Discretization of Expression Quantitative Trait Loci in Association Analysis Between Genotypes and Expression Data[§]

Andrés R. Masegosa[1], Rubén Armañanzas[2], María M. Abad-Grau[*,1], Víctor Potenciano[3], Serafín Moral[1], Pedro Larrañaga[4], Concha Bielza[4] and Fuencisla Matesanz[5]

[1]*CITIC, Universidad de Granada, Granada, Spain*

[2]*Krasnow Institute for Advanced Study, George Mason University, Fairfax, VA, USA*

[3]*Poten Dynamics, Granada, Spain*

[4]*Computational Intelligence Group, Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Madrid, Spain*

[5]*Instituto de Parasitología López Neyra, CSIC, Granada, Spain*

Abstract: Expression quantitative trait loci are used as a tool to identify genetic causes of natural variation in gene expression. Only in a few cases the expression of a gene is controlled by a variant on a single genetic marker. There is a plethora of different complexity levels of interaction effects within markers, within genes and between marker and genes. This complexity challenges biostatisticians and bioinformatitians every day and makes findings difficult to appear. As a way to simplify analysis and better control confounders, we tried a new approach for association analysis between genotypes and expression data. We pursued to understand whether discretization of expression data can be useful in genome-transcriptome association analyses. By discretizing the dependent variable, algorithms for learning classifiers from data as well as performing block selection were used to help understanding the relationship between the expression of a gene and genetic markers. We present the results of using this approach to detect new possible causes of expression variation of DRB5, a gene playing an important role within the immune system. Together with expression of gene DRB5 obtained from the classical microarray technology, we have also measured DRB5 expression by using the more recent next-generation sequencing technology. A supplementary website including a link to the software with the method implemented can be found at http: //bios.ugr.es/DRB5.

1. INTRODUCTION

Association between genotypes and mRNA transcript levels may help elucidating genetic basis of complex diseases by analyzing whenever genetic variants affect gene expression. Therefore, a genetic variant affecting a disease may be found also in association with the expression level of a gene. However, it is not straightforward to understand whether it may truly alter gene transcription or splicing, i.e., it may be an expression quantitative trait loci (eQTL), or just being in linkage disequilibrium (LD) with the real cause [1]. Moreover, because of small sample sizes and limited computational resources, regression models using several input variables have given results hardly reproducible and most successful association analyses only succeeded when testing a single polymorphic locus (SNP) against the expression of a gene instead of considering more than one SNP at a time [1-3]. We have used a different approach to measure association between SNPs and gene expression data which relies on a pre-discretization of expression data as a way to simplify input data and improve performance compared with standard regression models. Discretization of gene expression data is commonly performed when they are used as the input variables to predict different phenotypes such as cellular classification in cancer [4-6]. With this simplification, we are able to use data to learn a classifier, i.e. a model that relates how input variables, the SNPs, and their interactions, affect a discrete output variable, with values interpreted as high and low gene expression if it is the case of only two bins or high, regular and low gene expression if it is the case of three bins. By using classifiers instead of regression functions, other more complex analyses can be made, such as considering multiple SNPs at a time or haplotyping analysis [7]; it could be reduced the computational and statistical complexity; and, in consequence, to have a more affordable alternative. But most important, a different approach may help shed light about the main features of the data analyzed and thus about different interaction patterns between genes and regulatory proteins affecting their expression and about their association with SNPs within a block of high LD. Moreover,

*Address correspondence to this author at the Departamento Lenguajes y Sistemas Informáticos, C/ Periodista Daniel Saucedo Aranda s/n, Granada 18071, Spain; Tel: +34 958240634; Fax: +34 958243179; E-mail: mabad@ugr.es

[§]This work is an extension of the following conference paper: Andres R. Masegosa, María del Mar Abad-Grau, Serafín Moral, and Fuencisla Matesanz. Learning classifiers from discretized expression quantitative trait-loci. IWBBIO, page 427-436. Copicentro Editorial, (2013).

the use of different classifiers under different assumptions and the use of different learning algorithms under different approaches may help to increase chances of discovering new regulation patterns.

We focused on gene HLA-DRB5 (DRB5). We chose this gene because the expression pattern in the first two data sets (described at Section 2.8) analyzed -those obtained by using microarray technology- showed two non-overlapping distributions of low and high expression levels (see Fig. **1**) and it was easily translated to a binary variable.

DRB5 is one of the genes that encode β chains for the DR HLA class II receptor. The HLA genes are located on the short arm of chromosome 6 and are organized in three regions: MHC class I, MHC class II and MHC class III. HLA class II genes encode glycoproteins expressed primarily on antigen-presenting cells where they present processed antigenic peptides to CD4+ T cells. The DR β chain is encoded by 4 gens DRB1, 3, 4, and 5. There are also other pseudogenes that do not produce a protein: DRB2, 7, 8, and 9. Not everybody has a copy of each gene or pseudogene. There are 5 different haplotypes with different combinations of genes. DRB5 is only present in DR51 haplotype. This haplotype has been associated with immune related diseases susceptibility. In particular, the DRB5*0101- DRB1*1501-DQA1*0102- DQB1*0602 haplotype has been associated with Multiple Sclerosis (MS) in the North European population [8]. The strong linkage disequilibrium among the variants that integrate the mentioned haplotype, in the Caucasian population, makes very difficult to determine the primary associated variant. The polymorphisms at the DRB genes conferred different properties for antigen presentation and this has been postulated as the pathogenic mechanism. However, it could be not the only explanation for the HLA Class II association with Multiple Sclerosis. It has been described polymorphisms that alter HLA gene expression associated with Multiple Sclerosis susceptibility [9-11], which open the question of the role of the DRB gene expression levels in the pathology. In fact the ability to induce active experimental autoimmune encephalomyelitis (EAE), an animal model for MS disease, was increased in animals expressing higher levels of DRB5*01: 01, pointing to a role of the levels of expression of this gene in susceptibility [12].

The rest of the paper is divided in three main sections. In Section 2 we describe the proposed methods and the employed data sets. Results appear in Section 3 and a discussion can be read in Section 4.

2. MATERIALS AND METHODS

In Section 2.1, we start giving a motivation and a rough description of our approach. Section 2.2 contains the details of our discretization algorithm. Sections 2.3 and 2.4 present the basics of classification and regression models, respectively. Section 2.5 explains how these two different models can be compared between them. Section 2.6 details a preprocessing step for grouping SNPs. We show a flowchart in Section 2.7 with the steps followed by our proposed method. Finally in Section 2.8 we describe the data sets we used and the procedures we followed to obtain them.

2.1. Motivation

Our discretization step of mRNA transcript levels is empirically motivated. Particularly, it arose when we observed the histograms of the expression level of the gene DRB5 measured by using microarrays from the RNA of lymphoblastoid cell lines in two different populations (one from Utah, USA composed of individuals with ancestries in Northern and Western Europe (CEU), and the other (YRI) with individuals from Yoruba, Nigeria. These histograms are shown in Fig. (**1**).

As it can be seen, two different groups, within each population, neatly arose: one group with relative low expression levels and another group with relative high expression levels.

In this paper, we build on the identification of these two groups (i.e. under-expressed and over-expressed) to measure SNP-expression level associations. Our approach to measure

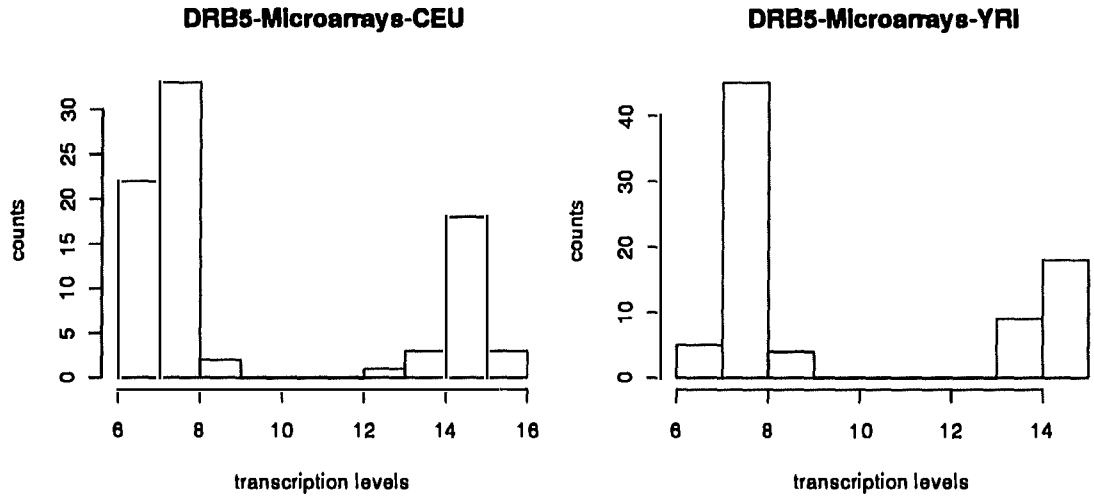


Fig. (1). Count-histograms of transcription levels (x-axis) for DRB5 in CEU-array data. set (**a**) and YRI-array data set (**b**).

these associations consists on employing statistical methods that considers the expression level as a binary variable (ternaries variables were later on explored in this work). This approach can be seen as an alternative, but not unique, methodology to those previously proposed statistical approaches which treat the expression level of a gene as a continuous real value [1-3] (e.g. the employment of a correlation test such as Spearman to measure the association between a single SNP and the expression of a given gene). More formally, we consider the following causal statistical models to explain the influence of one SNP over the expression of a target gene, the DRB5 gene in this work. These models are graphically described in Fig. (**2**) following the notation employed in Bayesian networks [13]. Under this notation, random variables are represented by round nodes and direct causal statistical influences are represented by directed edges. In this figure, the random variable S models the different values that a given SNP takes in a given population. In our case, we assume that one SNP can take three different values: 0, 1 and 2. Similarly, the random variable G models the expression level of any target gene in the population, which is assumed to be a random continuous value. Many previous approaches [1-3] for measuring SNP-expression levels associations implicitly employ the "Continuous Model": they test whether one SNP directly influences or not the continuous expression level of the target gene. In this paper we advocate for the "Discretization Model". This model assumes the existence of another hidden discrete variable, denoted by H in Fig. (**2**). This variable would represent a non-observable biological mechanism which is modulated by some SNP and, in turn, triggers the expression level, low or high, of the target gene. In result, we say that when one SNP regulates the expression level of a gene, it is not a direct cause of this regulation because it firstly affects this non-observable biological mechanism denoted by H. This mechanism would be the direct cause of the particular expression of the gene.

