Machine-learned analysis of the association of next-generation sequencing–based human \textit{TRPV1} and \textit{TRPA1} genotypes with the sensitivity to heat stimuli and topically applied capsaicin

Dario Kringel\textsuperscript{a}, Gerd Geisslinger\textsuperscript{a}, Eduard Resch\textsuperscript{b}, Bruno G. Oertel\textsuperscript{b}, Michael C. Thrun\textsuperscript{b}, Sarah Heinemann\textsuperscript{a}, Jöhn Lötsch\textsuperscript{a,b,*}

Abstract
Heat pain and its modulation by capsaicin varies among subjects in experimental and clinical settings. A plausible cause is a genetic component, of which TRPV1 ion channels, by their response to both heat and capsaicin, are primary candidates. However, TRPA1 channels can heterodimerize with TRPV1 channels and carry genetic variants reported to modulate heat pain sensitivity. To address the role of these candidate genes in capsaicin-induced hypersensitization to heat, pain thresholds acquired before and after topical application of capsaicin and \textit{TRPA1}/\textit{TRPV1} exomic sequences derived by next-generation sequencing were assessed in \( n = 75 \) healthy volunteers and the genetic information comprised 278 loci. Gaussian mixture modeling indicated 2 phenotype groups with high or low capsaicin-induced hypersensitization to heat. Unsupervised machine learning implemented as swarm-based clustering hinted at differences in the genetic pattern between these phenotype groups. Several methods of supervised machine learning implemented as random forests, adaptive boosting, k-nearest neighbors, naive Bayes, support vector machines, and for comparison, binary logistic regression predicted the phenotype group association consistently better when based on the observed genotypes than when using a random permutation of the exomic sequences. Of note, \textit{TRPA1} variants were more important for correct phenotype group association than \textit{TRPV1} variants. This indicates a role of the \textit{TRPA1} and \textit{TRPV1} next-generation sequencing–based genetic pattern in the modulation of the individual response to heat-related pain phenotypes. When considering earlier evidence that topical capsaicin can induce neuropathy-like quantitative sensory testing patterns in healthy subjects, implications for future analgesic treatments with transient receptor potential inhibitors arise.

Keywords: Data science, Machine learning, Next-generation sequencing, Genetics of pain, Human, Experimental pain models

1. Introduction
The perception of pain after noxious stimulation involves a complex pathophysiology\textsuperscript{13} processed in a large network of nociceptive molecular pathways.\textsuperscript{30} This complexity extends to the perception of apparently uniform stimuli such as heat shown to follow a multimodal distribution.\textsuperscript{30} This hints at interindividual differences in involved sensors of which the largest group belongs to the transient receptor potential (TRP) channels.\textsuperscript{31} In particular, TRPV1 is known as a thermosensitive channel involved in nociception,\textsuperscript{61} and in addition to heat, it is also gated by pungent chemicals such as vanilloids including capsaicin.\textsuperscript{15}

Synergistic effects of chemical and thermal gating are used to study mechanisms of thermal hyperalgesia in humans.\textsuperscript{56} Although hyperalgesia varies among patients with pain,\textsuperscript{42} in experimental settings topical capsaicin application induces hyperalgesia only in a fraction of subjects.\textsuperscript{42} This may point at a genetic background where \textit{TRPV1} as a primary candidate gene is playing a role in both, heat sensation and capsaicin hypersensitization. However, associations of genetic variants with the heat sensitivity or hypersensitization by capsaicin were only rarely reported. However, the only hint at an association of \textit{TRPV1} genetics with heat pain sensitivity in humans points at the rs8065080 single-nucleotide polymorphism (SNP),\textsuperscript{35,68} which was not replicated.\textsuperscript{35} Moreover, an unexpected role of a \textit{TRPA1} genetic variant rs11988795 in heat pain was reproduced.\textsuperscript{35,68} The coexpression of \textit{TRPV1} and \textit{TRPA1}\textsuperscript{72} and the flexibility of the TRP family channels raise the possibility that these channels might interact to influence the properties of one another.\textsuperscript{20} In this regard, recently reported heteromerization among the TRP channels is suggestive of the mechanism for interactions.\textsuperscript{28}

Based on this evidence and considering the unresolved role of \textit{TRPV1} variants for the modulation of human pain sensitivity, despite the molecular plausibility of an involvement, the present
analysis addressed the association between TRPV1 and TRPA1 genotypes with a human phenotype of capsaicin-induced hyperalgesia to heat stimuli. With the broader availability of next-generation sequencing (NGS), the limitation to known functional variants has fallen in favor of unrestricted access of TRP channel genetics. Therefore, it is not necessary to assess the phenotypic role of TRP channel genotypes for selected single variants. This accommodates increasing molecular evidence that noncoding variants can affect mRNA splicing, stability, and structure, resulting in a reduced transcriptional efficiency rendering them potentially functionally relevant. Hence, a recently developed genetic panel was used to address the role TRPV1 and TRPA1 genetic variants in the sensitivity to nociceptive heat and in the reaction to hypersensitization with topical capsaicin recently assessed in a cohort of healthy subjects.

2. Methods

2.1. Data sets, subjects, and study design

The phenotype data sets and DNA samples were available from a previous study enrolling n = 100 healthy volunteers (46 men) of Caucasian ethnicity by self-assignment, aged 19 to 42 years (mean ± SD 25 ± 3.5 years). In this data set, phenotypic measurements from n = 82 subjects were nonmissing and included in the present analysis. The study followed the Declaration of Helsinki and was approved by the Ethics Committee of the Goethe-University Medical Faculty, Frankfurt am Main, Germany. Informed written consent in the study procedures including the genotyping had been obtained from each participating subject.

Inclusion criteria were age between 18 and 50 years and no relevant current medical history. The subjects’ actual health had been ascertained by medical history and physical examination including vital signs. Exclusion criteria were drug intake during the previous week, except for oral contraceptives and vitamin or hormone-substituting drugs (eg, L-thyroxin), a current clinical condition involving pain, and current diseases according to questioning and medical examination. Alcohol was prohibited for 24 hours before the actual experiments. Before the experimental tests, all subjects completed training sessions with pain tests applied to an area different from the planned skin areas.

2.2. Assessment of heat pain sensitivity

In the capsaicin experimental pain model, chemical methods of nociceptor stimulation were used to produce stable and long-lasting hyperalgesia with a low potential for skin injury, in the original publication supplemented by heat stimulation. Topical application of 150 mg capsaicin cream (0.2%, manufactured by the local Hospital Pharmacy) onto a 3 × 3 cm² skin area was used. Subsequently, the area was covered with a plaster for 30 minutes.

Quantitative sensory testing (QST) was performed at baseline and after application of capsaicin. A clinically established QST test battery proposed by the German Research Network on Neuropathic Pain was used. For the present report, pain thresholds to noxious heat were selected. They were assessed using a 3 × 3 cm thermode (TSA 2001––II; Ramat Yishai, Israel) on a 9 cm² skin area at the inside of the forearm without any superficial veins or birth marks. Heat pain thresholds (HPTs) were measured by increasing the temperature of the thermode by 1°C/s, starting at 32°C, until the subject indicated pain, which triggered the reversal of the temperature ramp back to the baseline. According to the published instructions for the QST test battery, the HPT was defined as the mean of 3 measurement repetitions. During testing, the room temperature was kept at 20 to 25°C.

