Genes Genet. Syst. (2015) **90**, p. 343–356

Gene discovery for facioscapulohumeral muscular dystrophy by machine learning techniques

Félix F. González-Navarro[1*], Lluís A. Belanche-Muñoz[2], María G. Gámez-Moreno[1], Brenda L. Flores-Ríos[1], Jorge E. Ibarra-Esquer[3] and Gabriel A. López-Morteo[1]

[1]*Instituto de Ingeniería, Universidad Autónoma de Baja California, Mexicali 21280, México*
[2]*Dept. de Ciències de la Computació, Universitat Politècnica de Catalunya - Barcelona Tech, Barcelona 08034, Spain*
[3]*Facultad de Ingeniería Universidad Autónoma de Baja California, Mexicali 21280, México*

(Received 16 February 2015, accepted 8 August 2015; J-STAGE Advance published date: 10 March, 2016)

Facioscapulohumeral muscular dystrophy (FSHD) is a neuromuscular disorder that shows a preference for the facial, shoulder and upper arm muscles. FSHD affects about one in 20-400,000 people, and no effective therapeutic strategies are known to halt disease progression or reverse muscle weakness or atrophy. Many genes may be incorrectly regulated in affected muscle tissue, but the mechanisms responsible for the progressive muscle weakness remain largely unknown. Although machine learning (ML) has made significant inroads in biomedical disciplines such as cancer research, no reports have yet addressed FSHD analysis using ML techniques. This study explores a specific FSHD data set from a ML perspective. We report results showing a very promising small group of genes that clearly separates FSHD samples from healthy samples. In addition to numerical prediction figures, we show data visualizations and biological evidence illustrating the potential usefulness of these results.

Key words: Facioscapulohumeral muscular dystrophy, gene discovery, feature selection, machine learning, protein-protein association networks

INTRODUCTION

In recent years, machine learning (ML) has made significant inroads in the fields of bioinformatics and biomedicine; see, for example, Schölkopf et al. (2004). Specifically, in cancer research a variety of ML algorithms have been developed for tumor prediction by associating gene expression patterns with clinical outcomes for patients with tumors (Lukas et al., 2004). The majority of this research has focused on building accurate classification models from reduced sets of features. Some of the analyses also aimed to gain an understanding of the differences between normal and malignant cells and to identify genes that are differentially regulated during cancer development.

Facioscapulohumeral muscular dystrophy (FSHD) is an autosomal dominant neuromuscular disorder that shows a preference for the facial, shoulder and upper arm muscles, and is the third most common inherited muscular dystrophy (Flanigan, 2004; Tawil, 2008). Its incidence varies with geographic location and probably within different racial groups, recent estimates being one in about 400,000 to one in 20,000 (MDC, 2012). Patients usually become symptomatic in their twenties or later (Tawil and Maarel, 2006). The most common FSHD symptoms are progressive weakening of muscles and atrophy of the face, shoulder, upper arm and shoulder girdle, and lower limbs. These are usually accompanied by an inability to flex the foot upward, foot weakness, and an onset of right/left asymmetry (Tawil et al., 1998; Maarel et al., 2007). FSHD is considered a relatively benign dystrophy, despite the fact that around 20% of patients are eventually confined to a wheelchair (Tawil and Maarel, 2006), and an estimated 1% ultimately require breathing assistance (Wahl, 2007). Although no effective therapeutic strategies are known to either halt progression or reverse muscle weakness (or atrophy) (Rose and Tawil, 2004), a number of actions can provide symptomatic and functional improvement in many patients; the use of assistive devices such as braces, standing frames or walkers, and physical therapies such as exercises in water, helped by psychological support and speech therapy, may help to alleviate especially difficult life conditions.

