OXFORD
SOT | Society of Toxicology
academic.oup.com/toxsci

TOXICOLOGICAL SCIENCES, 176(2), 2020, 423–432

doi: 10.1093/toxsci/kfaa068
Advance Access Publication Date: 18 May 2020
Research article

The GARDpotency Assay for Potency-Associated Subclassification of Chemical Skin Sensitizers—Rationale, Method Development, and Ring Trial Results of Predictive Performance and Reproducibility

Robin Gradin, Angelica Johansson, Andy Forreryd, Emil Aaltonen, Anders Jerre, Olivia Larne, Ulrika Mattson, and Henrik Johansson[1]

SenzaGen AB, 22381 Lund, Sweden

[1]To whom correspondence should be addressed at SenzaGen AB, Medicon Village, 22381 Lund, Sweden. E-mail: henrik.johansson@senzagen.com.

ABSTRACT

Proactive identification and characterization of hazards attributable to chemicals are central aspects of risk assessments. Current legislations and trends in predictive toxicology advocate a transition from *in vivo* methods to nonanimal alternatives. For skin sensitization assessment, several OECD validated alternatives exist for hazard identification, but nonanimal methods capable of accurately characterizing the risks associated with sensitizing potency are still lacking. The GARD (Genomic Allergen Rapid Detection) platform utilizes exposure-induced gene expression profiles of a dendritic-like cell line in combination with machine learning to provide hazard classifications for different immunotoxicity endpoints. Recently, a novel genomic biomarker signature displaying promising potency-associated discrimination between weak and strong skin sensitizers was proposed. Here, we present the adaptation of the defined biomarker signature on a gene expression analysis platform suited for routine acquisition, confirm the validity of the proposed biomarkers, and define the GARDpotency assay for prediction of skin sensitizer potency. The performance of GARDpotency was validated in a blinded ring trial, in accordance with OECD guidance documents. The cumulative accuracy was estimated to 88.0% across 3 laboratories and 9 independent experiments. The within-laboratory reproducibility measures ranged between 62.5% and 88.9%, and the between-laboratory reproducibility was estimated to 61.1%. Currently, no direct or systematic cause for the observed inconsistencies between the laboratories has been identified. Further investigations into the sources of introduced variability will potentially allow for increased reproducibility. In conclusion, the *in vitro* GARDpotency assay constitutes a step forward for development of nonanimal alternatives for hazard characterization of skin sensitizers.

Key words: GARD; GARDpotency; *in vitro*; sensitization; potency; chemical sensitizers.

Adverse health effects brought on by exposure to chemicals is a regular occurrence (Prüss-Üstün *et al.*, 2016), influenced by the high incidence of contact with chemicals in modern society. One of the most typical adverse effects is allergic contact dermatitis, which is caused by repeated exposure to compounds known as skin sensitizers (Kimber *et al.*, 2011). Reducing the rate of such adverse effects, which would decrease economic burden and relieve human pain and suffering, is relying on the ability to accurately assess the risks associated with chemicals. A critical part of risk assessment includes the identification and characterization of the hazards attributable to chemicals (Gilmour *et al.*, 2019). For assessment of skin sensitizers, several *in vitro* assays have gained regulatory acceptance for the purpose of performing hazard identification (OECD, 2018a,b, 2019). However, because each of the assays target specific parts of the known sensitization mechanisms, such as protein reactivity (OECD, 2019), keratinocyte activation (OECD, 2018a), or dendritic cell activation (OECD, 2018b), applying a battery of tests targeting different mechanistic events is often recommended (Casati *et al.*, 2018; Daniel *et al.*, 2018). Furthermore, none of the regulatory accepted alternative assays do, on their own, provides information pertinent for hazard characterization, which is relevant for ranking chemicals according to their relative skin sensitization potency. In contrast, the regulatory accepted *in vivo* method, the murine local lymph node assay (LLNA) (OECD, 2010), can be used for both hazard identification and characterization (Loveless *et al.*, 2010). Indeed, it is the recommended method for determining skin sensitizer potency when other information sources are lacking, as described by the Classification, Labelling, and Packaging (CLP) guidance documents (ECHA, 2017). Therefore, with the continuous development and implementation of new legislations and regulations restricting the use of *in vivo* methods (EC, 2006; EU, 2003), and with the public opinion advocating replacement of animal methods, there is a need for further development of nonanimal alternatives that are also capable of providing information regarding the potency of skin sensitizers.

GARD (Genomic Allergen Rapid Detection) is a methodology platform for assessment of chemical sensitizers. It is based on an *in vitro* dendritic-like cell line, which is exposed to test chemicals of interest. Exposure-induced changes of transcriptional expression profiles are measured using state-of-the-art gene expression technologies. Generated high informational content data allows for machine-learning assisted classification of test chemical-specific hazards, eg, skin (Forreryd *et al.*, 2016; Johansson *et al.*, 2011, 2013) or respiratory (Forreryd *et al.*, 2015; Johansson, in preparation) sensitizing properties.

GARDskin was the first GARD platform application described and is consequently the most advanced in terms of regulatory acceptance. An interlaboratory ring trial was conducted in adherence with OECD guidance documents for acceptance of novel alternative methods (OECD, 2005, 2009), demonstrating that GARDskin is a powerful tool for assessment of chemical skin sensitizers, with a predictive accuracy of 93.8% and high reproducibility between laboratories (Johansson *et al.*, 2019).

Although it has been argued that the GARD platform indeed captures information of relevance for potency subcategorization (Johansson *et al.*, 2017), GARDskin is currently proposed for hazard identification. However, the concept of identifying and utilizing a subset of genomic biomarkers that are specific for sensitizing potency has been explored (Albrekt *et al.*, 2014). It was shown that strong and extreme sensitizers tend to induce increased regulation of a larger set of pathways, compared with moderate and weak sensitizers. Thus, it was hypothesized that there exist genes that are specifically regulated only if the sensitizing potency is sufficiently strong, and if found, such genes could be utilized as predictive tools. This hypothesis was further explored in work by (Zeller *et al.*, 2017), in which it was demonstrated that a complementary biomarker signature was able to subcategorize relatively strong and relatively weak sensitizers into potency categories 1A and 1B, respectively, according to the Globally Harmonized System (GHS)/CLP sensitizing potency classification system.

Based on this research, we here introduce the GARDpotency assay for subclassification of chemical sensitizers according to their relative sensitizing potency. We describe the prediction model rationale and the implementation of such a rationale on a technological platform for standardized gene expression measurements. We also propose a tiered approach in which GARDskin and GARDpotency are combined to provide complete risk assessment according to REACH requirements (EC, 2006). Finally, we provide descriptive parameters such as within- and between-laboratory reproducibility (WLR/BLR) and predictive performances of the GARDpotency assay and of the tiered approach, as obtained from an inter laboratory ring trial, the data of which is currently in review of validating bodies. The feasibilities of the approaches are demonstrated by comparisons with the existing state of the art.

