (based on nuclear counts) (Fig. 4c). Testing this algorithm, we found that the number of rosettes that were counted was stable for biological replicates as well as for technical replicates. The tool compound TSA (10 nM) inhibited the rosette formation, but did not affect the cell count (nuclear area) (Fig. 4d).

After the establishment of the rosette formation assay (RoFA) and the development of the quantification software, we started comparing the RoFA performance to the classical gene expression readout of the STOP-tox<sub>(UKN)</sub> assay. As

first approach, we quantified the concentration dependencies for VPA and TSA. For gene expression analysis, cells were harvested on DoD6, and the gene expression changes of marker genes were investigated. *MSX1*, *POU5F1* and *NANOG* were upregulated after TSA/VPA treatment, while *EMX2*, *OTX2*, and *PAX6* were downregulated. Gene expression changes required TSA concentrations  $\geq 2$ –8 nM and VPA concentrations of  $\geq 100$ –300  $\mu$ M (Fig. 5a). Rosette formation was inhibited in a similar concentration range. The BMC25 [benchmark concentration leading to a 25% decrease compared to control (Krebs et al. 2018)] for TSA was  $\approx 4$  nM and for VPA  $\approx 170$   $\mu$ M (Fig. 5b). Rosette formation was inhibited by  $> 50\%$  at 3 nM TSA and 220  $\mu$ M VPA. In summary, this first RoFA evaluation experiment showed a sensitivity of the new endpoint in the same range as the hitherto used gene expression. Moreover, an important advantage became obvious. Easily interpretable, quantitative inhibition data were obtained for the new endpoint. This was not the case for the gene expression data (the genes all behaved slightly different). For gene expression endpoints, a prediction model, taking into account the different weight of multiple expression markers [known to be affected differently by various classes of toxicants (Balmer et al. 2012; Rempel et al. 2015; Waldmann et al. 2014)], would be very difficult and resource intensive to establish. Notably, such an approach has been used for genome-wide expression data (however, without giving weights to individual genes) (Shinde et al. 2016a). Thus, gene expression endpoints and RoFA can be used in a complementary fashion. Where cost plays a major role, the RoFA may be preferable, at least as an initial filter.

### Correlation of rosette formation with sensitive time windows

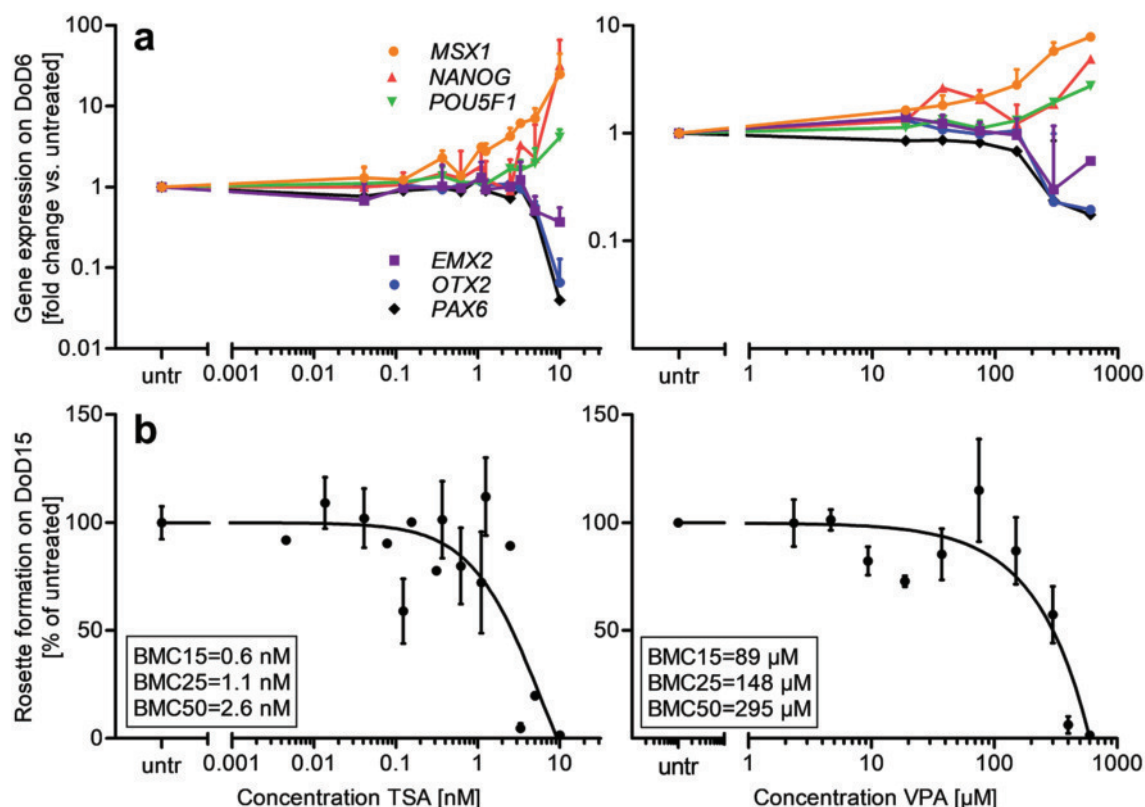
Neuroepithelial development is particularly sensitive to toxicants during certain time periods, and not at all sensitive during others (Balmer et al. 2012, 2014). We used this fact to further investigate the correlation of rosette formation and gene expression changes. Differentiating cells were treated with TSA (10 nM) during different time windows (e.g. DoD2–DoD3), and rosette formation was assessed for all conditions on DoD15. Cells treated with TSA from DoD0 to DoD6 (positive control) showed no rosette formation, while the short treatment during the first three hours of differentiation did not result in any inhibition of rosette formation (negative control). Six hours of treatment was sufficient to inhibit rosette formation, and all the other treatment schedules that exposed cells during the first three days of differentiation also inhibited rosette formation. In contrast, treatment from DoD3–4 did not decrease the number of rosettes significantly and compound exposure that started later than day 3 (DoD3–6, 4–9, 12–15) did not lead to any effect, even



**Fig. 4** Quantification of rosette formation. Neural rosettes were fixed, stained and imaged with an automated microscope. Then, a KNIME protocol was used for automated image recognition and quantification. **a** Working hypothesis for the rosette quantification: It is assumed that treatment of cells from DoD0-6 with neurodevelopmental toxicants leads to a concentration-dependent reduction in rosette number, and it may also lead to the presence of other cell types. **b** Control rosettes were stained as in Fig. 2d for ZO1 (red) and GM130 (green), and nuclei (blue). This exemplary micrograph is used to illustrate the quantification algorithm. **c** Quantification algorithm: The three imaging channels underlying the image in **b** are shown separately. A KNIME protocol is used to count the pixels of the nuclei

and from this the nuclear area is determined. ZO1 spots are recognized by their size and roundness. ZO1 spots that are surrounded by GM130 coronas (=pre-rosettes) are classified as rosettes. These “true rosettes” are counted and normalized to the nuclear area. **d** Cells were treated with TSA (10 nM) or solvent from DoD0-6. They were fixed and stained on DoD15 and used for rosette quantification as in **c**. The graph on the left shows the rosette counts, the graph in the middle shows the rosettes/nuclei ratio and the graph on the right shows the nuclear area. Each dot represents data from one well (technical replicate). Data are from three independent experiments with altogether  $n \geq 25$  technical replicates (color figure online)





**Fig. 5** Concentration-dependent changes in gene expression and rosette formation. Cells were differentiated towards neural rosettes as in Fig. 1a. During differentiation, cells were exposed to VPA or TSA from DoD0-6. **a** On DoD6, mRNA was isolated and gene expression of marker genes was assessed by RT-qPCR analysis. The relative expression of the marker genes *PAX6*, *OTX2*, *NANOG*, *POU5F1*, *EMX2* and *MSX1* (normalized to untreated controls) is shown for TSA- (left) and VPA-treated (right) cells. The cell viability was not