Data were preprocessed according to the QST test battery instructions, which included uniform direction along increasing stimulus intensity as $HPT_0 = HPT - 32°C$, where the subscript T denotes the data transformation. The values of $HPT_0$ were mapped onto the distribution of the reference group that consists of 180 healthy subjects, in whom a data set of 1080 QST parameter values has been obtained. This serves as the reference for all QST-based diagnoses. Therefore, according to the QST standard procedure, the individual QST parameter values were $z$-transformed as $z_{QST,individual} = (QST_{individual} - QST_{reference}) / SD_{reference}$, with QST reference values with regard to the sex, age, and tested body site of the actual subject taken from. The signs of the $z$-scores, $z_{HPT_0}$, were adjusted to denote that a $z$-score > 0 indicates high sensitivity and $z$-score < 0 indicates low sensitivity, according to the standardized instructions. The effect of capsaicin was quantified as the difference between the measurement after capsaicin application and the measurement without the presence of capsaicin, ie, $CapsEff = z_{HPT_{T, capsaicin}} - z_{HPT_{T, baseline}}$.

2.3. Transient receptor potential channel genotyping using next-generation sequencing

Next-generation sequencing of TRPA1 and TRPV1 genes was based on a custom AmpliSeq library and performed using a validated assay on an Ion Torrent personal genome machine as described in detail previously. In brief, genomic DNA was extracted from 200 µL venous blood on a BioRobot EZ1 workstation applying the blood and body fluid spin protocol provided in the EZ1 DNA Blood 200 µL Kit (Qiagen, Hilden, Germany). A multiplex amplification primer set for the exonic sequences of the TRP channel genes was designed online using a web tool (Ion Ampliseq Designer; Life Technologies, Darmstadt, Germany) provided by the manufacturer of the NGS device at http://www.ampliseq.com.

The present amplification design obtained coverage of 96% of target sequence. After sequencing, signal processing was performed using Torrent Suite software (version 5.2.2; Life Technologies), base calling and the generation of unmapped and mapped binary alignment map files (hg19 reference genomic sequence) were performed. Variant calling across the hg19 reference genomic sequence was performed with the Torrent Variant Caller Plugin (minimum quality = 10, minimum coverage = 20, and minimum coverage on either strand = 3) and variant annotation was performed using Ion Reporter Software (version 5.2.2; Life Technologies). Variant call format files containing the nucleotide reads were processed toward the individual genotypes using GenomeBrowse software (Version 2.0.4; Golden Helix, Bozeman, MT) and SNP and Variation Suite software (Version 8.7.1; Golden Helix).

2.4. Data analysis

To accommodate a large number of genetic variants expected to result from the NGS-based genotyping, the main genotype–phenotype association analysis was implemented using a novel approach based on machine-learned techniques (for an overview on machine learning in pain research, see 49). The main idea was to train an artificial intelligence, implemented as different types of machine learning, to learn the association of the genetic information with the pain-related phenotype, and to subsequently use the trained intelligence to predict a phenotype in new data
from genetic information. If this performed better than guessing the phenotype or than using genetic information unrelated to the phenotype, a genotype–phenotype association can be concluded as supported by the data. Machine learning was a priori preferred to the sole use of traditional approaches such as logistic regression analysis because of the expected high dimensionality and collinearity of the rich genetic information; indeed, the nevertheless included regression analysis was outperformed by several machine-learned methods (see Results section). The concept of training an artificial intelligence with genetic information to enable it to correctly associate an individual with a pain phenotype class required measures against overfitting, which are usually implemented as splitting the data set into a training subset that is provided to the artificial intelligence during the learning phase and a test subset which is not seen by the artificial intelligence during learning but provided when the learned algorithm is used for classification; usually, this procedure is repeated several times in a resampling design.\textsuperscript{54}

Data were analyzed using the R software package (version 3.4.1 for Linux; http://CRAN.R-project.org/)\textsuperscript{30} on an Intel Xeon computer running on Ubuntu Linux 16.04.3 64-bit. Supervised and unsupervised machine learning was used for genotype vs phenotype association. Machine learning addresses the so-called data space \( D = \{ (x_i, y_i) \mid x_i \in X, y_i \in Y, \ i = 1, \ldots, n \} \) including an input space \( X \) comprising vectors \( x_i = \langle x_{i1}, \ldots, x_{id} \rangle \) with \( d > 0 \) different parameters (here, the genetic information) acquired from \( n > 0 \) cases belonging to the output classes \( y_i \) (eg, a pain-related phenotype). In unsupervised learning, the class information is disregarded and only the so-called feature space comprising an unlabeled data set of \( D = \{ (x_i) \mid x_i \in X, \ i = 1, \ldots, n \} \), composed of values \( x_i \in X \subset \mathbb{R}^d \) comprising the \( d \) features, respectively, genetic markers is searched with the goal to find “interesting” structures, which can be associated subsequently with the phenotypes. By contrast, in supervised machine learning, an algorithm is trained on data for which the class labels of the cases are known that is able to assign future cases for which this class label information is unknown to the right class (prediction and generalization).\textsuperscript{18}

The analysis was performed in 4 main steps comprising (1) creation of a phenotype group structure, (2) preprocessing of the \( TRPV1 \) and \( TRPA1 \) NGS genetic information, (3) identification of a genetic marker pattern and its relation to the phenotype classes, and (4) finding a mapping of the genetic parameters to the phenotype classes.

### 2.5. Identification of capsaicin sensitivity phenotype classes

The first step of the data analysis aimed at establishing the output data space, ie, a phenotype class structure. Therefore, the distribution of the changes after capsaicin application, CapsEff, was investigated by analyzing the probability density function (PDF) as described previously.\textsuperscript{45,46} In brief, the Pareto density estimation (PDE), ie, a kernel density estimator particularly suitable for the discovery of groups in the data,\textsuperscript{81} was used. A multimodal distribution of the pain responses was assessed by fitting a Gaussian mixture model (GMM) to the PDEs as

\[
P(x) = \sum_{i=1}^{M} \omega_i N(x|m_i, s_i) = \sum_{i=1}^{M} \omega_i \frac{1}{\sqrt{2\pi} s_i} e^{-\frac{(x-m_i)^2}{2s_i^2}},
\]

where \( N(x|m_i, s_i) \) denotes Gaussian probability densities (components) with mean values \( m_i \) and SDs \( s_i \). The \( \omega_i \) denotes the mixture weights indicating the relative contribution of each Gaussian component to the overall distribution, which add up to a value of 1. \( M \) denotes the number of components in the mixture. Gaussian mixture model fitting was performed with our R package “AdaptGauss” (https://cran.r-project.org/package=AdaptGauss).\textsuperscript{86} To determine the optimum number of components, model optimization was performed for \( M = 1 \) to 5 components. The final model was selected based on likelihood ratio tests.\textsuperscript{73} In addition, the Kolmogorov–Smirnov test\textsuperscript{72} was applied to assess whether the observed distribution differed significantly from the expectation from the model, and the quality of the model to fit the distribution was assessed visually using a quantile–quantile (QQ) plot. Subject association to the identified subgroups was obtained using the Bayes’ theorem\textsuperscript{72} that provided the probability that an individual observation belongs to mode \( i \) calculated as the posterior probability. Thus, the output space \( Y \) was obtained, comprising \( y_i \in C = \{ 1, \ldots, c \} \), where \( c \) denotes possible unambiguous classes \( c \) where every \( y_i \) has a unique class label and the number of classes was equal to the number of Gaussian modes, \( M \).