MATERIAL AND METHODS

GARDpotency training dataset. The chemical constituents of the GARDpotency training dataset are listed in Table 1. All chemicals were purchased from Sigma Aldrich (St Louis, Missouri). Individual training dataset samples were created according to GARD cellular protocols. Samples were created in 3 independent cellular exposure experiment, thus generating biological triplicates for model training, including solvent (DMSO, Sigma Aldrich) controls and unstimulated cells.

GARD cellular protocols. All cellular protocols associated with GARDpotency are identical to those of GARDskin, which has been previously described (Forreryd *et al.*, 2016; Johansson *et al.*, 2019). For further details, the GARD SOP is available and attached to this publication as Supplementary material 1, including attachments amending the data analysis section for potency subclassification using GARDpotency.

In short, for generation of total RNA samples for downstream gene expression analysis, here utilized for both the generation of training dataset samples and the subsequent ring trial, cultivated SenzaCells (ATCC Depository PTA-123875), were exposed *in vitro* to (the) test chemical(s) for 24 h. Following dose-response measurements of induced cell toxicity, an appropriate and test chemical-specific input concentration was defined at non- to low-toxic levels. Genetic material (ie, mRNA) was isolated from cells exposed to the appropriate input concentration of (the) test chemical(s) in 3 biological replicates and stored at −80°C.

Gene expression acquisition platform transfer. NanoString nCounter probe pairs targeting endogenous transcripts were designed to match at exon level to the transcripts of the Affymetrix HuGene 1.0 ST arrays measurements of the 52 genes previously identified (Zeller *et al.*, 2017). Each probe pair was translated to a unique 100-bp target sequence, using Affymetrix transcript cluster IDs as identifier for a specific target transcript. Design priorities were given to reach the maximum number of variants associated with the target gene, as specified by the Affymetrix HuGene 1.0 ST array, while also maintaining kinetic parameters of the nCounter system and minimizing cross-reactivity with nontarget transcripts. Successful design of probe pairs was achieved for 51 out of the 52 genes (transcript cluster ID 7896697 was dropped because it could not be mapped to any RefSeq or Ensembl accession numbers). All probe pair design was performed by NanoString Technologies (Seattle, Washington). The final list of genomic biomarkers utilized in GARDpotency is attached as Supplementary material 2.

Table 1. Training Dataset Details and Prediction Model Results Summary

Chemical	CAS No.	CLP	C (μM)	Model 1[a]	Model 2[a]
2,4-Dinitrochlorobenzene	97-00-7	1A	5	1B	1A
2,4-Dinitrofluorobenzene	70-34-8	1A	16.8	1A	1A
2-Aminophenol	95-55-6	1A	100	1A	1A
2-Hydroxyethyl acrylate	818-61-1	1A	128	1A	1A
2-Nitro-1,4-phenylenediamine	5307-14-2	1A	200	1A	1A
3-Methylcatechol	488-17-5	1A	33	1A	1A
4-Methylaminophenol sulfate (metol)	55-55-0	1A	16.8	1B	1A
4-Nitrobenzylbromide	100-11-8	1A	5	1A	1A
Abietic acid	514-10-3	1B	100	1B	1B
Amylcinnamyl alcohol	101-85-9	1B	500	1B	1B
Aniline	62-53-3	1B	500	1B	1B
Anisyl alcohol	105-13-5	1B	500	1B	1B
Benzocaine	94-09-7	1B	500	1B	1B
Benzyl benzoate	120-51-4	1B	100	1B	1B
Bisphenol A-diglycidyl ether	1675-54-3	1A	100	1A	1A
Butyl glycidyl ether	2426-08-6	1B	500	1B	1B
Chloroaniline	106-47-8	1B	500	1B	1B
Cinnamaldehyde	104-55-2	1A	50	1B	1A
Cinnamyl alcohol	104-54-1	1B	500	1B	1B
Citral	5392-40-5	1B	50	1B	1B
Citronellol	106-22-9	1B	500	1B	1B
Diethanolamine	111-42-2	1B	500	1B	1B
Diethyl maleate	141-05-9	1B	200	1A	1B
Diphenylcyclopropenone	886-38-4	1A	10	1A	1A
Ethylenediamine	107-15-3	1B	500	1B	1B
Eugenol	97-53-0	1B	500	1B	1B
Formaldehyde	50-00-0	1A	100	1B	1B
Geraniol	106-24-1	1B	500	1B	1B
Glutaraldehyde	111-30-8	1A	26	1A	1A
Hexylcinnamic aldehyde	101-86-0	1B	50	1B	1B
Hydroquinone	123-31-9	1A	100	1A	1A
Hydroxycitronellal	107-75-5	1B	100	1A	1A
Imidazolidinyl urea	39236-46-9	1B	72.9	1A	1A
Iodopropynyl butylcarbamate	55406-53-6	1A	100	1A	1B
Isoeugenol	97-54-1	1A	500	1B	1B
Isopropyl myristate	110-27-0	1B	500	1B	1B
Lauryl gallate	1166-52-5	1A	5	1A	1A
Lilial	80-54-6	1B	150	1B	1B
Linalool	78-70-6	1B	500	1B	1B
Lyral	31906-04-4	1B	175	1B	1A
Methyl heptine carbonate	111-12-6	1A	100	1A	1A
p-Benzochinone	106-51-4	1A	100	1A	1A
Pentachlorophenol	87-86-5	1B	200	1B	1B
Phenyl benzoate	93-99-2	1B	100	1B	1B
Phenylacetaldehyde	122-78-1	1B	3.3	1A	1A
Potassium dichromate	7778-50-9	1A	75	1A	1A
p-Phenylenediamine	106-50-3	1A	100	1A	1A
Propyl gallate	121-79-9	1A	200	1A	1A
Pyridine	110-86-1	1B	500	1B	1B
Resorcinol	108-46-3	1B	500	1A	1A
Tetramethylthiuram disulfide	137-26-8	1B	0.17	1A	1A
Prediction accuracy (%)				78	82

[a] Model 1; Support Vector Machine (SVM)-based 51 genomic biomarkers.

[b] Model 2; SVM based on 51 genomic biomarkers and Genomic Allergen Rapid Detection input concentration.

Abbreviation: CLP, Classification, Labelling, and Packaging.

Gene expression analysis. All gene expression analysis using the NanoString GEN2 nCounter system, for both generation of the GARDpotency training dataset and subsequent application of the GARDpotency assay in the interlaboratory ring trial, was performed according to protocols provided by the supplier (NanoString Technologies). In short, isolated RNA samples were thawed on ice and quality controlled using the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, California). RNA samples were hybridized to the GARDpotency-specific probe pairs and assayed in the nCounter system, using recommended kits and reagents.