It is believed that FSHD is caused by deletion of a subset of D4Z4 macrosatellite repeat units in the subtelomere of chromosome 4q (Maarel et al., 2011), but this

Edited by Ali Masoudi-Nejad

* Corresponding author. E-mail: fernando.gonzalez@uabc.edu.mx

DOI: http://doi.org/10.1266/ggs.15-00017

modification needs to occur on a specific chromosomal background to cause FSHD. More than 95% of patients with clinical FSHD have an associated D4Z4 deletion on the 4q35 chromosome (Wahl, 2007). However, a small number of kindreds with typical FSHD do not display this deletion. Recent advances involve the *DUX4* gene, a retrogene sequence within D4Z4 that encodes a double homeodomain protein whose exact function is not known. FSHD is the first example of a human disease associated with the inefficient repression of a retrogene in a macrosatellite repeat array (Maarel et al., 2011). Although the mechanisms responsible for progressive muscle weakness remain unknown, the study of this gene may offer a therapeutic route (Maarel et al., 2011).

The simultaneous monitoring of expression levels for thousands of genes may allow the study of the effects of certain treatments, diseases and developmental stages on gene expression. In particular, microarray-based gene expression analysis based on statistical or ML methods can be used to identify differentially expressed genes (which act as features, to use ML terminology) by comparing expression in affected and healthy cells or tissues. A gene expression data set typically consists of dozens of observations and up to tens of thousands of genes. Predictive model construction and validation in this situation is difficult and prone to yield unreliable readings. As a result, dimensionality reduction and in particular feature subset selection (FSS) techniques may be very useful, as a way to reduce the problem complexity and facilitate expert medical diagnosis. In a practical medical context, the interpretability of the obtained solutions is also of paramount importance, limiting the applicability of methods such as Principal Components Analysis (PCA), which involves weighted combinations of many genes instead of individual genes. Moreover, data visualization in a low-dimensional representation space may become extremely important, as it helps doctors to gain insights into this medical area. Such contributions are scarce for the case of FSHD-associated gene expression data, probably due to a relative research bias toward more common diseases. This scenario is aggravated by the absence of publicly available scientific data, outside purely medical domains, although the situation is slowly changing. In contrast, there is now a vast body of available microarray gene expression data sets focused on cancer diseases.

The development of simple predictive models that can distinguish between healthy and FSHD samples with minimal error recognition rate is thus a clear research goal. This study explores a specific FSHD data set from a ML perspective. First, a dimensionality reduction stage is carried out. A feature selection algorithm is used as the main engine to select genes promoting the highest possible classification capacity to distinguish between healthy and FSHD samples. The combination of feature selection and classification aims at obtaining simple models (in terms of low numbers of genes) capable of good generalization. A prior selection stage is carried out, using a well-known statistical test, to filter out genes with negligible discrimination ability. To avoid selection biases, the full selection process is embedded into an outer Monte Carlo validation process.

We report experimental results supporting the practical advantage of combining robust feature selection and classification in the analyzed FSHD data set. The described method was able to unveil a consistent small group of genes that yields high mean test set accuracies. The practical utility of these results is reinforced by low-dimensional visualizations and supported by current biological evidence.

MATERIALS AND METHODS

Data set The database used in this study was obtained from the EMBL-EBI repository of the European Bioinformatics Institute (EMBL-EBI, 2014). Specifically, *Experiment E-GEOD-3307* uses the Affymetrix GeneChip Human Genome HG-133A and HG-U133B designs to analyze a group of muscle diseases for comparative gene expression profiling purposes. A total of 121 muscle samples of 11 muscle pathologies (plus several healthy samples) constitute the data: acute quadriplegic myopathy, juvenile dermatomyositis, amyotrophic lateral sclerosis, spastic paraplegia, fascioscapulohumeral muscular dystrophy, Emery-Dreifuss muscular dystrophy, Becker muscular dystrophy, Duchenne muscular dystrophy, calpain 3, dysferlin, and the FKRP using U133A and U133B array design. These are diseases with an extremely low incidence rate in the general population. Facioscapulohumeral Muscular Dystrophy (FSHD, HG-133A version), the targeted group in this work, consists of 14 people showing FSHD (hereafter referred to as FSHD samples or cases) and 18 people not showing FSHD (healthy cases), described by p = 22,283 genes or features.

