Journal of Biomedical Informatics 46 (2013) 921–928

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A machine-learned knowledge discovery method for associating complex phenotypes with complex genotypes. Application to pain

Jörn Lötsch [a,b,*], Alfred Ultsch [c]

[a] pharmazentrum frankfurt/ZAFES, Institute of Clinical Pharmacology, Johann Wolfgang Goethe University Hospital, Theodor-Stern-Kai 7, D-60590 Frankfurt am Main, Germany
[b] Fraunhofer Institute for Molecular Biology and Applied Ecology IME, Project Group Translational Medicine and Pharmacology TMP, Theodor-Stern-Kai 7, 60596 Frankfurt am Main, Germany
[c] DataBionics Research Group, University of Marburg, Hans-Meerwein-Straße, D-35032 Marburg, Germany

ARTICLE INFO

Article history:
Received 10 June 2013
Accepted 18 July 2013
Available online 27 July 2013

Keywords:
Machine-learning
Knowledge-generation
Genetics

ABSTRACT

Background: The association of genotyping information with common traits is not satisfactorily solved. One of the most complex traits is pain and association studies have failed so far to provide reproducible predictions of pain phenotypes from genotypes in the general population despite a well-established genetic basis of pain. We therefore aimed at developing a method able to prospectively and highly accurately predict pain phenotype from the underlying genotype.
Methods: Complex phenotypes and genotypes were obtained from experimental pain data including four different pain stimuli and genotypes with respect to 30 reportedly pain relevant variants in 10 genes. The training data set was obtained in 125 healthy volunteers and the independent prospective test data set was obtained in 89 subjects. The approach involved supervised machine learning.
Results: The phenotype–genotype association was reached in three major steps. First, the pain phenotype data was projected and clustered by means of emergent self-organizing map (ESOM) analysis and subsequent U-matrix visualization. Second, pain sub-phenotypes were identified by interpreting the cluster structure using classification and regression tree classifiers. Third, a supervised machine learning algorithm (Unweighted Label Rule generation) was applied to genetic markers reportedly modulating pain to obtain a complex genotype underlying the identified subgroups of subjects with homogenous pain response. This procedure correctly identified 80% of the subjects as belonging to an extreme pain phenotype in an independently and prospectively assessed cohort.
Conclusion: The developed methodology is a suitable basis for complex genotype–phenotype associations in pain. It may provide personalized treatments of complex traits. Due to its generality, this new method should also be applicable to other association tasks except pain.

1. Introduction

Human genotyping information elucidates pathogenetic mechanisms and provides clinical guidance for disease management. However, the association of genotyping information with common traits is not resolved satisfactorily [1]. Especially in complex traits emerging from multifactorial mechanisms, single genetic variants often produce only small effect sizes [2]. This weakens the utility of genotyping information [1,3].

One of the most challenging traits is pain. It involves a complex pathophysiology [4] underlying its sensory, affective, motor, vegetative and emotional components [5] reflected in the large network of underlying molecular nociceptive pathways [6]. The genetic basis of pain has been well established [7–9]. However, so far, association studies largely failed to provide reproducible predictions of phenotypes from genotypes in the average population [10]. Roughly this is caused by common genetic factors reciprocally canceling out their phenotypic consequences [11] and usually exerting only small effects [12]. To these poor results probably adds that current analytical methods for genotype phenotype association in pain are often insufficient. While the complexity of pain is increasingly accepted [13], its high-dimensional phenotypes [14] and underlying genotypes [11] are mainly subjected to low-dimensional analyses. Indeed, it becomes clear that it is advantageous to view pain as a complex phenotype when clustering individuals for their responses to different pain tests [15–17]. However, approaches applied so far have failed to provide a conclusive solution to pain genotype–phenotype association problems. This is probably due to a number of methodological shortcomings. **Firstly**, theoretical reasons suggest that the presently used clustering techniques should be revised in favor of those that make no prior

* Corresponding author at: pharmazentrum frankfurt/ZAFES, Institute of Clinical Pharmacology, Johann Wolfgang Goethe University Hospital, Theodor-Stern-Kai 7, D-60590 Frankfurt am Main, Germany. Fax: +49 69 6301 4354.
E-mail address: j.loetsch@em.uni-frankfurt.de (J. Lötsch).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.07.010

assumptions about the cluster structure, since the patterns of pain responses across different tests provide no indication of a particular cluster form. **Secondly**, clustering approaches used so far have been restricted to a mere description of the pattern of pain measures among individuals, without providing analyses of clinically relevant phenotypes that could be used for predictions by genotypes (for example, see [16], page 3, Table 1). **Thirdly**, genotype associations were mostly made in separate tests of single markers for phenotypic effects, without regard to the complexity of the genotypes [11].

Nowadays, more sophisticated bioinformatics tools are available to successfully approach this complex problem. Besides automated clustering of complex data, the bioinformatics toolbox also contains machine leaning methods for a subsequent knowledge-generation out of the clustering. In the present work, we aimed at developing a methodology that provides a basis for genotype–phenotype associations in complex traits. The methodology was developed to address several shortcomings of current approaches to genotype–phenotype associations and was presently applied to the complex trait of pain. It incorporates the complexity of both pain phenotypes and pain genotypes and is able to identify subgroups of individuals with similar pain phenotypes who share genotypic markers. We show that complex genotypes allow for correct prospective identification of up to 80% of the subjects who belong to a particular pain phenotype cluster. However, as a limited set of genotypes and phenotypes was used, the intention of this analysis rather was to pursue a clear methodological focus for the identification of complex genotypes and phenotypes and their associations than to identify new genotypes as a biological explanation of the observed pain phenotypes.

2. Methods

2.1. Data sources

2.1.1. Study cohorts

The investigations followed the Declaration of Helsinki on Biomedical Research Involving Human Subjects and were approved by the Ethics Committee of the Medical Faculty of the Goethe – University, Frankfurt am Main, Germany. All subjects gave written informed consent. Exclusion criteria employed were: drug intake less than seven days previously (except oral contraceptives), actual clinical pain, and concurrent diseases, based on questioning and a short medical examination.

Available data consisted of two independent data sets obtained in two independent study cohorts. The first data set, the **training data**, had been previously acquired from a random sample of 125 unrelated healthy young Caucasians (69 men, 56 women, mean age 25 ± 4.4 years) [12,14,18]. At this data set, the genotype–phenotype associations were established. To test their prediction, a new data set was acquired prospectively [19], the **test data** set, which was obtained in the same laboratory, from a random sample of 89 subjects of the same ethnicity and distribution (36 men, 53 women, mean age 25.6 ± 3.9 years).

Table 1
Decision rules (separated by lines) extracted from the CART classifier, providing a semantic description of the pain phenotypes found by the ESOM/U-matrix cluster analysis.