### 2.6. Preprocessing of the genetic information

The determination of single-nucleotide variants from the NGS data refers to the Software plugin “The Torrent Variant Caller” (TVC) provided by Life Technologies. A variant is defined as a nucleotide disagreeing with the nucleotide in the reference sequence. The TVC plugin calls SNPs, multinucleotide polymorphisms, insertions, and deletions in a sample across a reference (hg19). In the second step of the analysis, the genetic information (mainly SNPs) was curated by (1) eliminating non-informative variants and (2) creating of negative and positive genetic control data sets with respect to a possible association of the genotype with the phenotypes. Variants were eliminated for which the distribution of homozygous and heterozygous carriers differed from expectation according to the Hardy–Weinberg equilibrium.\textsuperscript{26} This was judged by means of Fisher’s exact tests\textsuperscript{51} using the R package “HardyWeinberg” (https://cran.r-project.org/package=HardyWeinberg).\textsuperscript{25} To avoid the inclusion of non-informative variants such as those carried by only very few subjects into the classifier, informative gene loci were detected based on the Shannon information\textsuperscript{69} computed as

\[
\text{Info} = -P_0 \ln(P_0) - P_1 \ln(P_1),
\]

where \( P_0 \) and \( P_1 \) are the observed probabilities of the nonobservation (0) or observation (1), respectively, of a variant allele in the ith gene locus. The precise limit of the Shannon information up to which a gene locus could be regarded sufficiently informative, was calculated by means of a computed ABC analysis.\textsuperscript{83} This is a categorization technique for the selection of a most important subset among a larger set of positive numerical items. It divides the set into 3 disjoint subsets “A,” “B,” and “C” and referred to in economic sciences where the method originates as “the important few” (set “A”) vs “the trivial many” (set “C”),\textsuperscript{31} whereas set “B” comprises items between the 2 extremes including elements where an increase in effort is proportional to the increase in yield. However, although earlier applications of ABC analyses parted the item set according to the so-called 80/20 rule, which sets the limit between sets “A” and “B” at 80% of the yield achieved with 20% efforts, this limit is based on mathematical calculations in computed ABC analysis\textsuperscript{65} implemented in our R package “ABCAnalysis” (http://cran.r-project.org/package=ABCAnalsysis).\textsuperscript{82} As subset “A” can be regarded as containing the most profitable features,\textsuperscript{21,55} it was chosen for classifier establishment. The limit to set “B” was found at Shannon information \( = 0.339 \). Furthermore, as implemented previously,\textsuperscript{38} further variants unlikely to provide a suitable basis for phenotype class assignment were excluded. In the present analysis, this was approached through the effect sizes of the allelic distribution between the phenotype classes used classic \( \chi^2 \) statistics.\textsuperscript{57} The
values of $\chi^2$ obtained for each gene locus were submitted to a computed ABC analysis described above. Here, only the clearly unsuitable variants were omitted, ie, ABC set “C” regarded as comprising “the trivial many.”

Genetic control data sets were created by rearranging the original genotype information. Specifically, a negative control feature set was obtained by random permutation of the genetic data. The expectation was that the association with the phenotypes was not better than guessing and should be consistently outperformed by the mapping of the true genotypes to the phenotypes using different machine-learned methods. In addition, a positive control feature set was obtained by sorting the original genotype information at each locus in descending order data of the number of variant alleles along the sorted phenotype classes (Fig. 1). The expectation was that the association with the phenotypes could be almost perfectly obtained by all machine-learning methods.

2.7. Identification of a genetic marker pattern and its relation to the phenotype classes

In the third step of the analysis, the genetic information was explored for data structures. Their existence would support that the TRPA1 and TRPV1 NGS genotypes were not homogenously distributed among the subjects but hinted at subgroups of subject based on the genetic information. This would be a first step to further explore the data for a possible relation of the genotype-based subgroups with the phenotype classes. Hence, the preprocessed genetic information was analyzed for patterns using unsupervised machine learning, which was implemented as a swarm of intelligent agents called “DataBots.” The data space $D = \{x_i, i = 1, \ldots, n\} \subset \mathbb{R}^d$, comprising $d$ genetic markers acquired in $n$ subjects was explored for distance-based structures using the cityblock (Manhattan) distance as used elsewhere for genetic data scaled $[0,1,2]$. To explore this feature space, topographic mapping was used, which provides data projection methods to create low-dimensional images from high-dimensional data. Specifically, topographic mapping was implemented as swarm intelligence, ie, an algorithm guided by the flocking behavior of numerous independent but cooperating the so-called DataBots, which are self-organizing artificial “life forms” identified with single data objects (subjects). These “DataBots” can move on a 2-dimensional grid, and their movements are either random or follow the attractive or repulsive forces proportionally to the (dis-)similarities of neighboring

Figure 1. Patterns of the TRPA1 (chromosome 8: X8) and TRPV1 (chromosome 17: X17) genotypes observed in $n = 75$ healthy volunteers of Caucasian ethnicity for whom phenotype data of the heat hypersensitization after capsaicin application were available. The heat plot shows the occurrence of variants (columns) per subject (lines). The genetic information is color coded as the number of nonreference alleles found at the respective locus in the respective sample as white, 0 nonreference alleles = wild type genotype; green, heterozygous; and blue, 2 nonreference alleles. Thus, the individual genotypes are given by the vectors (rows) associated with each subject (subjects count at the right of each panel). The bar plot at the left shows the phenotype group association, with gray indicating Gaussian #1 and black indicating Gaussian #2 in Figure 2. The original genotype information (left) was permuted to obtain a negative control data set for the association of genotypes with phenotypes, and sorted in descending order of alleles at each gene locus to obtain a positive control data set for the genotype-phenotype association. The figure has been created with the R software package (version 3.4.1 for Linux; http://CRAN.R-project.org) using the library “gplots” (Warnes et al., https://cran.r-project.org/package=gplots).
“DataBots.” Specifically, a parameter-free focusing projection method of a polar swarm, Pswarm, was used that exploits concepts of self-organization and swarm intelligence. During construction of this type of projection, which is called the learning phase and requires an annealing scheme, structure analysis shifts from global optimization to local distance preservation (focusing). Intelligent agents of Pswarm operate on a toroidal grid where positions are coded into polar coordinates allowing for a precise definition of their movement, neighborhood function, and annealing scheme. The size of the grid and, in contrast to other focusing projection methods, the annealing scheme is data based and therefore, the method does not require any parameters. During learning, each DataBot searched for the strongest “scent,” ie, for other agents that carried data with most similar features as it carried itself, by moving across the grid or staying in its current position, with a decreasing search radius.

After successful swarm learning, DataBots carrying items with similar features, ie, DataBots associated with similar data points, are placed in groups on the projection grid. The identification of emergent structures was enhanced on top of the learned structure. To this end, the distances between data points were calculated with the so-called U matrix shown previously to provide emergent structures corresponding to clusters and outperforming classic clustering methods. Every value (height) in the U matrix depicts the average high-dimensional distance of a prototype to all immediate neighboring prototypes regarding grid position. The corresponding visualization technique is a topographical map with hypsometric colors facilitating the recognition of data structures. The calculations were performed using the R library “DatabionicSwarm” (Thrun M, https://cran.r-project.org/package=DatabionicSwarm). Subsequently, clusters in the projected data were verified using the Ward method. Finally, a possible association of the genotype-based clusters with the phenotype classes was assessed using the Fisher exact statistics. In case of a positive association, this established that the genetic data were related to the pain phenotype, which was addressed in the next step of the data analysis.

2.8. Mapping of the genetic parameters to the phenotype classes

After establishment of a data structure in the genotype that reflected the phenotype structure, the association of the genotype with the phenotype was further analyzed. Therefore, in the fourth step of the data analysis, the question was pursued whether the phenotype can be predicted from the genotype. This was achieved by means of supervised machine learning, which addresses the data space \( D = \{ (x_i, y_i) | x_i \in X, y_i \in Y, i = 1, ..., n \} \) and tries to find a mapping of the input space \( X \), comprising vectors \( x_i = < x_{i,1}, ..., x_{i,d} > \) with \( d > 0 \) different parameters (here, the genetic information) acquired from \( n > 0 \) cases, to the output space \( Y \), comprising \( y_i \) classes, eg, a pain-related phenotypes obtained through GMM and subsequent calculation of the Bayesian decision limits used for class separation.

