



## Letter to the editor regarding the article “Is a combination of assays really needed for non-animal prediction of skin sensitization potential? Performance of the GARD™ (Genomic Allergen Rapid Detection) assay in comparison with OECD guideline assays alone and in combination.”

### To the editor:

Recently, a letter to the editor from Drs. Natsch and Landsiedel was published (Natsch and Landsiedel, 2019) raising concerns regarding the contents of an original research article authored by Dr Roberts (2018). The Roberts paper evaluated existing data to compare the performance of three validated *in vitro* assays for testing of skin sensitizing chemicals to the performance of the GARD assay on a coherent chemical dataset. Although SenzaGen AB supported preparation of the Roberts paper, the data that formed the basis of the comparison has previously been published in peer-reviewed scientific journals and are readily available for similar independent meta-analysis by anyone who so chooses. The GARD development team did not participate in the preparation of the manuscript, but since several of the comments in the letter by Natsch and Landsiedel concern the GARD platform, rather than the comparative analyses reported in the Roberts paper, we considered it appropriate to respond and address some of the points in this letter. In our pursuit of transparency and scientific rigour, all development stages of the GARD platform have been published over time through a series of publications to allow for full insight into the assay (Forreryd et al., 2014, 2016; Johansson et al., 2011, 2014; Zeller et al., 2017). We thank Natsch and Landsiedel for their feedback and welcome the opportunity to correct what appears to be some misconceptions and unfortunate misunderstandings regarding the GARD assay by referencing back to these historical publications.

### 1. Stable biological endpoints

Natsch and Landsiedel claim that the GARD prediction signature (GPS) is not a fixed and completely defined biological endpoint. This is not correct: the genes in the prediction signature have remained the same since the initial biomarker discovery in 2011 (Johansson et al., 2011). The identification of the GPS was initially performed from global microarray measurements on the Affymetrix microarray platform (Johansson et al., 2011). Due to high costs and time consuming protocols inherent to the microarray platform, gene expression measurements had to be transferred to another platform to prepare GARD for routine testing. The rationale for selecting the NanoString platform originated from an evaluation of a selection of gene expression platforms (Forreryd et al., 2014). This initial study was performed on a small subset of genes from the GPS (< 15%) and all platforms were used in their standard design with no attempts to optimize parameters to achieve high correlation to historical data.

In a follow up study after the selection of the NanoString platform (Forreryd et al., 2016), attention was directed from an overall assessment of technical parameters to focus on the biological model. Major

efforts were made by bioinformaticians and the NanoString design team to carefully match each NanoString probe pair to the specific target exon on the microarray platform to achieve high correlation to historical data. As evident from this publication, the microarray data could be readily reproduced on the NanoString platform.

In summary, the information published in the two papers by Forreryd et al. (2014, 2016) provides full transparency of the platform transfer process. As evident from the later publication, the structure of the microarray data was readily reproduced and validated on an independent gene expression platform, providing evidence that the GPS signature is a well-defined, stable and fixed biological endpoint, independent on technical platform. To further strengthen this argument, the GARD assay was recently evaluated in a blinded ring trial. A concordance of 92% was reported between the three participating laboratories, demonstrating that GARD is a highly reproducible assay, fulfilling requirements for formal validation (Johansson et al., 2019).

