GARD

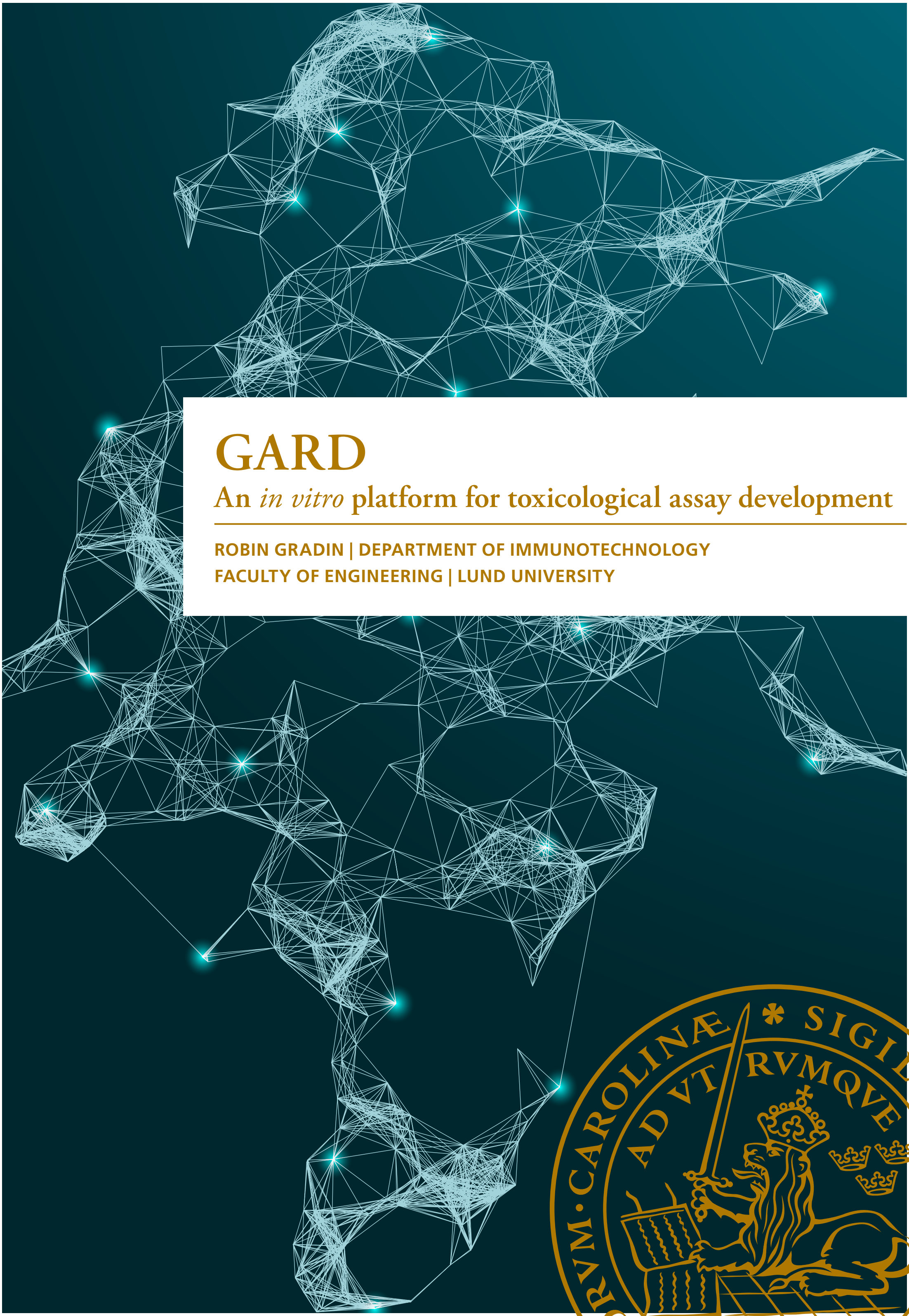
An in vitro platform for toxicological assay development

Gradin, Robin

2020

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00



GARD

An *in vitro* platform for toxicological assay development

ROBIN GRADIN | DEPARTMENT OF IMMUNOTECHNOLOGY
FACULTY OF ENGINEERING | LUND UNIVERSITY

Printed by Media-Tryck, Lund 2020 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

Faculty of Engineering
Department of Immunotechnology

ISBN 978-91-7895-650-0

9 789178 956500

GARD

An *in vitro* platform for toxicological assay development

Robin Gradin

LUND
UNIVERSITY

DOCTORAL DISSERTATION

by due permission of the Faculty of Engineering, Lund University, Sweden.
To be defended at Hörsalen, Medicon Village, Scheelevägen 2, Lund.

Friday November 27, 2020, at 09.00 am.

Faculty opponent
Associate Professor Mattias Öberg.
Karolinska Institute

Organization
LUND UNIVERSITY

Department of Immunotechnology
Medicon Village (building 406)
SE-223 81 Lund, Swede

Author Robin Gradin

Document name
DOCTORAL DISSERTATION

Date of issue
2020-11-27

Sponsoring organization
Senzagen AB

Title and subtitle
GARD - An in vitro platform for toxicological assay development

Abstract

Life in modern society is intricately intertwined with the results of continuous technological advances. This has given rise to daily routines where people are frequently exposed to a wide variety of chemicals. Though many of the chemicals do not generally induce adverse health effects upon normal exposure, several of them has the potential to negatively impact health. Indeed, in order to create products that are safe for the general population, potential hazards must be considered and characterized. However, many of the potential adverse health effects are challenging to assess due to the complexity of their underlying mechanisms.

Therefore, hazard assessment has traditionally been based on animal models, since the available knowledge has failed to reduce the complexity to a state that allows the creation of simpler yet effective alternative assays. However, animal models are problematic for several reasons, which makes the development of accurate non-animal alternatives for hazard assessment a sought-after endpoint.

The GARD platform has been developed towards this end, and it permits the development of effective non-animal assays, even for mechanistically complex endpoints. The method exploits the technological advances enabling transcriptomic analysis, following controlled exposure experiments, with machine learning techniques to identify predictive biomarkers and to define high-performing classification models.

This thesis presents the GARD platform and its technological constituents and highlights how it can be used to create assays aimed towards mechanistically complex endpoints, while simultaneously generating new information that can aid the understanding of said endpoints. Further, two of the developed GARD assays, GARDskin and GARDpotency, are described in more detail. Both assays have been developed toward the hazard endpoint of skin sensitization and provide cutting edge performances in hazard identification and hazard characterization, and represent viable alternatives for eventually eliminating the need for animal testing.

Key words *In vitro* assay, predictive modelling, skin sensitization, toxicology.

Classification system and/or index terms (if any)

Supplementary bibliographical information	**Language** English
ISSN and key title	**ISBN** 978-91-7895-650-0

Recipient's notes	**Number of pages** 123	Price
	Security classification	

Signature Robin Gradin

Date 2020-10-18

GARD

An *in vitro* platform for toxicological assay development

Robin Gradin

Lund University

Cover photo by Robin Gradin

Faculty of Engineering
Department of Immunotechnology

ISBN 978-91-7895-650-0 (print)
ISBN 978-91-7895-651-7 (pdf)

Printed in Sweden by Media-Tryck, Lund University
Lund 2020

Table of Contents

Acknowledgements

The making of this thesis has benefitted greatly from the aid of several people, to all of which I would like to express my gratitude.

First, I would like to thank my supervisors. Henrik, thank you for all your time, support, and guidance, and for having the confidence in me to allow me to explore and learn new subjects while also learning to accommodate their associated responsibilities. I am glad to have had the opportunity to learn from you. Malin, thank you for your assistance and understanding. For first introducing me to the Department of Immunotechnology and Senzagen and allowing me this experience. I am very grateful for your help, which has always been available whenever I have needed it. Kathrin, you have always made the department a welcoming place, and I truly appreciate your support and kindness and for your efforts in sharing your knowledge. Anders, thank you for offering your expertise and for your help, I am grateful.

Thank you to all my colleagues at Senzagen, both present and past, you have all taught me a lot and I am very appreciative for the opportunity to work with you. Maria, thank you for first introducing me to several of the concepts in the laboratory when I first started out, and for all your support since then. Your professionalism and knowledge are admirable. Olivia, thank you for your always kind demeanour and willingness to share your skill, it has been a pleasure to work with- and learn from you during these years. Angelica, your professional attitude and skill is encouraging, and I am thankful to have had the chance to collaborate with you in many projects. Andy, I appreciate your considerate manner, ever since my introduction at the department. Your enthusiasm and deep knowledge of the technology is inspiring. Anders, Ulrika, Emil, and Lisa, thank you all for making Senzagen a friendly workplace that is filled with fruitful and intriguing scientific discussions and ideas. To all my colleagues, thank you for making Senzagen what it is and for your continuous strive and effort towards making the GARD assays viable and attractive methods for hazard assessment. Your commitment and pursuit are motivating.

Thank you to everyone at the department of Immunotechnology. Though my time spent there during the last few years has been limited, it has always been a very welcoming and stimulating environment. Thank you Tim, for all your help and support, and for being a good friend throughout the studies. Thanks to Renato, Milad, Jakob, and Linnea for fruitful exchange of ideas during this project. Thanks to Mats, Sara, Fredrik, Kristina, and Mattias for all your feedback and tutoring during my time as a student, you have all helped me expand my understanding of the field.

I would also like to extend a special thanks to my family. You are all a constant source of inspiration and joy, and your selfless acts and willingness to aid seem boundless and I am very much humbled and immensely grateful.

Contributions to the papers in this thesis

I. Performed experimental work, analysed gene expression data, and participated in the interpretation of the results. Read and approved the manuscript.

II. Performed experimental work and statistical analysis. Contributed with figures and writing to the manuscript. Read and approved the manuscript.

III. Performed statistical analysis, including gene expression analysis and interpretation of results. Prepared figures and wrote the manuscript.

IV. Designed the study and performed the data analysis. Prepared figures and wrote the manuscript.

Original papers

I. 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.

II. Johansson H, Gradin R, Johansson A, Adriaens E, Edwards A, Zuckerstätter V, Jerre A, Burleson F, Gehrke H, Roggen EL. 2019. Validation of the GARD™skin assay for assessment of chemical skin sensitizers: Ring trial results of predictive performance and reproducibility. Toxicol Sci. 170(2):374-381.

III. Gradin R, Johansson A, Forreryd A, Aaltonen E, Jerre A, Larne O, Mattson U, Johansson H. 2020. The GARD™potency assay for potency-associated subclassification of chemical skin sensitizers - rationale, method development and ring trial results of predictive performance and reproducibility. Toxicol Sci.

IV. Gradin R, Lindstedt M, Johansson H. 2019. Batch adjustment by reference alignment (BARA): Improved prediction performance in biological test sets with batch effects. PLoS One. 14(2):e0212669.

Published papers are reproduced with the permission of the publisher.

Additional papers

I. 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.

II. Zeller KS, 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.

1 Introduction

An examination of the environment that is associated with our modern lifestyle reveals that most of us are frequently exposed to materials that could potentially give rise to adverse health effects. Yet, many are unconcerned of the risks and are rarely affected by their manifestations (OECD 2020a; OECD and EU 2018; WHO 2016). The apparent safety of our surroundings is not a circumstance that has arisen only by chance but stems from an understanding of the toxic properties that can be attributed to different substances and a deliberate appliance of that knowledge to enforce public safety. Nevertheless, the understanding of toxicological mechanisms is not complete, which when combined with a continuous strive towards producing novelty, presents several challenges.

A significant challenge concerns the identification and the characterization of potentially harmful effects that can be assigned to chemicals. This task is often aided by the utilization of tools and assays designed to assess specific properties or hazards, which can be based on e.g. computational analysis or by studies examining the exposure induced effects on test animals (Rowan and Spielmann 2019). The toxic effects that chemicals can induce are numerous and include complex endpoints such as reproductive and developmental toxicity, carcinogenicity, or skin sensitization (Parasuraman 2011). Traditionally, these endpoints have been evaluated using animal models (Council 2006). However, there are several issues associated with the application of *in vivo* assays for toxicological analysis, including ethical concerns, high costs, and low throughput. This has led to an increase in the scientific aspiration to develop novel non-animal assays, which is further fuelled by the growing general opposition towards animal experimentation.

The publication of the book "*The principles of humane experimental technique*" by Russel and Burch in 1959 (Russell and Burch 1959) constitutes an early and important event for conceptualizing guiding principles of animal experimentation. In the work, the authors introduce the principles of the 3Rs, which emphasize that new test methods should aim towards either Replacing existing animal methods, offer a Refinement of current methods (i.e. resulting in less painful and stressful experiments), or Reduce the number of animals required for an analysis. The 3Rs have been integrated into scientific, public, and regulatory fields, and have come to influence the handling of experimental animals, experimental design, and also spurred the development of novel test methods (Stephens and Mak 2014). In addition, several legislative incentives that promote the development of non-animal assays have been introduced, such as restrictions or bans on the use of animal methods within certain industries (Taylor and Rego Alvarez 2020).

Even so, the toxicological endpoints of interest are represented by complex biological mechanisms, several of which have yet to be effectively represented by alternative test methods, resulting in a persistent need to utilize animal models for many purposes. Thus, this should encourage the continued development of non-animal assays. Many interesting initiatives have been made to increase the understanding of toxicological testing using modern technologies and mechanism-based approaches. One of these is the U.S. Tox21 program (Council 2007; Krewski et al. 2010), which aims to improve the development and understanding of toxicological assays by employing high-throughput screening with computational techniques to enable assessment and prediction of chemicals' hazards. Similarly, another consortium has established a database, referred to as the connectivity map (CMAP), that contains a large number of gene expression profiles that have been acquired following exposure experiments using thousands of perturbagens, which can be used to e.g. compare similarities in induced expression profiles between chemicals (Lamb et al. 2006; Subramanian et al. 2017).

The approach and the design principles of the GARD (Genomic Allergen Rapid Detection) platform employ a similar mechanistic methodology, where induced molecular patterns are used to identify specific hazards of analysed chemicals. GARD is a platform that can be used to develop *in vitro* toxicological assays. It is characterized by transcriptomics-based analysis on data from cellular exposure experiments that have been performed on a purposely selected cell line, which enables the identification of genetic biomarkers that are predictive towards a specific hazard endpoint of interest. Identified biomarkers are processed using statistical learning to generate refined prediction models capable of optimizing discriminatory power. Currently, the GARD platform has been used to develop assays for hazard assessment of skin sensitizers, respiratory sensitizers, protein allergens, and respiratory irritants.

This thesis aims to describe the technical aspects of the GARD platform, including the technologies that it incorporates and discuss how they contribute to its performance. Further, two GARD assays will be presented in more detail. GARDskin was developed as an *in vitro* assay for hazard identification of skin sensitizers and papers I and II describe performance measures from two separate validation experiments; (I) an experiment where 72 chemicals were assayed blindly and (II) a ring trial experiment designed to assess predictive performance and reproducibility. The second assay is GARDpotency, which was developed for potency assessment of skin sensitizers. Paper III presents results from a GARDpotency ring trial experiment where predictive performance and reproducibility were examined. Finally, this thesis includes paper IV, which describes a novel normalization method that was designed for adjustment of batch effects that arise in predictive settings.