Case belongs to		IF (rule conditions)
Class 1	IF	Heat < 44.55 °C
	AND	Cold > 19.95 °C
	AND	Current < 2.65 mA
Class 2	IF	Heat < 44.55 °C
	AND	Pressure < 48.8 N/cm^2
Class 3	IF	44.5 ⩽ Heat < 45.5
	AND	Cold ⩽ 6.35 °C
	AND	Current < 2.25 mA
Class 4	IF	Heat ⩾ 44.45 °C
	AND	11.05 °C > Cold ⩽ 19.4 °C
Class 5	IF	Heat ⩾ 44.55 °C
	AND	Cold > 11.05 °C
	AND	Current ⩾ 2.65 mA
Class 6	IF	Heat < 44.5 °C
	AND	Cold ⩽ 7.95 °C
	AND	Pressure < 48.8 N/cm^2
Class 7	IF	Heat ⩾ 45.5 °C
	AND	2.25 mA ⩽ Current < 4.75 mA
Class 8	IF	Heat ⩾ 45.5 °C
	AND	Cold ⩽ 6.35 °C
	AND	Current ⩾ 4.75 mA

HPS: high-pain sensitivity phenotype, APS: average-pain sensitivity phenotype, LPS: low-pain sensitivity phenotype. *: Grouping according to the combined "Pain" variable calculated as the average of all z-transformed pain measures to model the overall sensitivity to any type of pain stimulus [32].

2.1.2. Phenotyping information

Pain thresholds to four stimuli, including heat, cold, mechanical and electrical pain, were measured as described previously [14,18]. In brief, **heat** stimuli were applied using a 3 × 3 cm thermode (Thermal Sensory Analyzer, Medoc, Ramat Yishai, Israel) placed onto the skin of the left volar forearm. While increasing temperature from 32 °C by 0.3 °C/s, the subject was requested to press a button when heat became painful, which was recorded as pain threshold and subsequently, the thermode was cooled down. **Cold** stimuli were applied similarly, however, with temperatures decreasing by 1 °C/s from 32 °C to 0 °C. **Blunt pressure** was exerted perpendicularly onto the dorsal side of mid-phalanx of the right middle finger using a pressure algometer (JTECH Medical, Midvale, USA) with a circular flat probe of 1 cm diameter. While increasing the pressure by 9 N/cm^2 per second, the threshold was reached when the subject indicated pain. **Electrical** stimuli employed were sine-wave stimuli at 5 Hz, applied via two gold electrodes to the medial and lateral side of the mid-phalanx of the right middle finger (Neurometer® CPT, Neurotron Inc., Baltimore, MD). As the intensity of the electrical stimulus was increased from 0 to 20 mA in 0.2 mA/s steps, the subjects were requested to interrupt the current by releasing a button when perceiving pain. The current at which this interruption occurred was the pain threshold.

2.1.3. Genotyping information

Genotyping was done for 20 single nucleotide polymorphisms (SNPs) [12]. These SNPs and resulting haplotypes, obtained *in silico* using PHASE software [20], have been reported previously to modulate pain [21]. The Hardy–Weinberg equilibrium was preserved in both cohorts (χ^2 goodness of fit tests); other details on SNPs and haplotypes have already been reported elsewhere [12] and are given in the Supplemental table to the present publication. Although restricted, in the light of the currently known >410 "pain genes" [22], the set nevertheless included some major players in nociception such as μ- and δ-opioid receptor genes (*OPRM1* [23] and *OPRD1* [24], respectively), transient receptor potential cation channel genes (*TRPV1* [25] and *TRPA1* [26]), catechol-O-methyl transferase (*COMT* [27,28]), fatty acid amide hydrolase (*FAAH* [27]), guanosine 5′-triphosphate cyclohydrolase 1 (*GCH1* [29]) and variants of the melanocortin-1 receptor gene (*MC1R*) associated with a red-head, -fair-skin phenotype [30,31]. Functional variants were diagnosed from genomic DNA by means of validated Pyrosequencing™ assays [12] on a PSQ 96 MA System (Qiagen, Hilden, Germany), with conventionally sequenced samples as controls.

2.2. Data analysis

Analyses were done using Matlab software (MathWorks, Natick, MS, USA) with functionality expanded by self-developed toolboxes

(publicly available at http://www.uni-marburg.de/fb12/datenbionik/software). Besides automated clustering of complex data, the bioinformatics toolbox also contains machine learning methods for a subsequent knowledge-generation from the clustering. Data analysis started with the identification of subjects who shared similar pain sensitivities to different stimuli. Subsequently, extreme pain phenotype subgroups (clusters) were analyzed for the underlying complex genetic architecture. Finally, the complex genotype was used to identify those subjects from the test data set who had belonged to a similar pain subgroup.

2.2.1. Analysis of the pain data cluster structure

2.2.1.1. Data exploration and preprocessing. Data was z-transformed and a combined "pain" variable, *zTotalPain*, was calculated as the sum of all rescaled pain measures to model the overall sensitivity to any type of pain stimulus [32]. This combined data was split into three classes of pain sensitivity (Fig. 1), i.e., "low pain" sensitivity (LPS; $zTotalPain \leqslant -s$), "average pain" (APS, *zTotalPain* in the interval $[-s, s]$) and "high pain" (HPS, $zTotalPain \geqslant s$). This corresponds to a classification with thresholds $\pm s$ at 20% and 80% of the distribution and reflects the previous classifications of LPS, APS and HPS phenotypes by Diatchenko et al. [32].

2.2.1.2. Projection and clustering pain data. This analysis focused on identifying subjects with similar pain phenotypes. The distributions of the phenotypical pain data turned out to be rather complex (Fig. 1). These distributions could be modeled with a mixture of three Gaussians $N(m_i, s_i)$, $i = 1, 2, 3$. The theorem of Bayes allows to calculate posterior probabilities $p(i|x)$, i.e., the probability that for a given pain value x the data belongs to Gaussian i. The value $B = p(3|x) - p(1|x)$ takes a value of 1 if x belongs to Gaussian 3, a value of 0, if x belongs to Gaussian 2 and -1 if x belongs to Gaussian 1. For the subsequent projection and clustering as measure of "similarity", the Euclidian distance of the B values was used. Each person's response to the pain stimuli ($n = 4$ dimensions), plus the overall sensitivity score ($n = 1$ dimension), was treated as a point in a five-dimensional Euclidean vector space (data space).

To obtain clusters in this vector space, the data was projected onto a two-dimensional plane. This space is called a map with a geographical interpretation in mind. As classical projection algorithms, such as principal component analysis or multidimensional scaling, cannot preserve complex cluster structures, the ESOM/U-matrix method was used [43]. ESOM clustering provides a number of advantages over alternative methods, such as K-means or Ward, the most relevant one being the lack of prior assumption about the cluster structure.