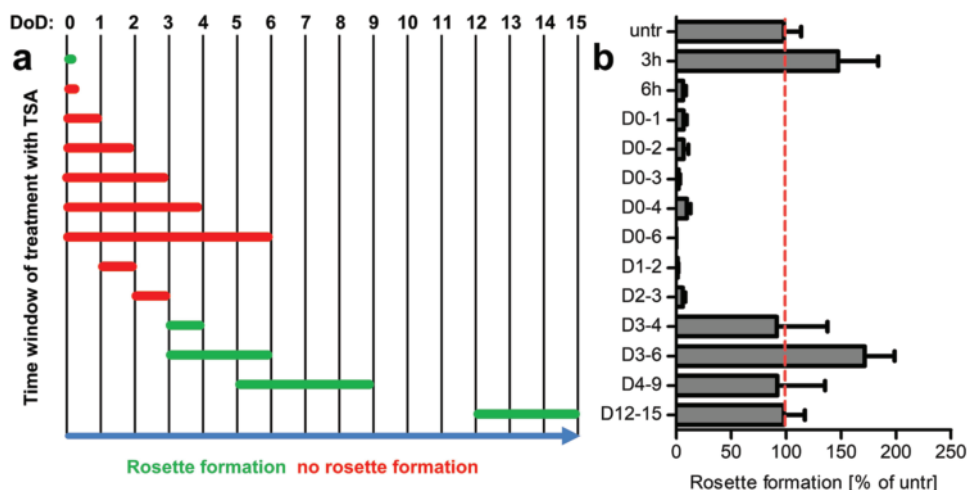
affected at the concentrations used (data not shown). **b** Cells (as in A) were continued to be cultured without toxicants until DoD15. Then, rosette formation was quantified as in Fig. 4b, c. Rosette count data were normalized to untreated controls. For all experiments, the means from 2–6 independent experiments are shown. For the rosette formation assay, eight out of the 15 concentrations were only included in a single experiment. The error bars  $\pm$  indicate the standard error of the means

if cells were exposed for up to 4 days (Fig. 6). Due to the re-seeding step on DoD11, treatment was never performed during this time (to avoid technical artefacts). These data suggest that the window of sensitivity of developing NEPs comprises the first three days of differentiation.

To clarify whether rosette formation on DoD15 was correlated with gene expression on DoD6, samples from the timed exposure to TSA were analysed by RT-qPCR for marker genes. The neuroepithelial genes *PAX6*, *OTX2* and *EMX2* were downregulated by TSA. At least 2 days of drug exposure were required for a ( $\geq$  twofold) reduction in gene expression, and this exposure window had to include the second day of differentiation. The pluripotency transcription factors *POU5F1* and *NANOG* are known to be strongly downregulated during neuroepithelial differentiation (Balmer et al. 2012; Chambers et al. 2009). Here, disturbance by TSA led to higher residual levels of these stem cell markers, and also of the early neural crest specifying gene *MSX1* (Fig. 7a, b). For such effects, a drug exposure of

6 (*MSX1*, *NANOG*) to 24 h (*POU5F1*) only was sufficient, if occurring early during differentiation (first day). Otherwise, the windows of sensitivity for dysregulation of *POU5F1*/*NANOG*/*MSX1* were similar to those for the neuroepithelial markers (Fig. 7a, c).

Comparison of the gene expression and the rosette formation endpoints revealed a good overall agreement. The rosette formation assay was slightly more sensitive, if our prediction model (described in Fig. 7c) for combinations of marker genes was applied. Notably, alternative prediction models could have been used, and they may have shown identical sensitivity. This comparison again demonstrates the practical difficulty of establishing a robust prediction model based on genetic markers. Prediction based on such markers would require a vast amount of test compounds to find out whether different markers should be considered as equally important, whether they should be assigned to weight factors and whether the average, median or, e.g. most sensitive set of changes is to be considered.



**Fig. 6** Dependence of rosette formation on the time window of exposure. Cells were differentiated towards neural rosettes according to Fig. 1a. **a** TSA (10 nM) was added to the cell cultures for different time periods, as depicted by the horizontal bars. At DoD15, cells were fixed, stained and rosettes were quantified. A qualitative overview of the outcome is given by color coding of the bars. Red bars indicate that rosette formation was disturbed; green bars indicate that

no inhibition of rosette formation was observed for the respective exposure condition. **b** Quantitative results for the rosette formation assay after exposure to TSA as specified in A. Rosettes were quantified as in Fig. 4b, c and normalized to untreated controls (untr). The grey bars indicate the means from 2–9 independent experiments ( $n=13$  for the untreated control); the error bars indicate the standard error of the means (color figure online)

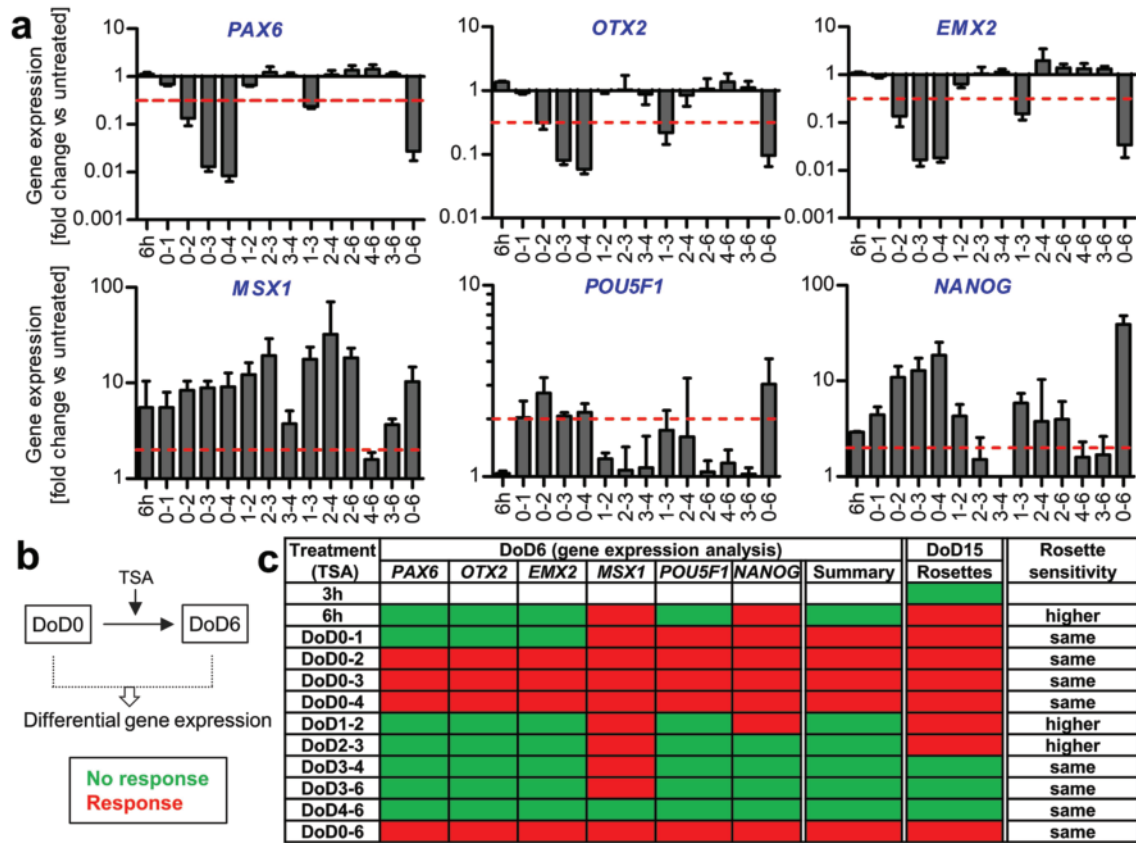
Altogether this set of experiments showed a high concordance of RoFA and gene expression ( $\text{STOP-tox}_{(\text{UKN})}$ ) endpoints (Fig. 7b, c), and we planned for further exploration of the predictivity of inhibited rosette formation.