2 Hazard assessment of skin sensitizers

2.1 Allergic contact dermatitis

2.1.1 Overview

Allergic contact dermatitis (ACD) is a delayed type hypersensitivity reaction that occurs in the skin upon repeated exposure to chemicals known as skin sensitizers (Kimber et al. 2002). The prevalence of ACD is high and figures close to 20% are typically reported to describe the fraction of individuals that are sensitized to chemicals (Alinaghi et al. 2019). Skin sensitizers can be found in common items such as personal care or household products, including fragrances, soaps, colouring dyes, jewellery, and cleaning agents (Jongeneel et al. 2018; Linauskienė et al. 2017; Minamoto 2010). ACD is also a common occupational health disease, where individuals within industries such as health care, cosmetology, cleaning, and certain types of manufacturing are at increased risk of becoming sensitized (Pacheco 2018). The symptoms associated with ACD include eczema and erythema, which typically occur at the site of exposure (Esser and Martin 2017), though affected areas can also extend beyond the exposure site (Martin et al. 2011). The acquisition of skin sensitization occurs in two phases: a sensitization phase and an elicitation phase. The sensitization phase occurs first and is a pre-requisite for a subsequent elicitation. During sensitization, a chemical skin sensitizer induces the engagement of both the innate and the adaptive immune system, leading to the generation of allergen-specific memory T-cells. The elicitation phase ensues upon repeated exposure to the same chemical and is characterized by effects induced by the previously generated memory T-cells that recognize the allergen, eventually giving rise to the associated symptoms.

ACD is a widespread ailment that is accompanied with a set of distressing symptoms. Further, once an individual has become sensitized to a chemical, avoidance from exposure to the compound is necessary since the condition cannot be cured. These issues make the objective of creating methods capable of accurately identifying chemicals with the potential to induce skin sensitization important, since it can proactively aid in the reduction of hazardous exposure.

2.1.2 The skin

The skin is a multifaceted organ responsible for providing protection against exogenous factors and is thus an important component when attempting to gain an understanding of the mechanisms influencing the induction of ACD. The skin comprises the initial obstacle that a potential skin sensitizer needs to penetrate to induce sensitization. In addition, it also constitutes the site where several important events transpire during both the sensitization and the elicitation phase.

The skin is typically divided into three distinct layers, the hypodermal-, the dermal-, and the epidermal layer (Agarwal and Krishnamurthy 2020). The epidermal layer is the outermost layer of the skin and provides the main barrier function against external agents (Boer et al. 2016). The epidermis can be further divided into sublayers, each with specific properties. The stratum corneum is the outward facing layer of the epidermis (Baroni et al. 2012). It is comprised of multiple layers of flattened corneocytes that are interspersed in a lipid-rich intercellular matrix (Proksch et al. 2008). Corneocytes are terminally differentiated keratinocytes that have lost their nucleus via a form of programmed cell death and mainly contain keratin and filaggrin enclosed within a cornified envelope (Proksch et al. 2008). The stratum corneum forms a hydrophobic protective layer that a potential skin sensitizer must penetrate to induce sensitization. Indeed, it has been generally believed that skin sensitizers must be comprised of small chemicals with molecular weights below 500 Daltons and with logKow coefficients above 1 (Bos and Meinardi 2000; Gerberick et al. 2004a; Smith Pease et al. 2003). However, recent evidence has rebutted such stringent thresholds for sensitizing chemicals, and examples of chemicals exceeding either limit have been identified (Fitzpatrick et al. 2017a; 2017b; Roberts et al. 2013). Further, evidence of sensitization towards proteins via route of skin exposure has also been observed (Izadi et al. 2015), suggesting that most compounds with an inherent ability to induce a sensitization reaction will probably do so in some cases. Additional layers of the epidermis are collectively referred to as the nucleated- or viable epidermis. These layers also mainly consist of keratinocytes, though other cell types can also be found. For example, Langerhans cells are present in the viable epidermis (Jaitley and

Saraswathi 2012) and are of interest when studying sensitization due to their functionality as a professional antigen presenting cell (APC). Indeed, Langerhans cells possess the ability to migrate from the epidermis into local lymph nodes to present sampled and processed antigens. However, their precise role in skin sensitization is still ambiguous (Deckers et al. 2018).

The dermis is located underneath the epidermis and contains high levels of collagen, blood- and lymphatic vessels, and is host to a larger number of cell types compared to the epidermis (Tsepkolenko et al. 2019). These cells include fibroblasts, dendritic cells, macrophages, and mast cells (Tsepkolenko et al. 2019). Further, additional cell types, such as neutrophils and T-lymphocytes, can migrate into the tissue during conditions where e.g. inflammatory and chemotactic molecules are secreted, which arise during exposure to skin sensitizers.

Finally, the hypodermal layer is located beneath the dermis and mainly consists of adipose tissues (Fenner and Clark 2016). It forms a protective and isolating layer (Fenner and Clark 2016) but is of limited relevance for understanding the mechanisms of ACD.

2.1.3 Mechanisms of skin sensitization

The major mechanisms underlying the sensitization phase of ACD are generally agreed upon and an adverse outcome pathway (AOP) describing these events has been defined (OECD 2014). The AOP contains four key events, which depict the known steps that are required for sensitization to transpire. The first key event (KE1), or the molecular initiating event, describes the covalent binding of a low molecular weight chemical, or hapten, to skin-residing proteins, forming an immunogenic protein-hapten complex. The second key event (KE2) describes the activation of keratinocytes, giving rise to the production and the release of inflammatory and chemotactic mediators. The third key event (KE3) consists of dendritic cell activation and maturation. During this step, dermal or epidermal dendritic cells become activated and mature, which leads to their migration from the skin tissue to draining local lymph nodes where they prime naïve T-cells. The fourth and final key event (KE4) of the AOP describes the activation, differentiation, and proliferation of T-cells, resulting in the generation of a repertoire of allergen-specific memory T-cells that can recognize the specific chemical upon renewed exposure (OECD 2014). However, even though a general understanding of the processes required for skin sensitization is established, there are still many details that remain elusive.

In order for sensitization to occur, the skin sensitizer needs to be able to engage both the innate and the adaptive immune system (Kimber et al. 2002). As a prerequisite for this, given that chemical sensitizers are generally too small to be immunologically active in themselves, the chemicals must typically be able to react with proteins and form conjugates. Most skin sensitizers are electrophilic compounds that can form covalent bonds with skin-residing proteins (Karlberg et al. 2008). However, some sensitizers do not initially possess this reactive property but can acquire it via either biotic activation (pro-haptens), which requires the metabolic machinery of the cells, or abiotic (pre-haptens) activation (Karlberg et al. 2013). Mechanisms that rely on the organic anion transport polypeptide family for transportation of pro-haptens into the cells, thereby enabling access to metabolic enzymes such as the cytochrome P450 enzymes, have been suggested (Martin 2012; Schiffer et al. 2003). Further, it has also been suggested that some pro-haptens can activate the aryl-hydrocarbon receptor, which upregulates the expression of multidrug resistance proteins that could potentially facilitate release of the formed reactive metabolites (Martin 2012). In fact, the aryl hydrocarbon receptor pathway was identified as significantly affected during the discovery experiments of GARDskin (Johansson et al. 2011).

It has been shown that the activation of the innate immune system depends on the chemicals' capacity to generate endogenous danger signals, which usually arise as a product of their irritancy properties (Martin et al. 2011). Some of these danger signals, referred to as danger-associated molecular patterns (DAMPs), can interact with innate pattern recognition receptors (PRRs), which are present on multiple cell types including dendritic cells, macrophages, and keratinocytes, and trigger downstream signalling cascades. Common DAMPs include reactive oxygen species (ROS), low-molecular weight hyaluronic acid (HA) fragments, biglycan, and extracellular ATP (Ainscough et al. 2013). The mechanisms by which skin sensitizers induce DAMP formation are diverse and not fully recognized. For example, the mechanism by which ROS is generated is not completely established, but it has been hypothesized that skin sensitizers could have a direct effect on the cells' antioxidant defence systems, i.e. abrogation of its function by depletion of intracellular levels of glutathione, which could facilitate the accumulation of ROS (Ferreira et al. 2018). Once formed however, the presence of ROS can contribute to the degradation of the extracellular matrix, giving rise to the production of additional DAMPs such as low-molecular weight HA-fragments and potentially also biglycan (Esser et al. 2012). Both of these DAMPs have been shown to function as endogenous ligands for TLR2 and TLR4 (Esser and Martin 2020; Moreth et al. 2014; Schaefer et al. 2005), which induce downstream signalling cascades via e.g.

nuclear factor kB (NF-κB), and can lead to the production of pro-inflammatory molecules such as pro-IL1B and pro-IL18. Increased extracellular levels of ATP is another recognized DAMP, which is thought to arise as a consequence of the irritant effect of skin sensitizers. Increased extracellular concentrations of ATP, which can be secreted by damaged cells, can be detected by the P2X7 receptor (Savio et al. 2018). This interaction activates the NLRP3 inflammasome, which in turn activates caspase-1 that cleaves the pro-inflammatory molecules pro-IL-1B and pro-IL-18 to generate their active forms (Cassel and Sutterwala 2010; Di Virgilio et al. 2017).

Another well-recognized pathway that has been shown to be activated by exposure to chemical skin sensitizers is the nuclear factor erythroid 2 related factor (NRF2) pathway, which constitutes a conserved cytoprotective mechanism (Baird and Dinkova-Kostova 2011). NRF2 activation can be attributed to the protein-reactive property of skin sensitizers (Helou et al. 2019). In steady state, the repressor protein Kelch ECH associating protein 1 (KEAP1) promotes the degradation of the transcription factor NRF2 via the ubiquitin proteasome system (Villeneuve et al. 2010). However, several accessible cysteine residues on the KEAP1 protein are sensitive to electrophilic attack, making it a potential target for modification by skin sensitizers (Natsch 2009). The reaction between skin sensitizers and the keap1 protein disables its repressor functionality, enabling NRF2 to translocate and accumulate in the nucleus. There it associates with MAF-proteins and initiates transcription of several genes containing an antioxidant response element (ARE) domain in their promotor sequences, including NAD(P)H quinone oxidoreductase 1 (nqo1), heme-oxygenase 1 (hmox1 or ho-1), superoxide dismutase (sod) and thioredoxin reductase 1 (txnrd1) (Helou et al. 2019; Natsch 2009). Interestingly, nqo1, hmox1 and txnrd1 are all present in the GARDskin prediction signature (Johansson et al. 2011). Activation of the NRF2 pathway has many downstream effects that regulate the inflammatory progression and occurs in multiple cell types during sensitization, including keratinocytes and dendritic cells. Its activation in keratinocytes are associated with improved proliferation and differentiation whereas it downregulates activation and maturation in dendritic cells (Helou et al. 2019).

Following activation of dendritic cells, which is dependent on the generation of DAMPs and the succeeding interaction with PRRs, the phenotype of the cells is altered, which includes the upregulation of chemokine receptors, upregulation of major histocompatibility complex (MHC) II, upregulation of costimulatory molecules such as CD80 and CD86, and secretion of interleukins such as IL-12 and IL-6 (Blanco et al. 2008; Ryan et al. 2007). Activated dendritic cells migrate to local lymph nodes where the sampled and processed antigens are presented to naïve T-lymphocytes (Joffre et al. 2009). A currently unresolved mechanism of the sensitization phase relates to the distinct differences in T-cell polarization that are observed between the responses induced by skin sensitizers and respiratory sensitizers. While both skin- and respiratory sensitizers share several attributes, such as protein reactivity and irritancy properties, they seem able to induce distinct T-cell populations (Kimber et al. 2014). Skin sensitizers are mainly thought to induce Th1/Th17 and Tc1/Tc17 polarization, whereas respiratory sensitizers mainly induce Th2 and Tc2 polarization (Kimber et al. 2018; Martin 2012; Sullivan et al. 2017). The T-cell polarization is dependent on the characteristics of the inflammatory environment and the presence of specific cytokines. For example, it is recognized that polarization towards type 1 T-cells is induced by e.g. IL-12 and IFN-γ and towards type 2 T-cells by e.g. IL-4, cytokines which can be secreted by cells such as dendritic cells (Paul 2013). However, the detailed molecular mechanisms responsible for inducing the skewness in T-cell polarization are still not well-established.

2.2 Test methods for assessment of skin sensitizers

2.2.1 *In vivo* methods

Traditionally, *in vivo* methods have been utilized to perform hazard assessment of skin sensitizers, and they are still of relevance (Daniel et al. 2018). Of these, guinea pig assays were among the earliest to be regularly used, some of which were initially proposed more than 50 years ago (Buehler 1965). Two notable guinea pig assays are the Buehler test (Buehler 1965) and the guinea pig maximization test (Magnusson and Kligman 1969). Both methods evaluate the skin sensitizing potential of chemicals by studying the elicitation phase of the condition. The Buehler test performs the induction phase (i.e. sensitization) by topical application of the test substance. In contrast, the guinea pig maximization test performs the sensitization by intradermal injections with or without a Freund’s adjuvant and occluded topical application of the test chemical. Both assays assess the elicitation reaction by the application of closed-patch tests (OECD 1992). However, today, they are typically considered superseded by the local lymph node assay (LLNA; (Dean et al. 2001)).

The LLNA shifts the evaluation of the endpoint to the sensitization phase by measuring the proliferation of T-cells in local lymph nodes following topical application of a test chemical (Kimber and Basketter 1992; Kimber et al. 1994). The LLNA provided a step in the direction of the principles of the 3Rs by acting as a refinement of existing methodologies while also providing a more unbiased detection of positive test cases. The assay is performed by topically exposing mice to a test chemical. Following repeated exposures, draining lymph nodes are excised and the proliferation of T-cells are measured and expressed as a stimulation index (SI; (Frank Gerberick et al. 2007)). The SI is defined as a proliferation ratio where challenged samples are compared to vehicle controls, and a test is generally considered positive when the SI is above 3 ($SI \geq 3$). Another advantage of the LLNA, which is also one of the attributes that enforces its current relevance, is that it allows for the assessment of chemicals' relative potencies (Frank Gerberick et al. 2007). The relative potency of chemicals can be generated from the LLNA by running the assay in a dose-dependent manner. The dose-response relationship (response is represented by SI) can then be examined to estimate a minimum dose capable of rendering a positive outcome, i.e. an EC3 value. The attained EC3 value reflects the chemical's potency, which can be informative for risk assessment.

2.2.2 Non-animal alternatives

A surge in the development of non-animal methods have been observed over the last decades with the aim of eventually replacing the need for *in vivo* methods (Ezendam et al. 2016). Today, several assays have reached the sophistication and level of performance that are required to become successfully formally validated and regulatory accepted for the hazard identification of skin sensitizers. These assays can be mapped to the key events in the AOP by their test principle, and the validated methods currently comprise the direct peptide reactivity assay (DPRA; (Gerberick et al. 2004b; Gerberick et al. 2007)) and the amino acid derivative reactivity assay (ADRA; (Fujita et al. 2014; Yamamoto et al. 2015)) for KE1, KeratinoSens (Emter et al. 2010) and LuSens (Ramirez et al. 2014) for KE2, the human cell line activation test (h-CLAT; (Ashikaga et al. 2006; Sakaguchi et al. 2006)), the U937 cell line activation test (U-SENS; (Piroird et al. 2015)) and the interleukin-8 reporter gene assay (IL-8 Luc assay; (Takahashi et al. 2011)) for KE3. However, it is generally believed that none of the assays are in themselves sufficiently informative for adequately representing the complete system partaking in the processes of skin sensitization. Therefore, it is often recommended that several non-animal methods should be combined to ensure that several key events are queried for hazard assessment (Kleinstreuer et al. 2018; Strickland et al. 2016). Of note, the validated assays are only recommended for hazard identification of skin sensitizers, i.e. classification of chemicals as either skin sensitizers or non-sensitizers, and cannot be readily employed for potency assessment (Barentsen et al. 2019).

Examining the different methods, DPRA and ADRA are both in chemico methods that were developed to assess the protein reactivity of test chemicals (OECD 2020b). For DPRA, synthetic heptapeptides containing cysteine or lysine residues are incubated with a test chemical, following which the amount of peptide depletion is measured using an HPLC system (Gerberick et al. 2004b; Gerberick et al. 2007). The evaluation procedure for the more recent assay ADRA is similar but it evaluates residual concentrations of the cysteine derivative NAC (N-(2-(1-naphthyl)acetyl)-L-cysteine) and the lysine derivative NAL (α-N-(2-(1-naphthyl)acetyl)-L-lysine) post incubation, also using HPLC (Fujita et al. 2014). ADRA was proposed as an alternative to DRPA to mitigate certain limitations, including the oxidative sensitivity of the heptapeptides and the requirement for using high concentrations of test chemicals to enable detection of peptide depletion, which makes it difficult to assess chemicals with low solubility (Fujita et al. 2014; OECD 2020b; Yamamoto et al. 2015).