Using ESOM, data was projected onto a two-dimensional borderless grid (map space) of $50 \times 82 = 4200$ units ("neurons"; Fig. 2). The map space is toroid [33] and the projection is neighborhood preserving [34]. I.e., points that are neighbors in the high dimensional data space are also neighbors on the map space. Each neuron holds a vector of "weights" of the same dimension as the five input dimensions (four z-transformed pain thresholds plus the combined pain variable) in the input space. The weights initially were randomly drawn from the range of the data variables. In an adaptation process, (learning phase, 100 epochs) they were adapted to the data (training cohort, $n = 125$). As learning and test data had been obtained in the same laboratory with the same equipment, it was expected that test data could be included in the map, which was verified by constructing the ESOM with all the data. The Adjusted Rand Index (ARI) [35] of a clustering using only the training data set compared to a clustering of all data is

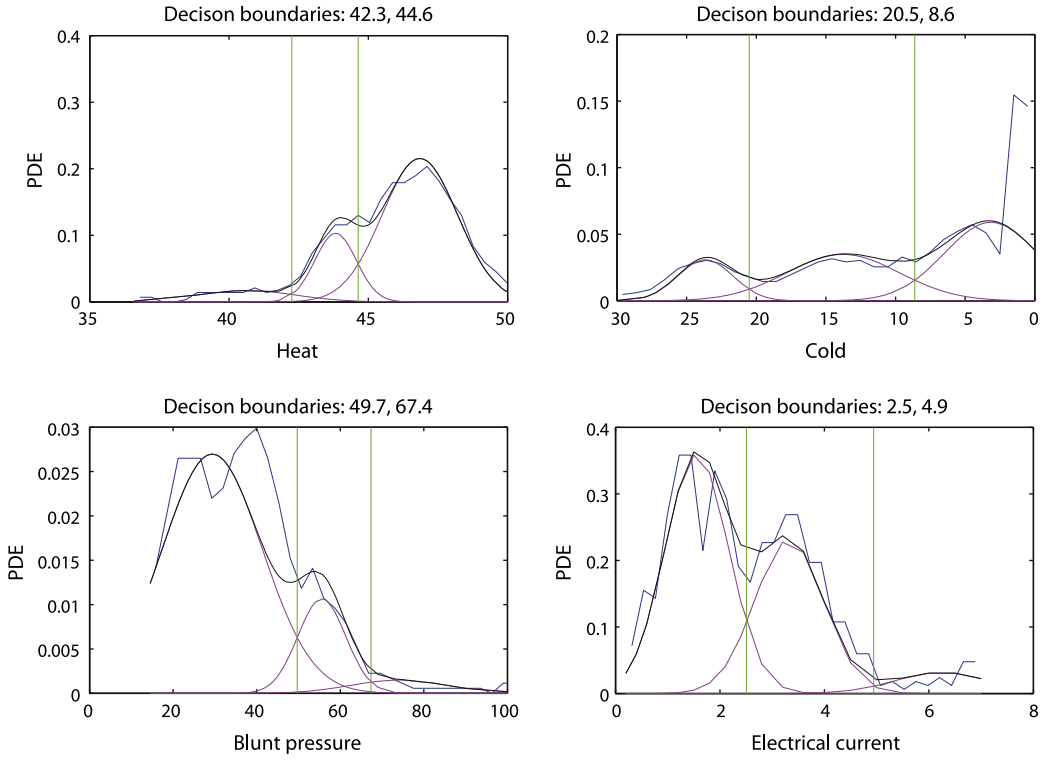

Fig. 1. Gaussian Mixture Models (GMM) of the pain variables. Blue is the measured empirical probability density using Pareto Density Estimation (PDE). Black indicates the GMM as sum of the three Gaussians shown in magenta. The vertical green lines indicate the decision boundaries according to the theorem of Bayes. An additional observation is that these boundaries are similar to the CART decision rule boundaries of Table 1. The measured probability densities (blue lines) could be modeled with a mixture of three Gaussians $N(m_i, s_i)$, $i = 1, 2, 3$ (black line including the three individual Gaussians in magenta). The theorem of Bayes allows to calculate posterior probabilities $p(i|x)$ i.e. the probability that for a given pain value x the data belongs to Gaussian i (green lines). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

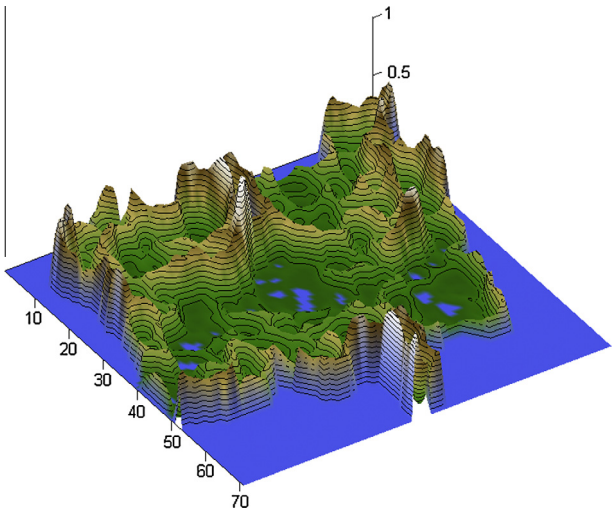

Fig. 2. U* matrix three-dimensional view, with a geographical interpretation in mind (for technical details of the presentation, see http://www.uni-marburg.de/fb12/datenbionik/forschung/esom), showing the clustering of subjects with comparable pain phenotypes as obtained using emergent self-organizing map (ESOM). The clusters were visualized using a U-matrix [36], which is a representation of the distances in data space on top of a map space. The U-matrix was cut from a tiled display of the ESOM to remove duplicate representation of the data. It was colored as a geographical map with brown or snow-covered heights and green valleys. High "walls" in a U-matrix indicate large distances between the pain responses of the 125 persons. Points represent persons and their coordinates in the toroid map space are used to address them when querying information. Points, or "persons", that lie together in a valley of the U-matrix indicate that these persons have a common response type pattern in all five stimuli. Thus, these valleys indicate clusters of pain types. The watersheds of the U-matrix indicate borderlines of clusters. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

96%. For the test data set the ARI is 89%. This shows the similarity of the cluster structure in the training and the test data.