### Correlation of RoFA and $\text{STOP-tox}_{(\text{UKN})}$ for a test set of toxicants

As our initial experiments convinced us that the RoFA shows high sensitivity towards the tool compounds TSA and VPA, we initiated a broader evaluation of altogether 24 test chemicals (see Fig. S1 for compounds and their concentrations used here). Cells were exposed to these from DoD0–6, and rosette formation was assessed on DoD15. Whole transcriptome data for DoD6 were obtained from the published literature [our legacy data on the same compound concentrations from (Balmer et al. 2014; Krug et al. 2013b; Rempel et al. 2015; Shinde et al. 2016a; Waldmann et al. 2014, 2017)]. Based on these results, we evaluated the reciprocal relationship of RoFA data and transcriptional changes using various correlation and classification strategies. In a first approach, we used an index of toxicity related to gene expression, the developmental potency ( $D(P)$ ) that we had developed and evaluated earlier (Shinde et al. 2016a). This correlated with rosette formation, but only to a moderate extent ( $\rho=0.61$ ) (Fig. S4). Correlations in a similar range were found when other relatively crude measures of transcriptome alterations were used (e.g. number of genes deregulated). Therefore, more complex classification methods were explored.

Several prediction models based on linear correlations and on random forest were tested and various sets of genes (e.g. top100 probe sets, top200, top1000, most variable, etc.) were used as input. Good data were obtained using the top 200 most variant probe sets from the microarray results (Fig. S5). The correlation for RoFA vs transcriptome changes (based on these 200 probe sets) was  $R^2=0.97$ , when a random forest-based prediction model was employed (Fig. 8a). The correlation plot furthermore indicates that predictions of RoFA outcomes are quite reliable, if they are in the range of 0–33% inhibition or 66–100% inhibition. For such compounds, transcriptome data of the 200 probe sets (Fig. S5) appear to be sufficient to predict phenotypic deficits in self-organization of the differentiating neuroepithelial cells, i.e. for a hazard classification of test compounds. If a reduction of rosette formation of  $>33\%$  and  $<66\%$  is predicted, we would suggest to initiate further testing in the RoFA (Fig. 8b).

The setup of initial/preliminary prediction models during assay development, using a limited number of compounds, necessarily leads to overfitting. Ideally, further validation with larger numbers of new reference chemicals would be performed, but this requires large resources and time. Alternatively, an internal validation can be performed to obtain at least some measure of reliability of the prediction model. We used this strategy and chose the leave-one-out cross-validation (LOO) approach: 23 compounds were used to predict the remaining compound, and this process was repeated for each of the compounds. The LOO data highly correlated ( $R^2=0.82$ ) with the full model ( $n=24$ ). Moreover, the LOO



**Fig. 7** Sensitivity comparison of gene expression analysis and rosette formation assay. Cells were differentiated for 15 days according to the scheme depicted in Fig. 1a. Treatment with TSA (10 nM) was performed for different time periods (specified as DoD on the x-axes). At DoD6, RNA was extracted and RT-qPCR was performed for *PAX6*, *OTX2*, *EMX2*, *MSX1*, *POU5F1* and *NANOG*. **a** Relative gene expression (compared to untreated controls) is depicted for the indicated genes. The dotted red line depicts a  $FC \geq 2$ . Regulations beyond this threshold were classified as positive. Data are means from 2 to 5 independent experiments. The error bars represent the standard error of the means. **b** Classification strategy: Different exposure scenarios

were investigated. Gene expressions were classified as toxicant regulated, if the relative expression was altered by  $\geq$  twofold. **c** Comparison of gene expression and rosette formation. Summary data for gene expression changes were produced by the following rule: If  $\geq$  three out of the six marker genes were regulated, gene regulation was considered to be disturbed (red). Rosette stage cells at DoD15 were fixed, stained and analysed with the rosette formation assay (Fig. 4). Disturbance of rosette formation was classified as specified in Fig. 6. No gene expression data were available for exposure of cells only during the first 3 h of the assay (color figure online)

prediction data showed again very clearly that there was particularly high uncertainty about compounds predicted to be 33–66% rosette inhibiting (Fig. S6). This confirmed that such compounds would require experimental testing, while compounds with very high or low prediction values agreed to a high degree (Fig. 8b).

Altogether, our comparison of the transcriptome endpoints and the RoFA suggested that both tests agree well in predicting a potential developmental hazard. On this basis, we undertook some further efforts to set up a binary prediction model for the RoFA, which works without requiring transcriptome data. For this purpose, positive and negative controls needed to be defined. This could theoretically be done using legacy data from a gold standard test (e.g. human clinical data or animal experimental data). However, this approach has weaknesses related to large data uncertainties,

data gaps and problems of species extrapolation. An additional problem is that the definition of positives and negatives refers to specific concentrations reached in vivo. Such data are in many cases not available in standardized form. As alternative approach, we decided to use an internal biological measure, as suggested for strategies of the so-called mechanistic validation (Aschner et al. 2017; Hartung et al. 2013; Leist et al. 2012b, 2014). A list of 21 marker genes (corresponding to 46 probe sets) was selected, based on the literature data indicating the link of these genes to different developmental stages expected in our test, or known to be reached upon disturbance of the system (Fig. S7a). We used mean relative expression levels for these probe sets (ratio of Affymetrix array readouts in the absence or presence of test compound) to calculate a neurodevelopmental distance measure (NDD) (Fig. S7b), i.e. a score defining how far



the system deviated from its normal state in the presence of a test compound. Compounds with a score of  $\geq 1$  were classified as positive hits (= positive controls). Compounds with a score  $< 1$  were regarded as negative controls. This approach resulted in 12 hits (belinostat, entinostat, retinoic acid, PMA, SAHA, CsA/FK506, panobinostat, LiCl, CHIR, VPA, BIO and TSA) and 12 negative compounds (DMSO, BPA, estradiol, HgCl<sub>2</sub>, IFN $\beta$ , MeHg, geldanamycin, thimerosal, gleevec, PCMB, galnon and HgBr<sub>2</sub>) according to their gene expression signature (Fig. 8c).

This set of reference compounds was then used to define a threshold for hit classification in the RoFA. For this purpose, the average for rosette inhibition of negatives was determined (101% rosette formation). We then determined the standard deviation (SD) for all negatives (26.8%) and used the value of  $2 \times \text{SD}$  (53.5%) as proxy for the noise band of negative compounds (Fig. 8d). Thus, rosette formation of  $< 47.6\%$  ( $101.1 - 53.5\%$ ) was considered as definitely disturbed rosette formation in this final prediction model (Fig. 8e). The mean value of positive hit compounds was  $10.6 \pm 17.8\%$  and thus separated generally very well from the negatives.

Based on this hit definition, the RoFA prediction model identified 12 compounds as positives and 12 compounds as negatives (Fig. 8e). All compounds that were positive in the NDD distance measure were also classified as hits in the RoFA. In the consideration of the prediction model, we also propose a further refinement: it has been suggested that binary models may work better if a borderline range (BR) is defined (Leontaridou et al. 2017, 2019). We adopted this suggestion and added the BR around the threshold  $T$  ( $\text{BR} = T \pm 24.5\%$ ) and found that three compounds fall into this range (Fig. 8e). Such borderline compounds may easily be classified one or the other way, due to their unclear test results. For practical purposes, such compounds would be mild alerts (compared to the full hits as serious alerts) and they may be prioritized for further testing in complementary assays (Leontaridou et al. 2017, 2019). Altogether, the NDD and RoFA endpoints correlated very well (correlation of  $R^2 = 0.84$ , if continuous measures were taken), and the hit classification agreed in all cases (Fig. S8).