The two methods assigned to KE2, i.e. keratinocyte activation, both monitor the activation of the NRF2 pathway by the utilization of a luciferase reporter gene (OECD 2018a). Following NRF2 activation, which as described can follow from KEAP1 incapacitation due to the interaction between protein reactive skin sensitizers and a number of accessible cysteine residues on KEAP1, the reporter luciferase gene is transcribed since it contains ARE-domains in its promotor sequence (OECD 2018a). Produced luciferase can then be detected using luminescence analysis.

Finally, for KE3, U-SENS and h-CLAT evaluate dendritic cell activation and maturation by monitoring the expression of dendritic cell maturation markers using flow cytometry (OECD 2018b). Specifically, U-SENS measures the expression of the costimulatory molecule CD86 (Piroird et al. 2015), and h-CLAT measures CD86 and the intercellular adhesion molecule CD54 (Ashikaga et al. 2006; Sakaguchi et al. 2006). Identification of skin sensitizers are made when a significant (above pre-specified thresholds) upregulation of the surface markers is observed following chemical exposure (OECD 2018b). The IL-8 Luc assay uses a THP-1-derived cell line (same as h-CLAT) but monitors IL-8 expression following exposure to a test substance (Kimura et al. 2015; Takahashi et al. 2011). IL-8 is a chemotactic factor and cytokine that has been identified as a dendritic cell maturation marker following exposure to skin sensitizers (Toebak et al. 2006). It has been described as transcriptionally regulated by several transcription factors including NF-κB and AP-1 (Hoffmann et al. 2002) and post-translationally by NRF2 (Zhang et al. 2005). The IL-8 Luc assay uses a luciferase reporter gene that is under regulatory control of IL-8 promotors (Takahashi et al. 2011).

In addition to the individual assays, several approaches with varying complexity for combining the outcomes of two or more assays have been examined and proposed, and guidelines for regulatory acceptance of defined approaches (DA) are currently under investigation (Kolle et al. 2020).

2.2.3 GARDskin

The GARDskin assay was the first assay to be developed using the principles of the GARD platform (Johansson et al. 2011). In fact, it was the creation of the assay that generated the main procedural steps that are currently associated with GARD. Therefore, it has also reached the furthest in terms of external validation and regulatory acceptance of the available GARD assays. Paper I describes the results from a performance assessment comprising the evaluation of 72 blinded test chemicals, and Paper II describes the results from the ring trial study that was performed to generate the results required for submission to be formally reviewed and validated by the European Centre for Validation of Alternative Methods (ECVAM).

The development and implementation of GARDskin followed the methodology currently associated with GARD. An initial discovery study was carried out and genes capable of providing information that were predictive of skin sensitizing hazard were identified. Prior to the initiation of the experiment, a cell line was selected that was deemed suitable for the problem at hand, i.e. discernment between non-sensitizers and skin sensitizers. Given the current understanding of the sensitization phase, a dendritic-like cell line was selected to enable modelling of the dendritic cell activation and maturation step, which is key to link the innate response to the subsequent generation of allergen-specific T-lymphocytes (Benvenuti 2016). Next, the set of reference chemicals destined to form the training dataset was selected. Attention was made to ensure that the set of chemicals was evenly balanced, i.e. that the dataset comprised a similar number of skin sensitizers and non-sensitizers. Further, the reactivity mechanisms of the included chemicals were examined to avoid selection bias, and it was made sure that both pre-haptens and pro-haptens (ethylenediamine, resorcinol, 1,4-phenylenediamine, 2-aminophenol, eugenol) were included among the skin sensitizers. In addition, skin sensitizers were selected to include a wide variety of potencies, ranging from relatively weak compounds such as resorcinol or hexylcinnamic aldehyde to relatively strong sensitizers such as p-phenylenediamine or dinitrochlorobenzene. Finally, of the non-sensitizers, chemicals with known irritancy properties (e.g. tween 80, sodium dodecyl sulfate, octanoic acid, and phenol) were selected to reduce the risk of introducing confoundment among the experimental sources of variation in the dataset.

Transcriptomic profiling was then conducted, using microarray analysis on samples acquired from cellular exposure experiments. Quantified data was mined for predictive genes and, after having combined a filter method based on the significance of evidence of differential expression with a backward elimination method to minimize the resampling error based on the Kullback-Leibler divergence, a biomarker signature comprising 200 genes was identified (Johansson et al. 2011). Examining the genes in the signature, the relevance of some of the entities could be verified based on already established knowledge of the mechanisms associated with the sensitization phase. For example, cd86 was one of the genes that had a well-established link to dendritic cell activation and maturation, since it constitutes a co-stimulatory signal required for T-cell activation. Indeed, as already discussed, CD86 expression comprise an important biomarker also for other non-animal methods (Ashikaga et al. 2006; Piroird et al. 2015; Sakaguchi et al. 2006). In addition to cd86, other examples include genes associated with NRF2 activation, such as nqo1, hmox1, and txrnd1. However, due to the nature of the experimental design, i.e. hypothesis generating, several other genes without clear prior implication in the sensitization mechanisms were also selected. In addition to studying individual genes, pathway analysis on the transcriptomic data was performed, and several pathways reflecting xenobiotic response mechanisms were identified (including the aryl hydrocarbon receptor pathway previously mentioned), which are relevant when considering the cellular perturbations used to generate the data.

Thus, following establishment of the biomarker signature and initial internal validation of the genes' predictive performance, the assay acquisition technology was transferred to the NanoString platform, the data pre-processing steps were optimized and defined, and it was shown that the assay was able to retain its discriminatory proficiency (Forreryd et al. 2016). Paper I describes the first extensive validation, which was performed with the aid of an external party, of the assay's performance on a set of blinded chemicals following the transfer to the NanoString platform, and paper II describes the ring trial that was performed in accordance with the OECD guidance documents to generate results required for submission to OECD for formal validation of GARDskin.

Examining the predictive performance estimates obtained in the aforementioned papers, the GARDskin assay seems to stand as an efficient assay for hazard identification of skin sensitizers. In fact, the performance of GARDskin was recently compared to the, at the time, formally validated assays and the results indicated that the GARDskin assay performed favourably with generally higher performance figures (Roberts 2018). In addition, the assay was successfully transferred to two external laboratories and the reproducibility measures estimated from the ring trial were found to be similar to those of already validated assays (OECD 2018a; 2018b; 2020b), suggesting that the experimental protocols and the processing pipeline enable robust assessment of test chemicals.

2.2.4 GARDpotency

The pursuit of non-animal methods for hazard identification of skin sensitizers has led to the proposal and the development of several assays (OECD 2018a; 2018b; 2020b), as already described. Despite this progress, there is a much smaller number of non-animal assays that have been developed for the purpose of hazard characterization (Ezendam et al. 2016), i.e. classification of compounds' relative potencies. Because of this, *in vivo* methods are still required and the LLNA is currently recommended for generating potency information when other data sources are lacking, when assessing chemicals for the purpose of regulatory registration (Daniel et al. 2018). This makes the development of non-animal alternatives for potency assessment an important objective.

The development and the design of GARDpotency emerged, in part, from observations made on data acquired during the discovery study of GARDskin. It was found in an extended data analysis that strong sensitizers tended to induce engagement of a larger number of pathways and signalling molecules compared to the relatively weak skin sensitizers (Albrekt et al. 2014), which gave rise to the hypothesis that this mechanism could potentially be exploited to develop an assay for potency assessment.

The development of GARDpotency slightly strayed from the general procedures of the GARD platform. Specifically, the classification endpoint was adjusted between the discovery study and the final definition of the assay on the NanoString platform. The initial goal, by which the discovery study was designed, was directed towards the development of an assay that could discriminate between compounds of different potencies as defined by the three potency categories of the Classification Labelling and Packaging regulation (CLP), i.e. non-sensitizers (CLP category: No Cat), weak skin sensitizers (CLP category: 1B), and strong skin sensitizers (CLP category: 1A) (Zeller et al. 2017). To this end, a set of 86 chemicals were collected, while being careful to consider the chemical's properties as described above for GARDskin, and their induced transcriptomic profiles were acquired using microarray analysis. The identification of the genetic biomarker signature was performed using a backward elimination algorithm that was based on the random forest classification algorithm (Diaz-Uriarte 2007; Díaz-Uriarte and Alvarez de Andrés 2006). Briefly, the genes were initially ranked by variable importance scores of a random forest model that was trained using all available genes. Then, the lowest ranking genes were recursively removed while the predictive performance was monitored. Eventually, an optimal performance was observed when 52 genes were retained in the prediction signature. The performance of the proposed model was assessed on a previously unseen test set comprising 18 chemicals resulting in an estimated classification accuracy of 78%, which was also very similar to the resampling error rates observed during model definition (Zeller et al. 2018).

Paper III describes the work that was performed following the discovery study, which was aimed towards finalizing the definition of the assay and to assess its performance in terms of predictability and reproducibility. During the technological transfer, it was decided that the modelled endpoint would be altered to remove the apparent redundancy of modelling the No Cat category given the existence of GARDskin, which would potentially also simplify the predictive modelling by allowing the discrimination between two categories instead of three. Therefore, a tiered approach was suggested as an alternative to the initial design, where a test chemical with unknown skin sensitizing hazard property would first be tested in the GARDskin assay to classify it as either a non-sensitizer or a skin sensitizer. Whereas non-sensitizers would be given the class label No Cat without further testing, sensitizers would enter the second tier comprised of assessment in the GARDpotency assay to classify it as either a weak skin sensitizer or a strong skin sensitizer. Importantly, the combination of the two assays would be simple since the assays share experimental protocols and the same RNA sample could be analysed twice, i.e. only one experiment with cellular exposures would be required.

Gene expression levels of samples in the training set were acquired on the NanoString platform and the signature was deemed to have retained its predictive capability, displaying significant ability to separate the weak from the strong skin sensitizers in the dataset using both unsupervised dimensionality reduction techniques and resampling methods. However, another optimization was implemented during this stage of the development, which consisted in the incorporation of an additional predictor, i.e. a feature describing the chemicals' exposure concentrations. Prior to the development of GARDpotency, it had already been observed that the chemicals' exposure concentrations seemed to carry information of relevance for potency assessment. More specifically, when considering data from historical analyses in GARDskin, it was noticed that strong sensitizers tended to be assayed at lower concentrations compared to relatively weaker sensitizers (Johansson et al. 2017). Thus, with this information at hand, and with the familiar association between exposure amount and potency (e.g. EC3 values in the LLNA), it was added as an additional predictor for the modelling. Finally, a prediction model was defined using a support vector machine (SVM) and the assay was considered ready for a more rigorous performance assessment.

A ring trial experiment was carried out to assess the performance of the GARDpotency assay and of the tiered approach comprising the combination of GARDskin and GARDpotency. The results revealed predictive accuracies that ranged between 76.5% to 94.4% between the participating laboratories when attempting to discriminate between weak and strong skin sensitizers (i.e. GARDpotency only), and between 75.0% and 92.6% for the tiered approach over the three CLP categories. The within laboratory reproducibility ranged between 62.5% to 88.9% and the between laboratory reproducibility was estimated to 61.1%. In conclusion, the performance estimates suggest that the GARDpotency assay could provide a useful tool for hazard characterization of skin sensitizers, and that potential future optimizations could favourably aim to reduce discrepancies between experiments.

2.2.5 GARDskin dose-response

One additional GARD assay has been developed for hazard characterization of skin sensitizers. It can be considered as an extension to the existing GARDskin assay. However, the objective of the method is, in contrast to either GARDskin or GARDpotency, to provide a quantitative endpoint measurement that can be used to infer the relative potency of individual chemicals' potency. The proposed assay, termed GARDskin dose-response, is based on the protocols of GARDskin but performs the analysis in a concentration-dependent manner. The end goal of the analysis is to estimate the smallest chemical-specific concentration capable of rendering a positive GARDskin prediction. This approach is in line with common toxicological practices and share the methodological principles employed by the LLNA to estimate EC3-values (Frank Gerberick et al. 2007).

The development of the GARDskin dose-response assay comprised both the evaluation of the endpoint's relevance and the adaptation of GARDskin's protocol to enable efficient concentration-dependent acquisition. The protocol optimization follows from the need to make the assay viable in terms of time and cost expenditures that are associated with the experimental procedure. To facilitate the analysis of both queries, a dose-response experiment comprising approximately 30 chemicals of varying potency categorization were designed. For each chemical, several concentrations were run and GARDskin predictions were generated. It was found that linear interpolation between concentration levels predicted on adjacent sides of the GARDskin decision border (i.e. the SVM's hyperplane) was effective for estimating the minimum concentration where a chemical could induce a positive prediction, henceforth referred to as cDV_0 concentrations. Further, comparisons between the estimated cDV_0 concentrations with the chemicals' expected potency measures, comprised of both LLNA EC3 values and human potency categories, showed statistically significant correlation. These results indicated that cDV_0 values, estimated from the GARDskin dose-response assay, could indeed be informative of chemicals' skin sensitizing potencies.

In conclusion, the GARDskin dose-response assay is a method proposed for hazard characterization of skin sensitizers. Estimated cDV_0 concentrations for the chemicals in the study showed significant correlation with existing potency properties, suggesting that the method could be a useful tool for generating information pertinent for e.g. subsequent risk assessments. Further, though additional data is required to substantiate the observed potency-associated information generated by the assay, it could provide a potential alternative to currently available *in vivo* assays. At the time of writing, details regarding the method, including the above-mentioned results, are being prepared in a separate manuscript.

3 The GARD platform

3.1 The principles of GARD

Today, the GARD platform encompasses a methodological framework for the development of toxicological assays. It has been designed via the ongoing application of ideas to solve aspects of, and contribute with information aimed towards aiding understanding of, relevant toxicological problems. The major procedural aspects of the framework were established during the development of the GARDskin assay, which was the first assay to be created and associated with the term GARD. During its development, many endeavours associated with the creation of novel experimental technologies were encountered and, in some regards, solved, which has allowed for the continuous forwarding of information, facilitating the development of novel assays targeting other toxicological endpoints, such as respiratory sensitization (Forreryd et al. 2015) or assessment of protein allergens (Zeller et al. 2018). Nevertheless, the field is vast and varied, and a multitude of opportunities remain to be explored, and the GARD platform could be a useful tool for such tasks.

As discussed previously, the overarching aim of the development of toxicological assays is to provide tools enabling the gathering of knowledge regarding the properties associated with a compound, that can be utilized for evaluating the risks and hazards attributable to it, thus facilitating the end goal of reducing health and environmental hazards. In such a sense, the major tasks involved in establishing an assay include the creation of a system that models relevant parts for the experimental question at hand. This system can take on a variety of appearances and range from models based on animals to purely computational systems. However, generally, the understanding and the complexity of the endpoint towards which a model is defined does, in part, dictate the form that a model can take. For example, this can be observed in the progression of toxicological assays for assessment of skin sensitizers, where test methods originated as *in vivo* models (Basketter et al. 2012), only to be refined and replaced as acquired knowledge was incorporated into novel methods that possess advantageous properties while enabling accurate hazard assessment of the same endpoint (Ezendam et al. 2016; Kleinstreuer et al. 2018).

The state of understanding of the biological mechanisms underlying the adverse outcomes associated with several complex conditions, such as chemical sensitization, are previously unsurpassed and has in numerous instances been described in some detail (Brys et al. 2020; Esser and Martin 2017; Shane et al. 2019; Silvestre et al. 2018). However, there are still many mechanisms that remain elusive and not completely understood, though great efforts have been made to elucidate them. Therefore, when developing assays towards hazard endpoints induced from complex underlying mechanisms, a certain amount of ignorance must be admitted, and the model system should preferably allow for such levels of ignorance, and perhaps simultaneously provide opportunities for improving and furthering the state of current understanding.