Feature selection algorithm The first action taken was to filter out those genes that do not possess a minimum discriminative capacity. In particular, we use Welch's t-test (Pan, 2002) computed as follows:

$$t_j = \frac{\overline{x}_j^+ - \overline{x}_j^-}{\sqrt{\dfrac{s_j^{2+}}{N^+} + \dfrac{s_j^{2-}}{N^-}}}, \qquad 1 \leq j \leq p \tag{1}$$

where the j subscript runs through all genes $G = \{g_1, \ldots, g_p\}$; x_j^- and s_j^2 denote sample mean and variance of a gene, and N is the sample size; the symbols + and − indicate positive (healthy) and negative (FSHD) cases, respectively. For each g_j, a large and positive (resp. negative) t_j score indicates high expression in favor of the positive (resp. negative) class. Hence, absolute values $|t_j|$

are considered and sorted in descending order to identify the top k genes; k = 100 in our study.

A simple yet very effective forward-backward FSS algorithm is then fed with the k genes previously selected. This algorithm follows the wrapper idea, i.e., the feature selection algorithm uses a learner as a subroutine in the search for good subsets (John et al., 1994). In this general setting, when features are added to or removed from the current subset, the algorithm resorts to some performance measure; for example, in classification problems, it may be the resampled recognition rate.

Wrappers are often criticized because they are computationally very expensive (Guyon and Elisseeff, 2003). Moreover, feature selection is badly affected by small sample sizes, producing overly optimistic results and introducing an excess of variance in the readings. This is aggravated in the presence of very sophisticated search algorithms (Reunanen, 2003). On the other hand, greedy search strategies seem to be particularly computationally advantageous and may alleviate the problem of overfitting (Guyon and Elisseeff, 2003). Nevertheless, traditional pure forward selection and backward elimination search algorithms are ill-advised in that they cannot rectify their decisions and may ultimately deliver poor solutions in terms of both quality and size.

Therefore, a forward-backward search is developed, looking for an improvement in performance of the chosen performance measure. The algorithm is presented as **Algorithm 1**. Given a performance measure L to be maximized (in this case, the resampled performance evaluation of a classifier), the algorithm searches the space of subsets by adding/removing features in an interleaved hill-climbing fashion. Specifically, in every iteration of the outer loop, one feature is added to the current best solution BEST, as long as this step improves on current performance L^{cur}. A variable number of feature removal steps is then carried out as long as the same condition of improved performance is met, a scheme oriented to favor solutions with low numbers of features. The outer iteration also ends when no further improvement is observed. The strategy generalizes pure forward or pure backward search, and bears some resemblance to floating search methods (Pudil et al., 1994); however, it has a far lower computational cost given that discarded features are not considered again for another inclusion round. Note also that, unlike floating methods, current subset performance is not compared specifically against the best performance achieved for the same size of the current subset. It should be mentioned that the algorithm itself needs no parameter specification, although the chosen performance measure L may.

Among the possible choices for two-class classifiers, we selected the linear discriminant classifier and the linear support vector machine. These methods are attractive because they are fast, because their limited complexity may be a solid guard against overfitting the data, and because they need no parameter tuning (except the cost in the SVM; see section 2.4). Moreover, the number of coefficients they estimate does not grow with dimension, which is very important in high-dimensional situations like the present one. Finally, their linear character allows the use of the obtained weights as a measure of the importance of the respective genes.