Natsch and Landsiedel furthermore state that there is a fine-tuning of the prediction signature in GARD during testing of external samples, and that the prediction model is not fixed. This incorrect observation seems to originate from two assay development publications, where we have suggested that a set of internal positive and negative control samples (i.e. benchmark samples) could be used to compensate for technical variations (i.e. batch effects) between samples processed in separate batches (Forreryd et al., 2016; Johansson et al., 2014). Variations between experiments is not an issue attributable only to the GARD assay, but rather a general issue present in many (if not all) biologic systems, as well as in technologies measuring biological signals. It is important to understand how such variations affects these signals and to take appropriate action to compensate for their potential impacts. In this aspect, we consider it to be an advantage of GARD that such effects have been carefully characterized, and a robust and statistically sophisticated method has been introduced to eliminate unwanted effects. Importantly, control samples that are included in each batch of unknown test samples do not affect the prediction signature nor the prediction model, they are simply used for calibration. In both assay development publications, we have made significant efforts to carefully describe this process, and the fact that it does not affect the *a priori* defined biomarkers nor the model (which are both fixed). While GARD operates on more biomarkers, the interpretation of the calibration procedure is analogous to the use of standard curves in single biomarker assays (e.g. protein measurements in the Bradford assay), or for example to the blank sample used during calibration prior to luminescence readings in LuSens and Keratinosens. Thus, the argument made by Natsch and Landsiedel regarding batch effects in GARD is perhaps more appropriately reversed: why do no other *in vitro* assay describe, or take appropriate actions against batch effects that may

affect the outcome of classifications?

## 2. Predictivity

Natsch and Landsiedel correctly observed that the balanced accuracy of GARD in a blinded study performed in collaboration with Cosmetics Europe (CE) (Johansson et al., 2017) was 75%, while a previous publication on GARD assay performance has reported a higher accuracy of 88% (Johansson et al., 2014). They also suggested that the lower accuracy in the CE data may partly be explained by the low number of non-sensitizers present in the test set (54S, 18NS). We completely agree that this is part of the explanation. In addition, and as also discussed in the CE publication, the non-sensitizers in the dataset were highly biased towards acknowledged “hard-to-predict” compounds: more than half of the non-sensitizers belonged to human potency group 5 as defined in Basketter et al. (2017), and are thus assumed to possess at least some inherent sensitizing capacity. Furthermore, at least three of the GARD false positive predictions are frequently classified as sensitizers in other *in vitro* assays, as apparent from the data presented in the recent publication by Hoffmann et al. (2018). Off note, as also acknowledged by Natsch and Landsiedel, all evaluated assays in this publication demonstrated a major reduction in performance on the compounds overlapping with the CE publication. Moreover, all chemicals from the CE publication on GARD are also included in the publication by Roberts, generating a larger and more diverse dataset. This obviously reduced the skewing of the specificity towards the lower figures observed on the smaller CE dataset, further strengthening the selection bias present in the CE dataset. Thus, to avoid any bias associated with testing of different subsets of chemicals, predictive performances are only directly comparable on coherent datasets. The comparative study by Roberts reporting assay performances on the same subset of chemicals should therefore be encouraged and considered as a great first step towards a more comprehensive and less biased evaluation of available *in vitro* assays.

## 3. Lack of transparency

We would firmly like to reject Natsch and Landsiedel's statement that GARD is more of a black box than any of the OECD validated assays. The term black box is more appropriately used to describe a process with known inputs and outputs, but whose internal structure or function is not described or understood. As both the genes of the predictive biomarker signature (GPS) and the Support Vector Machines (SVM), which constitute the basis of the prediction model in GARD, are well described and well understood, labeling GARD as a black box is inaccurate. The full list of transcripts in the prediction signature has been disclosed (Johansson et al., 2011), and describes many biological processes involved in skin sensitization. Interestingly, one of the most populated pathways was shown to involve the NRF-2 mediated oxidative response, which for example represents the single endpoint measured in the KeratinoSens and LuSens assays. In addition, genes are represented in many pathways known to be involved in DC maturation and activation, involving co-stimulatory molecules (e.g. CD86), as measured in h-CLAT, as well as pattern recognition receptors (e.g. TLRs), pro-inflammatory cytokines and chemokines. As evident from available data on assay performance, the complex endpoint of skin sensitization cannot accurately be captured by simplified assays monitoring single and isolated endpoints, nor by a 2 out of 3 ITS where the information regarding the target key event of a certain assay is not passed on to the final prediction model (see discussion in Johansson and Gradin (2017)). Thus, for the purpose of improving hazard predictions and advancing the field of toxicology, it may be necessary to utilize advanced mathematical and statistical tools and embrace novel approaches that are far more holistic than single endpoints.