When developing assays in accordance with the GARD platform, the initial step is to select a cellular system that can be used to model the relevant biological aspect of the hazard. All GARD assays hitherto described has been *in vitro* assays aimed towards immunological endpoints, and a cell line acting as a surrogate for dendritic cells has been deemed the most appropriate, rationalized from the dendritic cell's central importance in these events. However, future assays do not necessarily need to be restricted to these types of cells but can be altered to best reflect the understanding of the examined endpoint. Having selected a system from which it is expected that relevant signals can be generated, the method for monitoring the induction of said endpoints should be established. Again, it is often the case that the precise mechanisms leading up to the induction of an outcome, or even that the precise signals that can be expected from the selected model, is not fully understood. Some understanding of the molecular mechanisms might be known, which has been successfully utilized to develop several assays (OECD 2018a; 2018b; 2020b), but it is possible that relevant information is still to be discerned. Therefore, the GARD assays have been developed by initially performing hypothesis generating discovery experiments on the examined cell system. For example, transcriptomic profiling has been employed to identify suitable genetic biomarkers that enable accurate discrimination between substances of different hazard properties (Forreryd et al. 2015; Johansson et al. 2011; Zeller et al. 2017; Zeller et al. 2018). Following identification of promising candidate biomarkers, i.e. genes, the development of GARD assays is generally finalized by the definition of a prediction model that can, based on the values of the observed biomarkers, classify an examined test chemical into appropriate hazard categories. For the current GARD assays, the prediction models were developed by using machine learning techniques, which have facilitated the establishment of relevant prediction rules without the need for manually inspecting the identified biomarkers and their intrinsically complex relationships (if even possible) for the purpose of identifying an optimal discriminatory prediction rule. However, the exact techniques used to derive the prediction models can be open for optimization to allow for models of varying complexity, depending on the nature of examined problem.

Thus, in general, a GARD assay is developed by the deliberate combination of a relevant *in vitro* test system with exploratory analysis and data mining, on data originating from high-throughput acquisition techniques, with machine learning assisted derivation of prediction rules for hazard classification.

3.2 The cellular system

As noted above, an initial step when designing a GARD assay is the selection of the cellular system, i.e. selection of a system from which necessary signals can be derived for interpretation and subsequent assessment of the toxicological endpoint of analysis. Hitherto, the available GARD assays have been designed with the SenzaCell cell line (ATCC Depository PTA-123875), which has been chosen to act as a surrogate for dendritic cells, a relevant cell type for studying the induction of chemical hypersensitivity and protein allergenicity (Steinman and Hemmi 2006). But as also discussed, the GARD framework does not enforce any particular restrictions on the cell line that is utilized as the cellular system in an assay, and it seems plausible that future versions of the assay could incorporate other cell types that better mimic the *in vivo* system under investigation.

3.3 The acquisition technologies

Given that an appropriate cell system, assumed capable of generating signals informative for evaluating the hazard properties of interest, has been selected, the technology used for the monitoring of these signals must be determined. For the GARD assays, the most appealing technologies have been those based on transcriptomic analysis, and microarrays have been readily applied for the task of quantifying the genetic perturbations induced by the stimulating agents during the discovery phase of the experiments (Forreryd et al. 2015; Johansson et al. 2011; Zeller et al. 2017; Zeller et al. 2018). The reasons for targeting the transcriptome when evaluating monitoring technologies have been many. For example, the methods available for querying the expression levels of genes of cell suspensions has reached a state of maturity that enables consistent acquisition (Shi et al. 2006; Su et al. 2014). Further, obtained data offers a generally unbiased view of the transcriptome, i.e. the data does not strongly depend on a limited selection of targets defined from existing knowledge of the examined condition. And finally, methods for processing and analysing generated data are widely available, leveraging state-of-the-art methods for tackling common obstacles encountered with high-dimensional data, such as multiple hypothesis testing from limited sample sizes (Love et al. 2014; Ritchie et al. 2015). However, though the advantages are several, some difficulties are also evident. The cost of quantifying transcription levels of a sample has decreased but it is still noticeable, which limits the number of samples that can be examined in an experiment (Wheelan et al. 2008; Xiong et al. 2017). Further, depending on the acquisition technique, the obtained data must be adequately pre-processed and normalized to permit analysis (Li et al. 2014). Finally, the major technologies available for transcriptomic analysis has been observed to be somewhat sensitive to the inclusion of non-biological variance, originating from e.g. differences in experimental parameters, in the data (Goh et al. 2017). This can increase the risk of introducing potential confounders, emphasizing the importance of careful experimental design in these studies. Despite these potential drawbacks of the transcriptomic technologies, which in many scenarios can be mitigated, the potential insights that could be extracted from the acquired data makes it the preferred methodological option of the GARD platform.

The earliest technology that enabled widespread analysis of thousands of transcriptomic targets were the microarrays (Bumgarner 2013), which has been frequently used during the development of the GARD assays (Forreryd et al. 2015; Johansson et al. 2011; Zeller et al. 2017; Zeller et al. 2018). Microarrays consist of small chips onto which oligonucleotides that are complementary to known gene sequences have been attached. RNA samples to be analysed are first converted to complementary DNA (cDNA), which is then typically labelled and allowed to hybridize with the oligonucleotides. Following hybridization, a signal representing the amount of bound cDNA can be estimated (Govindarajan et al. 2012). Currently though, the microarray technology is often considered surpassed by the more recent method of RNA sequencing (RNAseq), which has been shown to provide better dynamic range, be better equipped to detect low-abundance genes, and to be more efficient at discriminating between different isoforms (Zhao et al. 2014). Instead of printing pre-specified oligonucleotides to a chip, RNAseq allows for parallelized high-throughput sequencing of nucleotides in an examined sample. Processed sequence data can then be used to infer the expression levels. Whereas the quantification of transcript levels in the microarrays are performed by studying intensity levels on the gene chip, expression levels from RNAseq can be established by mapping the obtained sequences of nucleotides to a suitable reference genome (Wang et al. 2009). Thus, quantified microarray data consist of gene-specific intensity levels and RNAseq data comprise gene-specific discrete counts.

Though both microarrays and RNAseq can be used to study similar endpoints, the analytical pipelines employed are fairly different. For microarrays, the output from an analysis comprises expression values represented by signal intensities. The observed expression values for individual probes are often normalized and summarized into gene expression representations, since several probes can target different sites of the same gene. Having acquired the gene expression values, microarray data can generally be modelled using common statistical methods, since it is assumed that the data can be modelled as normally distributed. In contrast, the pre-processing of RNAseq can be a bit more diverse and computationally more intensive. The output from the analysis is typically a sequencing file containing nucleotide calls for individual sequences with their associated certainty scores. Several approaches can be employed to generate gene expression values from the generated sequences. Initial steps can include quality assessments, adaptor sequence trimming, and quality trimming to remove low-quality nucleotide calls (Conesa et al. 2016a). Depending on the pipeline, the quantification of gene expression can be obtained by first mapping acquired reads to a suitable reference genome, followed by counting of the number of reads that map to known genes. However, these methods are considered fairly computationally intensive, and recently, alternative methods have been designed for RNA quantification that are faster and requires significantly less memory (Bray et al. 2016; Patro et al. 2017). Independent of pipeline, once expression levels have been obtained, the methods used in downstream analysis also usually differ compared to those used for microarrays. This is because the data can no longer be assumed to be adequately modelled by the normal distribution (Li et al. 2012). However, despite this, several popular methods have been developed to account for this. For example, methods have been designed to model the counts using alternative distribution assumptions (Love et al. 2014; Robinson et al. 2010), while other employ suitable transformations to the data to justify the subsequent assumption of modelling it as normally distributed (Law et al. 2014; Ritchie et al. 2015).

3.3.1 Extracting biological understanding

Due to the nature of the data obtained from transcriptomic profiling experiments, there is a possibility to extract information that could provide valuable information of the mechanisms underlying the condition of examination. Further, the information generated from such experiments does not only have to contribute with new knowledge to the field but can also be used to enhance the validity of the experimental approach by confirming *a priori* established findings. For example, during the development of the GARDskin assay, though 200 genes were included in the final prediction signature, several of the identified targets had previously been implicated in the mechanistic pathways associated with skin sensitization (Johansson et al. 2011).

The most common method for identifying genes or transcripts that could be of particular interest in a study is the assessment of differentially expressed genes. This can be achieved using a wide variety of methods but one of the most common approaches comprise the modelling of the observed expression levels using linear regression (Ritchie et al. 2015). By wrapping the procedure of differential expression analysis within the framework of linear regression analysis, efficient incorporation of e.g. covariates, which can be comprised of experimental factors capable of inducing variance in the observed data such as cell batch or reagent factors, can be made. This can increase the power to detect biological differences of interest (Leek and Storey 2007). Additionally, linear models are flexible and can be used to handle certain departures from the assumptions generally made by several statistical methods. For example, in some situations, experimental factors can be nested within the groups of other biologically interesting groups, which could potentially invalidate the assumption of independence. However, this issue can be mitigated when applying linear regression techniques to the experimental data by allowing for the incorporation of factors as random effects (Hoffman and Roussos 2020; Yu et al. 2019). In addition to these advantages, several statistical methods have been developed on the linear regression framework specifically for the analysis of transcriptomic data, which is often characterized by containing thousands of measured variables while consisting of a limited number of samples. These methods, with the most prominent probably being limma (Ritchie et al. 2015), leverage Bayesian statistics to borrow information between genes, making estimates required for the statistical inference more stable while also increasing the effective degrees of freedom in the statistical comparisons (Ritchie et al. 2015).

While differential expression analysis is often the first step that is performed when aiming to extract biological information from a transcriptomic experiment, obtained results can be difficult to interpret since the number of identified genes can be very large. Therefore, results are often used as input for subsequent analyses intended for generating a higher-level understanding of the perturbations induced by the experimental conditions, by e.g. examining pathway engagement. Several methods with varying complexity have been proposed for examining the effects on known pathways or gene sets (Beißbarth and Speed 2004; Castillo-Davis and Hartl 2003; Mootha et al. 2003; Subramanian et al. 2005; Tarca et al. 2009; Wu and Smyth 2012). A simple but common method that utilize the list of identified genes for this purpose is the overrepresentation analysis (Beißbarth and Speed 2004; Castillo-Davis and Hartl 2003). Here, the enrichment of genes associated with particular biological mechanism is assessed by evaluating the statistical significance of the overlap between the differentially expressed genes and the gene set, compared to what could be expected by chance, which can be tested by e.g. applying suitable statistical tests such as the hypergeometric test (Castillo-Davis and Hartl 2003). Another type of analysis that have been widely employed is the gene set enrichment analysis (GSEA; (Mootha et al. 2003; Subramanian et al. 2005)). Instead of using the differentially expressed genes as input, the method requires gene identifiers with an associated value reflective of the evidence of the genes' quantified differential expression, which is usually comprised of some signal to noise measure (Subramanian et al. 2005; Zyla et al. 2017). Genes are then ranked by their associated numerical values and the significance of the ranked position of genes in specific gene sets can be assessed. Different methods for evaluating the significance of the obtained rank positions for genes have been proposed (Maciejewski 2013), but the method attributed to the original description of GSEA was a Kolmogorov-Smirnov-based statistic (Mootha et al. 2003). Though both hitherto mentioned methods have been readily employed, the area of pathway analysis is still an active field of research and new methods are frequently described (Nguyen et al. 2019). Today, most novel methods assess pathway perturbations by including additional information when examining the significance of observed expression changes. For example, it is common to include prior knowledge of protein interactions in the analyses (Tarca et al. 2009) or examining perturbations in sub-networks of pathways (Hidalgo et al. 2017). Additionally, some methods have also been proposed for causal reasoning, i.e. the identification of probable sources of perturbations that could have induced the observed expression levels (Bradley and Barrett 2017).

3.4 Biomarker discovery and model definition

3.4.1 Identification of relevant predictors

Having acquired transcriptomic data from an experiment, which has been adequately pre-processed and quality controlled, a major step in the process of developing predictive assays is to identify the most important features that can be effectively used as a biomarker signature. The general term for this process is feature selection, which encompasses a wide variety of techniques (Guyon and Elisseeff 2003). The overarching goal of feature selection is to identity predictive features and simultaneously reduce the dimensionality of the modelled dataset thereby decreasing its complexity. The task of feature selection can for the purpose of the predictive assay development be both a requirement for making the assay feasible to run, but also for contributing with additional information of the examined endpoint (as discussed more broadly in the context of extracting biological information in the section above). Secondly, reducing the dimensionality of the original data can act as an important step towards achieving optimal prediction performance when attempting to model a hazard endpoint using statistical learning methods. This is in part due to the risk of overfitting to uninformative features of very high-dimensional datasets (Hira and Gillies 2015). Furthermore, transcriptomic data also contain highly correlated features, due to co-expression of certain genes (Michalak 2008), which could have a negative impact on predictive performances (Toloşi and Lengauer 2011).

Feature selection techniques can broadly be divided into three distinct categories depending on their mode of action: filtering methods, wrapper methods and embedded methods (Chandrashekar and Sahin 2014). All can be effective for finding candidate biomarkers, but constraints imposed by the properties of transcriptomic data can skew the selection of which method to apply. Filter methods operate independent from any machine learning algorithm, and features are selected based on some evaluation criteria that can be comprised of, for example, correlation measures or evidence of statistical association with the modelled endpoint, or metrices extracted from information theory such as mutual information or information gain (Lazar et al. 2012). Commonly, these methods are univariate, i.e. the performance of the individual features are considered, which can lead to the selection of a suboptimal subset of features (Chandrashekar and Sahin 2014). However, methods have been proposed where the relationship between features are also taken into account during the selection procedure (Bommert et al. 2020; Lazar et al. 2012).

Additionally, these methods are usually computationally inexpensive and can readily be applied to reduce the dimensionality of large data spaces. Nevertheless, it has also been argued that their model-independent approach can lead to suboptimal selection of feature subsets (John et al. 1994).

Wrapper methods also include a large set of techniques, which are based on the concept of selecting features that optimize the predictive performance of a specific machine learning algorithm. The technique does not generally inflict any restrictions on the type of machine learning algorithm that is used. Instead iterative approaches are combined with resampling methods to establish performance estimates for different feature subsets (Aboudi and Benhlima 2016). Commonly described iterative approaches include backward elimination and forward selection (Khaire and Dhanalakshmi 2019). As an example, the backward elimination starts by estimating a performance measure for classifications when all available features are utilized. Next, based on some feature performance measure (such as feature weights in an SVM, or perhaps more generally from some permutation-based performance score), a set of the worst performing features are dropped from the original set and the prediction performance is recalculated. This approach can be continued until no more variables exists in the set of active features. Then, the optimal feature subset can be identified from the elimination path of the algorithm. Due to the nature of these types of optimization algorithms, they are generally more computationally expensive compared to either filter methods or embedded methods. But because they identify a set of features that are specific for a given classifier, they are often thought to outperform filter methods (Aboudi and Benhlima 2016).

Finally, embedded methods include techniques where a machine learning algorithm performs feature selection internally during training. One of the most recognized embedded methods is the lasso, which is a regression technique where a regularization term has been added to the loss function (a function that describes how well a given model performs on the training set during fitting), which automatically shrinks some of the coefficient estimates to zero, effectively excluding their contribution to the final prediction model (Tibshirani 1996). On a side note, regularization is a general concept in machine learning that impacts an algorithm's solution to a given problem by providing measures to inflict e.g. bias into the final solution. For example, the regularization term in lasso introduces a penalty that is proportional to the absolute value of the coefficient estimates in the model, thereby restricting their values (Hastie et al. 2009; Tibshirani 1996). Other methods frequently associated with embedded feature selection are tree-based algorithms. During their construction, features are evaluated by their ability to introduce effective separation between class labels in the training set, often resulting in the utilization of only the most efficient features in the final model (Lal et al. 2006).