2.2.1.3. Visualizing pain phenotype clusters. Data on the trained ESOM was presented on the two-dimensional toroid map where a cluster structure was to be visualized. This was obtained by adding a third dimension consisting of the average distance of the weight vector of a neuron to the weight vectors of its direct neighbors, which is known as the U-matrix [36]. Specifically, the U-matrix is a representation of the distances in data space on top of the map space (Fig. 2), with a geographical map analogy in mind. The watersheds of the U-matrix show borderlines of these pain clusters, i.e., high "walls" between data points indicate large distances between individual pain responses. By contrast, points lying together in a valley of the U-matrix represent persons who have a common pain response pattern with respect to the four stimuli and the overall sensitivity score. Cluster visualization was further enhanced by calculating the U* matrix, which results from the combination of the U-matrix distances with the P-matrix. The latter also uses the ESOM map as a floor space layout, however, instead of the local distances, density values in data space measured at the neuron's weights are used as height values [33]. The process was performed using the ESOM toolbox [37], publicly available at http://www.uni-marburg.de/fb12/datenbionik/software (accessed on March 17, 2013).

2.2.2. Identification of classifiers for pain phenotypes

The ESOM/U-matrix clustering identified the subjects who had similar pain sensitivities. However, the result was still a "black box" with respect to the pattern of particular pain thresholds shared by the cluster members. Therefore, the subsequent analysis aimed at obtaining clear descriptions, in terms of pain threshold markers, of the sensitivity patterns, thus obtaining pain phenotype groups (clusters) in the training data cohort. The cluster structure obtained with the U-matrix was interpreted by extracting the decision rules, in terms of measured stimulus intensities at the pain thresholds, from a classification and regression tree (CART) classifier that assigned each subject of the training set to a particular pain cluster (Table 1). A random sub-sampling validation with 125 data sets for the construction of a CART decision tree and 81 data sets for testing repeated 50 times resulted in a mean accuracy of 95.5% with a standard deviation of 1.4% of the classifier. This is consistent with the CART classification accuracy of 95% of the split into training and test data set as given above. CART provides a simple and easy understandable form of the classification rules and effectively uses the conditional information of the GINI index [38] to find optimal (local) dichotomic decisions. Furthermore, CART is invariant under transformations of the variables, robust with respect to outliers and allows estimation of the misclassification rate [38]. The requirements for this classifier are that it is sufficiently able to perform the classification task and that the classification is based on rules that a human reader can understand (knowledge). As shown below the decision rules can be used for a very precise definition of the cluster's "meaning".

2.2.3. Identification of genotypic associations

2.2.3.1. Marker pre-selection. Associations were sought that would identify a positive or negative combination (i.e., a rule) for the 29 genetic variables (labels) that described extreme pain phenotype clusters as well as possible. Specifically, the 29 genetic markers for each subject were labeled with a "Yes" (=1), "No" (=−1) or "Don't care" (=0), modeling in the presence, absence or insignificance of the reportedly functional allele. In addition, sex was coded as +1, if female and −1 if male. To search for such rules in all variables means regarding a space of size $R = 3^{30} > 2.0 \times 10^{14}$. A complete search in this space is out of the processing time range of current personal computers. Therefore, the genotypes had to be preselected. This was achieved in a 100-fold repeated cross-validation experiment (80/20 split), on three blocks consisting of 10 genetic markers each. In these blocks those combination of markers were selected that were useful to predict the HPS/LPS dichotomy with an accuracy of at least 71%. This greedy search reduced the search space to $3^{14} > 4.7 \times 10^{6}$ potential candidates for rules, which is possible to explore in feasible computing time (ca. 2 h on a typical PC). The reduced set of 14 genotype markers, thus identified, was subsequently used to construct predictive complex genotypes.

2.2.3.2. Associating complex genotypes with pain phenotypes of interest. The preselected genetic markers and the subject's gender were submitted to a machine learning procedure called "Unweighted Label Rule generation" (ULR) [39]. This tested all possible additive combinations of the 29 genetic markers and of gender for the best prediction of the membership of the extreme pain clusters in the training cohort. The possible sum of the markers consisted of the genetic markers (present, 1, absent, 0) multiplied by 1, −1 or 0. ULR first tested all single-label rules ($R = 60$) and measured the performance of the rule to predict the membership of the subject of the selected pain cluster as the area (AUC) under the receiver operating characteristic (ROC) curve [40]. It then combined the best performing one-variable rules. All possible sums of variables were tested.

For all resulting ULR rules, AUC, sensitivity and specificity were calculated. The performance of each rule was compared with that of the best guessing rule, which means just assigning all cases to the larger class of the dichotomy. If a ULR provided at least 5% better prediction performance than the best guess, it was included as a

candidate for a classification rule. Among candidates, the rule with the best AUC using the least number of labels was finally selected.

2.2.4. Assessment of the predictive performance of phenotype and genotype markers

After having identified both pain phenotype clusters and the underlying genetics in the training data set, the obtained knowledge was applied to the test data set.

2.2.4.1. Predicting pain cluster membership from phenotypic markers. The decision rules obtained with the CART classifier (Table 1) provided characteristics of pain cluster membership of the training data. The rules were subsequently applied to the test data to assess their performance to assign an individual subject to a particular pain phenotype, based on the information of four individual pain thresholds and the overall pain sensitive score.

2.2.4.2. Predicting pain cluster membership from genetic markers. Pain phenotype groups of major interest for genetic predictions were considered to consist of those containing subjects at the extreme of the distribution, with either very high or very low pain sensitivity. The predicted genetic causes of high or low pain sensitivity can then be considered as reasonable drug targets for pain therapy. To emphasize extreme pain phenotypes, ESOM clusters were intersected with the HPS/APS/LPS providing groups which were homogeneous with respect to their sensitivity to single stimuli grouping, while at the same time sharing the same extreme overall pain sensitivity (Table 2).

Subsequently, the composed genotype obtained by means of ULR of the training data set was searched in the test data set. The performance assessed whether this correctly identified the subjects belonging to either the low or high pain threshold groups in the test data set.

3. Results

3.1. Pain phenotypes

The ESOM clustering of the pain data and subsequent visualizing provided a U-matrix (Fig. 2) in which eight pain phenotypes could be observed (Fig. 3B), comprising individuals who shared complex pain threshold patterns across the four different noxious stimuli (i.e., heat, cold, electrical current and blunt pressure) and the combined pain score (average of the *z*-values). The specific properties characterizing each phenotype, as derived from the CART classifier (Table 1), could be interpreted clinically. For example, those belonging to cluster 7 and 8 were both stoical towards thermal pain. Individuals in cluster 7 were average sensitive to current pain stimuli while those belonging to cluster 8 were neither cold nor current sensitive. On the other side, subjects belonging to cluster 1 were generally highly temperature sensitive. Those in cluster 2 were also very pressure sensitive. The CART decision rules allowed for the prediction of the cluster membership in the test data set with an accuracy of 95%. The cluster structures were similar in the training and test data sets. Specifically, the CART classifier created from the training data predicted the cluster membership of the test data set with an accuracy of 95%.