The prediction model used in the GARD assay is based on SVM, which is a well-defined algorithm (Cortes and Vapnik, 1995). We have

described in detail the parameters of the GARD SVM model, making it possible for anyone interested and having minimal knowledge in machine learning to interpret the prediction model and reconstruct the data using our training dataset (see for example Forreryd et al. (2016)). Regardless of whether a prediction model comprises a simple or a more complex set of rules it does not *per se* make the model less transparent to add parameters to the read out, although it may demand a bit more effort and knowledge to understand the detailed mechanisms resulting in the final predictions. Importantly, for the expertise, there is indeed full transparency on what is determining the predictions, and for the end user, the results from the GARD assay are as easily interpreted as any other assay: a positive SVM decision value is classified as a sensitizer and a negative value as a non-sensitizer.

## 4. GARDpotency

In this section, Natsch and Landsiedel discuss the GARD potency model described in Zeller et al. (2017), which was not considered in Roberts publication. While we agree with Natsch and Landsiedel that dose-response and concentrations are relevant in the context of developing a continuous regression model of LLNA EC<sub>3</sub>-values, we strongly wish to emphasize that GARD potency is neither a regression model generating a continuous output nor a model aiming to predict an arbitrary threshold value on a curve: classifications in GARD potency are based on whether the gene expression profile induced by the test substance is most similar to gene expression profiles induced either by strong (1A) or weak (1B) sensitizers in the training dataset. Of note, the GARD potency publication describes an early assay development initiative motivated by current legislation and the lack of reliable *in vitro* alternatives for potency assessment. As demonstrated on external test data in the GARD potency publication, and recently also in a blinded ring trial involving external laboratories (Manuscript in preparation), this model is superior in performance to other *in vitro* alternatives, such as for example those reported in Kleinstreuer et al. (2018).

## 5. Conclusions

We appreciate the efforts made by Natsch and Landsiedel to review and comment on the GARD assay, and we are grateful to the editor for the opportunity to address their concerns. Our intention with this response is to clarify a number of misunderstandings regarding the GARD platform, and clearly underline that all our conclusions are based on published data, transparently available for all readers to scrutinize.

## Funding sources

This work was supported by grants from the Swedish Foundation for Strategic Research (IB13-0118). The funding sources had no involvement in the study design; in the collection, analysis, and interpretation of data; in the writing of the manuscript, or in the decision to submit the article for publication.

## References

- Basketter, D.A., Alepee, N., Ashikaga, T., Barroso, J., Gilmour, N., Goebel, C., Hibatallah, J., Hoffmann, S., Kern, P., Martinuzzi-Teissier, S., Maxwell, G., Reisinger, K., Sakaguchi, H., Schepky, A., Tailhardat, M., Templier, M., 2014. Categorization of chemicals according to their relative human skinsensitizing potency. *Dermatitis* 25, 11–21.
- Cortes, C., Vapnik, V., 1995. Support vector networks. *Mach. Learn.* 20, 273–297.
- Forreryd, A., Johansson, H., Albrekt, A.S., Lindstedt, M., 2014. Evaluation of high throughput gene expression platforms using a genomic biomarker signature for prediction of skin sensitization. *BMC Genomics* 15, 379.
- Forreryd, A., Zeller, K.S., Lindberg, T., Johansson, H., 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.
- Hoffmann, S., Kleinstreuer, N., Alepee, N., Allen, D., Api, A.M., Ashikaga, T., Clouet, E., Cluzel, M., Desprez, B., Gellatly, N., Goebel, C., Kern, P.S., Klaric, M., Kühnl, J., Lalko, J.F., Martinuzzi-Teissier, S., Mewes, K., Miyazawa, M., Parakhia, R., van Vliet, E.,