Table 2. Assayed Chemicals During Validation Phase

Chemical	CAS No.	CLP
4-Nitrobenzyl bromide	100-11-8	1A
2-Bromo-2-glutaronitrile	35691-65-7	1A
Cinnamal	104-55-2	1A
Formaldehyde	50-00-0	1A
Lauryl gallate	1166-52-5	1A
4-(Methylamino)phenol sulfate	55-55-0	1A
Methylisothiazolinone	2682-20-4	1A
Propyl gallate	121-79-9	1A
Toluene diamine sulfate	615-50-9	1A
Diethyl maleate	141-05-9	1B
3-Dimethylaminopropylamine	109-55-7	1B
Ethylenediamine	107-15-3	1B
Isoeugenol	97-54-1	1A
2-Mercaptobenzothiazole	149-30-4	1A
Benzyl benzoate	120-51-4	1B
Cinnamyl alcohol	104-54-1	1B
Citral	5392-40-5	1B
Ethylene glycol dimethacrylate	97-90-5	1B
Eugenol	97-53-0	1B
Dextran	9004-54-0	NC
Glycerol	56-81-5	NC
Hexane	110-54-3	NC
Isopropanol	67-63-0	NC
Kanamycin	70560-51-9	NC
Lactic acid	50-21-5	NC
Propylene glycol	57-55-6	NC
Salicylic acid	69-72-7	NC
Vanillin	121-33-5	NC

Abbreviation: CLP, Classification, Labelling, and Packaging.

Training dataset exploration and prediction model optimization and finalization. NanoString raw data were imported into the R statistical programming environment (R Core Team, 2019) and RNA-content normalized using a single-sample counts per total counts algorithm, as previously described (Forreryd *et al.*, 2016). Tentative prediction models were trained using a Support Vector Machine (SVM) (Cortes and Vapnik, 1995), as implemented in the e1071 package (Meyer *et al.*, 2019). A linear kernel was utilized, and the cost parameter was set to 1. The remaining parameters were kept as default as described in Supplementary material 3. The predictive performances of tentative prediction models were evaluated using an iterative 10-fold cross-validation strategy. Two different prediction models were explored, both of which used the GHS/CLP category label of the training dataset (1A/1B) as the dependent (ie, modeled) variable. Model 1 utilized only the 51 genomic biomarkers as independent variables (ie, predictors), whereas Model 2 complemented the genomic biomarkers with a concentration parameter, ie, the GARD input concentration, given as μM. A finalized prediction model for potency subclassification was established based on Model 2, as trained by the entire training dataset, and frozen prior to any application of test data generated by the subsequent interlaboratory ring trial.

Tiered prediction approach. A tiered prediction approach for skin sensitizer potency assessment according to the 3 GHS/CLP categories was defined as follows: in an initial tier, chemicals were classified using the GARDskin assay. Chemicals classified as nonskin sensitizers were assigned the GHS/CLP class label No Cat. Chemicals classified as skin sensitizers were assessed in a second tier, using the herein described GARDpotency assay, for assignment to either of the GHS/CLP class labels 1B or 1A. This combined utilization of GARDskin and GARDpotency is referred to as the GARD tiered approach.

Interlaboratory ring trial. The design of the interlaboratory ring trial was to a great extent based on the structure of a previously described interlaboratory ring study performed with the purpose of validating the GARDskin assay (Johansson *et al.*, 2019), which was conducted in accordance with OECD guidance documents (OECD, 2005). Briefly, SenzaGen AB (SenzaGen, Lund, Sweden) initiated the study and acted as lead laboratory throughout the experiments. A validation management group (VMG), comprising 3Rs management and consulting ApS (Lyngby, Denmark), Triskelion (Zeist, the Netherlands) and Adriens Consulting (Aalter, Belgium), was assembled with the purpose of guiding and facilitating the validation process, and evaluating the results. Furthermore, the VMG was responsible for ensuring the study's blinded nature and was solely responsible for strategic decisions, such as the selection and approval of test chemicals. The study included 2 additional participating laboratories, Eurofins BioPharma Product Munich GmbH (Eurofins, Planegg, Germany) and Burleson Research Technologies, Inc (BRT, Morrisville, North Carolina).

Due to the identical experimental protocols of the GARDskin and the GARDpotency assays (Supplementary material 1), and because both participating laboratories had already shown proficiency in running the GARDskin assay (Johansson *et al.*, 2019), neither transfer nor training phases were performed.

For assessment of GARDpotency's predictive performance and reproducibility, the VMG determined that the 28 chemicals that were assayed for the purpose of validating GARDskin (Johansson *et al.*, 2019) would be used, see Table 2. Only chemicals that were classified as skin sensitizers by the GARDskin assay (by respective laboratory in each experiment) were assessed in the GARDpotency assay, as defined by the GARD tiered approach. The expression of the genes in the GARDpotency signature was quantified using the same RNA samples that were generated in the GARDskin validation study but using a GARDpotency-specific NanoString codeset. Of note, the identity of the 28 chemicals was not revealed to any of the participating laboratories until predictions from both assays had been reported to the VMG, ensuring the blinded nature of the study. Each participating laboratory assessed the chemicals in 3 independent experiments. Chemicals were distributed, uniquely encoded in every experiment, by the VMG.

All chemicals assayed during the validation phase were purchased by Triskelion from Sigma Aldrich Chemie N.V. (Zwijndrecht, the Netherlands).

Statistical analysis. Data submitted to the VMG was independently analyzed and summarized without interference or input from any of the participating laboratories, with the purpose of assessing the reproducibility and the predictive performance of the assay. Prior to the study initiation, it was decided that missing values would be excluded from the analysis. WLR was calculated as the fraction of substances that received consistent predictions over 3 independent experiments within a laboratory. For the calculation of BLR, a consensus prediction was established for each chemical and laboratory by majority voting over the 3 experiments. Substances that failed to generate a majority prediction due to inconsistencies were given the label "inconsistent." BLR was then calculated as the fraction of consistent consensus predictions between the laboratories. The predictive performance was evaluated by accuracy and class-