Though feature selection methods can be very effective, and in certain cases necessary for enabling the development of a functional assay, great care must be taken to avoid pitfalls that can be encountered during the process. A common mistake is to combine feature selection with the overall performance estimation in a fashion that leads to overly optimistic performance estimates (Krawczuk and Łukaszuk 2016). This can be achieved by using all available data for feature selection, and then apply a performance estimation technique to the reduced dataset. This can be especially troublesome when the initial dataset has a large feature space, which is the case in transcriptomic experiments. In fact, randomly generated data with proportions between the number of samples and the number of features similar to those observed in transcriptomic experiments can produce good performance measures if the estimation is inappropriately conducted. In any type of modelling setting, it is important to be careful when interpreting performance measures generated from data that has in any form been used for some optimization task during the modelling, such as hyperparameter optimization or feature selection. Therefore, when including feature selection in a pipeline, it should be ensured that any data that is used for the performance estimation (either external data or resampled) should not have participated in the selection of the features.

3.4.2 Learning from data

The task of defining classification heuristics from data is a general problem in machine learning, and one which has been highly relevant during the development of the GARD assays. The type of learning algorithms deployed during the development of the GARD assays are generally termed supervised algorithms, meaning that the algorithms learn from annotated data examples (Singh et al. 2016). The goal of the learning algorithm thus becomes the identification of some function that maps input features to output values. The features can comprise a variety of different data types that are generally selected for their relevance to the prediction problem at hand and can be comprised of e.g. gene expression values. The output values can consist of either quantitative or qualitative (including ordinal categories) variables and can be e.g. hazard labels such as *skin sensitizer* or *non-sensitizer*.

Though the overall aim of different supervised learning algorithms is the same, the way by which how a given task is solved vary between algorithms (Breiman 2001; Cortes and Vapnik 1995; Cover and Hart 1967). Further, there is not an obvious solution to the process of identifying an optimal model for a specific problem. Given the fact that the performance of different machine learning algorithms vary and depend on the properties of the examined data, the identification of an accurate classifier for a specific dataset can be a laborious task, including the construction and comparison of several different models.

To conceptualize the problem, a simple classification case can be examined. For example, given the data in Figure 1, which represent data containing two variables that can be used for classification and one output variable (colour of the points), the task would be to develop a prediction function that produces the most probable category given values for features X1 and X2. For this specific dataset and prediction problem, the task seems trivial and a line drawn manually would probably constitute a well-performing classification rule. In fact, given that the two categories are linearly separable, an infinite number of lines that separates the groups can be drawn (Han et al. 2012). Thus, the question arises: which line is the optimal line for enabling discrimination between the two categories when new data generated from the same process is assessed? Or even, is a line the optimal rule or is a non-linear classification rule superior? In a situation such as the one depicted in Figure 1, where the categories are easily separable and seems well-confined, a simple classification heuristic is often preferable compared to a complex model. Nonetheless, the definition of classification rules can be achieved by a variety of machine learning models, each possibly generating slightly different solutions by optimizing classifier-specific functions. The functions optimized when constructing classifiers are often referred to as loss functions, which aims to describe how well a given classifier performs (Hastie et al. 2009). For example, in logistic regression, the loss function can be characterized as penalising uncertainties in the predicted probabilities of the samples in the training set (Hastie et al. 2009; Hosmer and Lemeshow 2000). In contrast, an SVM bases its loss by only considering the samples closest to the separating line while attempting to maximize their separation (Burges 1998). Thus, though both classification algorithms identify a separating line, the slope and intercept can differ and therefore also potentially their performance as the classification rule is challenged with previously unseen test samples.

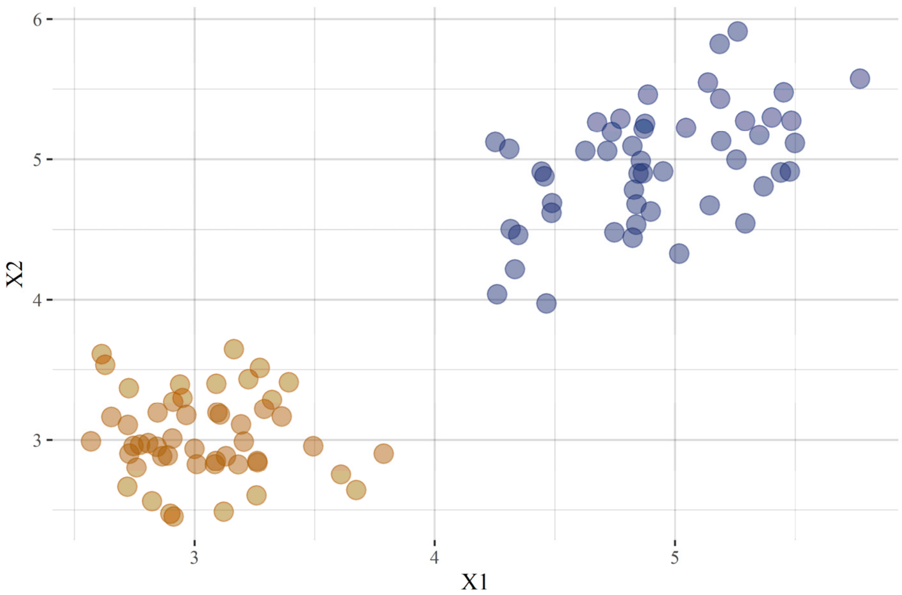


Figure 1. Data illustrating the problem of identifying a prediction rule given two measured variables (X1 and X2) and one output variable (colour of the points).

This methodology of optimizing a function for the identification of a classification model is not limited to simple scenarios like the one discussed above. Indeed, the true effectiveness of statistical learning arises when methods are generalized to encompass problems that would not be manageable for manual analysis. However, the ability to extend such algorithms to much more complex datasets is also associated with certain difficulties. Some are perhaps obvious, for example, how does one assess the performance or suitability of a model when the classification heuristic can no longer be easily visualized or understood? Whereas other difficulties, though related, are more inconspicuous. One of the most common and pressing issues when attempting to model data is the risk of overfitting. Overfitting is the act of creating a classifier that learns to represent the training data, i.e. the data from which the model defines its prediction rule, with high fidelity but is unable to classify previously unseen samples, which makes the model ineffective for subsequent application (Karystinos and Pados 2000). This issue is also further exacerbated in datasets with high dimensions, i.e. datasets with many features (Clarke et al. 2008).

Therefore, it is necessary to be able to monitor the presence to tendencies of overfitting during modelling and to be able to assess and estimate a model's performance. Several methods exist for assessing the performance of a prediction model, where resampling techniques including jack-knife, bootstrap and cross validation (Efron and Gong 1983) are common. However, several aspects should be taken into consideration when evaluating results from such techniques. First, it is not necessary that the resampling techniques are unbiased estimators of the performance (Kohavi 1995). Further, the obtained results are associated with variance, and parameters of the resampling can affect both the bias and the variance of the estimates. Cross validation is one of the most common methods for assessing classifier performance by resampling, and the main adjustable parameter comprises the number of folds, i.e. the number of splits that should be created and evaluated. However, there is no real consensus on how the size of the folds affect the variance and the bias. For example, in Hastie et al. it is suggested that leave one out cross validation constitute an approximately unbiased estimator but that it can simultaneously be associated with larger variance (Hastie et al. 2009). In contrast, Zhang and Yang argue that such variance inflation for the leave one out cross validation is generally not true. However, it has also been suggested that repeated cross validation procedures can be employed for model selection to attenuate the issue of high variance (Krstajic et al. 2014).

Even when data is readily available, results from resampling methods are important and often comprise, at least, important intermediate results. However, in these situations, data can also be split into proper training, validation, and test sets. From these, all model fitting steps can be employed on the training set, the hyperparameter tuning and optimization on the validation set, and performance estimation using the test set (Ripley 2007). This is a relatively common approach, but it is still necessary to ensure that the splits are made in such a way to ensure proper and unbiased sample selection. Despite the possibility to use data in this fashion, many algorithms can consume very large sample sizes to identify complex relationships and patterns (Touvron et al. 2019), making it attractive to use as much data as possible for modelling.

Overfitting is a real issue when attempting to model data, and a concern that should always be considered when attempting to create prediction models using machine learning (Hawkins 2004; James et al. 2013). Different learning algorithms have different affinities to overfit to a training dataset, which is usually associated to the amount of complexity that an algorithm can model (Hastie et al. 2009). This issue is often discussed in terms of a bias-variance trade-off, where some models infer more bias compared to others on a specific dataset. For example, linear models are generally considered fairly biased as they enforce a linear solution to the optimization. However, for certain prediction problems, they are too simple and cannot efficiently model complex structures in the data. In these situations, where training examples are available in sufficient quantities, complex models such as deep neural networks can generally outperform simpler models (Emmert-Streib et al. 2020). For many learning algorithms, the optimized solution can further be influenced by regularization, which are usually adjustable hyperparameters. For example, in linear regression, L1 or L2 penalization terms can be added to infer penalties on large coefficient estimates, thereby increasing the bias in the estimation procedure to further control the modelling (Tibshirani 1996). Similarly, other learning algorithms use their own regularization parameters, but generally towards the same end.

Finally, the assessment of a model's appropriateness, in terms of how it interprets the input variables and how it generates classifications, is not always trivial. For certain types of models, it is quite straightforward to assess how a model interprets the different variables. For example, in linear models and in certain simple classification trees, it is possible to examine how the individual features are utilized for generating a prediction. However, for other more complex models, including deep neural networks or gradient boosting algorithms, it is more difficult to obtain an unambiguous understanding of how the features are utilized. Further, it is not always clear or possible to explain why a complex model produces a specific classification. Nevertheless, it is an important and interesting question and methods have been developed to help shed light on the predictions of complex models. One such model is LIME (Local Interpretable Model-Agnostic Explanations, (Ribeiro et al. 2016)), which tries to create a local approximation of the original classifier close to the predicted sample, by modelling simpler linear relationships that can be more readily interpreted. This technique has, for example, been demonstrated on convolutional neural networks to explain how features in images contribute to specific predictions (Palatnik de Sousa et al. 2019; Ribeiro et al. 2016). Another notable method that has been developed towards the same end goal, i.e. model explainability, is the SHAP-values, which can also be used to e.g. identify the most important features that drive a specific prediction (Lundberg and Lee 2017).

3.4.3 Application to GARD

During the development of the GARD assays, data is generally scarce, and care must first be taken when generating the datasets. A crucial design aspect, which can influence a potential assay, is the selection of the samples that will constitute the training set, i.e. selection of chemicals from whose induced expression patterns the models are created. Given that the sample size is limited, attention must be taken to generate a sample that can be generalized to the remainder of the sample space. Of note, this is not an issue that is restricted to the GARD assays but is a vital design decision for most experiments. If the sample is not adequately representative of the sample space, there is a possibility that the model learns a pattern that is not effective for classifying other test substances. In the case of the GARDskin assay, several properties of the chemicals were evaluated prior to their inclusion in the training dataset including the chemicals' reaction mechanisms, their skin sensitizing potencies, their necessity of activation prior to being protein reactive (i.e. pre- and pro haptens), and other potential confounders such as the chemicals' abilities to induce irritation.

Once an experiment has been designed and carried out and the data has been acquired, the most predictive genes are identified using methods previously discussed and a prediction model is defined. Again, immense care must be taken to avoid negative effects of selection bias and overfitting. Generally, the way both are assessed in the GARD assays are by the utilization of resampling techniques and subsequent assessment by independently acquired test sets. For example, when a pipeline for feature selection and model definition has been defined, the entire pipeline is often evaluated using cross validation. As previously discussed, the full pipeline should be included in the cross validation, including the steps of the feature selection, in order to be able to generate a performance estimate that is not overly optimistic due to flaws in the estimation procedure. The machine learning algorithms that are employed in the GARD assays are often considered relatively simple but are nonetheless selected to decrease the risk of overfitting. Though the learned models are simplistic, the learning algorithms are fairly sophisticated. In the case of the SVM, which has been utilized to define all hitherto described GARD prediction models, the algorithm seeks to identify a hyperplane that discriminates between the categories in the output label by maximizing the margin between samples from respective category. The generalized solution to the SVM was proposed in 1995 by Cortes and Vapnik and is thus a relatively recent learning algorithm that has shown proficiency in many dataset applications (Cortes and Vapnik 1995; Fernández-Delgado et al. 2014).

3.5 The standardization and validation of assays

3.5.1 Standardization of data acquisition

The technologies for acquisition of global transcriptional profiles has, as already noted, reached a state of maturity enabling them to be routinely used in experimental studies (Shi et al. 2006; Su et al. 2014). However, they are still associated with high costs and experimental protocols that render them suboptimal for routine implementation when aiming to quantify a specific set of genes of interest. Therefore, alternative methods enabling more rapid, cheaper, and simpler acquisition were sought for the GARD assays (Forreryd et al. 2014). Of note, this is not an issue that is isolated to the GARD assays, but other research groups and consortia have also evaluated and developed methods for rapid acquisition of specific sets of genes. For example, the L1000 assay was developed specifically for high-throughput and cost-effective acquisition. It quantifies the expression levels of approximately 1000 genes and has been used to generate more than 1.5 million expression profiles (Subramanian et al. 2017).

For GARDskin, different quantification platforms were examined in Forreryd et al. 2014, and it was found that the NanoString platform fulfilled the requirements in terms of efficiency and correlation with historical data obtained from the previous acquisition platform. The NanoString platform allows for the acquisition of expression levels via direct quantification of mRNA levels, i.e. no cDNA generation is required which reduces the complexity of the experimental protocols. Further, acquired data consist of digital counts which require little pre-processing in order to be analysed. The acquisition is enabled by the utilization of sets of probes, which are complementary to target genes, which then binds and labels individual mRNA sequences, enabling their subsequent quantification.

3.5.2 Definition of a processing pipeline

The processing pipeline is here defined as the steps taken from having acquired data, in the form of expression levels, to having generated a prediction with an already defined prediction model. In the GARD assays, the pre-processing steps include quality control of the quantified data, normalization of individual samples, and adjustment of batch effects. One of the most important tasks of the pre-processing pipeline is to enhance the signal to noise ratio by reducing variation introduced by technical variance, i.e. non-biological variance that is not of interest for the assessment of the biological condition. In genomic technologies, non-biological variance is fairly common and can arise from small differences incorporated during the experimental procedure, including minor variations in pipetting, utilization of different reagents, or by having different laboratory persons performing the experiments (Goh et al. 2017; Leek et al. 2010). Many of these potential sources of undesired variance should be mitigated during the design phase of the experiment, which can consist of taking steps to ensure that e.g. necessary reagents are available in sufficient quantity so that the experiment can be completed with materials from a single lot. However, some residual variation attributable to experimental parameters will certainly always be present. Therefore, the processing pipelines in the GARD assays are designed with consideration to alleviate such instances of variance. First, during the initial transfer experiment, it was shown that a single-sample normalization procedure was effective at reducing technical variance, enabling accurate predictions (Forreryd et al. 2016). Today, this procedure has been further examined and its effectiveness has been confirmed by additional experiments aimed at assessing variation between technical replicates, which should be minimized, and between different test chemicals. Though this normalization procedure is effective at reducing technical variance that arise during experiments, larger sources of variation that can be observed between experiments separated by time cannot be completely accounted for, requiring the use of more rigorous correction methods.

Batch effects are another general issue that has been observed in data acquired from multiple genomic technologies including microarrays, RNAseq, qPCR, and NanoString (Goh et al. 2017; Talhouk et al. 2016). As already noted, batch effects can have a variety of sources and many can and should be mitigated by careful experimental design (Nygaard et al. 2016). However, some sources of variation cannot be removed by such interventions and will affect the generated data. Therefore, it is of particular importance to try to anticipate such sources of variation to ensure that the design is sufficiently robust to facilitate biological discovery despite their inclusion. Otherwise, batch effects can become confounded with the experimental details of interest, making the identification of the biologically relevant effects difficult and potentially bias downstream results (Nygaard et al. 2016; Soneson et al. 2014).