Algorithm 1 Forward-Backward gene selection (FBGS)

```
Input: $G = \{g_1, \ldots, g_p\}$: gene set;
       $C$: Class feature (Healthy, FSHD)
       $L: 2^G \to \mathbb{R}$: performance measure, to be maximized
BEST ← $\operatorname{argmax}_{g_i \in G} L(\{g_i\})$
$L^{cur} \leftarrow L(\{\text{BEST}\})$
$G \leftarrow G \setminus \{\text{BEST}\}$
repeat
    ***Begin Forward Stage***
    $g^{new} \leftarrow \operatorname{argmax}_{g_i \in G} L(\text{BEST} \cup \{g_i\})$
    $L^{new} \leftarrow L(\text{BEST} \cup \{g^{new}\})$
    if $L^{new} > L^{cur}$ then
        BEST ← BEST $\cup \{g^{new}\}$
        $L^{cur} \leftarrow L^{new}$
        $G \leftarrow G \setminus \{g^{new}\}$
    end if
    ***End Forward Stage***
    ***Begin Backward Stage***
    repeat
        $g^{new} \leftarrow \operatorname{argmax}_{g_i \in \text{BEST}} L(\text{BEST} \setminus \{g_i\})$
        $L^{new} \leftarrow L(\text{BEST} \setminus \{g^{new}\})$
        if $L^{new} \geq L^{cur}$ then
            BEST ← BEST $\setminus \{g^{new}\}$
            $L^{cur} \leftarrow L^{new}$
        end if
    until BEST does not change
    ***End Backward Stage***
until BEST does not change
Output: BEST : Optimized feature subset
```

Linear discriminant analysis Linear discriminant analysis or LDC (Duda et al., 2001) is a widely used parametric method which assumes that the class distributions are multivariate Gaussians. In LDC, all c classes are assumed to have the same covariance matrix. The QDC quadratic version does not make such an assumption; however, the number of parameters to be estimated from the data available for each class is much higher, entailing lower statistical significance. In both methods, classification is achieved by assigning an example to that class ω_k for which the posterior probability $P\ (\omega_k \,|\, \boldsymbol{x})$ is greater or, equivalently, for which ln $\{P\ (w_k)p(\boldsymbol{x} \,|\, w_k)\}$ is greater. In LDC, we assume that all class-conditional distributions $p(\boldsymbol{x} \,|\, w_k)$ have the same covariance matrix Σ, to obtain a linear discriminant function $d_k\ (\cdot)$ for class w_k expressed as:

$$d_k(\boldsymbol{x}) = \ln P(w_k) + \boldsymbol{\mu}_{\boldsymbol{k}}^{\boldsymbol{T}} \Sigma^{-1} \boldsymbol{x} - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k, \quad 1 \le k \le c$$

In practice, only an i.i.d. data sample S is available. When means, covariances and priors for every class are not available, maximum-likelihood estimates on S can be used, although in this case the Bayesian optimality properties are no longer valid. Letting $S_k \subset S$ be the subset of observations known to belong to class w_k, unbiased estimates for the class priors and vector means $\hat{\mu}_k$ can be obtained in the usual way, whereas a pooled covariance matrix is used (Ripley, 1996):

$$\Sigma \approx \hat{\Sigma}_{pooled} = \frac{1}{|S| - c} \sum_{k=1}^{c} \sum_{\boldsymbol{x} \in S_k} (\boldsymbol{x} - \hat{\mu}_k)(\boldsymbol{x} - \hat{\mu}_k)^T$$

Linear support vector machines The support vector machine (SVM) is a machine learning method solidly based on statistical learning theory (Vapnik, 1998). Intuitively, given a set of examples labeled into one of two classes, the linear SVM finds their optimal linear separation: this is the hyperplane that maximizes the minimum orthogonal distance to a point of either class (this distance is called the margin of the separation).