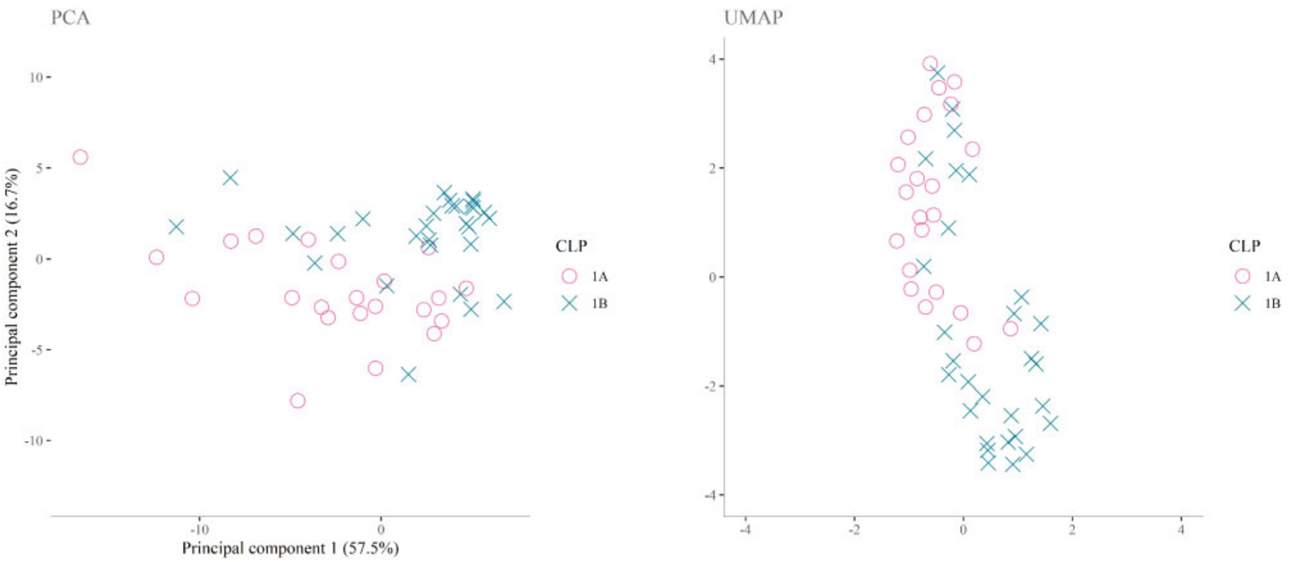

Figure 1. Training dataset visualized using principal component analysis (PCA) (left) and uniform manifold approximation and projection (UMAP) (right). Both plots were constructed using the gene expression values of the 51 genes in the biomarker signature acquired on the NanoString nCounter platform. For the PCA, the significance of the observed separation between the groups were assessed using Hotelling's 2 sample T^2 test, which indicated a significant separation with $p = 2.9*10^{-7}$. Abbreviation: CLP, Classification, Labelling, and Packaging.

based sensitivity metrices based on the consensus prediction from respective laboratory. Cumulative equivalents of the performance metrices were calculated from all consensus predictions established by all 3 laboratories. For example, the cumulative accuracy was calculated as the number of correct consensus predictions (over all laboratories) divided by the total number of established consensus predictions.

Visualization of data. Graphical illustrations were created in R v3.6.1 (R Core Team, 2019) with the assistance of the R-packages: ggplot2 v3.2.1 (Wickham, 2016), ggridges v0.5.1, and gridExtra v2.3. The uniform manifold approximation and projection (UMAP) map was created using the R-package umap v0.2.3.1.

Statistical analysis of dimensionality reduction. For the principal component analysis (PCA), the significance of the observed separation between strong and weak skin sensitizers in the first 2 principal components were estimated using Hotelling's 2 sample T^2 test (Hotelling, 1931). Prior to calculating the test statistic, the assumption of equal covariance matrices was tested using Box's *M* test (BOX GEP, 1949). The Hotelling's T^2 and the Box's *M* tests were utilized as implemented in the R packages Hotelling v1.0-5 and biotools v3.1 (da Silva *et al.*, 2017), respectively.

RESULTS

GARDpotency NanoString Technology Transfer

The concept of utilizing genomic biomarkers specific for sensitizing potency in the GARD platform has been previously explored and a biomarker signature consisting of 52 transcripts has been proposed (Zeller *et al.*, 2017). To verify these findings, data were reproduced in repeated GARD exposure experiments, aiming to produce a training dataset for subsequent predictive modeling on the NanoString nCounter platform. Novel total RNA samples were generated following cellular exposure to a reference panel of chemical sensitizers (Table 1) and gene expression levels of 51 of the 52 previously identified genes were quantified. The transcriptional profiles of these samples were investigated using the dimensionality reduction techniques PCA and UMAP, as presented in Figure 1. It was concluded that discriminatory capabilities of the proposed biomarkers were maintained also in the NanoString nCounter system, as distinct separation between samples of different sensitizing potencies were clearly detectable.

To estimate the predictive capacity of the proposed biomarker signature, tentative prediction models based on SVM were trained and evaluated in an iterative cross-validation exercise. In each iteration, 10-fold data were left out, whereas an SVM was trained on retained data and used to classify the left-out data. The predictions of left-out samples were recorded, and iterations proceeded until each sample had been left out once. Results are listed in Table 1 (Model 1) and a cumulative predictive accuracy was calculated to 78%.

GARDpotency Prediction Model Rationale and Optimization

It has previously been observed that information relating to sensitizing potency of chemicals correlates with the concentration in which chemicals are assayed (Johansson *et al.*, 2017). Table 1 includes information of the chemical-specific GARD input concentration for each of the 51 sensitizers included in the training dataset. These concentrations were further explored and correlated with known sensitizing potency parameters, ie, LLNA categories, LLNA EC3, and human potency categories, as defined by Basketter *et al.* (2014). Results are presented in Figure 2. Indeed, for all studied parameters, it was observed that strong sensitizers are typically assayed at a lower concentration, as compared with weak sensitizers. Although the choice of input concentration is based on a dose-response-dependent cytotoxicity profile, it is evident that strong sensitizers require smaller amounts of substance to trigger a response, one of the hallmarks of sensitizing potency.

Having confirmed the functionality of the 51 genomic biomarkers proposed by Zeller *et al.* and explored the linkage between used GARD input concentrations and sensitizing potency, the potential advantage of combining the 2 concepts