For discovery studies, where an experiment is carried out in a consistent and controlled manner, it is often possible to anticipate the major sources of technical variation. For example, it is common to observe batch effects in RNAseq and microarray studies that are attributable to either sequencing run or hybridization date, respectively (Conesa et al. 2016b; Luo et al. 2010). Because they are known, the experimental design can be made to ensure that they are not confounded with the biologically interesting sources of variation, and thereby facilitate downstream analysis. Non-confounded batch effects can be accounted for during differential expression analysis, which can increase the statistical power of detecting biologically relevant changes (Leek and Storey 2007).

In these types of studies, it is also possible to try to create a cleaned dataset that is relieved of the batch effects that can be used for other downstream analytical tasks, including predictive modelling. A variety of methods have been proposed to carry this out. One broad term of adjustment methods is the location and scale methods (Johnson et al. 2006). As the name suggest, these methods attempt to correct for batch effects by estimating the gene-wise location- and scale shifts that are attributable to different batches, and then to adjust for them. Due to issues similar to the ones associated with differential expression analysis of high-dimensional data, the small sample sizes can make the original location and scale methods instable, which enforces certain constraints on the experimental design in order for the methods to be applied. For example, batches should comprise a moderate number of samples to allow more stable estimates (Johnson et al. 2006). However, also in this case, the empirical Bayes method has facilitated the development of procedures to alleviate these issues. ComBat is one of the most well-known batch correction techniques (Johnson et al. 2006). By borrowing information between genes, it enables more stable estimates of batch-associated variation even in small sample sizes. Despite the popularity of ComBat, this is also still an active field of research and methods are frequently being proposed to alleviate the issue of batch effects in discovery studies (Hornung et al. 2016; Li et al. 2019; Oytam et al. 2016).

Even though batch effects can be adjusted for in well-designed discovery studies, a larger issue arises when batch effects are considered in prediction problems, i.e. when a prediction model has been defined on a fixed set of data that is used to classify subsequently acquired test samples. In these scenarios, the biological condition of the test sample is generally unknown and simultaneously confounded with batch. This confoundment can, if left unadjusted for, lead to severely reduced predictive performances. Depending on the size of the test set, the application of certain types of methods similar to the ones used in discovery studies could perhaps be justified. If the test set and the training set are large and if it is also assumed that the proportions of biological groups within the datasets in both datasets are similar, global location and scale adjustment methods could perhaps be employed (Luo et al. 2010). However, in most situations, such methods are not applicable. Further, to allow for routine testing, the normalization method should preferably not infer too stringent restrictions on experimental designs. For a testing platform such as GARD, the ideal normalization procedure should allow for testing of individual test substances (with the appropriate assay controls).

Batch correction in predictive modelling is a pressing and difficult issue. Due to the complexities associated with the confoundment, one of the prevailing methods for alleviating the negative impact on the predictive performance is to include reference samples in the test sets from which estimates of batch effects can be obtained (Luo et al. 2010). We have proposed a method called BARA, which is described in paper IV, for the adjustment of batch effects in prediction problems. The method performs the adjustment in a latent data space spanned by the training dataset, by adjusting the test set via alignment of the reference samples that are present in both datasets. The latent data space where the correction is performed is defined by the principal components of the training dataset. The method was designed to leverage the fact that the variance important for establishing accurate predictions can often be explained by transforming the original data space into a few latent factors (for datasets like the ones used in the GARD assays where feature selection has been used to select the most predictive biomarkers). The latent data space in BARA is obtained by decomposing the training dataset using singular value decomposition (Eckart and Young 1939; Golub and Reinsch 1970), thus enabling the identification of the right singular vectors, which can be used to obtain the principal components of the training dataset. By projecting both the training set and the test set into the data space (spanned by the right singular vectors), the batch correction can be made in a few components, and the normalized data can be reconstructed to the original feature space. The number of right singular vectors retained during the normalization is an adjustable hyperparameter that can be optimized using, for example, a validation set. In paper IV we show that the predictive performance of data normalized with BARA is competitive with state-of-the-art methods for batch adjustment in prediction problems.

3.5.3 Validation

One of the final phases of assay development comprises the validation of the assay's performance, which should constitute an unbiased opportunity to assess the predictive performance and to evaluate the assay's robustness.

Depending on the intended use and future plans for the assay, the rigorousness of the validation experiments can vary. For example, for the predictive assays GARDskin and GARDpotency, there exists incentives for having the assays evaluated in accordance with the OECD guidelines for validation of non-animal methods. However, the formal process of having an assay validated and accepted by the OECD is extensive and not always a requirement. Nevertheless, validation of an assay's performance should be carefully performed to ensure the possibility of assessing the aspects of the assay that are important for its implementation. This should at least include assessment of predictive performance in terms of both accuracy and class-specific metrices, and also the assessment of the reproducibility of the assay. These aspects can be examined by e.g. ring trial experiments where individual laboratories evaluate sets of blinded compounds that can provide a fair description of the assay's performance. Therefore, the selection of the reference chemicals used for validation must also be made with careful consideration. For example, attributes of the chemicals that can be considered during the design of validation experiment include whether or not they were part of the definition of the prediction model (i.e. the training set or validation sets for hyperparameter tuning), the quality and certainty of the annotations for them, and that they allow the estimation of representative performance figures.

Papers II and III describe ring trial experiments carried out to assess the performance and reproducibility of the GARDskin assay and the GARDpotency assay, respectively. To allow for such estimations, a set of well-annotated chemicals were assayed blindly in three independent experiments at each of three participating laboratories. To ensure the blinded nature of either study, a validation management group was created and declared responsible for selection, encoding, and distribution of the chemicals to respective laboratory. Further, predictions were reported from respective laboratory to the validation management group to ensure unbiased handling of the result and impartial reporting of the performance. The results from the studies indicated that GARDskin was both predictive of skin sensitizer hazard and reproducible. Further, GARDpotency was found to be predictive of skin sensitizer potency but a potential for improving the reproducibility was also observed.

4 Prospects and concluding remarks

The availability of tools that enable accurate hazard assessment of specific toxicological endpoints are important to ensure and improve human and environmental health. However, the development of such tools remains challenging, especially when a toxicological hazard endpoint of interest is constructed of complex biological mechanisms. Indeed, in some cases, the mechanistic underpinnings are not even well understood, which introduces additional challenges for reducing the complexity into a viable test method. This challenge has historically been met by the utilization of *in vivo* methods, but several concerns associated with these methods aspire researchers towards developing non-animal methods capable of eventually replacing them.

The GARD platform has been developed as a framework for the development of toxicological *in vitro* assays by monitoring the exposure-induced transcriptional changes in a suitable cell line by machine learning methods to provide accurate hazard assessments. The test methods described in this thesis have all been concerned with the identification and characterization of skin sensitizers. But as previously mentioned, the concepts of GARD are generalizable and can be applied to larger suits of problems. For example, assays for assessing respiratory sensitization, protein allergenicity, and respiratory irritation have been proposed. The discovery study for the respiratory sensitization assay was published in Forreryd et al. 2015, which suggested that a genetic biomarker signature could accurately identify respiratory sensitizers and discriminate them from both skin sensitizers and non-sensitizers. Respiratory sensitization is a serious condition with potentially fatal effects. However, the toxicological endpoint currently lacks any generally accepted test method, which makes the continued development and availability of the GARD assay important. Since the discovery study, additional efforts have recently been made to make it into an accessible assay, which included transfer experiment to the NanoString platform (to be published). Similarly, there is a lack of methods that can be used to infer the ability of novel proteins to induce allergenic potential. Zeller et al. recently described an initial discovery study in accordance with the GARD procedures where transcriptomic alterations were identified that could be used for predictive purposes (Zeller et al. 2018). Though the results from this study remains to be further explored and validated, the findings are interesting. Finally, attempts have been made to develop an assay for the recognition of respiratory irritants. The ability to identify respiratory irritants is an important objective for certain industries, including in pharmaceutical development where a large fraction of inhaled therapeutic projects is closed due to toxicity (Cook et al. 2014). The GARD approach was also recently applied to this toxicological endpoint, where a dataset comprising 19 chemicals with known effects were used to stimulate the SenzaCell cell line. Following gene expression acquisition using microarray analysis, several strong candidate genes predictive of the examined endpoint were identified. These results were only recently obtained, and though work remains to advance the evolution and validation of these assays, it lies beyond the scope of this thesis.

In summary, the GARD platform has been applied to a variety of toxicological endpoints but several opportunities for optimization of assays that are currently in early developmental stages and for further refinement of methodological choices remain. For example, it seems reasonable to assume that data that are acquired when the existing assays are run could eventually provide a large database of information for exploratory analysis. This could aid the understanding of xenobiotics' effects on the cell line and subsequently be incorporated in the development of future assays to increase robustness and performance. However, though a thrilling prospect, current assays acquire data in distinct sets of genes that comprise their respective biomarker signature, which makes it impossible to merge data acquired from the different assays on distinct chemicals. Furthermore, by selectively querying sets of genes, the majority of the genes in the transcriptome are ignored. Though these design choices were deliberately made to facilitate routine acquisition, it is possible that continued development of techniques such as RNAseq could eventually make it reasonable to acquire transcriptomic data for all assays, further harmonizing the acquisition and possibly making it conceivable to acquire a coherent expression database. In addition, given that the expression levels obtained using RNAseq possess similar discrete count distributions as the NanoString data, a potential transition phase could potentially allow for acquisition on either platform. In fact, unpublished data generated in-house seem to suggest that the current prediction models defined on NanoString data could be used to classify expression levels acquired using RNAseq.

In conclusion, the aim to create a safer environment by using test methods that are ethically justifiable but still accurate spurs the development of non-animal assays. The GARD platform comprises a framework of methodological choices that can be applied to aid the development of such assays.

Populärvetenskaplig sammanfattning

Möjligheten att kunna identifiera och karakterisera hälsopåverkande effekter associerade med specifika kemikalier är en viktig komponent i arbetet med att säkerställa individers välmående i samhället. Detta grundar sig till viss del i att vi frekvent exponeras mot ett stort antal kemikalier genom interaktioner med vardagliga produkter som till exempel kosmetika, tvål, eller städprodukter, eller kemikalier vars egenskaper gör dem nödvändiga inom vissa industrier. Eftersom somliga kemikalier kan ge upphov till negativa hälsoeffekter är det viktigt att kunna identifiera och karakterisera dessa, så att produkter kan utformas som är säkra att användas av befolkningen.

Vanliga tillstånd som kan uppstå vid kontakt med potentiellt skadliga kemikalier inkluderar irritation eller allergiska reaktioner i hud eller luftvägar. Då mekanismerna som utgör dessa hälsotillstånd är komplexa, introduceras vissa svårigheter när man vill testa huruvida en viss substans har potentialen att orsaka en sådan åkomma. Traditionellt har ofta djurmodeller används för att representera de biologiska funktionerna och de effekter som en kemikalie kan åstadkomma vid exponering. Dessa tester är dock problematiska ur flera aspekter, inte minst de etiska. De är också ineffektiva sett ur kostnads- och tidsperspektiv och inte alltid heller representativa för den effekt som induceras i människor. Detta har gett upphov till en strävan att utveckla alternativa testmetoder som effektivt, etiskt, och träffsäkert kan ersätta djurmodeller som vi tidigare varit beroende av för att utvinna säkerhetsinformation från kemikalier.

GARD är en teknologisk plattform som har utvecklats för att främja skapandet av sådana testmetoder. Konceptet bygger på att biologisk förståelse för ett särskilt hälsotillstånd driver valet av ett experimentellt system som förväntas kunna förmedla relevanta signaler vid exponering mot en testkemikalie. Detta kan, till exempel, bestå av celler som har en viktig funktion när åkomman framkallas i människor. Det som skiljer GARD från många andra testmetoder är att nästa steg i utvecklingen av ett test inte kräver fullständig förståelse för de interna stegen som sker vid induceringen av hälsotillståndet. Istället utnyttjas modern tekniks förmåga att i ett enda experiment kunna avläsa tusentals geners uttryck. Genom att exponera cellerna mot flera olika kemikalier, där vissa är kända att kunna orsaka den särskilda hälsopåverkande effekten, kan specifika genuttryck som har förmågan att särskilja mellan de undersökta grupperna av kemikalier identifieras. Sedan används maskininlärningsmetoder, för att baserat på de observerade nivåerna av uttryckta gener, definiera klassifikationsmodeller. För att undersöka den hälsopåverkande effekten hos en tidigare okänd kemikalie, exponeras samma celler i det experimentella systemet mot testkemikalien. Därefter avläses uttrycket av de tidigare identifierade generna och resultatet matas vidare till den definierade klassifikationsmodellen. Denna avgör huruvida det är troligt att kemikalien utgör en fara och potentiellt kan orsaka den undersökta hälsoeffekten.

GARD har använts för att skapa ett antal tester, vilka inkluderar metoderna GARDskin och GARDpotency. Båda dessa tester har utformats för att undersöka kemikaliers egenskaper med avseende på hudsensibilisering. GARDskin har skapats för att ge ett Ja- eller Nej-svar med avseende på ifall en kemikalie kan orsaka hudsensibilisering, medan GARDpotency konstruerats för att undersöka styrkan hos kända hudsensibiliserande kemikalier. Båda dessa metoder har funnits effektiva för respektive ändamål. Under en valideringsstudie för GARDskin, där både träffsäkerhet och robusthet undersöktes, fann man att testet var stabilt med en tillförlitlighet på 94%. Samma siffra för GARDpotency uppskattades i en liknande studie till 88%. Dessa värden tyder på att båda testerna är väl lämpade för sina ändamål och presterar mycket bra i jämförelse med både existerande djurmodeller och andra alternativa metoder. Detta visar också att GARD är en effektiv plattform för utveckling av alternativa testmetoder och att redan framtagna tester utgör attraktiva alternativ för att slutligen avlägsna behovet av djurtester.

References

Aboudi N, Benhlima L. Review on wrapper feature selection approaches. 2016 International Conference on Engineering & MIS (ICEMIS); 22-24 Sept. 2016.

Agarwal S, Krishnamurthy K. 2020. Histology, skin. Statpearls. Treasure Island (FL): StatPearls Publishing. Copyright © 2020, StatPearls Publishing LLC.

Ainscough JS, Frank Gerberick G, Dearman RJ, Kimber I. 2013. Danger, intracellular signaling, and the orchestration of dendritic cell function in skin sensitization. Journal of Immunotoxicology. 10(3):223-234.

Albrekt AS, Johansson H, Börje A, Borrebaeck C, Lindstedt M. 2014. Skin sensitizers differentially regulate signaling pathways in mutz-3 cells in relation to their individual potency. BMC pharmacology & toxicology. 15:5.

Alinaghi F, Bennike NH, Egeberg A, Thyssen JP, Johansen JD. 2019. Prevalence of contact allergy in the general population: A systematic review and meta-analysis. Contact Dermatitis. 80(2):77-85.

Ashikaga T, Yoshida Y, Hirota M, Yoneyama K, Itagaki H, Sakaguchi H, Miyazawa M, Ito Y, Suzuki H, Toyoda H. 2006. Development of an in vitro skin sensitization test using human cell lines: The human cell line activation test (h-clat). I. Optimization of the h-clat protocol. Toxicology in vitro : an international journal published in association with BIBRA. 20(5):767-773.

Baird L, Dinkova-Kostova AT. 2011. The cytoprotective role of the keap1–nrf2 pathway. Archives of Toxicology. 85(4):241-272.

Barentsen HM, Jonis SU, Pelgrom SMGJ, Rijk JCW, Westerink WMA, Paulussen JJC. 2019. Reach alternative testing strategy for skin sensitization in practice: Fact or fiction? Regulatory Toxicology and Pharmacology. 106:292-302.

