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