Table 2
Grouping of all subjects according to similarities in pain perception. The columns show the grouping according to the combined "Pain" variable, calculated as the average of all z-transformed pain measures to model the overall sensitivity to any type of pain stimulus [32]. This defined three subgroups of pain sensitivity: high-pain sensitivity (HPS) phenotype, average-pain sensitivity (APS) phenotype and low-pain sensitivity (LPS) phenotype [32]. The lines show pain groups (clusters) from the ESOM/U-matrix clustering (Fig. 3). Subjects with extreme phenotypes (high pain sensitivity, i.e., the intersection of the HPS group with ESOM clusters 1 and 2; low pain sensitivity, i.e., the intersection of the LPS group with ESOM clusters 7 and 8) were chosen to demonstrate the prediction of extreme phenotypes by complex genotypes (italicized table cells).

ESOM cluster (class) #	HPS	APS	LPS	*n* (%)
Pain subgroup				
1	*20*	2	0	22 (10.7)
2	*12*	1	0	13 (63.1)
3	2	47	2	51 (24.8)
4	2	28	0	30 (14.6)
5	3	11	0	14 (68)
6	2	9	0	11 (53.4)
7	0	18	*15*	33 (16)
8	0	7	*25*	32 (15.3)
n (%)	41 (19.9)	123 (59.7)	42 (20.4)	206 (100)

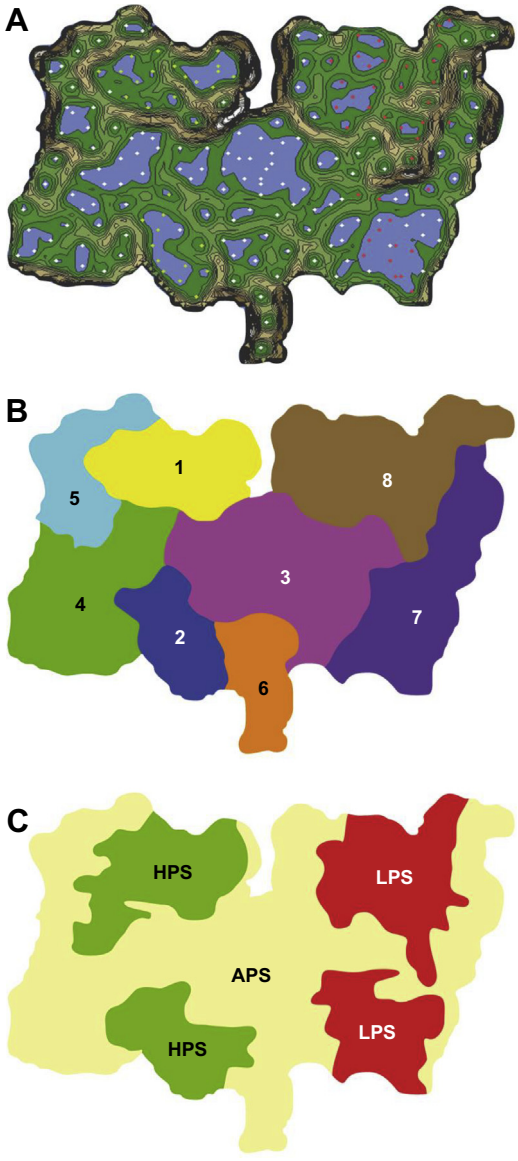

Fig. 3. Bird's view of the U matrix shown in Fig. 2. In the top panel (A), the map space is presented analogously to a physical map with a geographical map analogy in mind. Dots indicate the single individuals, with green and red dots emphasizing the HPS and LPS groups, respectively, as emphasized in panel C. Below this map (B), in the middle panel, the eight ESOM/U-matrix clusters 1–8 (Table 1) are indicated analogously to a political map. In the bottom panel (C), another "political map" emphasizes the HPS/APS/LPS pain groups [32]. A comparison with map B shows that the HPS and LPS were mainly composed of subjects belonging to ESOM clusters 1/2 or 7/8, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

ESOM clusters overlapped, although only incompletely (Table 2), with the classification into three major phenotypes [5], namely, "low pain sensitivity" LPS ($n = 41$), "average pain" (APS, $n = 123$) and "high pain" (HPS, $n = 42$). For example, the LPS group was formed mainly of subjects belonging to ESOM classes 7 and 8. However, while all subjects of classes 7 and 8 were stoical towards heat, only subjects of class 8 were also stoical towards cold (Table 1), emphasizing the greater complexity of ESOM clustering. Similarly, the HPS group consisted mainly of subjects belonging to ESOM classes 1 and 2. However, while all HPS subjects were sensitive to heat, those belonging to cluster 1 were also sensitive to cold, i.e., completely temperature sensitive, whereas those belonging to cluster 2 were pressure but not cold sensitive.

3.2. Genotype–phenotype associations

In a cross-validation experiment, repeated 100 times, 14 genetic variables including "gender" were identified that had sufficiently high accuracy (> 71%) for the training data set. These preselected genetic markers and the subject's gender were submitted to a ULR machine learning procedure aimed at identifying the marker combinations that best predicted the subjects' pain phenotype. Thus, a rule space of size $R = 3^{14} > 4.7 \times 10^6$ was searched in order to predict membership of the low or high pain thresholds clusters, which were defined, respectively, as the intersections between (i) ESOM clusters 1/2 and the HPS phenotype, or (ii) ESOM clusters 7/8 and the LPS phenotype (Table 2).

ULR generated the following rule for membership of the low pain threshold clusters: If (+*COMT* rs6269G + *TRPA1*rs 13255063A/rs11988795G + *TRPA1* rs13255063T/rs 11988795G + *FAAH*rs324419C/rs2295633A + $\text{Sex}_{(+1)\text{ if female, }(-1)\text{ if male}}$ − *COMT* rs6269A/rs4633T/4818C/rs4680A − *COMT* rs4646312T/ rs 165722 C/rs6269A/rs4633C/rs4818C/rs4680G − *OPRM 1* rs 1799971G − *FAAH* rs324419T/rs2295633A − $GCH1_{\text{"pain protecting haplotype"}}$ rs800 7267A/rs3783641T/rs10483639G[1] − $MC1R_{\text{redhead genotype}}$) > 0.5 then the case belongs to the intersection of ESOM/U-matrix cluster 1 and 2 with the HPS phenotype, i.e. to the pain phenotype subgroup with a very high pain sensitivity and not to the opposite phenotype clusters. Each component of this rule denotes presence or absence of the respective functional genotypic markers (0, 1) carried by the subject. In the training data set, this rule predicted membership for an individual of either low or high pain threshold clusters with an accuracy of 78%. The rule uses – apart from gender-only eight genetic markers.