In the present analysis, the mapping of the input space to the output space was performed using different methods of supervised machine learning, ie, (1) random forests, (2) adaptive boosting, (3) k-nearest neighbors (kNNs), (4) naive Bayesian classifiers, (5) support vector machines, and (6) logistic regression, which provided an internal validation of the results without the intention to compare the performances between machine-learning methods. The machine-learning methods were applied on the original data set and on the negative and positive control data set created as described above. The expectation was to observe a prediction of the phenotypes that were consistently better across several methods when using the original genotypes than when using the permuted genotypes, which should provide a classification performance not superior to guessing. In all data sets, the classifiers were trained at training data subsets comprising 2/3 of the data, and subsequently their performance was estimated on the test data subset consisting of the remaining 1/3 of the data. This was repeated in 1000 cross-validation runs using Monte-Carlo resampling and random splits of the original training data set into new training and test data subsets, using the R library “sampling” (https://cran.r-project.org/package=sampling).

Random forests create sets of different, uncorrelated, and often very simple decision trees with conditions on features as vertices and classes as leaves. The splits of the features are random and the classifier relates on the majority vote for class membership provided by a large number of decision trees. In the present analysis, 1000 decision trees were built containing \( \sqrt{d} \) features, respectively, to nucleotide positions as the standard setting implemented in the R library “randomForest” (https://cran.r-project.org/package=randomForest).

The number of trees was heuristically based on visual analysis of the relationship between the number of decision trees and the classification accuracy, which indicated that beyond 100 trees, the classification balanced accuracy remained stable and a larger number merely consumed available computation time (Supplementary Fig. 1, available online at http://links.lww.com/PAIN/A561).

Boosting approaches classification through a set of weak learners from which a single strong learner is created. As weak classifiers served small classification and regression trees, which provide a simple form of classification rules using the Gini impurity to find optimal (local) dichotomic decisions. In the present analysis, adaptive boosting as a successful algorithm for binary classification was used, in which during the learning phase, subsequent weak learners are tweaked in favor of those data instances that had been misclassified by previous classifiers. Initially, each of \( n \) data point is associated with the same weight \( w_i = 1/n \). A learner was trained to assign the correct class to each data point. Iteratively, the weights of misclassified data points were increased such that the subsequent learner gave more focus on the misclassified items. The final model combined all models using a weighted sum of the outputs that reflect the accuracy of all the constituent models. The number of iterations was heuristically based on the classification accuracy, which indicated no improvement beyond 500 runs, from which 1000 iterations were considered to provide robust results. These calculations were performed using the R package “ada” (http://cran.r-project.org/package=ada) with the partitioning and classification package “rpart” (https://cran.r-project.org/package=rpart).

The kNN classification is a nonparametric method that belongs to the most frequently used algorithms in data science, although it is one of the basic methods in machine learning. During kNN model building, the entire labeled training data set is stored while a test case is placed in the feature space in the vicinity of the test cases at the smallest high-dimensional distance. The test case receives the class label according to the majority vote of the class labels of the k-training cases in its vicinity. In the present implementation, the size of \( k \) was established in resampling experiments with \( k \) set at 3 or 5. Even numbers for \( k \) intuitively make a majority vote on which the class assignment is based difficult when one of the nearest neighbors belongs to class 1 and the other to class 2. We tested 3 and 5 because these are often used and the default in various implementations of kNN. A silhouette plot would show the quality
of a clustering and to compare alternatives, eg, with different numbers of clusters. However, here, we used kNN as a classifier for a predefined number of classes (c = 2), not to obtain clusters or to reassess the number of classes in the data that had been obtained by means of GMM. At k = 3 and using the Manhattan distance32 as used elsewhere34 for genetic data, the best classification accuracy of the classifier was observed in 100 runs on randomly resampled data. Other distances such as the Euclidian, Jaccard, or Bray–Curtis distances, or more sophisticated implementations of nearest neighbor–based class assignment such as weighting or the use of kernel of different shapes were tried but did not provide any improvements regarding the basic version. These calculations were performed using the R package “KernelKnn” (Mouselimis L, https://cran.r-project.org/package=KernelKnn).

Bayesian classifiers were used that provide the probability that a data point being assigned to a specific class calculated by application of the Bayes’ theorem.2 In naive Bayesian classifiers, the oversimplified assumption is included that all features are conditionally independent of each other, which is a widely used technique to assign class labels to the samples from the available set of features, describing a special case of the more general Bayesian network model. The calculations were performed using the R package “e1071” (Meyer D, https://cran.r-project.org/package=e1071).

Support vector machines are supervised learning methods that classify data mainly based on geometrical and statistical approaches used for finding an optimum decision surface (hyperplane) that can separate the data points of 1 class from those belonging to another class in the high-dimensional feature space.10 Using a kernel function, the hyperplane is frequently selected in a way to obtain a tradeoff between minimizing the misclassification rate and maximizing the distance of the plane to the nearest properly classified data point. In the present analysis, a Gaussian kernel with a radial basis was used. The analyses were performed using the R library “kernlab” (https://cran.r-project.org/package=kernlab).32

Finally, logistic regression91 was used to map the genotype information to the 2 phenotype classes. This accommodated the inclusion of a more classic data analysis method well known from statistics. Logistic regression estimates the probability of falling into a certain level of the categorical response given a set of predictors. The calculations were performed using the “glm” command and the “family = binomial” switch as implemented in the R “stats” package60 provided with the basic installation of the software core package (http://www.R-project.org/). The performances of all classifiers were assessed on the test data subsets created during cross-validation and are reported as the median of the resampling runs. Finally, a classic $\chi^2$–test–based genotype vs association was performed.

3. Results

3.1. Capsaicin sensitivity phenotype classes

Phenotype data (HPTs acquired before and after topical application of capsaicin) were complete from 82 subjects. For technical reasons, data from 18 subjects were incomplete and therefore, these subjects were excluded from all analyses. After capsaicin application, a right shift in the pain thresholds to heat stimuli, calculated as stimulus intensity $HPT_{T} = HPT - 32^\circ C$ (Fig. 2), was observed. The shift was pronounced enough to place the cohort in the range of HPT values typical for neuropathic patients according to the reference values of the QST test battery.52 That is, while at baseline, only 6 pathological values were observed; after capsaicin application, 78 of the 82 subjects displayed pathological HPT values.

Visual inspection of the probability density distribution (PDF) of the capsaicin effects, CapsEff, suggested a multimodal distribution (Fig. 2). This was statistically supported by a significant likelihood ratio test ($P = 1.87 \times 10^{-10}$) comparing the goodness of the fits of the PDF, estimated using the PDE, between a single Gaussian mode and a GMM using $M = 2$ modes. No more significant improvement of the fit was obtained when a further Gaussian was added, based on likelihood ratio tests ($P = 0.9403$ for $M = 3$ vs $M = 2$). A satisfactory fit by a GMM with $M = 2$ was also supported by the nonsignificant result of the Kolmogorov–Smirnov test ($P = 0.962$) and the visual inspection of the QQ plot (Fig. 2). The parameter values of the final GMM are provided in Table 1. Thus, the output space was structured into 2 classes containing $n = 24$ and 58 subjects with low or high hypersensitization response to heat after topical application of capsaicin, respectively.