### Outlook on the definition of better transcriptional markers for the prediction of developmental toxicity and of RoFA outcomes

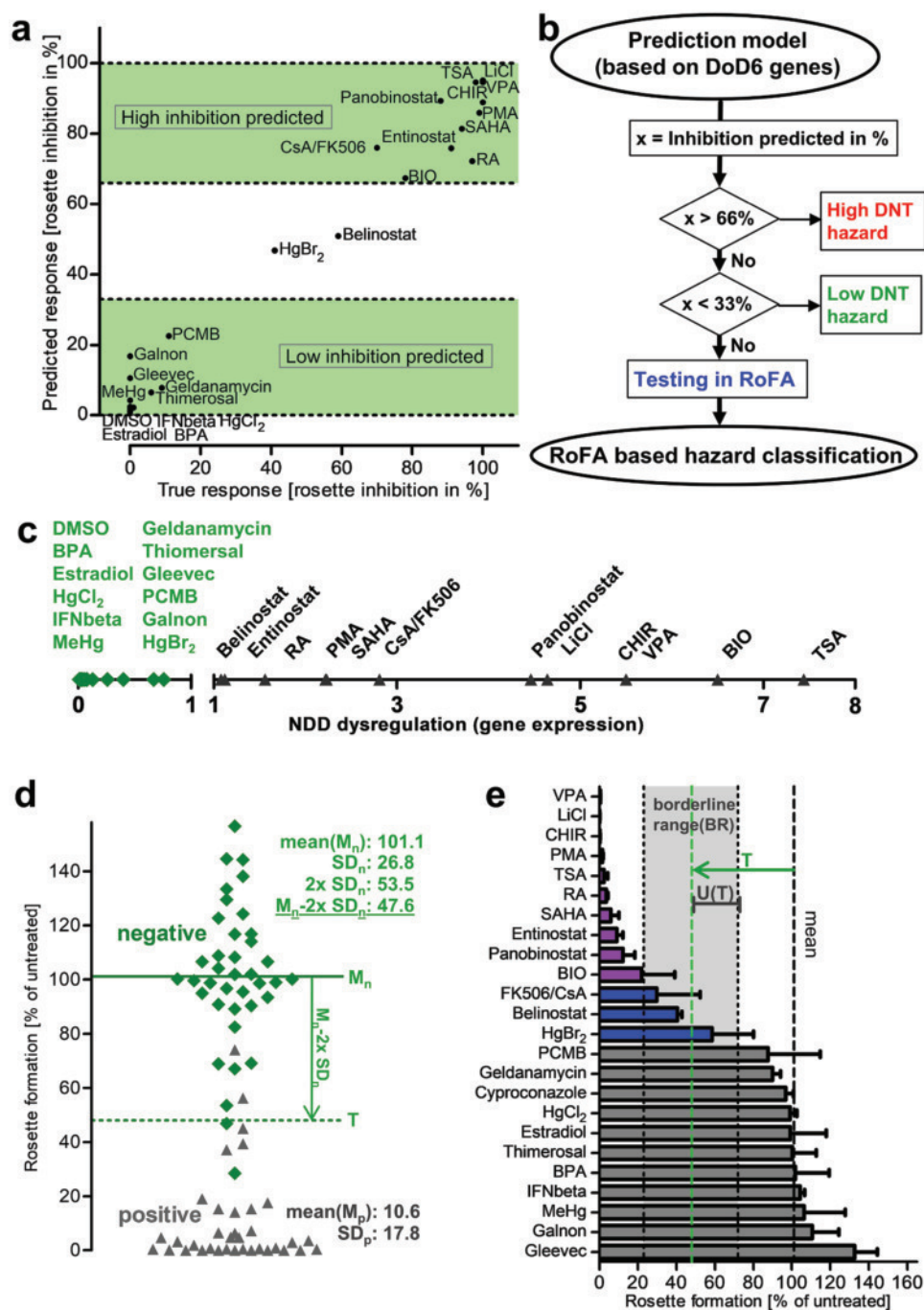
Despite the encouraging data presented above, further improvements of the RoFA prediction model will be required. The next steps should for instance include some form of validation against human hazard data. In case the validity of the RoFA is confirmed as predictive assay for human DNT (or general developmental defects), it may become a gold standard itself. In this function, it would

help to define and calibrate other prediction models. For instance, data from the RoFA could be used to define better transcriptional biomarkers for the traditional STOP-tox<sub>(UKN)</sub> test (Fig. S9). The advantage of this approach would be that one could perform a 6-day assay (STOP-tox<sub>(UKN)</sub>) instead of a 15 day assay (RoFA), but obtain similar predictive capacity. We used this approach here to identify such a set of transcriptional biomarkers. In follow-up work, they may be refined further and be validated against new sets of control compounds. The following approach was taken here.

For all 24 compounds and all probe sets measured with gene symbol annotation available ( $> 40,000$ ), the mean differential (control-adjusted) expression levels were correlated with the mean control-adjusted rosette formation results (using Pearson's correlation coefficient). Genes that were most positively correlated with rosette formation were identified, and they included *CHAMP1*, *CDH2*, *MYLK*, *FGF9* and *CDK5RAP2*. Also genes, correlating negatively with rosette formation were identified and they comprised *DDIT4*, *MEOX1*, *TFAP2A*, *GADD45B* and *DACT1* (Fig. S9a, S10). The performance of a prediction model based on these genes will be tested in the future, when data on new (different) compounds become available.

A second, complementary strategy to identify biomarkers was also followed (Fig. S9b). The starting point for this was that both HDAC inhibitors and Wnt activators were identified as hits in the RoFA, while several mercurials were identified as negatives (not affecting the transcriptome). We used this situation to identify a subpool of genes that were specifically regulated by rosette formation inhibiting compound groups. First, transcripts differing between HDACi (VPA, TSA, panobinostat) vs. three negatives (HgBr<sub>2</sub>, MeHg, thimerosal) were identified (Fig. S11, S12). Then, transcripts differing between Wnt activators (LiCl, CHIR, BIO) vs. negatives (HgBr<sub>2</sub>, MeHg, Thimerosal) were compiled (Fig. S13, S14). Finally, the overlap of the negatively correlated genes was determined. They included *EDNRA*, *DACT1*, *SLIT2*, *BMP5*, *ANXA2*, *SEMA3C* and *NPY* (Fig. S15, S9b). These transcripts were dysregulated by six different compounds (falling into at least two classes of largely differing toxicity mechanisms), and they may thus take a more general role in indicating disturbed differentiation. Similarly, an overlap of positively correlated genes was identified and they included *DUSP4*, *LHX2*, *SIX3*, *CAPN6* and *FZD5* (Fig. S15, S9b).

The two pools of biomarkers will be a valuable starting point to reduce the costs and efforts associated with the STOP-tox<sub>(UKN)</sub>, as they may be determined by a focused approach (e.g. PCR), independent of expensive whole gene transcriptomics approaches. Moreover, further exploration of these markers may give hints on biological mechanisms underlying early neurodevelopmental defects.