Subsequently, the combined genotype was identified in subjects belonging to the test cohort. It predicted membership for an individual of either low or high pain threshold clusters with an accuracy of 78%. For comparison, among single genetic markers and gender, only the latter provided a prediction better than guessing, i.e., assigning all subjects just to the larger group of two alternatives. However, gender provided only a predictive accuracy of 67%.

4. Discussion

In this study, we employed the whole cycle of machine-supported generation of presumably new knowledge (hypothesis generation) from complex high-dimensional data. Firstly, the inspection involved clustering of high dimensional phenotype data (ESOM/U-matrix). In a second step, this clustering of data was used to generate a (cross-validated) classifier (CART) which uses explicit rules for the classification. These rules are understandable and interpretable by humans. The rules were found to be consistent with previous work (LPS/APS/HPS [32]) and gave hints to a more complex structure of human pain sensation types (8 classes). This classification was used for the generation of a classifier of the genotype data. The methodology mapped the genetic architecture of subjects sharing a particular pain phenotype and predicted, on the basis of a complex genetic marker, their membership of this phenotype in an independent cohort with an accuracy of almost 80%. This indicates that ESOM/U-matrix-based clustering, with subsequent rule generation, is suitable to provide pain phenotypes that are carried by subjects sharing a common pain-relevant genetic background [11].

ESOM clustering makes no prior assumption about the cluster structure. This is a major advantage of the present method over previous attempts to cluster pain phenotypes [15–17] that may have missed relevant pain clusters or provided wrong individual cluster associations by superimposing a possibly inadequate cluster structure on the data. In addition, the proposed method exceeds previous approaches towards genotype–phenotype association in several further ways. Whereas other approaches stopped after the phenotype classification had been obtained, the present method continued at this stage with a machine learning algorithm that generated decision rules for each cluster, presently implemented using CART, C4.5 [42]. The aim of this approach was to gain a suiting ROC AUC (>0.8) and an understandable semantic description of a cluster. This provides a suitable basis for clinically relevant phenotype clustering and subsequent genotype associations. This was not included in previous approaches, even though, from a bird's eye view of the comprehensive display of the cluster averages (Fig. 2 of [17]), it can be seen that cluster number 2 could possibly be determined by the single variable "high sensitivity to spontaneous pain attacks". Similarly, five distinct clinical phenotypes of neuropathic pain [17] were obtained by means of hierarchical cluster analysis, in a classical approach from the patients' self-estimations of spontaneous and evoked pain. None of the clusters was described in any way. The same applies to the findings of Hastie et al. [15] who found that pain has distinct dimensions, leading to complex phenotype clusters. Four principal factors were derived from heat, pressure, ischemic pain and temporal summations of pain stimuli. Using hierarchical cluster analysis of 188 cases, four distinct groups were found, based on patterns of responses across multiple pain stimuli [15]. Cluster (i), "high pain sensitivity", partly coincides with the present ESOM clusters 1 and 2 [28].

Previous attempts to describe pain phenotypes that acknowledged the complexity of pain only incompletely extended this to genotyping information [43]. However, introducing complex genotypes is required [11] to obtain a successful genotype–phenotype association, which single genetic markers cannot provide [10,44]. This is because of the small effect size that was exerted through the genetic markers on the phenotype measures, which had already been shown in the present training data set [12]. Specifically, the values of Cohen's d for the effects of the 30 genotype markers on single pain threshold measures were lower than $d = 0.2$ (small) in 80% of the marker-threshold associations and only in 2% of the associations greater than $d = 0.5$ (medium). Using single markers on low versus high pain threshold clusters already provided increased effect sizes with 50% of the Cohen's d values >0.2 (small) and 12% > 0.5 (medium). However, a dramatic increase in effect sizes was obtained when using combined rather than single genetic markers. Thus, the composed ULR genotype provided values of Cohen's $d = 1.25$, 1.35, 0.63 and 1.33 for heat, cold, pressure and electrical stimuli, respectively, for subjects belonging to either low or high pain threshold clusters (Table 2). An effect size is usually considered to be large, if Cohen's $d > 0.8$.

[1] As discussed previously [29], although the pain protective haplotype originally comprised 15 *GCH1* SNPS, it can be identified with 100% accuracy by genotyping only the three variant alleles rs8007267A, rs3783641T and rs104836c39G [41], at least in caucasian subjects with the present origin [29].

The present selection of genetic markers originates from previously reported positive findings which, however, may implicitly be influenced by other non-accounted for variants [11], contain never-reproduced associations and often be based on weak biological bases. Therefore, the interpretation of variants expected to be present or absent in highly pain-sensitive subjects should be done with caution. Consistent with these expectations, the negatively influential *GCH1* haplotype was originally reported to be pain protective [45]. This is biologically plausible as this haplotype impedes GCH1 up-regulation, followed by reduced availability of the pronociceptive tetrahydrobiopterin [29]. The fact that the *COMT* rs6269A/rs4633T/rs4818C/rs4680A haplotype, associated previously with average, but not with high pain sensitivity (APS) [28], received a negative connotation in our study is also consistent with expectations. A further agreement with expectations is the positive influence of female gender, as in women, higher pain sensitivity than in men has been identified in most gender studies on pain (for reviews, see [46–48]). However, the role *MC1R* genotype is less clear, as subjects carrying non-functional *MC1R* were reported to exhibit an increased tolerance to electrical pain stimuli [30], but this was not reproduced [31] and occasionally contradicted [31]; the latter outcome fitted best to the high-pain rule found in our study. However, the biological, mechanistic basis of *MC1R* pain regulation is not yet completely clear. In contrast to expectations, *OPRM1* rs1799971 A > G, which has been associated with decreased nociception [49,50], was allocated a positive influence. However, the molecular basis of this association is questionable since the increased affinity of endorphin at N40D μ-opioid receptors [51], related to this genotype, with subsequent reduced activity of the endogenous nocifensive opioid system, could not be reproduced [52–54]. Moreover, the variant is well known to reduce opioid receptor signaling efficiency [52] and expression [53–55], including a genetic–epigenetic interaction impeding receptor up-regulation [56]. Therefore, its inclusion in lower pain threshold cluster predictions is biologically plausible. Further components of the genotypes are difficult to interpret on basis of previous findings, as they have been reported to be associated with changed pain sensitivity [27], without specifying the direction of that change.