The assumption of the above model leaded us to use the alternative methods for measuring the SNP-expression levels associations presented in this work. We assume we have a data set with M members of a given population and for each member in our data set we can measure the value of a given set of SNPs, which are denoted by $S = \{S_1, ..., S_N\}$. Under these settings, s_{ij} will denote the value of the SNP Si for the j-th individual in the data set, with $i \in \{1, ..., N\}$ and $j \in \{1, ..., M\}$. Additionally, we also have measures for the expression level of a target gene G for each member of the data set. The expression level of the j-th member of the data set will be denoted by g_j with $j \in \{1, ..., M\}$ and the range of values that G can take will be denoted by Val(G). The set of expression values for the whole data set will be denoted by g. Our goal is to measure the association level between a given SNP Si and the expression of the gene G, assuming the "Discretization Model". We carried out this evaluation performing the two following steps:

Step 1: In this step we inferred, for each member of the data set, the values of the hidden variable H, denoted by h_j with $j \in \{1, ..., M\}$. We will also denote by h to the whole set of h_j values. In that way, we evaluated whether the expression level of the j-th member of the data set is low or high. In the case of data sets with count-histograms such as the one shown in Fig. (**1**), it is straightforward to infer these values: if $g_j < 10$ then h_j= low, otherwise h_j = high (the cut-off point could be any value between 9 and 12). For data sets where the two groups overlap, and for the cases in which there are reasons to consider more than two groups, we present in Section 2.2 an automatic approach based on the EM algorithm [14] and the Gaussian mixture model [15] to infer the $h_j$values using the expression level values g_j.

Step 2: Once we computed the values of the variable H for the data set from the population under study (i.e. h), we measured the association between SNPs and gene expression by using the variable H instead of the variable G. Because the variable H is discrete, a different family of statistical approaches became available for this purpose. In Section 2.3, we detail our proposal to accomplish this step.

2.2. The EM Algorithm and the Gaussian Mixture Model (Step 1)

In this section we give details about the "Step 1" of our proposal, as detailed in the previous section. The goal of this step was to infer the set of values h using the values of the gene expression g. For this purpose, we assumed that the gene expression level followed a Gaussian mixture model (GMM) [15]. Under this model the expression level of a gene is normally distributed conditioned to H: $P(G|H = h) \sim$ Gaussian(μ_h, σ_h), where μ_h and $\sigma_h$denote the mean and the standard deviation of the Gaussian distribution when H = h; and, then, the distribution of G follows a weighted mixture of Gaussian distributions, or GMM, $P(G) = \Sigma_{h \in Val(H)} w_h \cdot$ Gaussian(μ_h, σ_h), where w_h is the weight of the h-th component of the mixture, $w_h = P(H = h)$. Although the variable H is not observed, we can employ the EM algorithm [14] to infer the parameters which define this mixture: $w = \{w_1, ..., w_K\}$, $\mu = \{\mu_1, ..., \mu_K\}$ and $\sigma = \{\sigma_1, ..., \sigma_K\}$, where K denotes the number of values of H. Once these parameters are estimated with the EM algorithm, the h_j values were computed as follows:

$$h_j = \underset{h \in Val(H)}{argmax} P(g_j | H = h) P(H = h)$$

$$= \underset{h \in Val(H)}{argmax} \frac{w_h}{\sigma_h \sqrt{2\pi}} e^{\frac{-(g_j - \mu_h)^2}{2\sigma_h^2}}$$

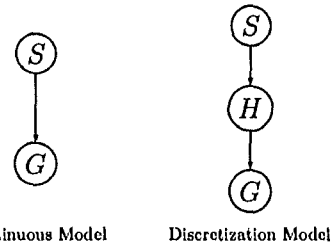


Fig. (2). The two causal statistical models evaluated in this work. Random variables are represented by round nodes and direct causal statistical influences are represented by directed edges. S models the different values that a given SNP takes in a given population; G models the expression level of any target gene, which is assumed to be a random continuous value. H is a hidden variable which models a non-observable biological mechanism which is modulated by some SNP and, in turn, triggers the expression level G of the target gene.

Algorithm 1: The EM Algorithm

Data: The vector of expression values **g**.
Result: An estimation of the parameters **w**, μ and σ.
t=0;
Start with a random initialization of the parameters: $\mathbf{w}^t$, μ^t and σ^t;
repeat
  for $j=1,...,M$ **do**
    for $h=1,...,K$ **do**

$$P(h_j^t = h|g_j) = \frac{\frac{w_h^t}{\sigma_h^t\sqrt{2\pi}} exp(-\frac{(g_j-\mu_h^t)^2}{2(\sigma_h^t)^2})}{\sum_{h'=1}^{K} \frac{w_{h'}^t}{\sigma_{h'}^t\sqrt{2\pi}} exp(-\frac{(g_j-\mu_{h'}^t)^2}{2(\sigma_{h'}^t)^2})};$$

  for $h=1,...,K$ **do**
    $\mu_h^{t+1} = \frac{1}{M}\sum_{j=1}^{M} g_j \cdot P(h_j^t = h|g_j)$;
    $\sigma_h^{t+1} = \sqrt{\frac{1}{M}\sum_{j=1}^{M}(g_j \cdot P(h_j^t = h|g_j) - \mu_h^{t+1})^2}$;
    $w_h^{t+1} = P(h_j^t = h|g_j)$;
  t=t+1;
until *convergence*;

In Algorithm 1 we give a pseudo-code description of the EM algorithm. This algorithm starts with a random initialization of the parameters w, μ and σ (when initializing w, it must be satisfied that $\sum_h w_h = 1$). The algorithm iterates until convergence (i.e., when the parameters in the iteration t + 1 are equal to the parameters in iteration t). However, it is well known that different executions of the algorithm can lead to different estimations of the parameters, where each different estimationcorresponds to alternative local maximum of the likelihood function. To avoid low quality convergence points, the EM algorithm was run 100 times as done and the solution with the highest likelihood was the one finally chosen.

2.3. Supervised Classification Models (Step 2)

We tried to answer the following question: given a subset of SNPs, denoted by $Q \subseteq S$, to which extent are they associated to the gene whose expression level is modeled by the random variable G, but using the variable H?. More precisely, we used supervised classification models [16] for this purpose. A supervised classification model learns a classification function from a pool of data samples, which is called the training data set or the training population. For this specific case, the classification function, f: Val(Q) → Val(H), maps any possible joint assignment, denoted by q, of the variables or SNPs in $Q \subseteq S$ to one value of the variable H.

In this work, we define the association degree between the set of SNPs in Q for a given person and H based on the quality with which the function f predicts the value of H when a joint assignment q is known. In other words, we want to measure how well f predicts the discretized expression level (i.e. high or low) of the target gene knowing only the genotype of the SNPs in Q for that person.

A possible measure would be the so-called classification accuracy, which has to be estimated from a test data set or a test population, which is different of the training data set:

$$\text{Accuracy}(Q,M) = \frac{1}{T}\sum_{j=1}^{T} I[f(q_j) = h_j],$$

where T is the count of samples in the test data set; q_j and h_j denote the joint value of the SNPs in Q and the value of H for the j-th member of the test data set, respectively; and I is the indicator function, which is equal to 1 if f (q_j) is equal to h_j (i.e., the prediction is correct) and 0 otherwise. Let us note, that the classification accuracy depends of the particular classification model M and the subset of SNPs Q we are evaluating. In this work, we also computed the Area Under the ROC Curve (AUC) [17] as a complementary and robust measure of the performance of a classifier.

Because our association measure depends of the particular classification model, several state-of-the-art and computationally affordable classifiers were evaluated to get a more robust estimate of the association degree between SNPs and gene expression:

Naive Bayes [16]: It is a simple probabilistic classifier which works under the assumption that all input variables are conditionally independent given the output variable. As usual, predictions with this model are made by choosing the most probable class value given the genotype of an individual.

C4.5 [18]: It is a classification tree model, in which the data set is divided in structured hypercubesof those individuals sharing values. This is the most competitive algorithm of this family. It is called J48 in an open-source version implemented in Weka [19]. In this model, predictions are made by choosing the most frequent class in the leaf of the inferred tree where the genotype of the individual to be classified falls.

Support Vector Machines (SVM): In this model the input variables are transformed in a higher-dimension space so that a classifier is learned from the set of transformed variables by using a kernel function. We chose the default implementation of support vector machines in Wekausing the LibSVM java library [20].

One advantage of the first two approaches is that they build white-box models, i.e., models are directly readable and interpretable by human experts.

2.4. Regression Models

While a classification model predicts discrete or categorical values, a regression model makes continuous predictions. Similarly, it learns a regression function from a training data set. In this case, this regression function is defined as follows: g: Val(Q) → Val(G), where Val(G) corresponds to the real interval where the expression level of our target gene lies. We used this model to test the performance of the previously described "Continuous Model" (see Fig. **2**), where no hidden variable it is assumed, for measuring SNP-expression level associations. Along the quality of the classification models, we used the quality of the continuous predictions of the g function as a degree of association between the SNPs in Q and the gene G. A possible measure to evaluate the quality of a regression model is the root mean square error (RMSE) and the Pearson's correlation between the predicted and the real values [16], which has to be computed over a different test data set.

As it happened with the classification accuracy, this association measure depends on the particular regression model used. For this reason, we also considered a broad set of different regression models in our analysis:

SVM-reg [21]: It is based on support vector models and kernel methods, as its supervised classification counterpart.

Gaussian processes [22]: It is a Bayesian approach that employs Gaussian process priors over regression functions to improve their generalization capacity.

Lasso [23]: It applies regularization -a process of introducing additional information in order to solve an ill-posed problem or prevent overfitting- in the form of a penalty term for complexity, which performs variable selection by driving a number of regression coefficients to zero.

k-nearest neighbor [24]: It is an extension of the k-nn classifier which computes the output of a case by averaging the values of its k-nearest neighbors. In the simplest approach, where k = 1, the assigned value is exactly the same as the one from the closest case. It is therefore advisable to use a k value bigger than one. Rightly so, it can be also useful to weight the contributions of the neighbors based on their distances to the case under evaluation. In our experiments, we used two values for k, namely k = 3 and k = 5, and three different configurations to compute the output.

Let $\hat{y}$ be the regression estimator for a case x, and let y_i be the value of the independent variable in the i-th case. The three approaches for the kNN-reg can be defined as follows:

1. Average value of the k nearest neighbors (kNN-reg-avg): $\hat{y} = \frac{1}{k}\sum_{i=1}^{k} y_i$
2. Weighted estimation of the k nearest neighbors (kNN-reg-wgt): $\hat{y} = \frac{\sum_{i=1}^{k} \frac{1}{D(\mathbf{x},\mathbf{x}_i)} y_i}{\sum_{i=1}^{k} \frac{1}{D(\mathbf{x},\mathbf{x}_i)}}$ where $D(\mathbf{x}, \mathbf{x}_i)$ is the distance value between the case under evaluation, $\mathbf{x}$, and $\mathbf{x}_i$, one of its k nearest neighbors.
3. Kernel density estimation (kNN-reg-krn): the output corresponds to the value for which a kernel density smoothing reaches its maximum density. The density function is estimated based on a normal kernel function, which takes the values of the k nearest neighbors as input data.

As for measuring the distance between two genotypes, we made use of the Hamming distance. "The Hamming distance between two strings of equal length is the number of positions at which the corresponding characters are different" [25]. In genetic terms, it can be seen as the number of variations that transformed one genotype into the other.