Consider an i.i.d. data sample $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ of training patterns (in p dimensions), labeled in $c = 2$ classes w_1, w_2 by $z_1, \ldots, z_N$, with $z_i = +1$ if $\boldsymbol{x}_i \in w_1$ and $z_i = -1$ if $\boldsymbol{x}_i \in w_2$. The optimal separating hyperplane can be found as the solution of the 1-norm quadratic programming (QP) problem:

$$\min_{w,\xi} \frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{N} \xi_i$$

$$s.t. \quad z_i\left(\langle \boldsymbol{w}, x_i \rangle + b\right) \ge 1 - \xi_i, \quad i = 1, \ldots, N$$

The solution to this optimization problem corresponds to the saddle point of its associated Lagrangian:

$$\frac{\|\boldsymbol{w}\|^2}{2} - \sum_{i=1}^{N} \alpha_i \left(z_i\left(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b\right) - 1 + \xi_i\right) + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \mu_i \xi_i$$

where α_i, $\mu_i \ge 0$ for $i = 1, ..., N$, and C is called the cost or complexity parameter, which allows the end-user to control the trade-off between margin size and error. Once this QP problem is solved, the solution vector $\boldsymbol{w}^*$ can be expressed as a linear expansion over the support vectors:

$$\boldsymbol{w}^* = \sum_{i=1}^{N} \alpha_i^* z_i \boldsymbol{x}_i \qquad (2)$$

the support vectors being those $\boldsymbol{x}_i \in S$ for which $\alpha_i^* > 0$.

Experimental setup As mentioned above, the performance measure L to be maximized in **Algorithm 1** is the accuracy rate of a classifier. Due to the low number of observations and the resampling, ties among the performance measure can happen easily. How these ties are broken is non-trivial and should be addressed specifically and explicitly (Zhou and Mao, 2006). In the literature, univariate methods such as entropy-based measures (Bell and Wang, 2000; Furlanello et al., 2003), the Fisher F-test or some other statistical test may have been preferred for their simplicity (Liu and Motoda, 1998; Liu et al., 2002). Instead, a multivariate feature ranking method seems more adequate to measure the relevance of a group of tied genes. In this work, tie-breaking strategies are developed taking advantage of the used performance function L in the main selection process. Specifically, for LDC the Fisher discriminant ratio F_s, included in **Algorithm 2**, can be computed by projecting the data using only the current subset BEST plus a tied gene g_j onto Fisher's linear discriminant[1] and then computing the optimal separability Fisher ratio in the projected space (Duda et al., 2001; Sotoca et al., 2005) as:

$$F_S(G_j) = \frac{\left(\boldsymbol{w}^T m^+ - \boldsymbol{w}^T m^-\right)^2}{\boldsymbol{w}^T S_w \boldsymbol{w}} \ge 0$$

where G_j = BEST $\cup$ $\{g_j\}$ and S_W, $\boldsymbol{m}$ stand for the mean and within-class scatter matrix of the data using the genes in G_j only. It can be verified that $F_s(G_j)$ is maximized when $\boldsymbol{w} = S^{-1}(\boldsymbol{m}^+ - \boldsymbol{m}^-)$ (Fukunaga, 1990). Among the tied genes g_j, the one maximizing $F_s(G_j)$ will be selected; this tie-breaking procedure is incorporated into the main FSS algorithm and used every time an evaluation of the performance measure may incur one or more ties (lines 2, 7 and 17 in **Algorithm 1**).

For the SVM, the tie-breaking method is given by the weight vector of the optimal separability maximum- margin solution as given by eq. (2) (Hamel, 2009). Indeed,

[1] We speak of the Fisher linear discriminant, since in the present case there is only one, provided there are two classes (healthy, FSHD).

the numbers $(w^*)^2$ in eq. (2) have been used as a surrogate for the relevance of the j-th gene since the pioneering work of Guyon et al. (2002). Notice, however, that our approach is different in that predictive performance is the main criterion for optimization; only in case of ties is the magnitude of the SVM weight vector being used. This is because the relation between this magnitude and final performance is indirect. Two linear SVMs are used, one with $C = 1$ and another with $C = 20$ (see section 2.4).