## Conclusions and overall outlook

Gene expression profiling was initially chosen as endpoint for the STOP-tox<sub>(UKN)</sub> assay (Krug et al. 2013b) as it was considered a proxy for a comprehensive description of the differentiation pattern and may in addition offer the possibility to derive mechanistic information. During the practical use of the test, it became clear that too few well-characterized control compounds exist (Aschner et al. 2017) to validate the assay by a correlation approach. For

mechanistic validation (Hartung et al. 2013; Leist et al. 2012a), the knowledge on transcriptional dysregulation is too limited. Therefore, a definite prediction model has not yet been established, and the assay has been mainly used to develop procedures for generating classifiers and toxicity indices. These worked very well under defined assay conditions within a narrow chemical space applicability domain (Shinde et al. 2016a). In the present work, we developed a different strategy to further refine the STOP-tox<sub>(UKN)</sub> test. For this purpose, we developed

**Fig. 8** Relationship between toxicant-induced transcriptome changes and rosette formation. Cells were differentiated for 15 days according to the scheme depicted in Fig. 1a. During differentiation, cells were treated with the indicated compounds from DoD0–6. At DoD15, rosette formation was quantified and adjusted for controls. In parallel cultures, RNA was extracted on DoD6, and differential gene expression was measured (compared to untreated controls). **a** A random-forest prediction model was built based on all 24 compounds and used to predict rosette inhibition from DoD6 transcriptome data (for gene list see Fig. S5). The predicted rosette responses for each compound are plotted against the % inhibition of rosette formation actually measured by RoFA. For faster orientation, the areas for which low (< 33%) or high (> 66%) inhibition of rosette formation was predicted, were shaded in green. **b** Based on the findings in a, a test procedure is suggested, in which DoD6 gene expression data are used to initially predict inhibition of rosette formation, and thus to assess developmental neurotoxicity (DNT) hazard. For the data range of 33–66%, at which the prediction model has high uncertainties, actual testing in the RoFA is necessary to determine functional effects of chemicals predictive of DNT. **c** To establish a preliminary prediction model for the experimental RoFA, the gene expression data on DoD6 were used to categorize DNT-negative and DNT-positive chemicals. For this, a panel of 21 hand selected biomarker genes (46 probe sets, for details see Fig. S7A) was employed: The expression changes by the test chemicals of the corresponding 46 probe sets were used to calculate a neurodevelopmental staging defect distance (NDD) measure (i.e. a deviation from normal development, as detailed in Fig. S7b). Twelve test compounds (shown in green) had an NDD score of < 1, and were therefore considered negative, those with a score > 1 were considered positive (black). **d** Cultured cells treated as in A were differentiated until DoD15 (RoFA as in Fig. 4). Each symbol shows a data point of an independent experiment. Data from compounds with an NDD < 1 are shown in green, with an NDD > 1 in grey. Data on the statistical average and spread of the data of positives (black) and negatives (green) are indicated on the right-hand side. A value separating positive and negative compounds was defined based on the variability of the negatives ( $=\text{noise band}$ ); threshold  $T = M_n(2 \times \text{SD}_n)$ . **e** Rosette formation was quantified for each compound as in d. Each bar represents the mean percentage of rosette formation (compared to untreated controls) from two to four independent experiments per compound. Error bars represent the SEM. The data from d were used for a preliminary prediction model of the RoFA, so that compounds that affected rosette formation for more than the noise band ( $T: 2 \times \text{SD}$  of the negatives) were considered positive (purple), while the other compounds were considered negative (grey). Compounds in the borderline range (grey band) are labelled blue;  $U(T)$ : uncertainty of  $T$ . The borderline range was calculated based on the average variation of all values as described by Leontaridou et al. (2017) (color figure online)

a “phenotypic anchor”, i.e. a morphological phenotypic endpoint that could be correlated on the one hand to the gene expression profile in the in vitro test, and that would on the other hand plausibly link to in vivo morphological changes, i.e. teratogenicity. This approach avoids the need for in vivo (human) altered gene expression data for mechanistic or correlative evaluation of assay performance. Notably, such a strategy has been successfully applied in other toxicological areas. For instance, the GARD test for skin sensitization (Johansson et al. 2011, 2013; Li et al. 2019; van Vliet et al. 2018) uses a complex transcriptional profile as readout, while phenotypic data (observations in

animals and man on an allergic response) were used to define appropriate genes and their combination. Another example are the studies of the Piersma group on zebrafish embryo transcriptome changes related to developmental toxicity (Hermesen et al. 2013). These altered transcript patterns were first anchored to morphological malformations. Then, they were used for quantifications or mechanistic conclusion (Tonk et al. 2015; Weigt et al. 2010). Also, assay development based on recursive optimization of mechanistic and phenotypic endpoints is well established in toxicology. For instance in genotoxicology, the primary cell endpoint (various forms of DNA damage and mutation) is hard to directly correlate with genetic damage in vivo. Typical phenotypic endpoints used as bridges are micronucleus formation (Fenech 2000; Fenech and Morley 1985), the formation of bacterial colonies in the Ames test (Leist et al. 1992) or the fluorescence of reporters, e.g. in the GADD45 assay (Walmsley 2008). Also in other fields, complex patterns of tissue stress response activation or of nuclear receptor activation networks have been anchored to simple reporter responses (Legler et al. 1999; Piersma et al. 2013; van der Burg et al. 2013; van der Linden et al. 2014) to provide robust assays with highly quantitative readouts. In developmental toxicology research, such development often started from phenotypic endpoints. E.g. the EST (Seiler and Spielmann 2011) or the whole embryo culture [WEC (New 1978; Piersma et al. 2004; Tonk et al. 2013)] test has mainly relied on their phenotypic endpoint, and a full characterization of associated transcriptome changes is still ongoing (Corvi et al. 2016). With the combination of the new RoFA readout and the established transcriptome patterns, we have taken an important step in the development of a human cell based in vitro assay to predict developmental toxicity. Further research will be performed to put the prediction model on a broader basis. In this context, it is also important to note that different cell lines may perform differently. We observed that various hESC and iPSC lines required extensive adaptation for the standard STOP-tox<sub>(UKN)</sub> assay and some cells were not usable at all. The same may be true for the RoFA. Our own experience showed that various iPSC lines did form rosettes, and e.g. the Sigma iPSC line (IPSC0028) showed assay performances similar to the H9 hESC line used for the assay development. This may be an important consideration for setup of the RoFA for commercial testing. However, such issues apply to many assays based on cell lines. E.g. the cell transformation assay to predict carcinogenicity only works with few cells, sometimes even only specific clones of such cells.

Another important future development will be the incorporation of the test into a more comprehensive battery, e.g. also including a neurite outgrowth assay (Delp et al. 2018; Stiegler et al. 2011), a precursor cell migration assay



(Barenys et al. 2017; Nyffeler et al. 2017b), assays addressing non-neuronal development (Jagtap et al. 2011; Krug et al. 2013b) and additional tests using model organisms (Bal-Price et al. 2018b; Hunt et al. 2018; Scholz et al. 2008).

**Acknowledgements** This work was supported by the Land BW, the Doerenkamp-Zbinden foundation, the DFG (RTG1331, KoRS-CB) and the Projects from the European Union's Horizon 2020 research and innovation programme EU-ToxRisk (Grant agreement No 681002) and ENDpoiNTs (Grant agreement No 825759). We are grateful to H. Leisner and D. Fischer and the staff of the bioimaging center (BIC) for invaluable experimental support.