3.2. Association of TRPV1 and TRPA1 genotypes with capsaicin sensitivity phenotypes

Next-generation sequencing data were obtained from 75 subjects distributed across phenotype classes in a proportion of $n = 24$ and $n = 51$. The genetic information initially comprised 278 loci wherein at least 1 subject an allele differing from the hg19 reference genomic sequence was observed. The TRPA1 gene at chromosome 8 displayed 134 loci with variant alleles and the TRPV1 gene at chromosome 17 displayed 144 loci with variant alleles. All variant alleles were observed at frequencies corresponding to the expectations from the Hardy–Weinberg equilibrium (Fisher exact tests: $P$ always $> 0.05$). After feature selection based on the Shannon information criterion and the ABC analysis of the Chi2 statistics for phenotype group differences (Fig. 3), $d = 31$ genetic features remained in the data set comprising 25 variants in the TRPA1 gene and 6 loci in the TRPV1 gene (Fig. 1) with different putative molecular functional consequences (Table 2). This corresponded to the size of the genetic features used in a previous study with comparable data analysis.59 The frequencies of the minor alleles, ie, those disagreeing with the hg19 reference genomic sequence, in the analyzed data set ranged between 5% and 61%, with a median of 28%.

3.3. Genetic marker pattern and its relation to the phenotype classes

Unsupervised machine learning, aiming at data structure detection, was applied to analyze the $75 \times 31$–sized matrix comprising $d = 31$ genetic variants acquired in $n = 75$ subjects. Training of a swarm of intelligent data bots provided a structure-preserving projection of the high-dimensional data space $D = \{x_i | i = 1, \ldots, n\} \subset \mathbb{R}^d$ onto a 2-dimensional toroid projection grid (Fig. 4). After addition of the U matrix, a cluster structure emerged from the separation of the data bots carrying the genetic information into 2 distinct groups as visually indicated by a "mountain range" on the topographic map analogy (Fig. 4 top). This was verified by Ward clustering that indicated 2 clusters differing with respect to the pattern of genetic variants (Fig. 4). Finally, the cluster membership was found to be unequally distributed among the phenotypes (the Fisher exact test: $P = 0.01199$), ie, the swarm-based cluster #1 comprising subjects carrying few variant alleles was underrepresented in phenotype cluster (Gaussian) #1 comprising subjects with low heat
were obtained using our package "AdaptGauss" (https://cran.r-project.org/package=AdaptGauss). QST, quantitative sensory testing.

3.4. Mapping of the genetic parameters to the phenotype classes

After establishment of a relation between the TRPA1 and TRPV1 NGS-based genetic patterns with the phenotype classes, the genotype–phenotype association was further analyzed. The classic $\chi^2$-based genotype vs phenotype association analysis was negative, ie, only the 2 TRPA1 variants X8.72934391.SNV and X8.72969263.SNV differed in allelic distribution between phenotype groups, but only at the uncorrected $p$ level (Fig. 3) while when corrected according to Bonferroni, the $p$ level of 0.0016 resulting for the $d=31$ genetic variants was exceeded for all gene loci. Subsequently, supervised machine leaning was applied in cross-validation experiments using 1000 Monte-Carlo random resamplings of 2/3 vs (new training) 1/3 (new test) of the data.

Table 1. Parameter values of Gaussian mixture models (GMMs)

| GMM parameter | $i=1$ | $i=2$ |
|---------------|-------|-------|
| $m_i$ (2HPT$_T$-capsaicin − 2HPT$_T$-baseline) | 1.371 | 3.215 |
| $\sigma_i$ | 0.567 | 0.629 |
| $w_i$ | 0.307 | 0.693 |

Bayesian decision limit $\frac{[2HPT_T,capsaicin − 2HPT_T,baseline]}{[2HPT_T,capsaicin − 2HPT_T,baseline]} = 2.107$

A mixture of $M=2$ Gaussians (Fig. 2) was found to provide the best fits, as indicated by likelihood ratio tests. HPT, heat pain threshold.

Figure 2. Original heat pain thresholds (HPTTs) and distribution of the effects of capsaicin. Top: One-dimensional scatter plot of the observed individual heat pain sensitivity (dots; raw data). At the upper half (green dots), the values acquired at baseline are shown, whereas at the lower half, the values acquired after topical application of capsaicin are shown (blue dots). Bottom: The distribution of the capsaicin effects, obtained from the z-transformed HPTTs according to the QST standard procedure (CastEff = zHPT$_T$-capsaicin − zHPT$_T$-baseline and shown as probability density function (PDF) estimated by means of the Pareto density estimation (PDE)) black line) overlaid on a histogram could be fitted using a Gaussian mixture model (GMM) given as $P(x) = \sum_{i=1}^{M} w_i N(x|m_i, \sigma_i)$, with $M=2$ modes. The fit is shown as a red line and the $M=2$ mixes are indicated as differently colored dashed lines (G #1–#2). The Bayesian boundary between the Gaussians is indicated as a perpendicular magenta line. At the right side, a quantile–quantile (QQ) plot is shown comparing the observed distribution of cold pain data (ordinate) with the distribution expected from the GMM (abscissa). The blue dots symbolize the quantiles of observed data vs predicted data and the red line indicates identity, ie, the agreement between the data distribution expected from the model with the observed data distribution. The close vicinity of the dots to this line indicates satisfactory fits of the data by the respective GMM. The figure has been created using the R software package (version 3.4.1 for Linux; http://CRAN.R-project.org/) in particular, the dot plot was drawn using the R library "beeswarm" (Eklund A, https://cran.r-project.org/package=beeswarm) and the GMM plots were obtained using our package "AdaptGauss" (https://cran.r-project.org/package=AdaptGauss). QST, quantitative sensory testing.
whereas the first TRPV1 variants figured only at rank 7 among the classification-relevant gene loci. This observation accompanied the results of the classic χ²-based genotype vs phenotype association analysis, in which only the 2 TRPA1 variants X8.72934391.SNV and X8.72969263.SNV differed in allelic distribution between phenotype groups, however, only at the uncorrected α level (Fig. 3). These variants could also be used for phenotype class association; however, when eliminating them from the data set, a phenotype association was still consistently better than chance (Table 3 and Fig. 3B), which supports that a complex genotype rather than a single variant modulated the phenotype.

4. Discussion

In the present analysis, several different methods of data analysis pointed toward a contribution of human TRP channel genotypes to the individual susceptibility to capsaicin-induced hypersensitization to heat stimuli. This was firstly hinted at by a high-dimensional pattern that emerged in the genotypes and could be statistically significantly associated with the 2 generated phenotype classes. Subsequently and most importantly, an importance of TRP genotypes for the heat pain–related phenotypes could be supported by the consistently better prediction of phenotypes from the genetic information than by chance, which was similarly observed across all machine-learned methods applied that always outperformed the phenotype class prediction when using randomly permuted genetic markers. Thus, the results can be summarized as an association of a complex TRP channel–related NGS genotype with the phenotype of the individual sensitivity to heat pain–related phenotypes.

The 31 genetic variants in the TRPA1 and TPRV1 genes that after feature selection were included in the association analyses, comprised 4 missense, 3 synonymous, and 1 deletion/insertion variation (Table 2), whereas the majority was located in introns or untranslated regions of the genes. The 2 polymorphism that differed in allelic distribution between phenotype classes at the uncorrected α level, i.e., rs9298197 and rs3779752, and in addition, the rs2278654 variant that got the highest random-forest–based rank among all genetic loci, are located in noncoding areas of the TRPA1 gene. Although they cannot affect the protein structure directly, recent studies in cancer tissue have highlighted the importance of noncoding variants and indeed, the majority of variants, both somatic and germline, had biological consequences as queried from several publicly available databases (NCBI gene index database at http://www.ncbi.nlm.nih.gov/gene; GeneCards at http://www.genecards.org, Short Genetic Variations database [dbSNP] at https://www.ncbi.nlm.nih.gov/snp and the “1000 Genomes Browser” at https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes; all accessed in August 2017).