Baroni A, Buommino E, De Gregorio V, Ruocco E, Ruocco V, Wolf R. 2012. Structure and function of the epidermis related to barrier properties. Clin Dermatol. 30(3):257-262.

Basketter D, Crozier J, Hubesch B, Manou I, Mehling A, Scheel J. 2012. Optimised testing strategies for skin sensitization – the llna and beyond. Regulatory Toxicology and Pharmacology. 64(1):9-16.

Beißbarth T, Speed TP. 2004. Gostat: Find statistically overrepresented gene ontologies within a group of genes. Bioinformatics. 20(9):1464-1465.

Benvenuti F. 2016. The dendritic cell synapse: A life dedicated to t cell activation. Frontiers in Immunology. 7(70).

Blanco P, Palucka AK, Pascual V, Banchereau J. 2008. Dendritic cells and cytokines in human inflammatory and autoimmune diseases. Cytokine Growth Factor Rev. 19(1):41-52.

Boer M, Duchnik E, Maleszka R, Marchlewicz M. 2016. Structural and biophysical characteristics of human skin in maintaining proper epidermal barrier function. Postepy Dermatol Alergol. 33(1):1-5.

Bommert A, Sun X, Bischl B, Rahnenführer J, Lang M. 2020. Benchmark for filter methods for feature selection in high-dimensional classification data. Computational Statistics & Data Analysis. 143:106839.

Bos JD, Meinardi MM. 2000. The 500 dalton rule for the skin penetration of chemical compounds and drugs. Exp Dermatol. 9(3):165-169.

Bradley G, Barrett SJ. 2017. Causalr: Extracting mechanistic sense from genome scale data. Bioinformatics (Oxford, England). 33(22):3670-3672.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic rna-seq quantification. Nature Biotechnology. 34(5):525-527.

Breiman L. 2001. Random forests. Machine Learning. 45(1):5-32.

Brys AK, Rodriguez-Homs LG, Suwanpradid J, Atwater AR, MacLeod AS. 2020. Shifting paradigms in allergic contact dermatitis: The role of innate immunity. Journal of Investigative Dermatology. 140(1):21-28.

Buehler EV. 1965. Delayed contact hypersensitivity in the guinea pig. Arch Dermatol. 91:171-177.

Bumgarner R. 2013. Overview of DNA microarrays: Types, applications, and their future. Curr Protoc Mol Biol. Chapter 22:Unit-22.21.

Burges CJC. 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery. 2(2):121-167.

Cassel SL, Sutterwala FS. 2010. Sterile inflammatory responses mediated by the nlrp3 inflammasome. European Journal of Immunology. 40(3):607-611.

Castillo-Davis CI, Hartl DL. 2003. Genemerge—post-genomic analysis, data mining, and hypothesis testing. Bioinformatics. 19(7):891-892.

Chandrashekar G, Sahin F. 2014. A survey on feature selection methods. Computers & Electrical Engineering. 40(1):16-28.

Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, Gehan EA, Wang Y. 2008. The properties of high-dimensional data spaces: Implications for exploring gene and protein expression data. Nat Rev Cancer. 8(1):37-49.

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X et al. 2016a. A survey of best practices for rna-seq data analysis. Genome Biology. 17(1):13.

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X et al. 2016b. A survey of best practices for rna-seq data analysis. Genome biology. 17:13-13.

Cook D, Brown D, Alexander R, March R, Morgan P, Satterthwaite G, Pangalos MN. 2014. Lessons learned from the fate of astrazeneca's drug pipeline: A five-dimensional framework. Nature Reviews Drug Discovery. 13(6):419-431.

Cortes C, Vapnik V. 1995. Support-vector networks. Machine Learning. 20(3):273-297.

Council NR. 2006. Toxicity testing for assessment of environmental agents: Interim report. National Academies Press.

Council NR. 2007. Toxicity testing in the 21st century: A vision and a strategy. Washington, DC: The National Academies Press.

Cover T, Hart P. 1967. Nearest neighbor pattern classification. IEEE Transactions on Information Theory. 13(1):21-27.

Daniel AB, Strickland J, Allen D, Casati S, Zuang V, Barroso J, Whelan M, Régimbald-Krnel MJ, Kojima H, Nishikawa A et al. 2018. International regulatory requirements for skin sensitization testing. Regulatory toxicology and pharmacology : RTP. 95:52-65.

Dean JH, Twerdok LE, Tice RR, Sailstad DM, Hattan DG, Stokes WS. 2001. Iccvam evaluation of the murine local lymph node assay: Ii. Conclusions and recommendations of an independent scientific peer review panel. Regulatory Toxicology and Pharmacology. 34(3):258-273.

Deckers J, Hammad H, Hoste E. 2018. Langerhans cells: Sensing the environment in health and disease. Front Immunol. 9:93.

Di Virgilio F, Dal Ben D, Sarti AC, Giuliani AL, Falzoni S. 2017. The p2x7 receptor in infection and inflammation. Immunity. 47(1):15-31.

Diaz-Uriarte R. 2007. Genesrf and varselrf: A web-based tool and r package for gene selection and classification using random forest. BMC Bioinformatics. 8(1):328.

Díaz-Uriarte R, Alvarez de Andrés S. 2006. Gene selection and classification of microarray data using random forest. BMC bioinformatics. 7:3-3.

Eckart C, Young G. 1939. A principal axis transformation for non-hermitian matrices. Bull Amer Math Soc. 45(2):118-121.

Efron B, Gong G. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician. 37(1):36-48.

Emmert-Streib F, Yang Z, Feng H, Tripathi S, Dehmer M. 2020. An introductory review of deep learning for prediction models with big data. Frontiers in Artificial Intelligence. 3(4).

Emter R, Ellis G, Natsch A. 2010. Performance of a novel keratinocyte-based reporter cell line to screen skin sensitizers in vitro. Toxicology and Applied Pharmacology. 245(3):281-290.

Esser PR, Martin SF. 2017. Pathomechanisms of contact sensitization. Curr Allergy Asthma Rep. 17(12):83.

Esser PR, Martin SF. 2020. [extended understanding of pathogenesis and treatment of contact allergy]. Hautarzt. 71(3):174-181.

Esser PR, Wölfle U, Dürr C, von Loewenich FD, Schempp CM, Freudenberg MA, Jakob T, Martin SF. 2012. Contact sensitizers induce skin inflammation via ros production and hyaluronic acid degradation. PloS one. 7(7):e41340-e41340.

Ezendam J, Braakhuis HM, Vandebriel RJ. 2016. State of the art in non-animal approaches for skin sensitization testing: From individual test methods towards testing strategies. Arch Toxicol. 90(12):2861-2883.

Fenner J, Clark RAF. 2016. Chapter 1 - anatomy, physiology, histology, and immunohistochemistry of human skin. In: Albanna MZ, Holmes Iv JH, editors. Skin tissue engineering and regenerative medicine. Boston: Academic Press. p. 1-17.

Fernández-Delgado M, Cernadas E, Barro S, Amorim D. 2014. Do we need hundreds of classifiers to solve real world classification problems? The journal of machine learning research. 15(1):3133-3181.

Ferreira I, Silva A, Martins JD, Neves BM, Cruz MT. 2018. Nature and kinetics of redox imbalance triggered by respiratory and skin chemical sensitizers on the human monocytic cell line thp-1. Redox Biol. 16:75-86.

Fitzpatrick JM, Roberts DW, Patlewicz G. 2017a. Is skin penetration a determining factor in skin sensitization potential and potency? Refuting the notion of a logkow threshold for skin sensitization. J Appl Toxicol. 37(1):117-127.

Fitzpatrick JM, Roberts DW, Patlewicz G. 2017b. What determines skin sensitization potency: Myths, maybes and realities. The 500 molecular weight cut-off: An updated analysis. J Appl Toxicol. 37(1):105-116.

Forreryd A, Johansson H, Albrekt AS, Borrebaeck CA, Lindstedt M. 2015. Prediction of chemical respiratory sensitizers using gard, a novel in vitro assay based on a genomic biomarker signature. PLoS One. 10(3):e0118808.

Forreryd A, Johansson H, Albrekt AS, Lindstedt M. 2014. Evaluation of high throughput gene expression platforms using a genomic biomarker signature for prediction of skin sensitization. BMC Genomics. 15(1):379.

Forreryd A, Zeller KS, 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. Toxicology in vitro : an international journal published in association with BIBRA. 37:178-188.

Frank Gerberick G, Ryan CA, Dearman RJ, Kimber I. 2007. Local lymph node assay (llna) for detection of sensitization capacity of chemicals. Methods. 41(1):54-60.

Fujita M, Yamamoto Y, Tahara H, Kasahara T, Jimbo Y, Hioki T. 2014. Development of a prediction method for skin sensitization using novel cysteine and lysine derivatives. Journal of Pharmacological and Toxicological Methods. 70(1):94-105.

Gerberick GF, Ryan CA, Kern PS, Dearman RJ, Kimber I, Patlewicz GY, Basketter DA. 2004a. A chemical dataset for evaluation of alternative approaches to skin-sensitization testing. Contact Dermatitis. 50(5):274-288.

Gerberick GF, Vassallo JD, Bailey RE, Chaney JG, Morrall SW, Lepoittevin J-P. 2004b. Development of a peptide reactivity assay for screening contact allergens. Toxicological Sciences. 81(2):332-343.

Gerberick GF, Vassallo JD, Foertsch LM, Price BB, Chaney JG, Lepoittevin J-P. 2007. Quantification of chemical peptide reactivity for screening contact allergens: A classification tree model approach. Toxicological Sciences. 97(2):417-427.

Goh WWB, Wang W, Wong L. 2017. Why batch effects matter in omics data, and how to avoid them. Trends in Biotechnology. 35(6):498-507.

Golub GH, Reinsch C. 1970. Singular value decomposition and least squares solutions. Numerische Mathematik. 14(5):403-420.

Govindarajan R, Duraiyan J, Kaliyappan K, Palanisamy M. 2012. Microarray and its applications. J Pharm Bioallied Sci. 4(Suppl 2):S310-S312.

Gradin R, Johansson A, Forreryd A, Aaltonen E, Jerre A, Larne O, Mattson U, Johansson H. 2020. The gardtmpotency assay for potency-associated subclassification of chemical skin sensitizers - rationale, method development and ring trial results of predictive performance and reproducibility. Toxicol Sci.

Guyon I, Elisseeff A. 2003. An introduction to variable and feature selection. Journal of machine learning research. 3(Mar):1157-1182.

Han J, Kamber M, Pei J. 2012. 9 - classification: Advanced methods. In: Han J, Kamber M, Pei J, editors. Data mining (third edition). Boston: Morgan Kaufmann. p. 393-442.

Hastie T, Tibshirani R, Friedman J. 2009. The elements of statistical learning: Data mining, inference, and prediction. Springer Science & Business Media.

Hawkins DM. 2004. The problem of overfitting. Journal of Chemical Information and Computer Sciences. 44(1):1-12.

Helou DG, Martin SF, Pallardy M, Chollet-Martin S, Kerdine-Römer S. 2019. Nrf2 involvement in chemical-induced skin innate immunity. Front Immunol. 10:1004.

Hidalgo MR, Cubuk C, Amadoz A, Salavert F, Carbonell-Caballero J, Dopazo J. 2017. High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. Oncotarget. 8(3):5160-5178.

Hira ZM, Gillies DF. 2015. A review of feature selection and feature extraction methods applied on microarray data. Adv Bioinformatics. 2015:198363-198363.

Hoffman GE, Roussos P. 2020. Dream: Powerful differential expression analysis for repeated measures designs. Bioinformatics.

Hoffmann E, Dittrich-Breiholz O, Holtmann H, Kracht M. 2002. Multiple control of interleukin-8 gene expression. Journal of Leukocyte Biology. 72(5):847-855.

Hornung R, Boulesteix A-L, Causeur D. 2016. Combining location-and-scale batch effect adjustment with data cleaning by latent factor adjustment. BMC Bioinformatics. 17(1):27.

Hosmer DW, Lemeshow S. 2000. Applied logistic regression. Wiley New York.

Izadi N, Luu M, Ong PY, Tam JS. 2015. The role of skin barrier in the pathogenesis of food allergy. Children (Basel). 2(3):382-402.

Jaitley S, Saraswathi T. 2012. Pathophysiology of langerhans cells. J Oral Maxillofac Pathol. 16(2):239-244.

James G, Witten D, Hastie T, Tibshirani R. 2013. An introduction to statistical learning. Springer.

Joffre O, Nolte MA, Spörri R, Reis e Sousa C. 2009. Inflammatory signals in dendritic cell activation and the induction of adaptive immunity. Immunol Rev. 227(1):234-247.

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 AS, Borrebaeck CA. 2011. A genomic biomarker signature can predict skin sensitizers using a cell-based in vitro alternative to animal tests. BMC Genomics. 12:399.

John GH, Kohavi R, Pfleger K. 1994. Irrelevant features and the subset selection problem. In: Cohen WW, Hirsh H, editors. Machine learning proceedings 1994. San Francisco (CA): Morgan Kaufmann. p. 121-129.

Johnson WE, Li C, Rabinovic A. 2006. Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics. 8(1):118-127.

Jongeneel WP, Delmaar JE, Bokkers BGH. 2018. Health impact assessment of a skin sensitizer: Analysis of potential policy measures aimed at reducing geraniol concentrations in personal care products and household cleaning products. Environ Int. 118:235-244.

Karlberg A-T, Bergström MA, Börje A, Luthman K, Nilsson JLG. 2008. Allergic contact dermatitis—formation, structural requirements, and reactivity of skin sensitizers. Chemical Research in Toxicology. 21(1):53-69.

Karlberg A-T, Börje A, Duus Johansen J, Lidén C, Rastogi S, Roberts D, Uter W, White IR. 2013. Activation of non-sensitizing or low-sensitizing fragrance substances into potent sensitizers – prehaptens and prohaptens. Contact Dermatitis. 69(6):323-334.

Karystinos GN, Pados DA. 2000. On overfitting, generalization, and randomly expanded training sets. IEEE Transactions on Neural Networks. 11(5):1050-1057.

Khaire UM, Dhanalakshmi R. 2019. Stability of feature selection algorithm: A review. Journal of King Saud University - Computer and Information Sciences.

Kimber I, Basketter DA. 1992. The murine local lymph node assay: A commentary on collaborative studies and new directions. Food and Chemical Toxicology. 30(2):165-169.

Kimber I, Basketter DA, Gerberick GF, Dearman RJ. 2002. Allergic contact dermatitis. International Immunopharmacology. 2(2):201-211.

Kimber I, Dearman RJ, Basketter DA, Boverhof DR. 2014. Chemical respiratory allergy: Reverse engineering an adverse outcome pathway. Toxicology. 318:32-39.

Kimber I, Dearman RJ, Scholes EW, Basketter DA. 1994. The local lymph node assay: Developments and applications. Toxicology. 93(1):13-31.

Kimber I, Poole A, Basketter DA. 2018. Skin and respiratory chemical allergy: Confluence and divergence in a hybrid adverse outcome pathway. Toxicol Res (Camb). 7(4):586-605.

Kimura Y, Fujimura C, Ito Y, Takahashi T, Nakajima Y, Ohmiya Y, Aiba S. 2015. Optimization of the il-8 luc assay as an in vitro test for skin sensitization. Toxicology in Vitro. 29(7):1816-1830.

Kleinstreuer NC, 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. Critical Reviews in Toxicology. 48(5):359-374.

Kohavi R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai. 14(2):1137-1145.

Kolle SN, Landsiedel R, Natsch A. 2020. Replacing the refinement for skin sensitization testing: Considerations to the implementation of adverse outcome pathway (aop)-based defined approaches (da) in oecd guidelines. Regulatory Toxicology and Pharmacology. 115:104713.