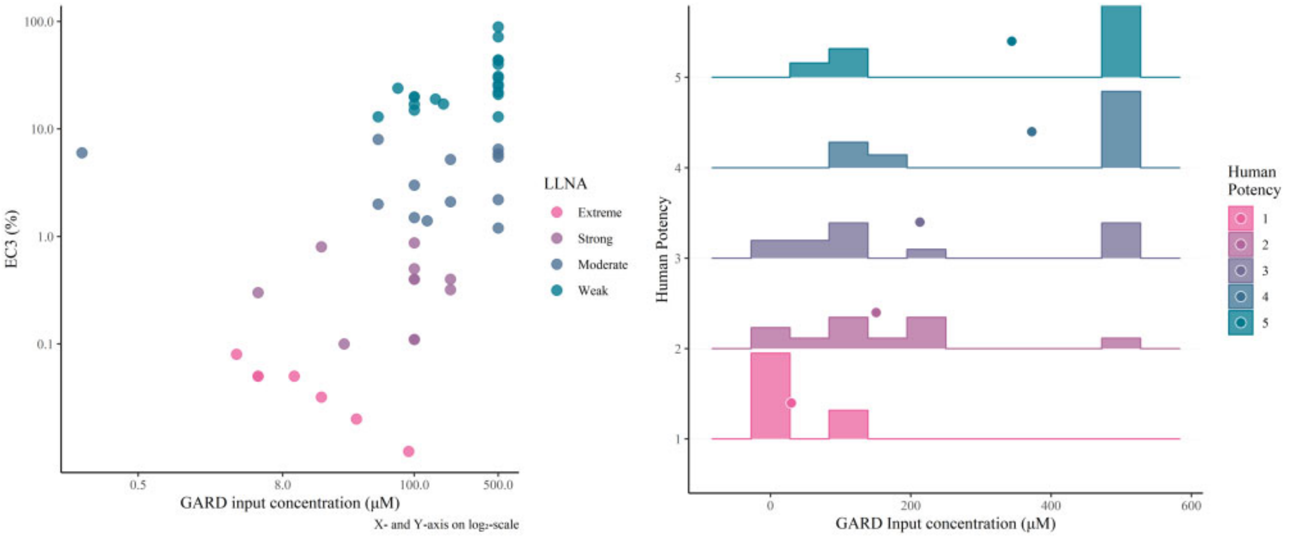

Figure 2. GARD (Genomic Allergen Rapid Detection) input concentrations for the substances in the training dataset visualized against local lymph node assay (LLNA) EC3 values and human potency categories. The left scatter plot visualizes the correlation between the GARD input concentrations and LLNA EC3 values. The right plot shows histograms of the GARD input concentrations for each human potency category. For each category, the mean GARD input concentration is described by the point.

was explored. To this end, the readout of the transcriptional expression levels of the genomic biomarkers was complemented with the experimentally derived concentration parameter in the form of chemical-specific GARD input concentrations, given as in-well concentration during stimulation (µM).

Using this extended dataset, 10-fold cross-validations were repeated as described above, with results presented in Table 1 (Model 2). The predictive accuracy of this updated model was 82%. Thus, based on available information in the internal training dataset, there is indeed an added value in complementing the readout of the genomic biomarker with the experimentally derived GARD input concentration.

Based on these observations, a finalized GARDpotency model was defined based on Model 2; An SVM that is trained on the 51 samples (true sensitizers, divided into 1A and 1B potency classes as presented), that uses the GHS/CLP class as the dependent variable (ie, the predicted parameter) and the expression levels of 51 genes (Supplementary material 2), as assessed by NanoString nCounter measurements, along with the GARD input concentration (µM) as independent variables (ie, predictors). As such, the model was frozen prior to any subsequent classifications of test chemicals assayed in an interlaboratory ring trial, as described below.

Interlaboratory Ring Trial

Predictive performance and assay robustness of GARDpotency and of the GARD tiered approach, for prediction of chemicals' relative skin sensitization potency according to the GHS/CLP categories, were evaluated in a blinded ring trial study comprising 3 laboratories, each assessing a set of 28 chemicals in 3 independent experiments. First, all chemicals were assessed by the GARDskin assay, the results of which have been reported (Johansson *et al.*, 2019). Substances classified as nonskin sensitizers were assigned the GHS/CLP category No Cat, and substances classified as skin sensitizers were assessed in a second tier comprising the GARDpotency assay for subcategorization by their relative potency. The experimental setup of the ring trial generated a total of 252 chemical assessments (28 encoded substances in 3 laboratories in 3 separated experiments). During the course of the experiments, 12 chemical instances failed to generate valid predictions. The missing values were due to; solubility issues in 6 instances (dextran for both Eurofins and BRT in all 3 experiments), interference in the flow-cytometry-based cell-viability assessment caused by autofluorescence (citral in 3 experiments by BRT), failure to meet the cell-viability requirements as specified in the SOP (2-bromo-2-glutaronitrile and 4-(methylamino)phenol sulfate in 2 and 1 experiments by BRT, respectively).

GARDpotency: Reproducibility

The reproducibility of GARDpotency was evaluated by examining the consistency between predictions generated within laboratories (WLR), as well as between laboratories (BLR). When evaluating BLR, consensus predictions for test substances were acquired from all laboratories by majority voting.

For binary potency subclassification, ie, classification of previously predicted skin sensitizers into the GHS/CLP category 1A or 1B using GARDpotency, the WLR was estimated to; 62.5% for BRT, 83.3% for Eurofins, and 88.9% for SenzaGen. The consistency of consensus predictions between laboratories resulted in a BLR estimate of 61.1%.

GARDpotency: Predictive Performance

The binary potency subclassifications of GARDpotency generated by the laboratories can be seen in Figure 3 and in the contingency tables in Table 3. Predictive accuracies were estimated to 93.3% for Burleson, 94.4% for Eurofins, and 76.5% for SenzaGen. Considering each GHS/CLP category in turn, the sensitivity for category 1A was 88.9% for BRT, 90.9% for Eurofins, and 90.9% for SenzaGen. Conversely, the sensitivity for category 1B substances was estimated to 100%, 100%, and 50% for BRT, Eurofins, and SenzaGen, respectively. The cumulative performance of binary potency subclassification was determined to 88.0% accuracy, 90.3% category 1A sensitivity, and 84.2% category 1B sensitivity. A closer examination of the predictions shows that none of the assayed chemicals were consistently