Algorithm 2 Tie-breaking criterion when L is LDC

```
Input: $G_j$ = BEST $\cup$ $\{g_j\}$: gene subset
       $C$: Class feature
$S_W \leftarrow$ WithinClassScatterMatrix($G_j$, $C$)
$\boldsymbol{w} \leftarrow S^{-1}(\boldsymbol{m}^+ - \boldsymbol{m}^-)$
$F_s(G_j) \leftarrow (\boldsymbol{w}^T\boldsymbol{m}^+ - \boldsymbol{w}^T\boldsymbol{m}^-)^2/(\boldsymbol{w}^T S_W \boldsymbol{w})$
Output: : Fisher ratio for $G_j$
```

Model assessment and selection Feature selection can often be considered part of model selection and become an important step, especially when the number of features clearly surpasses the number of observations. Performing model selection in the joint space of features and parameters in this situation can be a delicate task that entails a very high risk of overfitting. Many authors often perform variable selection only once using part or all of the available data, i.e., through a training set, delivering a final subset. By definition, a feature selection process must invariably render more than one subset, as a consequence of the random partitions in the validation process, among other factors. How to derive a single solution from a group of solutions (i.e., subsets) is a poorly addressed computational problem. Stability analysis for the outcome of feature selection is an incipient field, and there is no consensus yet on how to derive a single solution (Kalousis et al., 2007).

To avoid selection biases, the full selection process is embedded into an outer Monte Carlo resampling loop, in which the data set is repeatedly split into training and test sets, presented in **Algorithm 3**. The modeling process (both FSS and classification) is performed in each training part; the best model is then evaluated in the corresponding test parts, leading to a more robust estimation (Boulesteix, 2007).

A further advantage is the obtained support for those genes which consistently appear in selected subsets in every outer run. In this paper the splitting proportion was $P = 2/3$ and the process was repeated $B = 100$ times; note that the FBGS method refers to **Algorithm 1**.

Algorithm 3 Monte Carlo resampling for feature selection

```
Input: $D$: Data set
       $B$: No. of Montecarlo runs
       $P$ : Training set proportion
for $i$ = 1 to $B$ do
    $trndata \leftarrow$ RandomPartition($D$, $P$)
    $tstdata \leftarrow D \setminus trndata$
    BEST$_i \leftarrow$ FBGS($trndata$)
    $L \leftarrow$ TrainClassifier($trndata$(BEST$_i$))
    Accuracy$_i \leftarrow$ TestClassifier($L$, $tstdata$)
end for
Output: $\{$BEST$_i\}$, $\{$Accuracy$_i\}$ distributions
```

Comparison to other methods The proposed method is compared to a filter strategy, using Welch's t-test as suggested in Pan (2002), and to a well-known information-theoretic method, the minimum redundancy-maximum relevance (mRmR) criterion (Ding and Peng, 2005). For the former, eq. (1) is used to select the top ten genes. The mRmR measure defines both relevance and redundancy of genes using mutual information. Let G be a subset of features (e.g., genes) to be evaluated. The 'minimum-redundancy' condition is given by:

$$\frac{1}{|G|^2}\sum_{g_i \in G}\sum_{g_j \in G} MI\left(g_i; g_j\right) \tag{3}$$

where g_i and g_j represent two genes. This expression tries to select all genes that are not correlated with each other, eliminating unnecessary genes whose information could be explained by others. On the other hand, to measure the discriminative power of genes with respect to the class variable C ('maximum-relevance'), the following expression is used:

$$\frac{1}{|G|}\sum_{g_i \in G} MI\left(g_i; C\right) \tag{4}$$

Genes (features) are incrementally selected in order to optimize both (3) and (4) at the same time; in particular, given a set of already selected genes G_k, one needs to select $g \in G$ ¥ G_k as

$$g = arg \max_{g' \in G \setminus G_k}\left\{MI\left(g'; C\right) - \frac{1}{|G_k|}\sum_{\hat{g} \in G_k} MI\left(g'; \hat{g}\right)\right\} \tag{5}$$

These two methods are executed in the same conditions as **Algorithm 1** with respect to the resampling partitions and the three classifiers used.