After the successful development and verification of the present methodology for pain phenotype–genotype association, the analysis cannot be extended beyond the prediction of extreme phenotypes from combined genetic markers. A final characterization of human pain could not be expected as the data set was limited by the numbers of cases, pain markers and genetic markers. Currently, at least 410 pain genes have been established [9], for example, 390 "pain genes" are found in the PainGenes database [57] at http://www.jbldesign.com/jmogil/enter.html (accessed on April 4, 2013). The inclusion of a more comprehensive set of genetic markers is very likely to change the set. With genome-wide data, there is no impediment to the replacement of the ULR based generation of complex genotypes by ESOM/U-matrix based genotypes and this option is opened by the present methodology. Psychological factors also may be included, but again, the methodology is now available.

The present results show that machine-learned knowledge-generation from identified phenotype structures is suited for associating underlying biomarkers. The method exceeds the currently available pain genotype–phenotype association methods in several ways. Most importantly, the high-dimensionality of both pain phenotype and pain genotype is taken into account by the present ESOM/U-matrix-based cluster identification, CART rule-based phenotype extraction and ULR-based genotype association. Thus, the present methodology appears able to resolve the poor clinical utility of current genotyping information for pain management [10], by providing larger effect sizes of combined genetic as compared to single genetic markers and by preselecting subjects with similar pain phenotypes who are more likely to share pain-relevant genotypes. The key to successful approaches to personalized pain therapy seems to lie less in the identification of more and more factors but in the combination of the high-dimensional information by informatics methods. Provided accessibility to large pain phenotype and genotype data sets is possible, the present methodology may provide the basis for genetics-based personalized pain treatment. The approach thus satisfies the complexity of pain and exceeds by far the so far available pain genotype–phenotype association methods. This new method should be applied to larger data sets to ease the clinical utility of genotyping information for pain research and therapy. Due to its generality, this new method should also be applicable to other association tasks apart from pain.

Acknowledgments

This work has been supported by the "Landesoffensive zur Entwicklung wissenschaftlich-ökonomischer Exzellenz": "LOEWE-Schwerpunkt: Anwendungsorientierte Arzneimittelforschung" and TRIP (JL). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors have declared that no competing interests exist. We thank Dr. A. Doehring for her contribution to the genotyping, Dr. K. Flühr and Dr. T.J. Neddermeyer for their contribution to the acquisition of the training data set (which had been used in previous publications with non-redundant data analyses: Neddermeyer TJ et al., Pain 2008; 138(2): 286–291, Flühr K et al., Clin J Pain 2009; 25(2): 128–131, Doehring A et al., PloS One 2011; 6(3): e17724 and Heimann et al., PLoS One 2013), and M. Schütz for his contribution to the acquisition of the test data set. We thank Prof. Michael Parnham for manuscript editing and Wolf von Waldow for help with Fig. 3. A preliminary version of this work has been presented at the Joint Annual Conference of the German Association for Pattern Recognition (DAGM) and the German Classification Society (GfKl).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2013.07.010.

References

[1] Goldstein DB. Common genetic variation and human traits. N Engl J Med 2009;360:1696–8.
[2] Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. J Clin Invest 2008;118:1590–605.
[3] Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, et al. Genome-wide association analysis identifies 20 loci that influence adult height. Nat Genet 2008;40:575–83.
[4] Cross SA. Pathophysiology of pain. Mayo Clin Proc 1994;69:375–83.
[5] Tracey I, Mantyh PW. The cerebral signature for pain perception and its modulation. Neuron 2007;55:377–91.
[6] Julius D, Basbaum AI. Molecular mechanisms of nociception. Nature 2001;413:203–10.
[7] Godinova AM. Genetic analysis of migraine. Zh Nevrol Psikhiatr Im S S Korsakova 1965;65:1132–8.
[8] Diatchenko L, Nackley AG, Tchivileva IE, Shabalina SA, Maixner W. Genetic architecture of human pain perception. Trends Genet 2007;23:605–13.
[9] Lötsch J, Doehring A, Mogil JS, Arndt T, Geisslinger G, Ultsch A. Functional genomics of pain in analgesic drug development and therapy. Pharmacol Ther 2013;139(1):60–70.
[10] Mogil JS. Are we getting anywhere in human pain genetics? Pain 2009;146:231–2.
[11] Lötsch J, Flühr K, Neddermayer T, Doehring A, Geisslinger G. The consequence of concomitantly present functional genetic variants for the identification of functional genotype–phenotype associations in pain. Clin Pharmacol Ther 2009;85:25–30.
[12] Doehring A, Küsener N, Fluhr K, Neddermeyer TJ, Schneider G, Lötsch J. Effect sizes in experimental pain produced by gender, genetic variants and sensitization procedures. PloS One 2011;6:e17724.