2.5. Comparing Classification and Regression Models

As already commented, one of the main aims of this work was to evaluate how models which employ discretized expression values perform with respect to models which directly treat the expression as a real value. It is not straightforward to compare both approaches because they have different properties and, as commented in the above sections, the error measures usually employed to evaluate their performance are different. A possible approach to overcome this problem is to use the so-called relative absolute error (RAE). This metric evaluates a model by comparing the absolute error of the model itself with respect to the absolute error of a blinded model (i.e. the same model but not using any predictive variable: Q = ∅). This error metric is computed as follows:

$$\mathrm{RAE} = 100 \cdot \frac{\sum_{j=1}^{T}\sum_{l=0}^{K} |\mathrm{Prediction}(j,l) - \mathrm{ActualValue}(j,l)|}{\sum_{j=1}^{T}\sum_{l=0}^{K} |\mathrm{PriorPrediction}(l) - \mathrm{ActualValue}(j,l)|}$$

For the regression models, K is equal to 1; Prediction(j, l) corresponds to the real value prediction for the j-th test sample; ActualValue(j, l) corresponds to the actual value associated to the same instance; and PriorPrediction(l) is the average value of the expression level in the training data set; this prediction does not depend of the particular individual because it does not use any knowledge about the SNPs (i.e. Q = ∅).

For the classification models, K is equal to the number of values of the discrete expression level. Thus, ActualValue(j, l) = 1 if the discretized expression value of the j-th instance of the test data set is equal to l, and 0 otherwise. Prediction(j, l) with $l \in \{1, ..., K\}$ is the prediction vector. For the Naive Bayes model it corresponds to the probabilities that this model assigns to each value of the discretized expression, because this is a probabilistic classifier which makes soft predictions. On the contrary, C4.5 and SVM models make hard predictions and, then, Prediction(j, l) = 1 for the predicted class and 0 otherwise. In this case, PriorPrediction(l) is equal to the proportion of samples in the training population whose discretized expression value is equal to l. The above measure ranges from 0 to infinity. A zero value indicates a perfect prediction. Larger values indicate worse prediction capacity. A value over 100%

indicates strong overfitting because the model is performing worst than its blinded counterpart.

2.6. Block Processing

We were also interested to observe whether prediction accuracy changed when using models with a reduced number of SNPs, grouped by blocks of low recombination (i.e., SNPs with high linkage disequilibrium (LD) among them). We grouped SNPs by using a common approach based on pairwise computations of confidence intervals of LD [14]. Pairwise LD is measured by the normalized statistic of allelic association D'. The algorithm chooses the largest set of consecutive SNPs that reaches the requirements to be defined as a low recombination block, defined in terms of a minimum number of pairs of SNPs being in strong LD (one-sided upper 95 European ancestries than from Africa [21], we made blocks of chromosome 6 by using CEU, the data set of individuals with European ancestries (see Section 2.8) and used those blocks to group SNPs in YRI, the data set of individuals with African ancestries. Fig. (**S2**) shows the average number of SNPs by block in the data set used. Given a block, a classifier with only those SNPs in that block as input variables was learned. As a result, SNPs in chromosome 6 were grouped in 345 non-overlapping blocks of low recombination [26], which were learned from the CEU data set. DRB5 is a gene coded between 11 physical positions 32593098 and 32606042 in assembly NCBI36/hg18 or between 32485120 and 32498064 in assembly GRCh37/hg19. 6 SNPs has been genotyped in HapMap 3 within the gene DNA region. These SNPs belong to block 223. Tables **1** and **2** shows the SNPs within the block. Those in bold correspond to SNPs within the gene.

2.7. An Overview of the Procedures

Fig. (**3**) shows a flowchart with all the steps followed to conduct this study.

2.8. Data Sets Used

Expression data of gene DRB5 came from the mRNA of lymphoblastoid cell lines of 228 individuals from different populations. Details about the procedure followed to obtain expression data, including raw expression data normalization, population stratification correction andcorrection for known and unknown factors are described by Stranger *et al.* [3].

Several of these individuals were genotyped by the International HapMap project [27]. We used the third phase [28] of HapMap project to obtain genotypes in order to avoid large amounts of missing data, as in the other phases not all the individuals were genotyped. From all the CEU and YRI parental individuals in HapMap third phase, we only chose those CEU individuals (107) and those YRI individuals

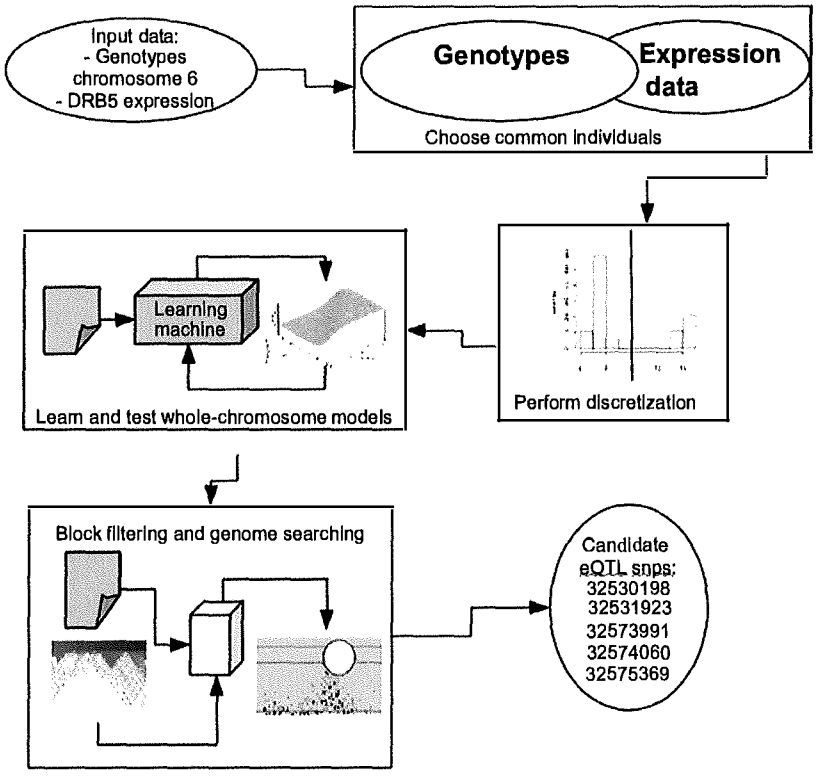


Fig. (3). Flowchart showing the steps followed in this study in order to obtain a minimal set of candidate eQTLs for expression of gene DRB5.

Table 1. List of SNPs within block number 223. In bold those within the gene DNA region.

rs Number	NCBI36/hg18	GRCh37/hg19
rs6901541	32550239	32442261
rs17209754	32550681	32442703
rs9269101	32550689	32442711
rs9269110	32551247	32443269
rs9378264	32551429	32443451
rs12194148	32552176	32444198
rs9405112	32553578	32445600
rs9378212	32553669	32445691
rs9269182	32555835	32447857
rs4999342	32556076	32448098
rs4410767	32556107	32448129
rs35465556	32556112	32448134
rs9378266	32556167	32448189
rs9378213	32556376	32448398
rs9269186	32556394	32448416
rs9269187	32556418	32448440
rs9269190	32556478	32448500
rs9391786	32556539	32448561
rs7748270	32556577	32448599
rs7748472	32556741	32448763
rs7749057	32556882	32448904
rs7749242	32557008	32449030
rs7749092	32557028	32449050
rs6911871	32557256	32449278
rs1964995	32557389	32449411
rs11752428	32557448	32449470
rs9269202	32557501	32449523
rs11757500	32557633	32449655
rs9269204	32557775	32449797
rs7754119	32557836	32449858
rs9269211	32558036	32450058
rs5020946	32558067	32450089
rs12191360	32559339	32451361
rs1557551	32560168	32452190
rs6916742	32561169	32453191
rs7754731	32561832	32453854
rs28877027	32574137	32466159
rs35847514	32587470	32479492
rs34507021	32587485	32479507
rs1137498	32593190	32485212
rs35085841	32597079	32489101
rs35056680	32601256	32493278
rs35739325	32601768	32493790
rs34328528	32601789	32493811
rs17203992	32604788	32496810
rs16870207	32606390	32498412
rs2157337	32609122	32501144
rs34249660	32609486	32501508
rs28772724	32617335	32509357
rs28760027	32617963	32509985
rs2157339	32619650	32511672
rs28495356	32620142	32512164
rs35571839	32620591	32512613
rs34369284	32622122	32514144
rs28490179	32626983	32519005
rs11759557	32628011	32520033
rs11757159	32628250	32520272
rs34781832	32628606	32520628
rs1064611	32630503	32522525
rs34569694	32632659	32524681
rs28656080	32633666	32525688
rs28530648	32635057	32527079
rs35464393	32638176	32530198
rs12661707	32639223	32531245
rs35366052	32639901	32531923

(107), for which we had the expression of gene DRB5. We chose the SNPs passing quality control which overlapped genes DRB5 and DRB1 or within a window of 1 million basis before the first gene position (28922491 in assembly GRCh37/hg19) and after the last gene position (33961785 in assembly GRCh37/hg19) in chromosome 6 (where genes DRB5 and DRB1 belong to). The total number of SNPs was 6593. Missing information was inferred by using familial information and the IMPUTE2 algorithm [29] and these data were downloaded from the HapMap project website (http://www.hapmap.org). The final two data sets were called CEU-array and YRI-array. For a second analysis pursuing to replicate results using expression data obtained by next-generation RNA sequencing (NGS) technology, we built other two data sets (CAU-RNASeq and YRI-RNASeq). Expression data for gene DRB5 was obtained from the Geuvadis project [30], a second experiment using NGS technology in which the RNA of 465 lymphoblastoid cell lines from the 1000 Genomes project [31] was sequenced. To accurately quantify the expression level of genes from RNA-seq reads we first applied a cleaning procedure on raw data followed by a common protocol TopHat-Cufflink [32] to obtain gene expression levels. Thus, from the cleaned data (files if FASTQ format) at the Geuvadis project we used TopHat software, "a read alignment program that allows alignments between a read and the genome to contain large gaps" [32], to perform read alignment using the human genome as reference. "To assemble individual transcripts from the RNA-seq reads that were aligned to the genome and obtain the expression level of genes" [32], we used Cufflinks. Out of these 465 cell lines, 259 belonged to Caucasian individuals (only 82 of them were included in the data used in our first analysis, the others correspond to British individuals) and 79 were the only African individuals in the study, all of them Yoruban included in the data used in our first analysis. To avoid reducing sample size of the CEU data from 259 to 79, we decided to make two data sets with the 259 Caucasian and the 79 African individuals by using genotypes from 1000 Genomes as the other individuals were not genotyped by the HapMap project. SNP selection was made following the same criterion as with HapMap data. The total number of selected SNPs, 97, 484, was much higher due to the higher genotyping density used by the 1000 Genomes project. Finally, and in order to understand the lower performance in results when usingRNASeq data, we also created two data sets, CEU-commonIndividuals, YRI-commonIndividuals with respectively only the 82 CEU and 81 YRI whose DRB5-expression was obtained by the two different gene expression technologies.

3. RESULTS

3.1. Microarray Expression Data[1]

3.1.1. Discretization Step

The result of the discretization step on the microarray gene expression *via* the EM algorithm was clear for the CEU

[1]A reduced version of these experiments with microarray expression data were presented in the conference paper's version of this work. We present again some of the results for the sake of completeness and readability of the paper.

and YRI populations. The EM algorithm defined two groups which could be easily identified by looking at the histograms shown in Fig. (**1**). Before comparing multivariate models, and in order to understand how discretization behaved when using the common single-SNP association, we first compared, for each one of the 6593 SNPs considered in CEU-array and YRI-array data sets, the Spearman correlation coefficients obtained when using the continuous expression of the DRB5 gene and when using the discretized variable. For this last case, the binary variable was assumed to take two real values: 0 for low-expression; and 1 for high-expression. In Fig. (**4**), we plot the Spearman correlation coefficients of this comparison for the two populations.