Krawczuk J, Łukaszuk T. 2016. The feature selection bias problem in relation to high-dimensional gene data. Artificial Intelligence in Medicine. 66:63-71.

Krewski D, Acosta D, Jr., Andersen M, Anderson H, Bailar JC, 3rd, Boekelheide K, Brent R, Charnley G, Cheung VG, Green S, Jr. et al. 2010. Toxicity testing in the 21st century: A vision and a strategy. J Toxicol Environ Health B Crit Rev. 13(2-4):51-138.

Krstajic D, Buturovic LJ, Leahy DE, Thomas S. 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. Journal of Cheminformatics. 6(1):10.

Lal TN, Chapelle O, Weston J, Elisseeff A. 2006. Embedded methods. In: Guyon I, Nikravesh M, Gunn S, Zadeh LA, editors. Feature extraction: Foundations and applications. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 137-165.

Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN et al. 2006. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. Science. 313(5795):1929-1935.

Law CW, Chen Y, Shi W, Smyth GK. 2014. Voom: Precision weights unlock linear model analysis tools for rna-seq read counts. Genome Biol. 15(2):R29.

Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C, Schaetzen Vd, Duque R, Bersini H, Nowe A. 2012. A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 9(4):1106-1119.

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 11(10):733-739.

Leek JT, Storey JD. 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 3(9):1724-1735.

Li H, Dogan H, Cui J. 2019. A new approach to batch effect removal based on distribution matching in latent space. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 423-430.

Li J, Witten DM, Johnstone IM, Tibshirani R. 2012. Normalization, testing, and false discovery rate estimation for rna-sequencing data. Biostatistics. 13(3):523-538.

Li S, Łabaj PP, Zumbo P, Sykacek P, Shi W, Shi L, Phan J, Wu P-Y, Wang M, Wang C et al. 2014. Detecting and correcting systematic variation in large-scale rna sequencing data. Nature Biotechnology. 32(9):888-895.

Linauskienė K, Malinauskienė L, Blažienė A. 2017. Metals are important contact sensitizers: An experience from lithuania. Biomed Res Int. 2017:3964045.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome Biology. 15(12):550.

Lundberg SM, Lee S-I. 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems; 4765-4774.

Luo J, Schumacher M, Scherer A, Sanoudou D, Megherbi D, Davison T, Shi T, Tong W, Shi L, Hong H et al. 2010. A comparison of batch effect removal methods for enhancement of prediction performance using maqc-ii microarray gene expression data. The Pharmacogenomics Journal. 10(4):278-291.

Maciejewski H. 2013. Gene set analysis methods: Statistical models and methodological differences. Briefings in Bioinformatics. 15(4):504-518.

Magnusson B, Kligman AM. 1969. The identification of contact allergens by animal assay. The guinea pig maximization test**from the department of dermatology, university of gothenburg, sahlgrenska sjukhuset, gothenburg, sweden and the department of dermatology, university of pennsylvania school of medicine, philadelphia, pennsylvania 19104. Journal of Investigative Dermatology. 52(3):268-276.

Martin SF. 2012. Allergic contact dermatitis: Xenoinflammation of the skin. Curr Opin Immunol. 24(6):720-729.

Martin SF, Esser PR, Weber FC, Jakob T, Freudenberg MA, Schmidt M, Goebeler M. 2011. Mechanisms of chemical-induced innate immunity in allergic contact dermatitis. Allergy. 66(9):1152-1163.

Michalak P. 2008. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics. 91(3):243-248.

Minamoto K. 2010. [skin sensitizers in cosmetics and skin care products]. Nihon Eiseigaku Zasshi. 65(1):20-29.

Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E et al. 2003. Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics. 34(3):267-273.

Moreth K, Frey H, Hubo M, Zeng-Brouwers J, Nastase M-V, Hsieh LT-H, Haceni R, Pfeilschifter J, Iozzo RV, Schaefer L. 2014. Biglycan-triggered tlr-2- and tlr-4-signaling exacerbates the pathophysiology of ischemic acute kidney injury. Matrix Biol. 35:143-151.

Natsch A. 2009. The nrf2-keap1-are toxicity pathway as a cellular sensor for skin sensitizers—functional relevance and a hypothesis on innate reactions to skin sensitizers. Toxicological Sciences. 113(2):284-292.

Nguyen T-M, Shafi A, Nguyen T, Draghici S. 2019. Identifying significantly impacted pathways: A comprehensive review and assessment. Genome Biology. 20(1):203.

Nygaard V, Rødland EA, Hovig E. 2016. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics (Oxford, England). 17(1):29-39.

OECD. 1992. Test no. 406: Skin sensitisation.

OECD. 2014. The adverse outcome pathway for skin sensitisation initiated by covalent binding to proteins.

OECD. 2018a. Test no. 442d: In vitro skin sensitisation.

OECD. 2018b. Test no. 442e: In vitro skin sensitisation.

OECD. 2020a. How's life? 2020.

OECD. 2020b. Test no. 442c: In chemico skin sensitisation.

OECD, EU. 2018. Health at a glance: Europe 2018.

Oytam Y, Sobhanmanesh F, Duesing K, Bowden JC, Osmond-McLeod M, Ross J. 2016. Risk-conscious correction of batch effects: Maximising information extraction from high-throughput genomic datasets. BMC Bioinformatics. 17(1):332.

Pacheco KA. 2018. Occupational dermatitis: How to identify the exposures, make the diagnosis, and treat the disease. Ann Allergy Asthma Immunol. 120(6):583-591.

Palatnik de Sousa I, Maria Bernardes Rebuzzi Vellasco M, Costa da Silva E. 2019. Local interpretable model-agnostic explanations for classification of lymph node metastases. Sensors (Basel). 19(13):2969.

Parasuraman S. 2011. Toxicological screening. J Pharmacol Pharmacother. 2(2):74-79.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 14(4):417-419.

Paul WE. 2013. Fundamental immunology. Paul WE, editor. Philadelphia: LIPPINCOTT WILLIAMS & WILKINS, a WOLTER KUWER business.

Piroird C, Ovigne JM, Rousset F, Martinozzi-Teissier S, Gomes C, Cotovio J, Alépée N. 2015. The myeloid u937 skin sensitization test (u-sens) addresses the activation of dendritic cell event in the adverse outcome pathway for skin sensitization. Toxicology in vitro : an international journal published in association with BIBRA. 29(5):901-916.

Proksch E, Brandner JM, Jensen JM. 2008. The skin: An indispensable barrier. Exp Dermatol. 17(12):1063-1072.

Ramirez T, Mehling A, Kolle SN, Wruck CJ, Teubner W, Eltze T, Aumann A, Urbisch D, van Ravenzwaay B, Landsiedel R. 2014. Lusens: A keratinocyte based are reporter gene assay for use in integrated testing strategies for skin sensitization hazard identification. Toxicology in Vitro. 28(8):1482-1497.

Ribeiro MT, Singh S, Guestrin C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 1135-1144.

Ripley BD. 2007. Pattern recognition and neural networks. Cambridge university press.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. Limma powers differential expression analyses for rna-sequencing and microarray studies. Nucleic Acids Res. 43(7):e47.

Roberts DW. 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. Regulatory toxicology and pharmacology : RTP. 98:155-160.

Roberts DW, Mekenyan OG, Dimitrov SD, Dimitrova GD. 2013. What determines skin sensitization potency-myths, maybes and realities. Part 1. The 500 molecular weight cut-off. Contact Dermatitis. 68(1):32-41.

Robinson MD, McCarthy DJ, Smyth GK. 2010. Edger: A bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 26(1):139-140.

Rowan A, Spielmann H. 2019. Chapter 6.2 - the current situation and prospects for tomorrow: Toward the achievement of historical ambitions. In: Balls M, Combes R, Worth A, editors. The history of alternative test methods in toxicology. Academic Press. p. 325-331.

Russell WMS, Burch RL. 1959. The principles of humane experimental technique. Methuen.

Ryan CA, Kimber I, Basketter DA, Pallardy M, Gildea LA, Gerberick GF. 2007. Dendritic cells and skin sensitization: Biological roles and uses in hazard identification. Toxicology and Applied Pharmacology. 221(3):384-394.

Sakaguchi H, Ashikaga T, Miyazawa M, Yoshida Y, Ito Y, Yoneyama K, Hirota M, Itagaki H, Toyoda H, Suzuki H. 2006. Development of an in vitro skin sensitization test using human cell lines; human cell line activation test (h-clat). Ii. An inter-laboratory study of the h-clat. Toxicology in vitro : an international journal published in association with BIBRA. 20(5):774-784.

Savio LEB, de Andrade Mello P, da Silva CG, Coutinho-Silva R. 2018. The p2x7 receptor in inflammatory diseases: Angel or demon? Frontiers in Pharmacology. 9(52).

Schaefer L, Babelova A, Kiss E, Hausser H-J, Baliova M, Krzyzankova M, Marsche G, Young MF, Mihalik D, Götte M et al. 2005. The matrix component biglycan is proinflammatory and signals through toll-like receptors 4 and 2 in macrophages. J Clin Invest. 115(8):2223-2233.

Schiffer R, Neis M, Höller D, Rodríguez F, Geier A, Gartung C, Lammert F, Dreuw A, Zwadlo-Klarwasser G, Merk H et al. 2003. Active influx transport is mediated by members of the organic anion transporting polypeptide family in human epidermal keratinocytes. Journal of Investigative Dermatology. 120(2):285-291.

Shane HL, Long CM, Anderson SE. 2019. Novel cutaneous mediators of chemical allergy. J Immunotoxicol. 16(1):13-27.

Shi L, Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES et al. 2006. The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements. Nature Biotechnology. 24(9):1151-1161.

Silvestre MC, Sato MN, Reis VMSD. 2018. Innate immunity and effector and regulatory mechanisms involved in allergic contact dermatitis. An Bras Dermatol. 93(2):242-250.

Singh A, Thakur N, Sharma A. 2016. A review of supervised machine learning algorithms. 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). 1310-1315.

Smith Pease CK, Basketter DA, Patlewicz GY. 2003. Contact allergy: The role of skin chemistry and metabolism. Clinical and Experimental Dermatology. 28(2):177-183.

Soneson C, Gerster S, Delorenzi M. 2014. Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. PloS one. 9(6):e100335-e100335.

Steinman RM, Hemmi H. 2006. Dendritic cells: Translating innate to adaptive immunity. Curr Top Microbiol Immunol. 311:17-58.

Stephens ML, Mak NS. 2014. Chapter 1 history of the 3rs in toxicity testing: From russell and burch to 21st century toxicology. Reducing, refining and replacing the use of animals in toxicity testing. The Royal Society of Chemistry. p. 1-43.

Strickland J, Zang Q, Kleinstreuer N, Paris M, Lehmann DM, Choksi N, Matheson J, Jacobs A, Lowit A, Allen D et al. 2016. Integrated decision strategies for skin sensitization hazard. J Appl Toxicol. 36(9):1150-1162.

Su Z, Łabaj PP, Li S, Thierry-Mieg J, Thierry-Mieg D, Shi W, Wang C, Schroth GP, Setterquist RA, Thompson JF et al. 2014. A comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium. Nature Biotechnology. 32(9):903-914.

Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK et al. 2017. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell. 171(6):1437-1452.e1417.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES et al. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 102(43):15545-15550.

Sullivan KM, Enoch SJ, Ezendam J, Sewald K, Roggen EL, Cochrane S. 2017. An adverse outcome pathway for sensitization of the respiratory tract by low-molecular-weight chemicals: Building evidence to support the utility of in vitro and in silico methods in a regulatory context. Applied In Vitro Toxicology. 3(3):213-226.

Takahashi T, Kimura Y, Saito R, Nakajima Y, Ohmiya Y, Yamasaki K, Aiba S. 2011. An in vitro test to screen skin sensitizers using a stable thp-1–derived il-8 reporter cell line, thp-g8. Toxicological Sciences. 124(2):359-369.

Talhouk A, Kommoss S, Mackenzie R, Cheung M, Leung S, Chiu DS, Kalloger SE, Huntsman DG, Chen S, Intermaggio M et al. 2016. Single-patient molecular testing with nanostring ncounter data using a reference-based strategy for batch effect correction. PLOS ONE. 11(4):e0153844.

Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim J-S, Kim CJ, Kusanovic JP, Romero R. 2009. A novel signaling pathway impact analysis. Bioinformatics (Oxford, England). 25(1):75-82.

Taylor K, Rego Alvarez L. 2020. Regulatory drivers in the last 20 years towards the use of in silico techniques as replacements to animal testing for cosmetic-related substances. Computational Toxicology. 13:100112.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 58(1):267-288.

Toebak MJ, Pohlmann PR, Sampat-Sardjoepersad SC, von Blomberg BME, Bruynzeel DP, Scheper RJ, Rustemeyer T, Gibbs S. 2006. Cxcl8 secretion by dendritic cells predicts contact allergens from irritants. Toxicology in Vitro. 20(1):117-124.

Toloşi L, Lengauer T. 2011. Classification with correlated features: Unreliability of feature ranking and solutions. Bioinformatics. 27(14):1986-1994.

Touvron H, Vedaldi A, Douze M, Jégou H. 2019. Fixing the train-test resolution discrepancy. Advances in Neural Information Processing Systems. 8252-8262.

Tsepkolenko A, Tsepkolenko V, Dash S, Mishra A, Bader A, Melerzanov A, Giri S. 2019. The regenerative potential of skin and the immune system. Clin Cosmet Investig Dermatol. 12:519-532.

Villeneuve NF, Lau A, Zhang DD. 2010. Regulation of the nrf2-keap1 antioxidant response by the ubiquitin proteasome system: An insight into cullin-ring ubiquitin ligases. Antioxid Redox Signal. 13(11):1699-1712.

Wang Z, Gerstein M, Snyder M. 2009. Rna-seq: A revolutionary tool for transcriptomics. Nat Rev Genet. 10(1):57-63.

Wheelan SJ, Martínez Murillo F, Boeke JD. 2008. The incredible shrinking world of DNA microarrays. Mol Biosyst. 4(7):726-732.

WHO. 2016. The public health impact of chemicals: Knowns and unknowns. World Health Organization.

Wu D, Smyth GK. 2012. Camera: A competitive gene set test accounting for inter-gene correlation. Nucleic Acids Res. 40(17):e133.

Xiong Y, Soumillon M, Wu J, Hansen J, Hu B, van Hasselt JGC, Jayaraman G, Lim R, Bouhaddou M, Ornelas L et al. 2017. A comparison of mrna sequencing with random primed and 3′-directed libraries. Scientific Reports. 7(1):14626.

Yamamoto Y, Tahara H, Usami R, Kasahara T, Jimbo Y, Hioki T, Fujita M. 2015. A novel in chemico method to detect skin sensitizers in highly diluted reaction conditions. J Appl Toxicol. 35(11):1348-1360.

Yu L, Zhang J, Brock G, Fernandez S. 2019. Fully moderated t-statistic in linear modeling of mixed effects for differential expression analysis. BMC Bioinformatics. 20(24):675.

Zeller KS, 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.

Zeller KS, Johansson H, Lund T, Kristensen NN, Roggen EL, Lindstedt M. 2018. An alternative biomarker-based approach for the prediction of proteins known to sensitize the respiratory tract. Toxicology in vitro : an international journal published in association with BIBRA. 46:155-162.

Zhang X, Chen X, Song H, Chen HZ, Rovin BH. 2005. Activation of the nrf2/antioxidant response pathway increases il-8 expression. Eur J Immunol. 35(11):3258-3267.

Zhao S, Fung-Leung W-P, Bittner A, Ngo K, Liu X. 2014. Comparison of rna-seq and microarray in transcriptome profiling of activated t cells. PLOS ONE. 9(1):e78644.

Zyla J, Marczyk M, Weiner J, Polanska J. 2017. Ranking metrics in gene set enrichment analysis: Do they matter? BMC Bioinformatics. 18(1):256.