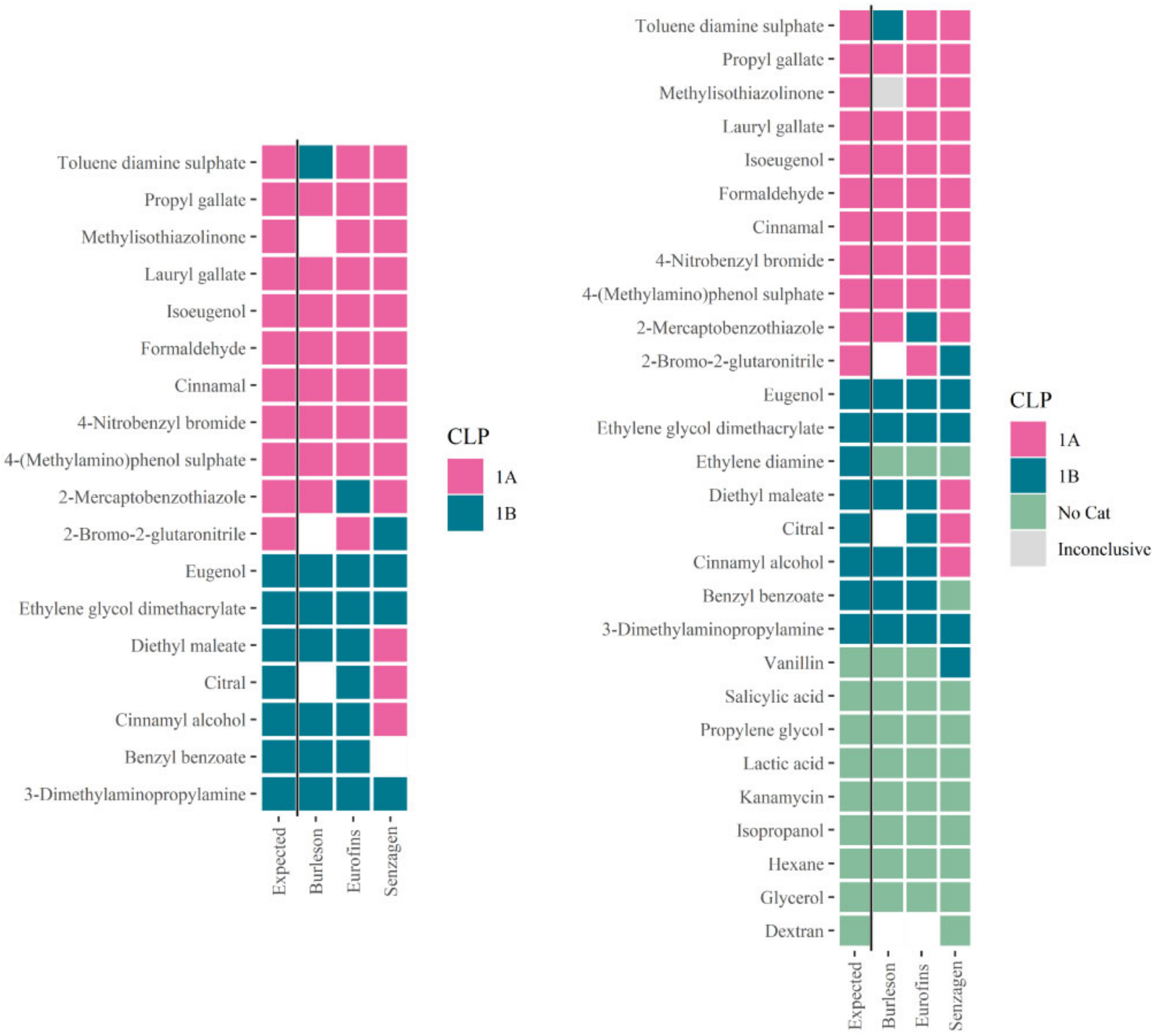

Figure 3. Consensus predictions established by each laboratory for the GARDpotency assay (left) and for the Genomic Allergen Rapid Detection tiered approach (right), described by the color of respective tile. For substances where no prediction could be obtained, the rectangle fill color is white. Abbreviation: CLP, Classification, Labelling, and Packaging.

misclassified by the laboratories. Nonetheless, misclassifications were made on toluene diamine sulfate by BRT, 2-mercaptobenzothiazole by Eurofins, and 2-bromo-2-glutaronitrile, cinnamyl alcohol, citral, and diethyl maleate by SenzaGen.

Table 3. Contingency Tables and Prediction Performances Achieved by Respective Laboratory in the Ring Trial for the GARDpotency Assay

	BRT		Eurofins		SenzaGen	
Reference	1A	1B	1A	1B	1A	1B
1A	8	1	10	1	10	1
1B	0	6	0	7	3	3
Accuracy (%)	93.3		94.4		76.5	
Sensitivity 1A (%)	88.9		90.9		90.9	
Sensitivity 1B (%)	100		100		50	

GARD Tiered Approach: Reproducibility

The reproducibility measures of the GARD tiered approach, ie, classification of substances into either of the 3 GHS/CLP categories—No Cat, 1B, or 1A by combining GARDskin and GARDpotency—resulted in WLRs of 60% for Burleson, 77.8% for Eurofins, and 75% for SenzaGen. The BLR for consensus predictions was estimated to 66.7%.

GARD Tiered Approach: Predictive Performance

Figure 3 shows the consensus predictions of the GARD tiered approach established by respective laboratory and Table 4 describes the contingency tables and the laboratories' individual performance metrices. Summarizing the results, the cumulative accuracy was estimated to 86.1%. The cumulative sensitivities for respective class (considering each category in turn to be the "positive" outcome) were 96.0% for No Cat, 69.6% for category 1B, and 90.3% for category 1A. Of the misclassified compounds, only ethylenediamine was consistently

Table 4. Contingency Tables and Prediction Performances Achieved by Respective Laboratory in the Ring Trial for the Genomic Allergen Rapid Detection Tiered Approach

	BRT			Eurofins			SenzaGen		
Reference	1A	1B	No Cat	1A	1B	No Cat	1A	1B	No Cat
1A	8	1	0	10	1	0	10	1	0
1B	0	6	1	0	7	1	3	3	2
No Cat	0	0	8	0	0	8	0	1	8
Accuracy (%)	91.7			92.6			75		
Sensitivity 1A (%)	88.9			90.9			90.9		
Sensitivity 1B (%)	85.7			87.5			37.5		
Sensitivity No Cat (%)	100			100			88.9		

misclassified as No Cat by all 3 laboratories. Additional misclassifications not previously described include the prediction on vanillin as a weak skin sensitizer (1B) rather than a nonskin sensitizer, and the prediction on benzyl benzoate as a nonskin sensitizer instead of a weak sensitizer, both predictions generated by SenzaGen. Furthermore, methylisothiazolinone was inconsistently classified by BRT in 2 experiments and therefore failed to generate a valid consensus prediction.

DISCUSSION

Skin sensitizer hazard identification and characterization are crucial aspects of chemical risk assessment. Though both endpoints have historically been acquired using *in vivo* models, several nonanimal alternatives have been developed in recent years. However, these approaches have mainly shown proficiency in hazard identification, and methods enabling the relative ranking of chemicals by their skin sensitizing potency are still lacking. Currently, weight-of-evidence approaches has been suggested, that utilizes regulatory accepted nonanimal alternative assays, for assessment of skin sensitizer potency (Casati *et al.*, 2018). Several such approaches were recently examined and compared with human data, and though performances were comparable with those of the LLNA (Kleinstreuer *et al.*, 2018), no strategy based on nonanimal methods has yet been recommended for regulatory purposes. Furthermore, considering the estimated prediction performances, it is the authors belief that continued development of nonanimal approaches for prediction of skin sensitizer potency will lead to additional improvements in correlation with human data.