As can be seen in this figure, the "2 Bins Discretization" series and the "Continuous Value" series are quite similar. Actually, if we compute the Spearman correlation coefficient

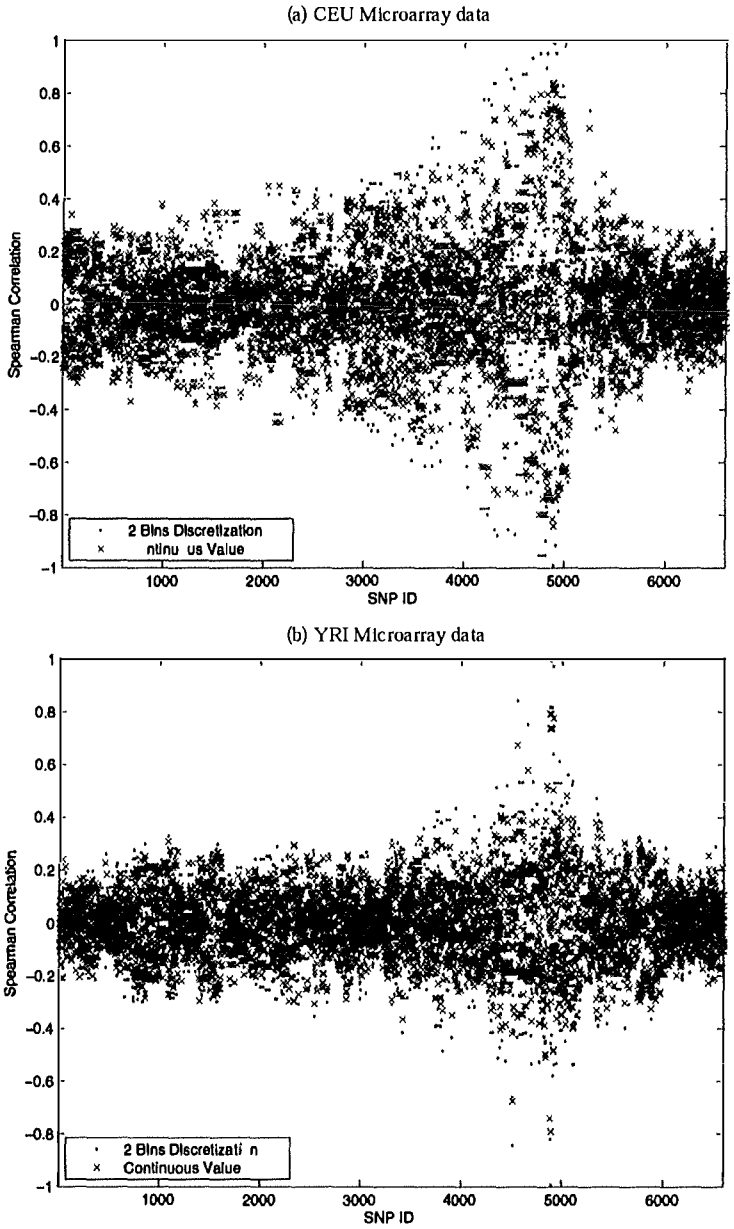


Fig. (4). Spearman correlation for each 6593 analyzed SNPs and the continuous expression level (blue points); and for the same SNPs and the binarized expression (red points). The Spearman correlation value between the continuous and discrete series of CEU-array was equal to 0.8938; for YRI-array was equal to 0.8798.

Table 3. The 10 SNPs with the highest Spearman correlation coefficients for the two populations computed when the expression of the gene is continuous and discretized. Those four SNPs among the 10 top SNPs in both populations are highlighted.

YRI-Array				CEU-Array			
Discretized		Continuous		Discretized		Continuous	
Pou	Corr.	Pou	Corr.	Pou	Corr.	Pou	Corr.
32638176	0.9981	32638176	0.7944	32639901	1	32681969	0.8413
32639901	0.9981	32639901	0.7944	**32638176**	0.995	32682038	-0.8413
32682038	-0.9946	32681969	0.792	**32681969**	0.9905	32683347	-0.8413
32683347	-0.9946	32682038	-0.7904	**32682038**	-0.9905	32638176	0.8373
32681969	0.9931	32683347	-0.7904	**32683347**	-0.9905	32639901	0.8337
32694832	0.9727	32694832	0.7759	32500884	0.9517	32648019	0.82
32396216	-0.8431	32666924	0.7411	32500959	-0.9517	32666924	0.8151
32417132	0.8431	32678817	-0.7411	32512355	-0.9517	32678817	-0.8151
32666924	0.8193	32648019	0.7372	32518965	-0.9517	32684456	0.8151
32678817	-0.8193	32684456	0.7372	32521029	-0.9517	32685867	0.8151

for these two pair of series (i.e. treating the Spearman coefficients between SNPs and gene expression as two independent series of real values) we find that the '2 Bins Discretization" series and the "Continuous Value" series were highly correlated. For CEU population the correlation was equal to 0.8938 while for YRI population was equal to 0.8798. That means that those SNPs that are highly correlated with the continuous expression of the gene are also similarly correlated with the discretized expression of the gene. In Table **3** we detail the 10 SNPs with the highest Spearman correlation coefficients for the two populations computed when the expression of the gene is continuous and discretized. As it can be seen, with the discretized expression of the gene the correlation notably increased in both populations.

As summary, we find a positive effect on the Spearman correlation coefficients when discretizing the DRB5 microarray expression level.

3.1.2. Whole Models

We later evaluated classification and regression models using all SNPs at a time (i.e. Q = S) to study the extent to which the expression level using microarrays of DRB5 is controlled by the selected SNPs. To correctly evaluate the performance of classification and regression models, we employed the so-called 10-fold cross-validation methodology (10-cv) [33] to build different training and test data sets: firstly, the members of one data set are randomly divided in 10 groups; then, ten different test data sets are created by selecting each time a different group; the other ten different training sets are built with the remaining 9 groups; finally, the models are trained and tested ten times with each train/test pair and the averaged performance measures of these ten validations are reported.

Tables **4** and **5** show respectively results for data sets CEU-array and YRI-array. The relative absolute error (see Section 2.5) is defined for both classification and regression models. The application of a paired t-test reveals that RAE for the algorithm with best results under the discretization approach (C4.5) is significantly lower than the best algorithm among the common regression approach (Lasso): p values are 1.0e-6 and 1.0e-4 for CEU and YRI data sets respectively. Again, a non-parametric test such as Wilcoxon could have been also applied.

Table 4. Generalization capacity of different classification and regression models using all SNPs as input variables in CEU-array data set.

Classification Models	Accuracy	AUC	Relat. Abs. Error
NB	91.64	0.89	18.58
C4.5	**100.0**	**1.00**	**0.00**
SVM	89.92	0.86	21.87
Regression Models	**RMSE**	**Correlation**	**Relat. Abs. Error**
SVM-Reg.	1.56	0.87	37.27
Gaussian Proc.	1.56	0.87	37.28
Lasso	**0.41**	**0.99**	**10.47**
3-NN-reg-avg	2.74	0.52	60.78
5-NN-reg-avg	2.66	0.54	62.47
3-NN-reg-wgt	2.73	0.53	60.55
5-NN-reg-wgt	2.65	0.55	62.17
3NN-reg-krn	3.12	0.57	60.12
5NN-reg-krn	2.99	0.61	54.62

Looking at these results, we can observe that C4.5 has an outstanding performance because it achieves perfect or almost perfect classification in CEU-array and YRI-array data sets. These are very good news because C4.5 is not a black-box machine learner. Its decision tree based nature allows to easily interpret the classification rules used for making predictions. So, this model could potentially help biomedical researchers to understand SNPs regulation of DRB5 expression. On the contrary, classifiers based on SVM and regression models had much worst predictive performance (i.e. higher relative absolute error). However, a much better performance was expected due to the well-known fact that expression regulation of DRB5 is controlled by genetic variants in chromosome 6 tagged by some of the SNPs from the genotype array used.

We see again how the discretization step built more accurate prediction models than considering the gene expression as a continuous value and, hence, using regression techniques.

However, it has to be noted that Lasso regression clearly outperforms the other regression methods and it also outperforms SVM classifiers. A further discussion of this result is given in Section 4.

Finally, we want to comment that when inspecting the tree model learnt with the C4.5 algorithm we can see that, either for CEU-array or for YRI data sets, we obtain a simple tree with one single attribute. This single attribute is different for each population but it belongs in both cases to the block 223. Further analyses at this respect are given in following section.

3.1.3. Block-Based Approach

In this new analysis we depart from the block partition described in Section 2.6. With this analysis, we tried to understand which blocks are more correlated with the microarray expression of DRB5. For this purpose, we evaluated the performance of the classification and regression models by using as input variables (i.e. Q) those SNPs contained in a single block. This is a biologically-inspired feature selection method. Would accuracy keep as higher as when all SNPs were used in classification/ regression models? Would it increase because some noise or redundant variables were eliminated?

Among the classification models, we picked the C4.5 model, the one which performed best, as shown in the previous section. The Lasso regression model was chosen from the same reason among the regression models. We built a total of 345 data sets for each population by selecting the SNPs within each block and used the 10-cv evaluation method to estimate the different performance measures.

In Fig. (**5**), we display the results of this analysis. In Table **6** we detail the 10 blocks with the lowest relative absolute error for the classification and the regression model in the two populations. It can be seen again that classification models performed in both populations better than regression models in the key SNPs blocks, which are highly associated to the expression of the DRB5.

Another question that arises in this analysis is whether the performance of a classification model using a set of SNPs is higher or not than the performance obtained using a single SNP. That is to say, can we predict better the expression level of DRB5 by aggregating multiple SNPs? To try to answer this question, we compared AUC of the C4.5 classification model using all SNPs within the same block with respect to the AUC obtained using only one single SNP. We selected the SNP with the maximum AUC within the same block. In Fig. (**6**) we plot both measures for all the 345 blocks. In Table **7** we detail the 10 blocks with the highest AUC using the C4.5 classifier as well as the AUC obtained with the best performing single SNP among all SNPs within this block. These results are deeply discussed in Section 4.

3.2. Replication with RNASeq Expression Data

3.2.1. Discretization Step

The discretization of the two RNASeq-based populations, CAU-RNASeq and YRI-RNASeq, were much less straightforward than the array-based populations. As can be

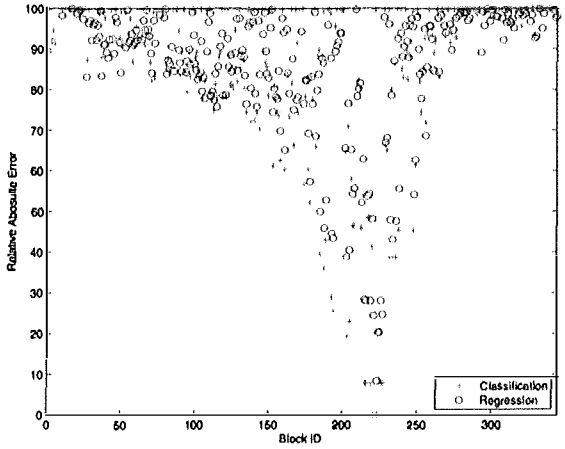
Table 5. Generalization capacity of different classification and regression models using all SNPs as input variables in YRI-array data set.