Table 2

| Gene | Variant | DNA change | Molecular consequence | dbSNP ID | Region score | TSS score | Unmatched score |
|------|---------|------------|-----------------------|----------|--------------|-----------|----------------|
| TRPA1 | X8.72933632.SNV | C>T | 3 prime UTR variant | rs6996723 | 0.34 | 0.63 | 0.78 |
| TRPA1 | X8.72934032.SNV | A>G | 3 prime UTR variant | rs7827617 | — | — | — |
| TRPA1 | X8.72934391.SNV | G>T | 3 prime UTR variant | rs9298197 | 0.34 | 0.7 | 0.77 |
| TRPA1 | X8.72936145.SNV | T>C | Missense variant | rs9659976 | 0.29 | 0.07 | 0.36 |
| TRPA1 | X8.72938332.SNV | G>A | Intron variant | rs2305017 | 0.25 | 0.19 | 0.22 |
| TRPA1 | X8.72938357.SNV | A>C | Intron variant | rs2305018 | 0.24 | 0.17 | 0.18 |
| TRPA1 | X8.72938359.SNV | T>C | Intron variant | rs2305019 | 0.23 | 0.2 | 0.17 |
| TRPA1 | X8.72987959.SNV | G>A | Synonymous variant | rs61757563 | 0.23 | 0.1 | 0.37 |
| TRPA1 | X8.72963135.SNV | A>G | Intron variant | rs1025927 | 0.23 | 0.28 | 0.07 |
| TRPA1 | X8.72965043.SNV | T>C | Intron variant | rs13271151 | 0.51 | 0.68 | 0.79 |
| TRPA1 | X8.72965973.SNV | G>A | Intron variant | rs3735942 | 0.21 | 0.2 | 0.49 |
| TRPA1 | X8.72966002.SNV | G>A | Synonymous variant | rs3735943 | 0.32 | 0.25 | 0.44 |
| TRPA1 | X8.72969263.SNV | A>C | Intron variant | rs3779752 | 0.49 | 0.31 | 0.31 |
| TRPA1 | X8.72971459.SNV | T>C | Intron variant | rs12541196 | 0.23 | 0.17 | 0.1 |
| TRPA1 | X8.72971473.SNV | C>T | Intron variant | rs71526150 | 0.15 | 0.19 | 0.11 |
| TRPA1 | X8.72971502.SNV | T>C | Intron variant | rs12541199 | 0.26 | 0.2 | 0.1 |
| TRPA1 | X8.72971506.SNV | T>G | Intron variant | rs12541200 | 0.14 | 0.17 | 0.15 |
| TRPA1 | X8.72971527.SNV | C>T | Intron variant | rs12548486 | — | — | — |
| TRPA1 | X8.72971570.SNV | C>T | Intron variant | rs9298198 | 0.12 | 0.12 | 0.18 |
| TRPA1 | X8.72971579.SNV | A>C | Intron variant | rs11423229 | 0.27 | 0.39 | 0.18 |
| TRPA1 | X8.72975801.SNV | T>A | Missense variant | rs7819749 | 0.27 | 0.37 | 0.49 |
| TRPA1 | X8.72981327.SNV | A>G | Synonymous variant | rs1811457 | 0.15 | 0.32 | 0.46 |
| TRPA1 | X8.72987699.SNV | C>T | Intron variant | rs2278654 | 0.16 | 0.46 | 0.87 |
| TRPA1 | X8.72987834.SNV | A>T | Intron variant | rs2278653 | 0.17 | 0.52 | 0.82 |

The putative functional consequences according to 65 are amino acid or protein changes for missense and deletion/insertion variants, and reduced transcriptional efficiency for UTR and synonymous exonic variants. At the right side of the tables, the values of 3 scores are provided by the genome-wide annotation of variants tool (GWAVA; at http://www.sanger.ac.uk/sanger/StatGen_Gwava) that generates 3 different so-called GWAVA scores, ie, the “region score,” the “TSS score,” and the “unmatched score,” all in the range [0, . . . , 1]. A high GWAVA score means more active functionality with respect to a low GWAVA score.

MIX, A mixture of variation types; SNV, single-nucleotide variation; TSS, transcription start site.
been observed in noncoding portions of the genome. This observation implies that variants that affect the risk of complex diseases often exert their effect by altering the regulation of genes rather than by directly affecting the gene and protein function. They act by affecting gene expression, eg, by disrupting a transcription factor–binding site or by affecting mRNA splicing, stability, and structure, which may result in a reduced transcriptional efficiency.

Along this line, to further assess the biological plausibility of a functional consequence of the present machine-learned derived selection of gene loci, the 31 selected variants were queried in the genome-wide annotation of variants tool (GWAVA, https://www.sanger.ac.uk/sanger/StatGen_Gwava). This web-based tool produces a prediction of the functional impact of noncoding genetic variants that are based on machine learning from a wide range of annotations of noncoding elements for which the functional consequences are known. For this task, it uses a tailored random-forest algorithm that builds 3 different classifiers, the so-called GWAVA scores named “region score,” “TSS score,” and “unmatched score” and all scaled in the range [0, 1], by using all available annotations to discriminate between disease variants and variants from 3 control data sets. Specifically, the “unmatched” classifier bases on a random selection of single-nucleotide variations (SNVs) from across the genome to get a reasonable sample of the background, the “TSS score” includes genome-wide variants matched for distance to a transcription start site and the “region score” is composed of all variants in the 1 kb surrounding each of the disease variants. The machine-learned algorithm is trained with a set of variants with known function and learns to predict the function of further variants from their location within the gene. A high GWAVA score means more active functionality with respect to a low GWAVA score. The quality of the prediction was addressed in the original publication where the authors showed that the classifier for each training set could usefully discriminate between disease and control variants. The area under the receiver operating characteristic curves were 0.97, 0.88, and 0.71, respectively, where a value of 0.5 denotes a bad classifier and 1 denotes an excellent classifier.

A GWAVA analysis for all the 134 gene loci in the TRPA1 yielded 58 hits; 76 of the present variants that had not been reported previously were not found. Interestingly, the GWAVA tool found all but 2 of the 25 TRPA1 variants included in the final analyses, which provides a first support for the potential importance, ie, for the successful of the applied machine-learned methods in selecting relevant gene loci for phenotype association (Table 2). Moreover, the 3 TRPA1 variants highlighted by the random-forest classifier as most important (Fig. 6), ie, x8.72934392.SNV (rs9298197), x8.72969263.SNV (rs3779752), and x8.72987369.SNV (rs2278654) figured at the first or second positions of at least 1 of the GWAVA prediction scores (Table 2). This supports (1) the present data analysis approach and (2) the functional role of variants, although located in noncoding regions of the genes. Further variants included in the selection of d = 31 gene loci in the TRPA1 and TRPV1 genes could be supported by previous evidence of a functional role. This includes the TRPA1 variants rs11988795, rs37735942, and rs37735943, which have been reported as associated with different sensitivity to pain, or the TRPA1 variant rs12548486, which has been associated with menthol preference among smokers. In addition, the ile685Val encoded by rs8065060 in the TRPV1 gene has been reported to be associated with genetic risk of painful knee osteoarthritis, and carriers of the TRPA1 variant rs8065060 had a 1.6 time longer cold withdrawal time than noncarriers. A further positive hit was the missense variant Lys186Asn (rs7819749) in TRPV1, which has been linked with glioblastoma multiforme.