In this study, an optimized GARDpotency prediction model was established, based on both gene expression analysis of genomic biomarkers, as well as an experimentally derived concentration parameter, as defined by dose-dependent cytotoxicity measurements. The relevance of gene expression measurements for characterization of skin sensitizer potency has previously been described (Albrekt *et al.*, 2014; Zeller *et al.*, 2017), and the genomic biomarkers' ability to provide discriminatory power between weak and strong skin sensitizers was here confirmed. Furthermore, the link between the concentration required to trigger a binary event, and the severity or frequency of such an event, is often explored within the field of toxicology, and provides an opportunity for an intuitive interpretation. In the field of sensitization assessment, examples include the *in vivo* LLNA, which determines sensitizing potency by linking the binary event of a 3-fold induction of T-cell proliferation, as compared with a vehicle control, to the concentration required to generate the response. Similarly, the No Observed Adverse Effect Level value describes the maximum concentration studied that did not induce sensitization in a clinical setting. Here, as well as in other studies (Johansson *et al.*, 2017), correlations between the GARD input concentration, ie, the single dose used for cell exposures prior to gene expression measurements, and both LLNA and Human Potency Categories were observed. Based on these observations, it was hypothesized that the experimentally derived concentration with which test chemicals are assayed provides predictive information related to sensitizing potency. The hypothesis was tested in cross-validation exercises within the training dataset. Indeed, inclusion of GARD input concentration information in the GARDpotency prediction model provides both an intuitive interpretation of data and an increased predictive capacity, as demonstrated.

To validate the GARDpotency prediction model and to allow for estimations of its predictive performance and robustness, a blinded interlaboratory ring trial was performed, encompassing 3 laboratories, each assessing a set of chemicals in 3 independent experiments, using a test dataset comprising positive classifications from the associated GARDskin validation study. The cumulative accuracy of GARDpotency was estimated to 88.0%, suggesting that GARDpotency is indeed functional, as it harbors a capacity to distinguish weak from strong skin sensitizers.

Considering GARDpotency classifications of the test dataset, observed misclassifications were only generated at individual laboratories. Of these, a set of weak sensitizers was overpredicted by the lead laboratory; citral, diethyl maleate, and cinnamyl alcohol. However, no systematic differences in experimental details capable of explaining the observed prediction discrepancies could be identified.

Having assessed the functionality of GARDpotency as a tool for potency-associated subclassification of skin sensitizers, a tiered approach was proposed for complete hazard assessment and characterization. In this GARD tiered approach, test items are subjected to skin sensitizing hazard assessment in a first tier by utilization of the GARDskin assay. Any test item classified as a skin sensitizer is passed to a second tier, in which GARDpotency allows for potency-associated subclassification. Taken together, the GARD tiered approach allows for risk assessment into 3 categories, similar to currently proposed testing strategies (Kleinstreuer *et al.*, 2018).

The predictive performance of the GARD tiered approach was also evaluated within the scope of this study, with an estimated accuracy of 86.1%, based on the common test dataset defined for the validation of GARDskin and GARDpotency, respectively. As evidenced by recent review articles, proposed *in vitro* and *in silico* assays and defined approaches (DAs) for potency assessment exhibit predictive accuracies ranging between 55% and 69% when predicting human sensitizing potencies into 3 discrete categories, similar to the GARD tiered approach (Kleinstreuer *et al.*, 2018). Similarly, when predicting sensitizing potency into 3 categories, *in vivo* counterparts (ie, the LLNA and guinea pig data) exhibit predictive accuracies of 59% (ICCVAM and NICEATM, 2010; Kleinstreuer *et al.*, 2018). Based on such estimates, it is generally agreed that developing approaches that can provide an understanding of potency and that facilitates risk assessment processes represents one of the most significant challenges in skin sensitization sciences today. Adhering the herein reported data to this context, we propose that the GARD tiered approach constitutes progress in the field and that it can positively contribute valuable information to already proposed testing strategies, DA:s, or more loosely incorporated into weight-of-evidence approaches.

Considering the reproducibility of the GARD tiered approach, the herein reported results shows that consistent predictions for potency categorization, in accordance with the 3 GHS/CLP categories, between laboratories were obtained for 66.7% of the assayed chemicals. Furthermore, the reproducibility within respective laboratory ranged between 60.0% and 77.8%. To the best of our knowledge, this is the first study presented to date, in which attempts have been made to evaluate the reproducibility of an *in vitro* potency assessment strategy in a ring trial, making a comparative evaluation of the herein obtained figures difficult. However, it is known that the reproducibility of *in vivo* counterparts, eg, the LLNA, drops from approximately 70%–80% for binary hazard, to 60%–70% for potency categorization in 3 categories, identical to the strategy presented here (Dumont *et al.*, 2016; Kleinstreuer *et al.*, 2018). Thus, the BLR results acquired in the above described ring trial are considered comparable with current regulatory accepted methods for potency categorization of skin sensitizers. Currently, no direct or systematic cause for the observed inconsistencies between the laboratories has been identified. Further investigations into the sources of introduced variability will potentially allow for increased reproducibility.

In conclusion, we have described the optimization of the GARDpotency assay, for discrimination between weak and strong skin sensitizers, and the transfer of the proposed model to a standardized gene acquisition platform. Furthermore, the validation results from a ring trial study performed in accordance with OECD guidance documents were reported. The results suggest that the assay is indeed able to categorize chemical skin sensitizers according to their relative potency, in compliance with the GHS/CLP categories. We suggest that the described testing strategy provides valuable data that, together with the work of others, will contribute toward realization of an ultimate solution that will allow, in the future, an assessment of skin sensitizing potency without recourse to animal experimentation.

SUPPLEMENTARY DATA

Supplementary data are available at *Toxicological Sciences* online.

ACKNOWLEDGMENTS

The authors wish to thank Els Adriens for evaluating and summarizing generated predictions; Amber Edwards, Florence Burleson, Veronika Zuckerstätter, and Helge Gehrke for facilitating and performing GARD assessments; Erwin Roggen for management and experimental directives; Astrid Reus for coding and distribution of blinded chemicals; Annika Eriksson and Matthew Thomas for their technical assistance with NanoString sample preparation and data generation.

FUNDING

SenzaGen AB.

DECLARATION OF CONFLICTING INTERESTS

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

REFERENCES

Albrekt, A.-S., Johansson, H., Börje, A., Borrebaeck, C., and Lindstedt, M. (2014). Skin sensitizers differentially regulate signaling pathways in MUTZ-3 cells in relation to their individual potency. *BMC Pharmacol. Toxicol.* **15**, 5.

Basketter, D. A., Alépée, N., Ashikaga, T., Barroso, J., Gilmour, N., Goebel, C., Hibatallah, J., Hoffmann, S., Kern, P., Martinozzi-Teissier, S., *et al.* (2014). Categorization of chemicals according to their relative human skin sensitizing potency. *Dermatitis* **25**, 11–21.