Classification Models	Accuracy	AUC	Relat. Abs. Error
NB	89.7	0.89	24.53
C4.5	99.0	1.00	1.66
SVM	79.5	0.66	48.43
Regression Models	**RMSE**	**Correlation**	**Relat. Abs. Error**
SVM-Reg.	1.99	0.76	55.06
Gaussian Proc.	1.99	0.76	55.05
Lasso	0.74	0.98	17.29
3-NN-reg-avg	2.66	0.54	69.28
5-NN-reg-avg	2.62	0.50	72.26
3-NN-reg-wgt	2.68	0.53	69.65
5-NN-reg-wgt	2.62	0.50	72.33
3NN-reg-krn	2.34	0.63	46.53
5NN-reg-krn	2.68	0.52	51.37

Table 6. The 10 blocks with the lowest relative absolute error (RAE) using a classification model (C4.5) learnt with all SNPs within this block. 7 blocks (in bold) out of the 10 are shared by the two populations.

YRI-Array				CEU-Array			
Discretized		Continuous		Discretized		Continuous	
Block Id	RAE	Block Id	RAE	Block Id	RAE	Block Id	RAE
224.00	0.00	223.00	12.85	**221.00**	0.00	223.00	8.34
225.00	0.00	225.00	23.62	**223.00**	0.00	224.00	20.27
223.00	1.56	224.00	24.84	**224.00**	0.00	225.00	20.37
227.00	4.39	227.00	26.72	**226.00**	7.79	221.00	24.44
226.00	9.08	226.00	31.38	215.00	7.86	227.00	24.72
214.00	25.48	213.00	45.16	216.00	7.86	226.00	28.08
203.00	28.49	203.00	45.62	219.00	7.86	216.00	28.11
205.00	28.49	205.00	48.15	**227.00**	7.86	219.00	28.12
221.00	30.22	198.00	54.17	**225.00**	8.50	215.00	28.42
213.00	32.62	221.00	54.63	**203.00**	19.26	203.00	38.81

(a) CEU Microarray data



(b) YRI Microarray data

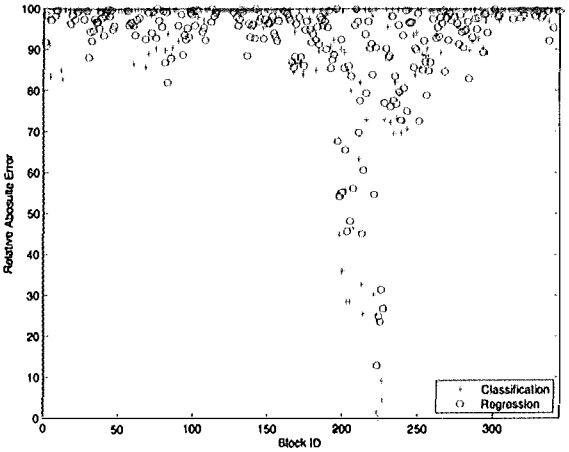


Fig. (5). Relative Absolute Error comparison between a classification model (C4.5, red plus sign) and a regression model (Lasso, blue circle). For the CEU-array population, the number of blocks where the classification model obtains a RAE < 10% is 9 and for RAE < 30% the number of blocks is 13. For the regression model in the CEU-array population, these numbers are 1 and 9 respectively. For the YRI-array population, the number of blocks where the classification model obtains a RAE < 10% is 5 and for RAE < 30% the number of blocks is 8. For the regression model in the YRI-array population, these numbers are 0 and 4 respectively.

Table 7. The 10 blocks with the highest AUC using a classification model (C4.5). "Single SNP AUC" column displays the AUC obtained with the best performing single SNP among all SNPs within this block. 7 blocks (in bold) out of the 10 are shared by the two populations.

YRI-Array			CEU-Array		
Block ID	Block AUC	Single SNP AUC	Block ID	Block AUC	Single SNP AUC
223	1.000	1.000	**221**	1.000	0.866
224	1.000	1.000	**223**	1.000	1.000
225	1.000	0.668	**224**	1.000	1.000
226	0.973	0.827	**225**	0.983	0.760
227	0.970	0.970	**226**	0.973	0.697
221	0.899	0.689	215	0.954	0.954
213	0.892	0.741	216	0.954	0.954
214	0.858	0.678	219	0.954	0.954
203	0.855	0.839	**227**	0.954	0.954
205	0.855	0.839	**203**	0.899	0.897

seen in the count-histograms shown in Fig. (**7**), there were not so clearly differentiated groups as in the case of array-based data sets. This has opened an issue still under research that will be discussed in Section 4.

The first decision we had to make to proceed with the application of the EM algorithm for discretizing these data sets was to choose the number of bins (i.e. the K value). The problem here is that EM is not able to directly indicate which is the optimal number of bins. Although some methods has been proposed to help EM to decide the number of bins [34], there is not clear solution to this problem and the best approach usually is a mixture of using some expert knowledge, if available, and trial-error tests. In our case, we evaluated two and three bins discretization configurations.

In a first analysis, we compared the Spearman correlation coefficients obtained between each one of the 97, 484 SNPs considered in this data set and the continuous expression of DRB5 gene, and between the same SNPs and the discretized expression in two and three bins obtained with the EM algorithm. For these last two cases, the 2 bins variable was assumed to take two values: 0 for low-expression; and 1 for high-expression; and the 3 bins variable was assumed to take three values: 0 for the lowest expression group; 1 for the middle expression group; and 2 for the highest expression group. In Fig. (**8**), we plot the Spearman correlation coefficients of this comparison for the two populations.

As it can be seen in this figure, the "2 Bins Discretization" series, the "3 Bins Discretization" series and the "Continuous Value" series are not very different among them. When we computed the Spearman correlation coefficient between "Continuous Value" series and the "3 Bins Discretization" series (i.e. treating the Spearman correlation coefficients between SNPs and gene expression as real values) we found that they were highly correlated: for CEU population the correlation was equal to 0.9636 while for YRI population was equal to 0.9315. When we performed this analysis with the "2 Bins Discretization" the Spearman correlation coefficients were 0.9319 and 0.9218 for CAU-RNASeq and YRI-RNASeq, respectively.

In Tables **8** and **9**, we detail the 10 SNPs with the highest Spearman correlation coefficients for the two populations computed when the expression of the gene is continuous and discretized in 2 and 3 bins. As can be seen, the discretization with 3 bins generates higher Spearman correlation between SNPs and gene expression than 2 bins discretization. We can also see like the correlation with the discretized expression does not strongly increase as happens with the microarray data.

In light of the above results we decided to continue with 3 bins discretization. In Fig. (**9**) we show the Gaussian mixtures inferred by the EM algorithm using 3 Gaussian components for the two data sets.

3.2.2. Whole Models

As we did with microarray data, we evaluated again the classification and regression models using all SNPs at a time (i.e. Q = S) to study the extent to which the RNASeq expression of DRB5 is controlled by the new selected SNPs.

Tables **10** and **11** show respectively results for data sets CAU-RNASeq and YRI-RNASeq, following the same evaluation methodology used in Section 3.1.2. The application of a paired t-test at 0.05 level reveals that SVM, the algorithm with the best results among those under the discretization approach, has RAE significantly lower than the one reached by the algorithm under the common regression approach (Lasso): p values are 0.0396 and 0.0068 for CEU and YRI respectively. A non-parametric test such as Wilcoxon could have been also applied.

Looking at these tables, we can see that both the classification and the regression models performs poorly than in the case of microarray data. In Section 4 we discuss about possible reasons of RNASeq underperforming microarrays, when it is supposed to be a more accurate technology [34]. However, the classification models still perform better than the regression models when comparing their relative absolute errors. Another unexpected result is that SVM outperform C4.5 in YRI, even when it does not perform any variable selection procedure. Again it may be

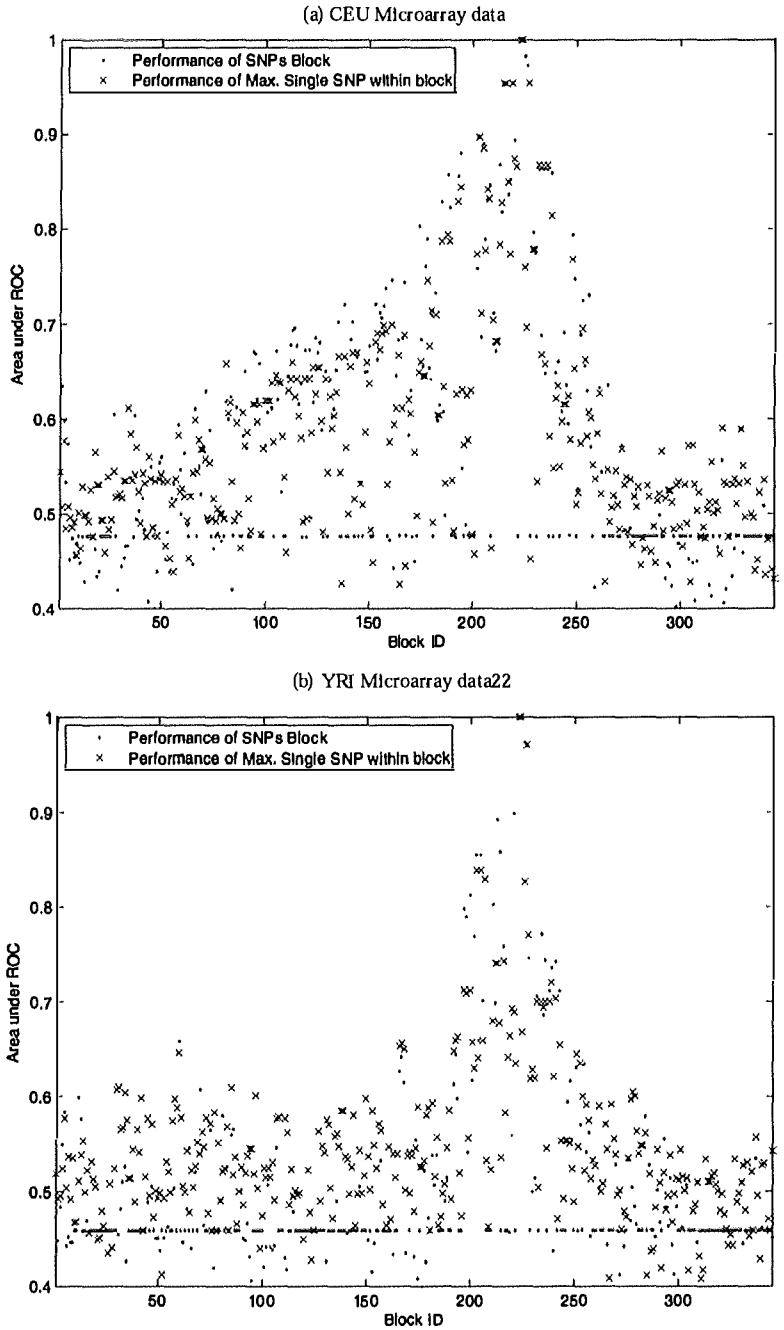


Fig. (6). AUC comparison between a classification model (C4.5) using all SNPs in a block (red filled circle); and the maximum performance obtained using a single SNP within the same block (blue cross).

due to a problem in the way expression data was obtained by RNASeq technology and will be discussed in Section 4. Like in the case of microarray data, we tried to inspect the tree models induced by C4.5 using the CAU-RNASeq and the YRI-RNASeq populations. However in this case trees are not as easily interpretable as before because they involved tens of different SNPs and, moreover, they tend to vary if these trees are induced with slightly different training data sets, as happens when using cross validation.