## Compliance with ethical standards

**Conflict of interest** The authors declare no conflict of interest.

## References


- Aschner M, Ceccatelli S, Daneshian M, Fritsche E, Hasiwa N, Hartung T, Hogberg HT, Leist M, Li A, Mundi WR et al (2017) Reference compounds for alternative test methods to indicate developmental neurotoxicity (DNT) potential of chemicals: example lists and criteria for their selection and use. *Altex* 34:49–74
- Baker N, Boobis A, Burgoon L, Carney E, Currie R, Fritsche E, Knudsen T, Laffont M, Piersma AH, Poole A et al (2018) Building a developmental toxicity ontology. *Birth Defects Res* 110:502–518
- Bal-Price A, Fritsche E (2018) Editorial: developmental neurotoxicity. *Toxicol Appl Pharmacol* 354:1–2
- Bal-Price A, Crofton KM, Leist M, Allen S, Arand M, Buetler T, Delrue N, FitzGerald RE, Hartung T, Heinonen T et al (2015) International STakeholder NETwork (ISTNET): creating a developmental neurotoxicity (DNT) testing road map for regulatory purposes. *Arch Toxicol* 89:269–287
- Bal-Price A, Hogberg HT, Crofton KM, Daneshian M, FitzGerald RE, Fritsche E, Heinonen T, Hougaard Bennekou S, Klima S, Piersma AH et al (2018a) Recommendation on test readiness criteria for new approach methods in toxicology: exemplified for developmental neurotoxicity. *Altex* 35:306–352
- Bal-Price A, Pistollato F, Sachana M, Bopp SK, Munn S, Worth A (2018b) Strategies to improve the regulatory assessment of developmental neurotoxicity (DNT) using in vitro methods. *Toxicol Appl Pharmacol* 354:7–18
- Balmer NV, Weng MK, Zimmer B, Ivanova VN, Chambers SM, Nikolaeva E, Jagtap S, Sachinidis A, Hescheler J, Waldmann T et al (2012) Epigenetic changes and disturbed neural development in a human embryonic stem cell-based model relating to the fetal valproate syndrome. *Hum Mol Genet* 21:4104–4114
- Balmer NV, Klima S, Rempel E, Ivanova VN, Kolde R, Weng MK, Meganathan K, Henry M, Sachinidis A, Berthold MR et al (2014) From transient transcriptome responses to disturbed neurodevelopment: role of histone acetylation and methylation as epigenetic switch between reversible and irreversible drug effects. *Arch Toxicol* 88:1451–1468
- Barenys M, Gassmann K, Baksmeier C, Heinz S, Reverte I, Schmuck M, Temme T, Bendt F, Zschauer TC, Rockel TD et al (2017) Epigallocatechin gallate (EGCG) inhibits adhesion and migration of neural progenitor cells in vitro. *Arch Toxicol* 91:827–837
- Baumann J, Barenys M, Gassmann K, Fritsche E (2014) Comparative human and rat "neurosphere assay" for developmental neurotoxicity testing. *Curr Protoc Toxicol* 59:11–24
- Baumann J, Gassmann K, Masjosthusmann S, DeBoer D, Bendt F, Giersiefer S, Fritsche E (2016) Comparative human and rat neurospheres reveal species differences in chemical effects on neurodevelopmental key events. *Arch Toxicol* 90:1415–1427
- Beccari L, Moris N, Girgin M, Turner DA, Baillie-Johnson P, Cossy AC, Lutolf MP, Duboule D, Arias AM (2018) Multi-axial self-organization properties of mouse embryonic stem cells into gastruloids. *Nature* 562:272–276
- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc* 57:289–300
- Berthold MR, Cebon N, Dill F, Gabriel TR, Kötter T, Meintl T, Ohl P, Sieb C, Thiel K, Wiswedel B (2007) KNIME: The konstanzt information miner. In: Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R (eds) *Data analysis, machine learning and applications*. Springer, Berlin, pp 319–326
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32
- Campbell CT, Sampathkumar SG, Yarema KJ (2007) Metabolic oligosaccharide engineering: perspectives, applications, and future directions. *Mol Biosyst* 3:187–194
- Chambers SM, Fasano CA, Papapetrou EP, Tomishima M, Sadelain M, Studer L (2009) Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. *Nat Biotechnol* 27:275–280
- Chambers SM, Mica Y, Studer L, Tomishima MJ (2011) Converting human pluripotent stem cells to neural tissue and neurons to model neurodegeneration. *Methods Mol Biol* 793:87–97
- Colleoni S, Galli C, Gaspar JA, Meganathan K, Jagtap S, Hescheler J, Sachinidis A, Lazzari G (2011) Development of a neural teratogenicity test based on human embryonic stem cells: response to retinoic acid exposure. *Toxicol Sci* 124:370–377
- Conti L, Cattaneo E (2010) Neural stem cell systems: physiological players or in vitro entities? *Nat Rev Neurosci* 11:176–187
- Corvi R, Vilardell M, Aubrecht J, Piersma A (2016) Validation of transcriptomics-based in vitro methods. *Adv Exp Med Biol* 856:243–257
- Delp J, Gutbier S, Klima S, Hoelting L, Pinto-Gil K, Hsieh JH, Aichem M, Klein K, Schreiber F, Tice RR et al (2018) A high-throughput approach to identify specific neurotoxicants/developmental toxicants in human neuronal cell function assays. *Altex* 35:235–253
- Dhara SK, Stice SL (2008) Neural differentiation of human embryonic stem cells. *J Cell Biochem* 105:633–640
- Dhara SK, Hasneen K, Machacek DW, Boyd NL, Rao RR, Stice SL (2008) Human neural progenitor cells derived from embryonic stem cells in feeder-free cultures. *Differentiation* 76:454–464
- Dreser N, Zimmer B, Dietz C, Sugis E, Pallocca G, Nyffeler J, Meisig J, Bluthgen N, Berthold MR, Waldmann T et al (2015) Grouping of histone deacetylase inhibitors and other toxicants disturbing neural crest migration by transcriptional profiling. *Neurotoxicology* 50:56–70
- Elkabatz Y, Panagiotakos G, Al Shamy G, Socci ND, Tabar V, Studer L (2008) Human ES cell-derived neural rosettes reveal a functionally distinct early neural stem cell stage. *Genes Dev* 22:152–165
- Fenech M (2000) The in vitro micronucleus technique. *Mutat Res* 455:81–95
- Fenech M, Morley AA (1985) Measurement of micronuclei in lymphocytes. *Mutat Res* 147:29–36
- Frank CL, Brown JP, Wallace K, Mundy WR, Shafer TJ (2017) From the cover: developmental neurotoxicants disrupt activity in cortical networks on microelectrode arrays: results of screening 86 compounds during neural network formation. *Toxicol Sci* 160:121–135
- Frank CL, Brown JP, Wallace K, Wambaugh JF, Shah I, Shafer TJ (2018) Defining toxicological tipping points in neuronal network development. *Toxicol Appl Pharmacol* 354:81–93