[13] Janal MN, Glusman M, Kuhl JP, Clark WC. On the absence of correlation between responses to noxious heat, cold, electrical and ischemic stimulation. Pain 1994;58:403–11.
[14] Neddermeyer TJ, Flühr K, Lötsch J. Principal components analysis of pain thresholds to thermal, electrical, and mechanical stimuli suggests a predominant common source of variance. Pain 2008;138:286–91.
[15] Hastie BA, Riley 3rd JL, Robinson ME, Glover T, Campbell CM, Staud R, et al. Cluster analysis of multiple experimental pain modalities. Pain 2005;116:227–37.
[16] Binder A, May D, Baron R, Maier C, Tolle TR, Treede RD, et al. Transient receptor potential channel polymorphisms are associated with the somatosensory function in neuropathic pain patients. PloS One 2011;6:e17387.
[17] Baron R, Binder A, Wasner G. Neuropathic pain: diagnosis, pathophysiological mechanisms, and treatment. Lancet Neurol 2010;9:807–19.
[18] Flühr K, Neddermeyer TJ, Lötsch J. Capsaicin or menthol sensitization induces quantitative but no qualitative changes to thermal and mechanical pain thresholds. Clin J Pain 2009;25:128–31.
[19] Heimann D, Lötsch J, Hummel T, Doehring A, Oertel BG. Linkage between increased nociception and olfaction via a SCN9A haplotype. PLoS One 2013.
[20] Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 2001;68:978–89.
[21] Lötsch J, Geisslinger G. Current evidence for a modulation of nociception by human genetic polymorphisms. Pain 2007;132:18–22.
[22] Lötsch J, Doehring A, Mogil JS, Arndt T, Geisslinger G, Ultsch A. Functional genomics of pain in analgesic drug development and therapy. Pharmacol Ther 2013.
[23] Wendel B, Hoehe MR. The human mu opioid receptor gene: 5′ regulatory and intronic sequences. J Mol Med 1998;76:525–32.
[24] Bzdega T, Chin H, Kim H, Jung HH, Kozak CA, Klee WA. Regional expression and chromosomal localization of the delta opiate receptor gene. Proc Natl Acad Sci U S A 1993;90:9305–9.
[25] Xue Q, Yu Y, Trilk SL, Jong BE, Schumacher MA. The genomic organization of the gene encoding the vanilloid receptor: evidence for multiple splice variants. Genomics 2001;76:14–20.
[26] Clapham DE, Julius D, Montell C, Schultz G. International union of pharmacology. XLIX. Nomenclature and structure–function relationships of transient receptor potential channels. Pharmacol Rev 2005;57:427–50.
[27] Kim H, Mittal DP, Iadarola MJ, Dionne RA. Genetic predictors for acute experimental cold and heat pain sensitivity in humans. J Med Genet 2006;43:e40.
[28] Diatchenko L, Slade GD, Nackley AG, Bhalang K, Sigurdsson A, Belfer I, et al. Genetic basis for individual variations in pain perception and the development of a chronic pain condition. Hum Mol Genet 2005;14:135–43.
[29] Tegeder I, Adolph J, Schmidt H, Woolf CJ, Geisslinger G, Lötsch J. Reduced hyperalgesia in homozygous carriers of a GTP cyclohydrolase 1 haplotype. Eur J Pain 2008;12:1069–77.
[30] Mogil JS, Ritchie J, Smith SB, Strasburg K, Kaplan L, Wallace MR, et al. Melanocortin-1 receptor gene variants affect pain and mu-opioid analgesia in mice and humans. J Med Genet 2005;42:583–7.
[31] Liem EB, Joiner TV, Tsueda K, Sessler DI. Increased sensitivity to thermal pain and reduced subcutaneous lidocaine efficacy in redheads. Anesthesiology 2005;102:509–14.
[32] Diatchenko L, Nackley AG, Slade GD, Bhalang K, Belfer I, Max MB, et al. Catechol-O-methyltransferase gene polymorphisms are associated with multiple pain-evoking stimuli. Pain 2006;125:216–24.
[33] Ultsch A. Maps for visualization of high-dimensional data spaces. Japan: WSOM Kyushu; 2003. p. 225–30.
[34] Heskes T. Energy functions for self-organizing maps. In: Oja E, Kaski S, editors. Kohonen maps. Amsterdam: Elsevier; 1999.
[35] Hubert L, Arabie P. Comparing partitions. J Classif 1985;2:193–218.
[36] Izenmann A. Modern multivariate statistical techniques. Berlin: Springer; 2009.
[37] Ultsch A, Moerchen F. Databionic ESOM tools 2005.
[38] Hill T, Lewicki P. STATISTICS: methods and applications. Tulsa, OK: StatSoft; 2007.
[39] Lötsch J, Hofmann WP, Schlecker C, Zeuzem S, Geisslinger G, Ultsch A, et al. Current evidence and predictive performance of single and combined IL28B, ITPA and SLC28A3 host genetic markers modulating response to anti-hepatitis C therapy; 2011.
[40] Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8:283–98.
[41] Lötsch J, Belfer I, Kirchhof A, Mishra BK, Max MB, Doehring A, et al. Reliable screening for a pain-protective haplotype in the GTP cyclohydrolase 1 gene (GCH1) through the use of 3 or fewer single nucleotide polymorphisms. Clin Chem 2007;53:1010–5.
[42] Ultsch A, Li C. Automatic acquisition of symbolic knowledge from sub symbolic neural networks. In: International conference of signal processing, II, Beijing, China; 1993.
[43] Kim H, Neubert JK, San Miguel A, Xu K, Krishnaraju RK, Iadarola MJ, et al. Genetic influence on variability in human acute experimental pain sensitivity associated with gender, ethnicity and psychological temperament. Pain 2004;109:488–96.
[44] Lötsch J, Geisslinger G. A critical appraisal of human genotyping for pain therapy. Trends Pharmacol Sci 2010;31:312–7.
[45] Tegeder I, Costigan M, Griffin RS, Abele A, Belfer I, Schmidt H, et al. GTP cyclohydrolase and tetrahydrobiopterin regulate pain sensitivity and persistence. Nat Med 2006;12:1269–77.
[46] Berkley KJ. Sex differences in pain. Behav Brain Sci 1997;20:371–80 [discussion 435–513].
[47] Riley 3rd JL, Robinson ME, Wise EA, Myers CD, Fillingim RB. Sex differences in the perception of noxious experimental stimuli: a meta-analysis. Pain 1998;74:181–7.
[48] Derbyshire SW. Gender, pain, and the brain. Pain: clinical updates; 2008.
[49] Fillingim RB, Kaplan L, Staud R, Ness TJ, Glover TL, Campbell CM, et al. The A118G single nucleotide polymorphism of the mu-opioid receptor gene (OPRM1) is associated with pressure pain sensitivity in humans. J Pain 2005;6:159–67.
[50] Lötsch J, Stuck B, Hummel T. The human mu-opioid receptor gene polymorphism 118A > G decreases cortical activation in response to specific nociceptive stimulation. Behav Neurosci 2006;120:1218–24.
[51] Bond C, LaForge KS, Tian M, Melia D, Zhang S, Borg L, et al. Single-nucleotide polymorphism in the human mu opioid receptor gene alters beta-endorphin binding and activity: possible implications for opiate addiction. Proc Natl Acad Sci U S A 1998;95:9608–13.
[52] Oertel BG, Kettner M, Scholich K, Renne C, Roskam B, Geisslinger G, et al. A common human mu-opioid receptor genetic variant diminishes the receptor signaling efficacy in brain regions processing the sensory information of pain. J Biol Chem 2009;284:6530–5.
[53] Beyer A, Koch T, Schroder H, Schulz S, Höllt V. Effect of the A118G polymorphism on binding affinity, potency and agonist-mediated endocytosis, desensitization, and resensitization of the human mu-opioid receptor. J Neurochem 2004;89:553–60.
[54] Kroslak T, Laforge KS, Gianotti RJ, Ho A, Nielsen DA, Kreek MJ. The single nucleotide polymorphism A118G alters functional properties of the human mu opioid receptor. J Neurochem 2007;103:77–87.
[55] Zhang Y, Wang D, Johnson AD, Papp AC, Sadee W. Allelic expression imbalance of human mu opioid receptor (OPRM1) caused by variant A118G. J Biol Chem 2005;280:32618–24.
[56] Oertel BG, Doehring A, Roskam B, Kettner M, Hackmann N, Ferreiros N, et al. Genetic–epigenetic interaction modulates mu-opioid receptor regulation. Hum Mol Genet 2012;21:4751–60.
[57] Lacroix-Fralish ML, Ledoux JB, Mogil JS. The pain genes database: an interactive web browser of pain-related transgenic knockout studies. Pain 2007;131. 3e1-4.