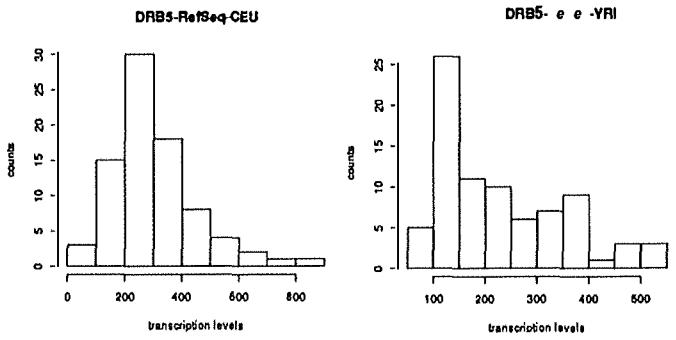


Fig. (7). Count-histograms of transcription levels (x-axis) for DRB5 in CAU-RNASeq data set (**a**) and YRI-RNASeq data set (**b**).

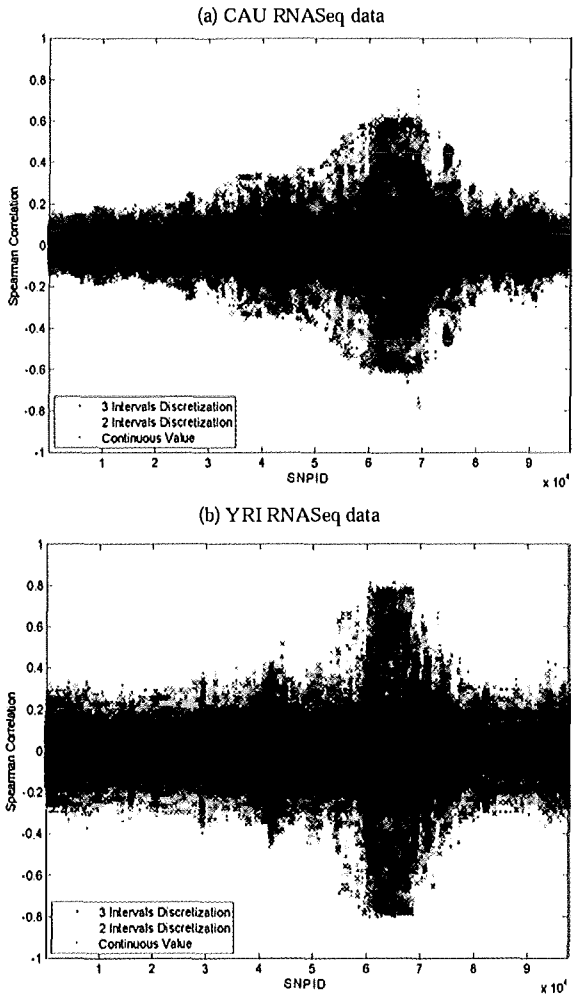


Fig. (8). Spearman correlation between 97, 484 analyzed SNPs and the continuous expression level (blue points); between the SNPs and the 3 bins discretization expression (red points); and between the SNPs and the 2 bins discretization expression (cyan points). For the CAU population, the Spearman correlation value between the continuous and the 2 bins series was 0.9319; and between the continuous and the 3 bins series was 0.9636. For this same CAU population, the averaged mean square error between the continuous and the 2 bins series was 0.0013; and between the continuous and the 3 bins series was 0.000055. For the YRI population, the Spearman correlation value between the continuous and the 2 bins series was 0.9218; and between the continuous and the 3 bins series was 0.9315. For this same YRI population, the averaged mean square error between the continuous and the 2 bins series was 0.0020; and between the continuous and the 3 bins series was 0.0018.

Table 8. The 10 SNPs with the highest Spearman correlation coefficients for the YRI-RNASeq population computed when the expression of the gene is continuous and discretized in two and three bins.

YRI-RNASeq					
Discretized 3 Bins		Discretized 2 Bins		Continuous	
Pou	Corr.	Pou	Corr.	Pou	Corr.
32458003	0.8132	32486115	-0.7741	32533567	0.8126
32533567	0.8124	32458228	0.7713	32561743	0.7973
32458007	-0.8007	32502393	-0.7682	32533813	-0.793
32486115	-0.7993	32491826	-0.7663	32537290	-0.793
32492845	-0.7971	32502507	-0.7657	32486115	-0.7927
32458228	0.7952	32502513	-0.7657	32481244	-0.7917
32561743	0.7894	32502522	0.7633	32486632	0.7908
32486639	0.7821	32492845	-0.7627	32533755	-0.7878
32533813	-0.7821	32533567	0.7587	32553705	-0.7864
32537290	-0.7815	32481244	-0.7528	32553531	-0.7864

Table 9. The 10 SNPs with the highest Spearman correlation coefficients for the CAU-RNASeq population computed when the expression of the gene is continuous and discretized in two and three bins.

CAU-RNASeq					
Discretized 3 Bins		Discretized 2 Bins		Continuous	
Pou	Corr.	Pou	Corr.	Pou	Corr.
32602872	-0.757	32602872	-0.7171	32602872	-0.7831
32602396	-0.7371	32602396	-0.7042	32602396	-0.7637
32601332	0.7195	32601332	0.6847	32601332	0.7523
32568292	-0.658	32568292	-0.6091	32568292	-0.6663
32540158	0.6251	32486683	-0.5634	32540158	0.6511
32546592	-0.6029	32489908	0.5615	32561424	-0.637
32561743	0.6005	32488193	-0.548	32561743	0.6307
32545106	-0.6005	32479606	-0.547	32489908	0.6267
32599071	0.5964	32501522	0.5411	32599071	0.6232
32534976	0.5949	32454968	-0.5392	32454968	-0.6231

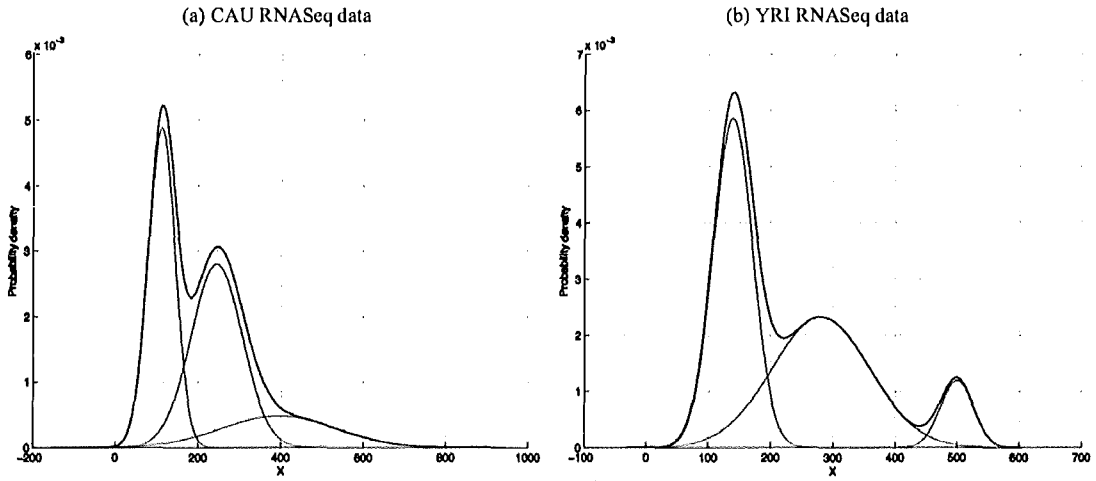


Fig. (9). Gaussian mixtures estimated by the EM algorithm. EM was run 100 times with different random starting points, the solution with best likelihood was chosen. For the CAU-RNASeq population, the mixture model defined two cut-off points: 166.00 and 361.97; and these cut-off points created 3 groups with 118, 110 and 31 individuals inside each group. For the YRI population, the mixture model defined two cut-off points: 194.39 and 451.11; these cut-off points created 3 groups with 39, 32 and 6 individuals inside each group.

Table 10. Generalization capacity of different learning machines using all SNPs as input variables in CAU-RNASeq data set.

Classification Method	Accuracy	AUC	Relat. Abs. Error
NB	56.38	0.67	72.83
C4.5	**68.71**	0.77	53.73
SVM	67.95	**0.81**	**53.53**
Regression Method	**RMSE**	**Correlation**	**Relat. Abs. Error**
SVM-Reg.	87.93	0.72	68.24
Gaussian Proc.	88.05	0.72	68.35
Lasso	**77.96**	**0.79**	**60.76**
3-NN-reg-avg	105.66	0.60	81.18
5-NN-reg-avg	100.62	0.63	76.85
3-NN-reg-wgt	105.81	0.60	81.34
5-NN-reg-wgt	100.62	0.63	76.90
3NN-reg-krn	111.80	0.57	86.78
5NN-reg-krn	113.00	0.55	84.94

Table 11. Generalization capacity of different learning machines using all SNPs as input variables in YRI-RNASeq data set.

Classification Method	Accuracy	AUC	Relat. Abs. Error
NB	62.68	0.63	64.84
C4.5	60.36	0.70	70.43
SVM	**78.21**	0.85	**37.78**
Regression Method	**RMSE**	**Correlation**	**Relat. Abs. Error**
SVM-Reg.	93.63	**0.64**	79.84
Gaussian Proc.	93.69	**0.64**	79.93
Lasso	**90.05**	0.61	82.69
3-NN-reg-avg	98.05	0.59	81.17
5-NN-reg-avg	93.31	0.62	78.08
3-NN-reg-wgt	97.93	0.59	80.98
5-NN-reg-wgt	93.19	0.62	**78.04**
3NN-reg-krn	99.99	0.62	84.50
5NN-reg-krn	93.67	0.60	78.41

3.2.3. Block-Based Approach

In this section we pursue the same analysis carried out in Section 3.1.3 for microarray data. In this case, among the classification algorithms we picked the SVM model because it discovered blocks with higher prediction capacity than C4.5 classifier. The Lasso regression model was chosen again because was one of the most competitive regressors. In Fig. (**10**), we display the results of this analysis. In Table **12** we detail the 10 blocks with the lowest relative absolute error for the classification and the regression model in the two populations. As it happened with the microarray data, classification algorithms performed better than regression algorithms because as made more accurate predictions with the key blocks associated with the RNASeq expression of DRB5. With these new data sets, we also sought whether the performance of a classification model using a set of SNPs was higher or not than the performance obtained using a single SNP. In that way, we compared AUC of the SVM classification model using all SNPs within the same block with respect to the AUC obtained using only one single SNP. We selected the SNP with the maximum AUC within the same block. In Fig. (**11**) we plot both measures for all the 345 blocks. In Table **13** we detail the 10 blocks with the highest AUC using the SVM classifier as well as the AUC obtained with the best performing single SNP among all SNPs within this block. In this case, it is curious to see how there were not the strong increments we observed with the same experiment using microarray data. In Section 4 we discuss this issue.