BOX GEP. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika* **36**, 317–346.

Casati, S., Aschberger, K., Barroso, J., Casey, W., Delgado, I., Kim, T. S., Kleinstreuer, N., Kojima, H., Lee, J. K., Lowit, A., *et al.* (2018). Standardisation of defined approaches for skin sensitisation testing to support regulatory use and international adoption: Position of the international cooperation on alternative test methods. *Arch. Toxicol.* **92**, 611–617.

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* **20**, 273–297.

da Silva, A. R., Malafaia, G., and Menezes, I. (2017). Biotools: An R function to predict spatial gene diversity via an individual-based approach. *Genet. Mol. Res.* **16**, doi:10.4238/gmr16029655.

Daniel, A. B., Strickland, J., Allen, D., Casati, S., Zuang, V., Barroso, J., Whelan, M., Régimbald-Krnel, M. J., Kojima, H., Nishikawa, A., *et al.* (2018). International regulatory requirements for skin sensitization testing. *Regul. Toxicol. Pharmacol.* **95**, 52–65.

Dumont, C., Barroso, J., Matys, I., Worth, A., and Casati, S. (2016). Analysis of the local lymph node assay (LLNA) variability for assessing the prediction of skin sensitisation potential and potency of chemicals with non-animal approaches. *Toxicol. In Vitro* **34**, 220–228.

EC. (2006). Regulation (EC) no 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) no 793/93 and Commission Regulation (EC) no 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. *Official J. Eur. Union* **396**, 1–849.

ECHA. (2017). *Guidance on the Application of CLP Criteria*: guidance to Regulation (EC) No 1272/2008 on classification, labelling and packaging (CLP) of substances and mixtures. European Chemicals Agency, Helsinki.

EU. (2003). *Directive, B Council of 27 July 1976 on the Approximation of the Laws of the Member States Relating to Cosmetic Products. Official Journal of the European Union* **66**, 26–35.

Forreryd, A., Johansson, H., Albrekt, A.-S., Borrebaeck, C. A. K., and Lindstedt, M. (2015). Prediction of chemical respiratory sensitizers using GARD, a novel *in vitro* assay based on a genomic biomarker signature. *PLoS One* **10**, e0118808.

Forreryd, A., Zeller, K. S., Lindberg, T., Johansson, H., and Lindstedt, M. (2016). From genome-wide arrays to tailor-made biomarker readout—Progress towards routine analysis of skin sensitizing chemicals with GARD. *Toxicol. In Vitro* **37**, 178–188.

Gilmour, N., Kimber, I., Williams, J., and Maxwell, G. (2019). Skin sensitization: Uncertainties, challenges, and opportunities for improved risk assessment. *Contact Dermatitis* **80**, 195–200.

Hotelling, H. (1931). The generalization of student's ratio. *Ann. Math. Stat.* **2**, 360–378.

ICCVAM, and NICEATM. (2010). *ICCVAM Test Method Evaluation Report on the Murine Local Lymph Node Assay: BrdU-Elisa, a Nonradioactive Alternative Test Method to Assess the Allergic Contact Dermatitis Potential of Chemicals and Products*. NIH Publication No. 10-7552. National Institute of Environmental Health Sciences, Research Triangle Park, NC.

Johansson, H., Albrekt, A.-S., Borrebaeck, C. A. K., and Lindstedt, M. (2013). The GARD assay for assessment of chemical skin sensitizers. *Toxicol. In Vitro* **27**, 1163–1169.

Johansson, H., Gradin, R., Forreryd, A., Agemark, M., Zeller, K., Johansson, A., Larne, O., van Vliet, E., Borrebaeck, C., and Lindstedt, M. (2017). Evaluation of the GARD assay in a blind cosmetics Europe study. *ALTEX* **34**, 515–523.

Johansson, H., Gradin, R., Johansson, A., Adriaens, E., Edwards, A., Zuckerstätter, V., Jerre, A., Burleson, F., Gehrke, H., and Roggen, E. L. (2019). Validation of the GARD^{TM}skin assay for assessment of chemical skin sensitizers: Ring trial results of predictive performance and reproducibility. *Toxicol. Sci.* **170**, 374–381.

Johansson, H., Lindstedt, M., Albrekt, A.-S., and Borrebaeck, C. A. (2011). A genomic biomarker signature can predict skin sensitizers using a cell-based *in vitro* alternative to animal tests. *BMC Genomics* **12**, 1–19.

Kimber, I., Basketter, D. A., Gerberick, G. F., Ryan, C. A., and Dearman, R. J. (2011). Chemical allergy: Translating biology into hazard characterization. *Toxicol. Sci.* **120**, S238–268.

Kleinstreuer, N. C., Hoffmann, S., Alépée, N., Allen, D., Ashikaga, T., Casey, W., Clouet, E., Cluzel, M., Desprez, B., Gellatly, N., *et al.* (2018). Non-animal methods to predict skin sensitization (II): An assessment of defined approaches (*). *Crit. Rev. Toxicol.* **48**, 359–374.

Loveless, S. E., Api, A. M., Crevel, R. W. R., Debruyne, E., Gamer, A., Jowsey, I. R., Kern, P., Kimber, I., Lea, L., Lloyd, P., *et al.* (2010). Potency values from the local lymph node assay: Application to classification, labelling and risk assessment. *Regul. Toxicol. Pharmacol.* **56**, 54–66.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2019). *E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071)*. TU Wien. R package.

OECD. (2005). *Series on Testing and Assessment, No 34: Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment*. OECD Publishing, Paris.

OECD. (2009). *Series on Testing and Assessment, No 1: Guidance Document for the Development of OECD Guidelines for Testing of Chemicals (as Revised in 2009)*. OECD Publishing, Paris.

OECD. (2010). *Test No. 429: Skin Sensitisation*. OECD Publishing, Paris.

OECD. (2018a). *Test No. 442D: In Vitro Skin Sensitisation*. OECD Publishing, Paris.

OECD. (2018b). *Test No. 442E: In Vitro Skin Sensitisation*. OECD Publishing, Paris.

OECD. (2019). *Test No. 442C: In Chemico Skin Sensitisation*. OECD Publishing, Paris.

Prüss-Üstün, A., Wolf, J., Corvalán, C., Bos, R., and Neira, M. (2016). *Preventing Disease Through Healthy Environments: A Global Assessment of the Burden of Disease From Environmental Risks*. World Health Organization, Geneva, Switzerland.

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, NY.

Zeller, K. S., Forreryd, A., Lindberg, T., Gradin, R., Chawade, A., and Lindstedt, M. (2017). The GARD platform for potency assessment of skin sensitizing chemicals. *ALTEX* **34**, 539–559.