- Fritsche E, Crofton KM, Hernandez AF, Hougaard Bennekou S, Leist M, Bal-Price A, Reaves E, Wilks MF, Terron A, Solecki R et al (2017) OECD/EFSA workshop on developmental neurotoxicity (DNT): The use of non-animal test methods for regulatory purposes. *Altox* 34:311–315
- Fritsche E, Barenys M, Klose J, Masjosthusmann S, Nimtz L, Schmuck M, Wuttke S, Tigges J (2018a) Development of the concept for stem cell-based developmental neurotoxicity evaluation. *Toxicol Sci* 165:14–20
- Fritsche E, Grandjean P, Crofton KM, Aschner M, Goldberg A, Heinonen T, Hessel EVS, Hogberg HT, Bennekou SH, Lein PJ et al (2018b) Consensus statement on the need for innovation, transition and implementation of developmental neurotoxicity (DNT) testing for regulatory purposes. *Toxicol Appl Pharmacol* 354:3–6
- Grandjean P, Landrigan PJ (2006) Developmental neurotoxicity of industrial chemicals. *Lancet* 368:2167–2178
- Grandjean P, Landrigan PJ (2014) Neurodevelopmental toxicity: still more questions than answers—authors' response. *Lancet Neurol* 13:648–649
- Grinberg M, Stober RM, Edlund K, Rempel E, Godoy P, Reif R, Widera A, Madjar K, Schmidt-Heck W, Marchan R et al (2014) Toxicogenomics directory of chemically exposed human hepatocytes. *Arch Toxicol* 88:2261–2287
- Harbron C, Chang KM, South MC (2007) RefPlus: an R package extending the RMA algorithm. *Bioinformatics* 23:2493–2494
- Harrill JA, Freudenrich T, Wallace K, Ball K, Shafer TJ, Mundy WR (2018) Testing for developmental neurotoxicity using a battery of in vitro assays for key cellular events in neurodevelopment. *Toxicol Appl Pharmacol* 354:24–39
- Hartung T, Hoffmann S, Stephens M (2013) Mechanistic validation. *Altox* 30:119–130
- Hermesen SA, Pronk TE, van den Brandhof EJ, van der Ven LT, Piersma AH (2013) Transcriptomic analysis in the developing zebrafish embryo after compound exposure: individual gene expression and pathway regulation. *Toxicol Appl Pharmacol* 272:161–171
- Hoeltig L, Klima S, Karreman C, Grinberg M, Meisig J, Henry M, Rotshteyn T, Rahnenfuhrer J, Bluthgen N, Sachinidis A et al (2016) Stem cell-derived immature human dorsal root ganglia neurons to identify peripheral neurotoxicants. *Stem Cells Transl Med* 5:476–487
- Hunt PR, Olejnik N, Bailey KD, Vaught CA, Sprando RL (2018) *C. elegans* development and activity test detects mammalian developmental neurotoxins. *Food Chem Toxicol* 121:583–592
- Jagtap S, Meganathan K, Gaspar J, Wagh V, Winkler J, Hescheler J, Sachinidis A (2011) Cytosine arabinoside induces ectoderm and inhibits mesoderm expression in human embryonic stem cells during multilineage differentiation. *Br J Pharmacol* 162:1743–1756
- Johansson H, Lindstedt M, Albrekt AS, Borrebaeck CA (2011) A genomic biomarker signature can predict skin sensitizers using a cell-based in vitro alternative to animal tests. *BMC Genom* 12:399
- Johansson H, Albrekt AS, Borrebaeck CA, Lindstedt M (2013) The GARD assay for assessment of chemical skin sensitizers. *Toxicol In Vitro* 27:1163–1169
- Krebs A, Nyffeler J, Rahnenfuhrer J, Leist M (2018) Normalization of data for viability and relative cell function curves. *Altox* 35:268–271
- Krug AK, Balmer NV, Matt F, Schonenberger F, Merhof D, Leist M (2013a) Evaluation of a human neurite growth assay as specific screen for developmental neurotoxicants. *Arch Toxicol* 87:2215–2231
- Krug AK, Kolde R, Gaspar JA, Rempel E, Balmer NV, Meganathan K, Vojnits K, Baquie M, Waldmann T, Ensenat-Waser R et al (2013b) Human embryonic stem cell-derived test systems for developmental neurotoxicity: a transcriptomics approach. *Arch Toxicol* 87:123–143
- Lancaster MA, Corsini NS, Wolfinger S, Gustafson EH, Phillips AW, Burkard TR, Otani T, Livesey FJ, Knoblich JA (2017) Guided self-organization and cortical plate formation in human brain organoids. *Nat Biotechnol* 35:659–666
- Legler J, van den Brink CE, Brouwer A, Murk AJ, van der Saag PT, Vethaak AD, van der Burg B (1999) Development of a stably transfected estrogen receptor-mediated luciferase reporter gene assay in the human T47D breast cancer cell line. *Toxicol Sci* 48:55–66
- Leist M, Ayrton AD, Ioannides C (1992) A cytosolic oxygenase activity involved in the bioactivation of 2-aminofluorene. *Toxicology* 71:7–20
- Leist M, Hasiwa N, Daneshian M, Hartung T (2012a) Validation and quality control of replacement alternatives—current status and future challenges. *Toxicol Res* 1:8–22
- Leist M, Lidbury BA, Yang C, Hayden PJ, Kelm JM, Ringeissen S, Detroyer A, Meunier JR, Rathman JF, Jackson GR Jr et al (2012b) Novel technologies and an overall strategy to allow hazard assessment and risk prediction of chemicals, cosmetics, and drugs with animal-free methods. *Altox* 29:373–388
- Leist M, Hasiwa N, Rovida C, Daneshian M, Basketter D, Kimber I, Clewell H, Gocht T, Goldberg A, Busquet F et al (2014) Consensus report on the future of animal-free systemic toxicity testing. *Altox* 31:341–356
- Leist M, Ghallab A, Graepel R, Marchan R, Hassan R, Bennekou SH, Limonciel A, Vinken M, Schildknecht S, Waldmann T et al (2017) Adverse outcome pathways: opportunities, limitations and open questions. *Arch Toxicol* 91:3477–3505
- Leontaridou M, Urbisch D, Kolle SN, Ott K, Mulliner DS, Gabbert S, Landsiedel R (2017) The borderline range of toxicological methods: quantification and implications for evaluating precision. *Altox* 34:525–538
- Leontaridou M, Gabbert S, Landsiedel R (2019) The impact of precision uncertainty on predictive accuracy metrics of non-animal testing methods. *Altox* 36:435–446
- Li H, Bai J, Zhong G, Lin H, He C, Dai R, Du H, Huang L (2019) Improved defined approaches for predicting skin sensitization hazard and potency in humans. *Altox* 36:363–372
- Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2<sup>-</sup>(Delta Delta C(T)) Method. *Methods* 25:402–408
- London L, Beseler C, Bouchard MF, Bellinger DC, Colosio C, Grandjean P, Harari R, Kootbodin T, Kromhout H, Little F et al (2012) Neurobehavioral and neurodevelopmental effects of pesticide exposures. *Neurotoxicology* 33:887–896
- Mundy WR, Robinette B, Radio NM, Freudenrich TM (2008) Protein biomarkers associated with growth and synaptogenesis in a cell culture model of neuronal development. *Toxicology* 249:220–229
- Murko C, Lager S, Steiner M, Seiser C, Schoefer C, Pusch O (2013) Histone deacetylase inhibitor Trichostatin A induces neural tube defects and promotes neural crest specification in the chicken neural tube. *Differentiation* 85:55–66
- New DA (1978) Whole-embryo culture and the study of mammalian embryos during organogenesis. *Biol Rev Camb Philos Soc* 53:81–122
- Nyffeler J, Dolde X, Krebs A, Pinto-Gil K, Pastor M, Behl M, Waldmann T, Leist M (2017a) Combination of multiple neural crest migration assays to identify environmental toxicants from a proof-of-concept chemical library. *Arch Toxicol* 91:3613–3632
- Nyffeler J, Karreman C, Leisner H, Kim YJ, Lee G, Waldmann T, Leist M (2017b) Design of a high-throughput human neural crest cell migration assay to indicate potential developmental toxicants. *Altox* 34:75–94