4. DISCUSSION

Results obtained for gene DRB5 when using single-SNP classifiers and microarrays have extensively been confirmed by common SNP-RNA expression correlation models. The levels of DR gene expression could condition the type of immune response. The high expression of DRB5 gene could increase the amount of DR receptor in the surfaces of the antigen presenting cell (APC) and as consequence increase

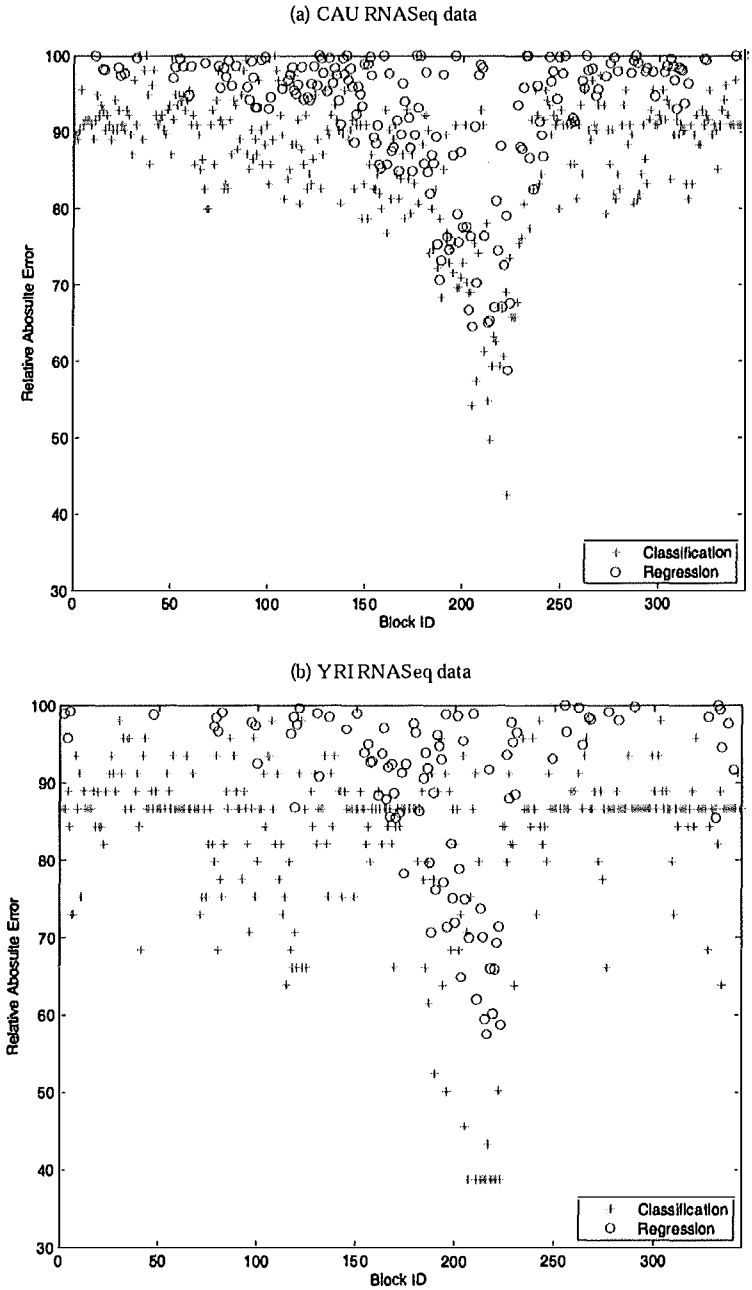


Fig. (10). Relative Absolute Error comparison between a classification model (SVM) and a regression model (Lasso). For the CAU-RNASeq population, the number of blocks where the classification model obtained a RAE < 50% was 2 and for RAE < 70% the number of blocks was 23. For the regression model in the CEU-array population, these numbers were 0 and 8 respectively. For the YRI-array population, the number of blocks where the classification model obtained a RAE < 50% was 13 and for RAE < 70% the number of blocks was 34. For the regression model in the YRI-array population, these numbers were 0 and 10 respectively.

the concentration of peptide-MHC complex and in turn affect the duration and specificity of the T cell-TCR with APC-HLA molecules interaction. The immunological synapse strength between APC and the T cell determines the fate of T cells into Th1 or Th2 types [35] favoring Th1 differentiation when a stronger TCR signal is produced.

Table 12. The 10 blocks with the lowest relative absolute error (RAE) using a classification model (SVM) learnt with all SNPs within this block. 7 blocks (in bold) out of the 10 are shared by the two populations.

CAU-RNASeq				YRI-RNASeq			
Discretized		Continuous		Discretized		Continuous	
Block ID	RAE	Block ID	RAE	Block ID	RAE	Block ID	RAE
223	42.57	223.00	58.88	**207**	38.75	216	57.47
214	49.66	205.00	64.55	**211**	38.75	223	58.70
205	54.18	213.00	65.08	**213**	38.75	215	59.44
213	54.82	214.00	65.33	**214**	38.75	219	60.14
207	57.40	203.00	66.76	**215**	38.75	211	62.00
215	59.34	216.00	67.08	216	38.75	203	64.90
219	59.34	220.00	67.20	218	38.75	220	65.93
221	60.63	224.00	67.69	**219**	38.75	218	66.00
211	61.27	207.00	70.27	220	38.75	221	69.27
217	62.56	188.00	70.68	**221**	38.75	207	69.98

Table 13. The 10 blocks with the highest AUC using a classification model (SVM). "Single SNP AUC" column displays the AUC obtained with the best performing single SNP among all SNPs within this block. 7 blocks (in bold) out of the 10 are shared by the two populations.

CAU-RNASeq			YRI-RNASeq		
Block ID	Block AUC	Single SNP AUC	Block ID	Block AUC	Single SNP AUC
223	0.771	0.797	**207**	0.789	0.788
214	0.733	0.695	**211**	0.789	0.776
205	0.709	0.612	**213**	0.789	0.797
213	0.704	0.707	**214**	0.789	0.788
207	0.690	0.686	**215**	0.789	0.787
219	0.680	0.698	216	0.789	0.789
215	0.679	0.684	218	0.789	0.749
221	0.677	0.694	**219**	0.789	0.761
211	0.670	0.703	220	0.789	0.816
217	0.662	0.654	**221**	0.789	0.772

Therefore, the combination of DRB expression levels with specific structure receptors produced by the variants in the region would determine the fate of the T-cell ant the immune response [36].

However, in this work we have built multivariate models able to outperform single-model results, and have shown how by discretizing gene expression, classification learning machines can be used as an alternative tool to regression learning, which are robust to redundancy and noisy variables, when learning and testing complex models composed of hundred or thousand input variables. Moreover, some of them such as NB or C4.5 learn white-box models that can be interpretable by human experts. In the case of binarized gene expression, as in expression data sets from microarrays, NB classifier can be also understood as a Genetic Risk Score [37], a logistic regression widely used to predict individual predisposition to have a disease, considered as a binary trait, in which the output is interpreted as the probability of having or not the disease or, in our study, of having a high/low expression of gene DRB5.

Results obtained for gene DRB5 showed that there was always a classification approach that outperformed all regression models tried. Some multivariate models used in this study show their robustness to redundant variables, an important feature that will help model replication in an independent data set. In fact, those classification or regression learning algorithms able to perform variable filtering or weighting, i.e. C4.5 among classifiers and Lasso among regression methods, showed higher performance in a cross-validation approach in the microarray data sets. By using all SNPs in the data sets most likely we are considering SNPs with no role in DRB5 expression that may introduce noise if not removed by the learning algorithms, or they may introduce redundancy if they are not causal but are in high LD with a causal SNP.

Multivariate models are very important whenever a single SNP may not completely explain the genetic effect on the expression level of a gene, either because the truly cause was not genotyped or because there is an epistatic effect among two or more causal loci. Our results conducted on microarray data showed how different blocks affected gene regulation and the multivariate models outperformed single SNP models in some of the blocks with best performance, showing again the importance of using robust multivariate

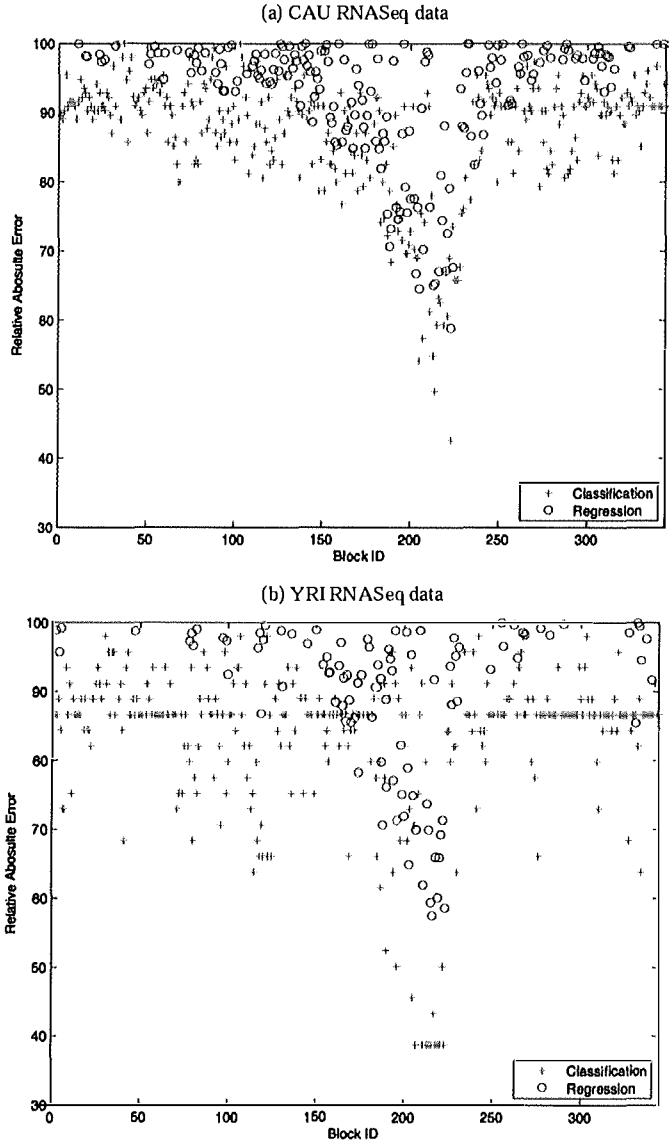


Fig. (11). AUC comparison between a classification model (SVM) using all SNPs in a block; and the maximum performance obtained using a single SNP within the same block.

models. Results from RNASeq data set show the same pattern between classifiers and regression functions, with always a classifier outperforming all the regression models. However, they deserve a deep discussion as they, in all experiments and against the supposed superiority in power of RNASeq over microarray technology, underperformed those obtained by microarray data.

The first important difference compared with microarrays was the distribution of expression values that showed a more complex pattern to be discretized. Perhaps sequencing errors explain this complex pattern, which may also explain why predictive capacity is never as high as the one reached by microarray data.