The pattern of variant alleles differed between phenotype groups in the direction that carriers of fewer variant alleles were underrepresented in the phenotype group with less pronounced changes of HPTs after topical application of capsaicin. Both directions of changes would seem biological plausible, and in particular, gain-of-function mutations in ion channels may lead to increased agonist sensitivity or altered gating properties, and may render the channel constitutively active. For example, an
autosomal-dominant hereditary form of high-pain sensitivity, the so-called familial episodic pain syndrome, FEPS1 (accession number 615040 in the Online Mendelian Inheritance in Man (OMIM) database; http://www.ncbi.nlm.nih.gov/omim), which is characterized by episodes of debilitating upper-body pain, triggered by fasting and physical stress, is caused by a gain-of-function SNP (rs398123010) in the \( \text{TRPA1} \) gene.\(^3\) In carriers, QST showed normal baseline sensory thresholds but enhanced secondary hyperalgesia to punctate stimuli after treatment with mustard oil.\(^3\) Accordingly, this mutation increases the chemical sensitivity of TRPA1, but leaves the voltage sensitivity unchanged. Other gain-of-function mutations, rs753375978 and rs575489206, located in the analogous region of the \( \text{TRPV1} \) gene, severely affect all aspects of channel activation and lead to spontaneous activity.\(^5\)

The more important role of TRPA1 as compared to that of TRPV1 in the sensitivity to heat or the hypersensitization response to capsaicin bears implications for the development of novel analgesic treatments involving TRP channel inhibitors. Specifically, a query of the Thomson Reuters “Drugs and Biologics Search Tool” (http://integrity.thomsonpharma.com) in August 2017 indicated that by far, the most frequently regarded TRP channel family member in analgesic drug development is TRPV1, for which are 29 agonists or antagonists currently under active development. TRPA1 agonists or antagonists figured with only 7 entries. If, based on the present results, the functional impact of TRPA1 variants exceeds TRPV1 variants, TRPA1 may play a greater role in pain including neuropathic pain when considering that topical capsaicin can induce a neuropathy-like QST results pattern in a small subgroup of healthy subjects.\(^4\)

**Figure 4.** Data structure found in the TRPA1/TRPV1 NGS genotypes and its relation with the phenotypes. Top: Visualization of high-dimensional data consisting of \( d = 31 \) gene loci analyzed in \( n = 75 \) subjects. The data were projected onto a 2-dimensional grid using a parameter-free projection polar swarm, \( \text{Pswarm} \).\(^7\) During the learning phase, the DataBots were allowed for adaptively adjusting their location on the grid close to DataBots carrying data with similar features, with successively decreasing search radius. When the algorithm ends, the DataBots become projected points. To enhance the emergence of data structures on this projection, a generalized U matrix displaying the distance in the high-dimensional space was added as a third dimension to this visualization.\(^7\) The U matrix was colored in hypsometric colors\(^7\) making the visualization appear as a geographical map with brown (up to snow-covered) heights and green valleys with blue lakes. Watersheds indicate borderlines between different groups of subjects according to the pattern of repeated cold pain measurements. The data points are colored according to the emerging 2-cluster structure. Bottom left: Ward clustering of the projected data clearly indicated 2 clusters using the Manhattan distance. Bottom center: Heat plot of the pattern of genetic variants (columns) per subject (lines), grouped for the data structure of the genetic information. The \( 75 \times 31 \) matrix is a visualization of high-dimensional data consisting of \( d = 31 \) gene loci analyzed in \( n = 75 \) subjects. The allele occurrences are shown color coded as the number of nonreference alleles found at the respective locus in the respective sample as white, 0 nonreference alleles = wild type genotype; green, heterozygous; and blue, 2 non-reference alleles. Bottom right: Subjects belonging to the different genotype clusters were unevenly distributed across the phenotype clusters, ie, assignment to the 2 Gaussian modes in the distribution of capsaicin effects (Fig. 2), at a statistical significance level of \( P < 0.05 \) (the Fisher exact test). The mosaic plots represent the contingency table of the genotype vs phenotype class structure (membership sizes given as numbers in the fields of the mosaic). The figure has been created using the R software package (version 3.4.1 for Linux; http://CRAN.R-project.org),\(^6\) in particular the libraries “DatabionicSwarm” (M. Thrun, https://cran.r-project.org/package=DatabionicSwarm)\(^7\) and “gplots” (Warnes G et al., https://cran.r-project.org/package=gplots). NGS, next-generation sequencing.
Table 3 shows the performance of classifiers obtained using different machine-learned methods (random forests, adaptive boosting, k-nearest neighbors [kNNs], naive Bayes, and support vector machines [SVMs]) on (1) the original data, (2) a reduced data set from which the 2 variants that differently distributed between phenotype groups at a noncorrected significance level (Fig. 3) and which alone provided a separation between phenotypes better than guessing (X8.72934391.SNV and X8.72969263.SNV) were left out, (3) a data set constructed to provide a negative control by permuting the original genotypes, and (4) a data set constructed to provide a positive control by sorting the genotype information in descending order of alleles at each gene locus (Fig. 1).

| Parameter | Random forests | SVM | Naive Bayes |
|-----------|----------------|-----|-------------|
| [%] Original data | Permuted data | Sorted data (negative control) | Sorted data (positive control) |
| Sensitivity | 88.2 | 94.1 | 100 | 76.5 | 70.6 | 100 | 82.4 | 76.5 | 100 | 62.5 | 100 | 64.7 | 88.8 | 100 |
| Specificity | 37.5 | 12.5 | 100 | 37.5 | 25 | 100 | 37.5 | 11.8 | 29.4 | 100 | 70.6 | 76.5 | 100 | 50 | 37.5 | 87.5 |
| Positive predictive value | 75 | 68 | 100 | 73.7 | 68 | 100 | 72.7 | 68.2 | 100 | 34.8 | 33.3 | 100 | 33.3 | 100 | 37.5 | 94.4 |
| Negative predictive value | 55.6 | 25 | 100 | 44.4 | 31.3 | 100 | 50 | 33 | 100 | 100 | 68.4 | 100 | 72.2 | 66.7 | 100 | 40 | 31.3 |
| Balanced accuracy | 62.5 | 50 | 97.1 | 59.9 | 50.4 | 100 | 55.9 | 50.7 | 100 | 57 | 50.4 | 100 | 57 | 50.4 | 100 | 40 | 31.3 |
| Area under the ROC curve | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |

Results represent the medians of the test performance measured from 1000 runs using Monte-Carlo resampling with splits into 2/3 of the data for training and 1/3 for testing (Fig. 1).

ROC: receiver operating characteristic.
The improvement of phenotype prediction provided by the TRPA1 and TRPV1 genotypes over a nonsense genotype was consistent yet modest when comparing the almost perfect phenotype association with an idealized arbitrary genotype. This points at further factors modulating the individual sensitivity to heat pain or the response to capsaicin, which is highly plausible and an  

gene regulatory heat pain sensitivity or its enhancement by capsaicin was not expected considering the current knowledge about the complex genetic architecture of pain and the role of competitive genotype effects not controlled for. Indeed, although the present assessment had an explicit focus on TRPV1 and TRPA1, further genetic variants are known to play a role in thermal pain sensitivity. For example, the third heat transducer, TRPM3, was not addressed in this study but may also contribute to heat pain sensitivity as shown in mice. Furthermore, variants implicated in the present phenotype have been found in the genes coding for GTP cyclohydrolase 1 (GCH1), for the melanocortin 1 receptor (MC1R) or for the vaspressin receptor 1A (AVPR1A). Furthermore, nongenetic factors play a role up to the estimate that only 26% of the variance in heat pain responses can be explained by genetic factors. The present data-driven analyses were based on machine learning, which in its unsupervised form was applied to detect structures in the genetic data that hinted at a group separation, and in its supervised form was applied to assess the question whether the genotype provides information suitable for correct pain phenotype assignment. The methods were selected heuristically; possible alternatives such as self-organizing maps as used previously, or multidimensional scaling or principal component analysis did not offer obvious advantages.
chosen to cover a variety of machine-learned classifiers previously used in pain research\textsuperscript{70} such as prototype based (eg, kNN) or collective decision based (eg, boosting and random forests), with the addition of classical methods such as logistic regression included for its vicinity to classic statistical approaches, or naive Bayes. Indeed, the agreement among results obtained using different kinds of machine-learning methods and the biological plausibility of the results did not indicate an immediate need to include further methods.