- Otsu N (1979) A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern* 9:62–66
- Pallocca G, Grinberg M, Henry M, Frickey T, Hengstler JG, Waldmann T, Sachinidis A, Rahnenfuhrer J, Leist M (2016) Identification of transcriptome signatures and biomarkers specific for potential developmental toxicants inhibiting human neural crest cell migration. *Arch Toxicol* 90:159–180
- Piersma AH, Genschow E, Verhoef A, Spanjersberg MQ, Brown NA, 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
- Piersma AH, Bosgra S, van Duursen MB, Hermesen SA, Jonker LR, Kroese ED, van der Linden SC, Man H, Roelofs MJ, Schulpen SH et al (2013) Evaluation of an alternative in vitro test battery for detecting reproductive toxicants. *Reprod Toxicol* 38:53–64
- Radio NM, Breier JM, Shafer TJ, Mundy WR (2008) Assessment of chemical effects on neurite outgrowth in PC12 cells using high content screening. *Toxicol Sci* 105:106–118
- Rempel E, Hoelting L, Waldmann T, Balmer NV, Schildknecht S, Grinberg M, Das Gaspar JA, Shinde V, Stober R, Marchan R et al (2015) A transcriptome-based classifier to identify developmental toxicants by stem cell testing: design, validation and optimization for histone deacetylase inhibitors. *Arch Toxicol* 89:1599–1618
- Schmidt BZ, Lehmann M, Gutbier S, Nembo E, Noel S, Smirnova L, Forsby A, Hescheler J, Avci HX, Hartung T et al (2017) In vitro acute and developmental neurotoxicity screening: an overview of cellular platforms and high-throughput technical possibilities. *Arch Toxicol* 91:1–33
- Scholz S, Fischer S, Gundel U, Kuster E, Luckenbach T, Voelker D (2008) The zebrafish embryo model in environmental risk assessment—applications beyond acute toxicity testing. *Environ Sci Pollut Res Int* 15:394–404
- Seiler AE, Spielmann H (2011) The validated embryonic stem cell test to predict embryotoxicity in vitro. *Nat Protoc* 6:961–978
- Shinde V, Klima S, Sureshkumar PS, Meganathan K, Jagtap S, Rempel E, Rahnenfuhrer J, Hengstler JG, Waldmann T, Hescheler J et al (2015) Human pluripotent stem cell based developmental toxicity assays for chemical safety screening and systems biology data generation. *J Vis Exp* 100:e52333
- Shinde V, Hoelting L, Srinivasan SP, Meisig J, Meganathan K, Jagtap S, Grinberg M, Liebing J, Bluthgen N, Rahnenfuhrer J et al (2016a) Definition of transcriptome-based indices for quantitative characterization of chemically disturbed stem cell development: introduction of the STOP-Tox and STOP-Tox tests. *Arch Toxicol* 91:839–864
- Shinde V, Perumal Srinivasan S, Henry M, Rotshteyn T, Hescheler J, Rahnenfuhrer J, Grinberg M, Meisig J, Bluthgen N, Waldmann T et al (2016b) Comparison of a teratogenic transcriptome-based predictive test based on human embryonic versus inducible pluripotent stem cells. *Stem Cell Res Ther* 7:190
- Sletten EM, Bertozzi CR (2009) Bioorthogonal chemistry: fishing for selectivity in a sea of functionality. *Angew Chem Int Ed Engl* 48:6974–6998
- Smirnova L, Hogberg HT, Leist M, Hartung T (2014) Developmental neurotoxicity—challenges in the 21st century and in vitro opportunities. *Altex* 31:129–156
- Spate AK, Busskamp H, Niederwieser A, Scharf VF, Marx A, Wittmann V (2014) Rapid labeling of metabolically engineered cell-surface glycoconjugates with a carbamate-linked cyclopropene reporter. *Bioconjug Chem* 25:147–154
- Stiegler NV, Krug AK, Matt F, Leist M (2011) Assessment of chemical-induced impairment of human neurite outgrowth by multiparametric live cell imaging in high-density cultures. *Toxicol Sci* 121:73–87
- Terron A, Bennekou SH (2018) Towards a regulatory use of alternative developmental neurotoxicity testing (DNT). *Toxicol Appl Pharmacol* 354:19–23
- Tonk EC, Robinson JF, Verhoef A, Theunissen PT, Pennings JL, Piersma AH (2013) Valproic acid-induced gene expression responses in rat whole embryo culture and comparison across in vitro developmental and non-developmental models. *Reprod Toxicol* 41:57–66
- Tonk EC, Pennings JL, Piersma AH (2015) An adverse outcome pathway framework for neural tube and axial defects mediated by modulation of retinoic acid homeostasis. *Reprod Toxicol* 55:104–113
- van der Burg B, van der Linden S, Man H, Winter R, Jonker L, van Vugt-Lussenburg B, Brouwe A (2013) A panel of quantitative calux<sup>®</sup> reporter gene assays for reliable high-throughput toxicity screening of chemicals and complex mixtures. In *High-throughput screening methods in toxicity testing*, pp 519–532
- van der Linden SC, von Bergh AR, van Vught-Lussenburg BM, Jonker LR, Teunis M, Krul CA, van der Burg B (2014) Development of a panel of high-throughput reporter-gene assays to detect genotoxicity and oxidative stress. *Mutat Res Genet Toxicol Environ Mutagen* 760:23–32
- van Thriel C, Westerink RH, Beste C, Bale AS, Lein PJ, Leist M (2012) Translating neurobehavioural endpoints of developmental neurotoxicity tests into in vitro assays and readouts. *Neurotoxicology* 33:911–924
- van Vliet E, Kuhn J, Goebel C, Martinozzi-Teissier S, Alepee N, Ashikaga T, Blomeke B, Del Bufalo A, Cluzel M, Corsini E et al (2018) State-of-the-art and new options to assess T cell activation by skin sensitizers: cosmetics Europe workshop. *Altex* 35:179–192
- Waldmann T, Rempel E, Balmer NV, König A, Kolde R, Gaspar JA, Henry M, Hescheler J, Sachinidis A, Rahnenfuhrer J et al (2014) Design principles of concentration-dependent transcriptome deviations in drug-exposed differentiating stem cells. *Chem Res Toxicol* 27:408–420
- Waldmann T, Grinberg M, König A, Rempel E, Schildknecht S, Henry M, Holzer AK, Dreser N, Shinde V, Sachinidis A et al (2017) Stem cell transcriptome responses and corresponding biomarkers that indicate the transition from adaptive responses to cytotoxicity. *Chem Res Toxicol* 30:905–922
- Walmsley RM (2008) GADD45a-GFP GreenScreen HC genotoxicity screening assay. *Expert Opin Drug Metab Toxicol* 4:827–835
- Weigt S, Huebler N, Braunbeck T, von Landenberg F, Broschard TH (2010) Zebrafish teratogenicity test with metabolic activation (mDarT): effects of phase I activation of acetaminophen on zebrafish *Danio rerio* embryos. *Toxicology* 275:36–49
- Weng MK, Zimmer B, Polt D, Broeg MP, Ivanova V, Gaspar JA, Sachinidis A, Wullner U, Waldmann T, Leist M (2012) Extensive transcriptional regulation of chromatin modifiers during human neurodevelopment. *PLoS ONE* 7:e36708
- Zhang X, Huang CT, Chen J, Pankratz MT, Xi J, Li J, Yang Y, Lavaute TM, Li XJ, Ayala M et al (2010) Pax6 is a human neuroectoderm cell fate determinant. *Cell Stem Cell* 7:90–100



## Affiliations

**Nadine Dreser<sup>1</sup>  · Katrin Madjar<sup>2</sup> · Anna-Katharina Holzer<sup>1</sup> · Marion Kapitza<sup>1</sup> · Christopher Scholz<sup>1</sup> · Petra Kranaster<sup>1,7</sup> · Simon Gutbier<sup>1,8</sup> · Stefanie Klima<sup>1</sup> · David Kolb<sup>3,9</sup> · Christian Dietz<sup>3,9</sup> · Timo Trefzer<sup>1,10</sup> · Johannes Meisig<sup>4</sup> · Christoph van Thriel<sup>5</sup> · Margit Henry<sup>6</sup> · Michael R. Berthold<sup>3</sup> · Nils Blüthgen<sup>4</sup> · Agapios Sachinidis<sup>6</sup> · Jörg Rahnenführer<sup>2</sup> · Jan G. Hengstler<sup>5</sup> · Tanja Waldmann<sup>1</sup> · Marcel Leist<sup>1</sup>**

<sup>1</sup> In Vitro Toxicology and Biomedicine, Department Inaugurated By the Doerenkamp-Zbinden Chair Foundation, University of Konstanz, Box 657, 78457 Konstanz, Germany

<sup>2</sup> Department of Statistics, TU Dortmund, 44221 Dortmund, Germany

<sup>3</sup> Department of Computer and Information Science, University of Konstanz, 78457 Konstanz, Germany

<sup>4</sup> Institute of Pathology, Charité-Universitätsmedizin, 10117 Berlin, Germany

<sup>5</sup> Leibniz Research Centre for Working Environment and Human Factors (IfADo), Technical University of Dortmund, 44139 Dortmund, Germany

<sup>6</sup> Center of Physiology and Pathophysiology, Institute of Neurophysiology, University of Cologne (UKK), 50931 Cologne, Germany

<sup>7</sup> Konstanz Research School Chemical Biology (KoRS-CB), University of Konstanz, 78457 Konstanz, Germany

<sup>8</sup> Present Address: Roche Pharma Development, Grenzacherstrasse, 4070 Basel, Switzerland

<sup>9</sup> Present Address: KNIME GmbH, 78467 Konstanz, Germany

<sup>10</sup> Present Address: Digital Health Center, Berlin Institute of Health (BIH), Charité-Universitätsmedizin, 10117 Berlin, Germany