When comparing in an individual basis two-bins discretization from RNASeq with that from microarrays, by using only common individuals -i.e. data sets CEU-commonIndividuals and YRI-commonIndividuals- an interesting result arose: differences in discretization results between the two technologies (microarrays and RNASeq) depended on the population. Therefore, YRI-RNASeq could be up-biased, as there were 18 individuals with different discretized expression level (low or high) between the two technologies but all of them had expression levels always

classified as low in YRI-array and as high in YRI-RNASeq. On the contrary, almost all individuals 14 out of 18 with different expression level (low or high) between the two technologies had expression levels always classified as high in CEU-array and as low in CEU-RNASeq. If we assume that DRB5 expression from arrays was correctly discretized because its simplicity (see Fig. **1**) and the perfect classification accuracy when estimation from genotypes reached by some classifiers, we may have found an interesting situation in which the theoretically more accurate RNASeq brought worse results and the bias depended on the population, so that CEU individuals tended to have lower expression of DRB5 and YRI tended to have higher expression of DRB5. Some issues in a still very novel NGS technique have been suggested before [32]. Moreover, predictive accuracy behaved very differently between populations in whole, block-based and single models. As an example, in whole models differences between microarray and RNASeq were large in Caucasian: SVM and C4.5 had same accuracy in Caucasian and much lower than with microarray data (69.5 in CAU-RNASeq *versus* 100 and 90.65 in CEU-RNASeq for C4.5 and SVM respectively). However, when using YRI data differences were outstanding: SVM clearly outperformed C4.5 (77.92 *versus* 40.26) in RNASeq when in microarrays it was clearly worse (79.5 *versus* 99). In presence of large amounts of errors in RNA sequencing, variable selection may not be that good, as the errors show up more clearly when only a few variables are selected than whenever all variables are used, given that the whole DNA region was selected to be close to where DRB5 gene is encoded. Some possible explanation of the lack of accuracy in RNASeq data sets is that genes may have several isoforms, many of them sharing exons so that some reads cannot be unequivocally assigned to a transcript [38]. This effect may even occur between genes, as among DRB1 and DRB5 genes. In our situation, even if DRB5 has several transcripts, only one was used when mapping reads to genes. It could be that other transcripts common in CEU and performing the same molecular effect were expressed but not captured. At the same time, it is already known that some reads map exons in both DRB5 and DRB1 gene. As only one isoform was used for DRB5, it could exist an isoform common in YRI very similar to the one used in the read-mapping phase, which caused the up-bias in YRI-RNASeq.

CONCLUSION

We have shown how several classification algorithms, which are robust to redundant and noise variables, show a high predictive accuracy. Some of the classification approaches have revealed to be very helpful for biomedical researchers, as they have learned white-box models easily interpretable by human experts.

Feature selection of SNPs based on LD criterion has helped to identify SNPs that may be candidate eQTLs in the predictive models. Because of their low computational complexity to the number of input variables, we have been able to use very robust classification algorithms under different approaches with all the SNPs within the vicinity of genes DRB1 and DRB5.

These conclusions are more difficult to obtain when using regression models, as the larger complexity of a regression model compared with a classifier translated into a reduction in robustness to redundancy and therefore in generalization capacity and interpretability. However, when they were able to perform variable filtering some way, such as in Lasso regression they performed much better.

Finally, although these results have been replicated with RNASeq data, they show very different and unexpected patterns, such as a lower performance compared with microarrays perhaps because a high level of noise introduced in the sequencing process, which is also biased depending on the population. This high degree of sequencing errors may explain a better performance reached by models that did not perform variable selection. All these results, discussed above, may be revealing several open issues in a very novel NGS technology and will need deeper research.

Given these results, one of our main short-term challenges is to determine the functional link between the SNPs and the phenotype at two levels, DRB5 expression and disease implication. We plan to focus on Copy Number Variations (CNV), since there are multiples CNV in the HLA region. We will try to connect the SNPs discovered in this work with insertions or deletions of the DRB5 gene. On the other hand, ENCODE database clusters information about genome-wide regulatory regions. This information can be cross with the data obtained in the present work with the objective to determine which is the ultimate cause of the DRB5 expression levels.

ABBREVIATIONS

AUC	=	Area Under the ROC Curve
cv	=	Cross Validation
CEU	=	Samples of Utah residents with Ancestry from Northern and Western Europe Used by HapMap and 1000 Genome Projects
NGS	=	Next Generation Sequencing
RNASeq	=	RNA Sequencing Technology
SNP	=	Single Nucleotide Polymorphism
YRI	=	Samples of Yoruba from Ibadan, Nigeria Used by HapMap and 1000 Genome Projects

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

ACKNOWLEDGEMENTS

The authors were supported by the second call program for research and development of the International Excelence Campus BioTic Granada under microproject CEI-mic2013-2, and under research project CEI-IDI-2013-15, the Spanish Research Program under projects TIN2010-20900-C04-1 and TIN2013-46638-C3-2-P, the Andalusian Research Program under project P08-TIC-03717 and the European Regional Development Fund (ERDF). RA is currently supported by NIH grant 1R01NS0860832 under the CRCNS program.

REFERENCES

[1] Wallace C, Rotival M, Cooper JD, *et al.* Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. Hum Mol Genet 2012; 21(12): 2815-24.

[2] Gat Viks I, Meller R, Kupiec M, Shamir R. Understanding gene sequence variation in the context of transcription regulation in yeast. PLoS Genet 2010; 6(1): e1000800.

[3] Stranger BE, Montgomery SB, Dimas AS, *et al.* Patterns of Cis Regulatory Variation in Diverse Human Populations. PLoS Genet 2012; 8(4): e1002639.

[4] Liu H, Li J, Wong L. A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns. Genome Inform 2002; 13: 51-60.

[5] Ross ME, Zhou X, Song G, *et al.* Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. Blood 2003; 102(8): 2951-9.

[6] Wanga Y, Tetko IV, Hall MA, *et al.* Gene selection from microarray data for cancer classification, a machine learning approach. Comput Biol Chem 2004; 29(1): 37-46.

[7] Abad-Grau M, Medina-Medina N, Montes-Soldado R, Matesanz F, Bafna V. Sam ple Reproducibility of Genetic Association using Different Multimarker TDTs in Genome wide Association Studies: Characterization and a New Approach. PLoS ONE 2012; 7(2): 29613.

[8] Fogdell A, Hillert J, Sachs C, Olerup O. The multiple sclerosis and narcolepsy associated HLA class II haplotype includes the DRB5*0101 allele. Tissue Antigens 1995; 46: 333-6.

[9] VincentR, PPL, GongoraC, Papal, *et al.* JCJ. Quantitative analysis of the expression of the HLA-DRB genes at the transcriptional level by competitive polymerase chain reaction. J Immunol 1996; 156: 603-10.

[10] Schadt EE, Molony C, Chudin E, *et al.* XY. Mapping the genetic architecture of gene expression in human liver. PLoS Biology 2008; 2008: 6: e107.

[11] Dixon AL, Liang L, Moffatt MF, Chen W, *et al.* SH. A genome-wide association study of global gene expression. Nat Genet 2007; 39: 1202-7.

[12] JA JQ, Huh J, M MB, *et al.* Myelin basic protein-specific TCR/HLA-DRB5*01: 01 transgenic mice support the etiologic role of DRB5*01: 01 in multiple sclerosis. J Immuno 2012; 189(6): 2897-908.

[13] Pearl J. Probabilistic Reasoning with Intelligent Systems. San Mateo: Morgan & Kaufman; 1988.

[14] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data *via* the EM algorithm. J Roy Stat Soc B Met 1977; 39 (1): 1-38.

[15] EverittBS. A finite mixture model for the clustering of mixed mode data. Stat Probabil Lett 1988; 6(5): 305-9.

[16] Duda RO, Hart PE. Pattern Classification and Scene Analysis. New York: John Wiley Sons; 1973.

[17] Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett 2006; 27(8): 861-74.

[18] Quinlan JR. C4.5: Programs for Machine Learning. San Francisco: Morgan Kaufmann; 1993.

[19] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update. In: SIGKDD Explorations 2009; 11: 1.

[20] Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Trans Intel Syst Technol 2011; 2: 1-27. Software available at http: //www.csie.ntu.edu.tw/_cjlin/libsvm.

[21] Shevade SK, Keerthi SS, Bhattacharyya C, Murthy KRK. Improvements to the SMO algorithm for SVM regression. IEEE Trans Neural Networks Learning Syst 2000; 11(5): 1188-93.

[22] Williams C, Rasmussen C. Gaussian Processes for Regression. In: Advances in Neural Information Processing Systems 1996; 8: 514-20.

[23] Tibshirani R. Regression shrinkage and selection *via* the Lasso. J Roy Stat Soc B 1996; 67(1): 91-108.

[24] Cover T. Estimation by the nearest neighbor rule. IEEE Transactions on Information Theory 1968; 14(1): 50-5.

[25] Hamming RW. Error detecting and error correcting codes. Bell System Technical Journal 1950; 29(2): 147-60.

[26] Gabriel S, Schaffner S, Nguyen H, *et al.* The Structure of Haplotype Blocks in the Human Genome. Science 2002; 296: 2225-9.

[27] Hap Map Consortium TI. The International Hap Map Project. Nature 2003; 426: 789-796.

[28] Hap Map Consortium TI. Integrating common and rare genetic variation in diverse human populations. Nature 2010; 467(7311): 52-58. Available from: http: //dx.doi.org/10.1038/nature09298.

[29] Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome wide association studies. PLoS Genetics 2009; 5(6): e1000529.

[30] Dermitzakis E. E-GEUV-1 - RNA-sequencing of 465 lymphoblastoid cell lines from the 1000 Genomes; 2013. http: //www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-1/.

[31] Consortium GP, Abecasis G, Altshuler D, *et al.* A map of human genome variation from population scale sequencing. Nature 2010; 467(7319): 1061-1073. Available from: http: //dx.doi.org/10.1038/nature09534.

[32] Trapnell C, Roberts A, Goff L, *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with Top Hat and Cufflinks. Nat Protocols 2012; 7(3): 562-578. Available from: http: //dx.doi. org/10.1038/nprot.2012.016.

[33] Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: Proceedings of the 15th International Joint Conference on Artificial Intelligence 1995; 114-119.

[34] Fraley C, Raftery AE. How many clusters? Which clustering method? Answers *via* model-based cluster analysis. The Comp J 1998; 41(8): 578-588.

[35] Corse E, RA RAG, Allison JP. Strength of TCR-peptide/MHC interactions and *in vivo* T cell responses. J Immuno 2011; 186: 5039-45.

[36] Alcina A, Abad-Grau MDM, Fedetz M, *et al.* Multiple sclerosis risk variant HLA-DRB1*1501 associates with high expression of DRB1 gene in different human populations. PLoS One 2012; 7(1): e29819.

[37] Sebastiani P, Solovieff N, Sun JX. Nave Bayesian classifier and genetic risk score for genetic risk prediction of a categorical trait: not so different after all. Front Genet 2012; 3(26): doi: 10.3389/fgene.2012.00026.

[38] Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods 2011; 8(6): 469-477. Available from: http: //dx.doi.org/10.1038/nmeth. 1613.