- Zang, Q., Petersohn, D., 2018. Non-animal methods to predict skin sensitization (I): the Cosmetics Europe database. *Crit. Rev. Toxicol.* 48 (5), 344–358.
- Johansson, H., Gradin, R., Forreryd, A., Agemark, M., Zeller, K., Johansson, A., Larne, O., van Vliet, E., Borrebaeck, C., Lindstedt, M., 2017. Evaluation of the GARD assay in a blind Cosmetics Europe study. *ALTEX* 34 (4), 515–523.
- Johansson, H., Lindstedt, M., Albrekt, A., Borrebaeck, C., 2011. A genomic biomarker signature can predict skin sensitizers using a cell-based in vitro alternative to animal tests. *BMC Genomics* 12, 399.
- Johansson, H., Rydnert, F., Kuhn, J., Schepky, A., Borrebaeck, C., Lindstedt, M., 2014. Genomic allergen rapid detection in-house validation—a proof of concept. *Toxicol. Sci.* 139 (2), 362–370.
- Johansson, H., Gradin, R., 2017. Skin sensitization: challenging the conventional thinking - a case against 2 out of 3 as integrated testing strategy. *Toxicol. Sci.* 159 (1), 3–5.
- Johansson, H., Gradin, R., Johansson, A., Adriaens, E., Edwards, A., Zuckerstätter, V., Jerre, A., Burleson, F., Gehrke, H., Roggen, E., 2019. Validation of the GARDskin assay for assessment of chemical skin sensitizers: ring trial results of predictive performance and reproducibility. *Toxicol. Sci.* <https://doi.org/10.1093/toxsci/kfz108>.
- Kleinstreuer, N.C., Hoffmann, S., Alepee, N., Allen, D., Ashikaga, T., Casey, W., Clouet, E., Cluzel, M., Desprez, B., Gellatly, N., Göbel, C., Kern, P.S., Klaric, M., Kühnl, J., Martinozzi-Teissier, S., Mewes, K., Miyazawa, M., Strickland, J., van Vliet, E., Zang, Q., Petersohn, D., 2018. Non-animal methods to predict skin sensitization (II): an assessment of defined approaches. *Crit. Rev. Toxicol.* 48 (5), 359–374.
- Natsch, A., Landsiedel, R., 2019. Letter to the editor regarding the article by Roberts, 2018. *Regul. Toxicol. Pharmacol.* 102, 115–116.
- Roberts, D.W., 2018. Is a combination of assays really needed for non-animal prediction of skin sensitization potential? Performance of the GARD (Genomic Allergen Rapid Detection) assay in comparison with OECD guideline assays alone and in combination. *Regul. Toxicol. Pharmacol.* 98, 155–160.
- Zeller, K.S., Forreryd, A., Lindberg, T., Gradin, R., Chawade, A., Lindstedt, M., 2017. The GARD platform for potency assessment of skin sensitizing chemicals. *ALTEX* 34 (4), 539–559.

Andy Forreryd\*

*SenzaGen AB, Medicon Village Bldg 401, Scheelevägen 2, 22381, Lund, Sweden*

*E-mail address: [andy.forreryd@senzagen.com](mailto:andy.forreryd@senzagen.com).*

Henrik Johansson

*SenzaGen AB, Medicon Village Bldg 401, Scheelevägen 2, 22381, Lund, Sweden*

*E-mail address: [henrik.johansson@senzagen.com](mailto:henrik.johansson@senzagen.com).*

Tim Lindberg

*Department of Immunotechnology, Lund University, Medicon Village Bldg 406, Scheelevägen 2, 22381, Lund, Sweden*

*E-mail address: [tim.lindberg@immun.lth.se](mailto:tim.lindberg@immun.lth.se).*

Kathrin S. Zeller

*Department of Immunotechnology, Lund University, Medicon Village Bldg 406, Scheelevägen 2, 22381, Lund, Sweden*

*E-mail address: [kathrin.zeller@immun.lth.se](mailto:kathrin.zeller@immun.lth.se).*

Malin Lindstedt

*Department of Immunotechnology, Lund University, Medicon Village Bldg 406, Scheelevägen 2, 22381, Lund, Sweden*

*E-mail address: [malin.lindstedt@immun.lth.se](mailto:malin.lindstedt@immun.lth.se).*

\* Corresponding author.