The present analyses used machine learning for knowledge discovery, ie, an association of the TRPV1 and TRPA1 genetics with the heat-related pain phenotype was sought rather than a clinical tool for diagnostics. The moderate classification performance strongly suggests to base such a diagnostic tool on further factors including demographic, psychological and clinical parameters and factors derived from "omics," ie, proteomics, lipidomics, or genome-wide based features. Moreover, the present methods produced subsymbolic\textsuperscript{71} classifiers where a better performance of machine-learned algorithm is sought by waiving the possibility to understand the details, ie, it is impossible to obtain complete biomedical explanations for the functioning of the algorithm. For example, random forests use hundreds or thousands of simple decision trees that escape interpretation; the classification is obtained through the complete set of trees, ie, the "forest."\textsuperscript{6} The subsequently applied ranking of the importance of single variants only partly provided a biological explanation. With other classifiers, this was even less possible or completely impossible. However, the purpose of the present analysis was to study whether or not the genetic information contained in the sequences of TRPV1 and TRPA1 may contribute to the prediction of the phenotype, which establishes a genotype–phenotype association as the main purpose of this analysis.

5. Conclusions

In a cross-validated scenario, several analytical paths supported a role of TRPA1 and, to a lower degree, TRPV1 NGS-based genotypes for a potentially clinically relevant pain phenotype. The analysis shows that the complexity of the genotype is a relevant factor and machine-learned methods provide biologically plausible results, outperforming classic statistical genotype vs phenotype association analyses. The results were biologically plausible and fit with evidence of function TRPA1 or TRPV1 variants. Moreover, the relative importance of the variants observed with the machine learning agrees with an independent computer-based prediction of the biological roles of noncoding gene variants obtained in a GWAVA analysis. From this, a role of TRPA1 or TRPV1 NGS genotyping in personalized approaches at analgesic therapy with the respective novel analgesics may be expected. However, the improvement of phenotype prediction over chance was consistent but small when compared with expected. However, the improvement of phenotype prediction over chance was consistent but small when compared with expected. However, the improvement of phenotype prediction over chance was consistent but small when compared with expected.

Conflict of interest statement

The authors have no conflict of interest to declare.

This work has been funded by the European Union Seventh Framework Programme (FP7/2013) under grant agreement no. 602919 (J.L., GLORIA) and by the Landesoffensive zur Entwicklung wissenschaftlich-ökonomischer Exzellenz (LOEWE).

Table 4

| Drug               | Action            | Company                     |
|--------------------|-------------------|-----------------------------|
| Capsaicin          | TRPV1 agonist     | Winston Laboratories        |
| Resiniferatoxin    |                   | Icos                        |
| Capsaicin          |                   | Perrigo                     |
| Cannabidivarin     |                   | GW Pharmaceuticals           |
| Etodolac           |                   | MEDRx                       |
| CGS-125            |                   | Vizuri Health Sciences      |
| Hyaluronan         |                   | Vizuri Health Sciences      |
| Diclofenac sodium  |                   | Boehringer Ingelhein        |
| Propofol           |                   | Aspen Pharmacare            |
| Axomadol           |                   | Grunenthal                  |
| Tivanisiran        | TRPV1 expression inhibitor | Sylents                     |
| Davasacin          | TRPV1 ligand      | Dong-A                      |
| DWP-05195          | TRPV1 antagonist  | Daewoong                    |
| TR-1               |                   | Daewoong                    |
| V-116517           |                   | Purdue Pharma               |
| JYL-1421           |                   | AmorePacific                |
| NGD-8095           | Ligand            | Ligand                      |
| NGD-8243           | Ligand            | Ligand                      |
| NGD-9611           | Ligand            | Ligand                      |
| Mavatrep           |                   | Johnson & Johnson           |
| JTS-653            |                   | Japan Tobacco (JT)          |
| DD-04107           | BCN Peptides      |                             |
| AMG-51             |                   | Amgen                       |
| AMG-628            |                   | Amgen                       |
| ABT-102            |                   | Abbott                      |
| SAR-115740         | Sanofi            |                             |
| AZD-1386           | AstraZeneca       |                             |
| GRC-6211           | Lilly             |                             |
| Catharanthine      | TRPA1 agonist     | University of Toronto       |
| KDS-4103           |                   | Kadmus Pharmaceuticals       |
| Cannabidivarin     |                   | GW Pharmaceuticals           |
| ODM-108            | TRPA1 modulator   | Orion (F1)                  |
| HC-030031          | TRPA1 antagonist  | Hydra Biosciences           |
| CB-625             |                   | Merck & Co                  |
| GRC-17536          |                   | Glenmark Pharmaceuticals      |

The information was queried on August 23, 2017, from the Thomson Reuters Integrity database at https://integrity.thomson-pharma.com.
[82] Ultsch A, Kringel D, Kalso E, Mogil JS, Lötsch J. A data science approach
to candidate gene selection of pain regarded as a process of learning and
neural plasticity. PAIN 2016;157:2747–57.

[83] Ultsch A, Lötisch J. Computed ABC analysis for rational selection of most
informative variables in multivariate data. PLoS One 2015;10:e0129767.

[84] Ultsch A, Lötisch J. Machine-learned cluster identification in high-
dimensional data. J Biomed Inform 2017;66:93–104.

[85] Ultsch A, Sieman HP. Kohonen’s self organizing feature maps for
exploratory data analysis. Proceedings of the INNC’90, Int Neural
Network Conference, International Neural Network Conference, July
9–13, 1990, Palais des Congres, Paris, France. Kluwer, 1990. p. 305–8.

[86] Ultsch A, Thrun MC, Hansen-Goos O, Lötisch J. Identification of molecular
fingerprints in human heat pain thresholds by use of an interactive mixture
model R toolbox (AdaptGauss). Int J Mol Sci 2015;16:25897–911.

[87] Valdes AM, De Wilde G, Doherty SA, Lories RJ, Vaughn FL, Laslett LL,
Maciewicz RA, Soni A, Hart DJ, Zhang W, Muir KR, Dennison EM,
Wheeler M, Leaverton P, Cooper C, Spector TD, Cicutti FM,
Chapman V, Jones G, Arden NK, Doherty M. The Ile585Val TRPV1
variant is involved in risk of painful knee osteoarthritis. Ann Rheum Dis
2011;70:1556–61.

[88] van der Maaten LJP, Hinton GE. Visualizing high-dimensional data using
t-SNE. J Mach Learn Res 2008;9:2579–605.

[89] Venna J, Peltonen J, Nybo K, Aidos H, Kaski S. Information retrieval
perspective to Nonlinear dimensionality reduction for data visualization.
J Mach Learn Res 2010;11:451–90.

[90] Vriens J, Owsianny G, Hofmann T, Philipp Stephan E, Stab J, Chen X,
Benoit M, Yue F, Janssens A, Kerselaers S, Oberwinkler J, Vennekens R,
Gudermann T, Nilius B, Voets T. TRPM3 is a nociceptor channel involved
in the detection of noxious heat. Neuron 2011;70:482–94.

[91] Walker SH, Duncan DB. Estimation of the probability of an event as
a function of several independent variables. Biometrika 1967;54:
167–79.

[92] Ward JH Jr. Hierarchical grouping to optimize an objective function. J Am
Stat Assoc 1963;58:236–44.

[93] Wild A. Best practice in inventory management. New York: Wiley, 1997.

[94] You FM, Deal KR, Wang J, Britton MT, Fass JN, Lin D, Dandekar AM,
Leslie CA, Aradhya M, Luo MC, Dvorak J. Genome-wide SNP discovery
in walnut with an AGSNP pipeline updated for SNP discovery in
allogamous organisms. BMC Genomics 2012;13:354.