RESULTS AND DISCUSSION

Experimental results Figure 1 shows the frequency distribution of the top consistently selected genes in the whole repeated resampling process, for each classifier. It can be observed that four genes participate in a signifi-

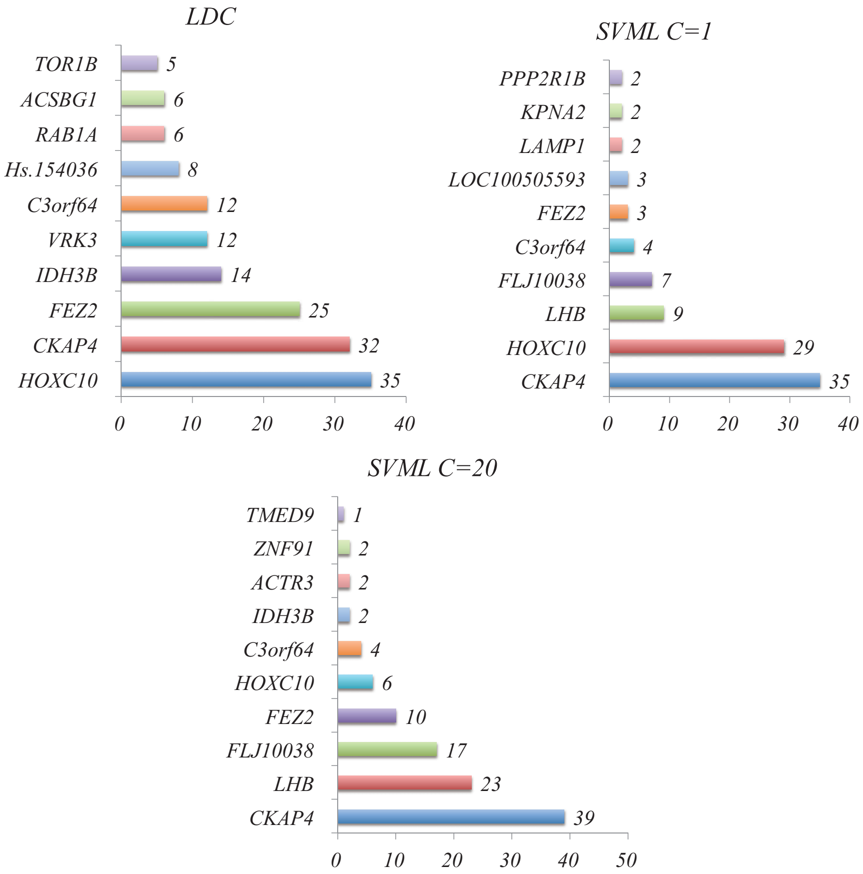

Fig. 1. Top selected genes by classifier. Top left: LDC classifier; top right: SVM with $C = 1$; bottom: SVM with $C = 20$.

Table 1. Top four most frequently appearing genes. Gene information is from GeneCards (2012)

HG U133 array Probe set ID	Gene	Name
218959_*at*	*HOXC10*	homeobox C10
200999_*s_at*	*CKAP4*	cytoskeleton-associated protein 4
215000_*s_at*	*FEZ2*	fasciculation and elongation protein zeta 2 (zygin II)
214471_*x_at*	*LHB*	luteinizing hormone beta polypeptide

cant fraction of the final subsets: for example, using LDC, the *HOXC10* gene belongs to 35% of the selections. Table 1 summarizes these findings and gives details of the genes.

Tables 2, 3 and 4 show test set accuracy distributional statistics for the three FSS methods. Focusing first on the results achieved by the proposed method (**Algorithm 1**) in Table 2, it is seen that, on average, the three classifiers perform similarly: around 91% mean accuracy and 92% median accuracy. However, in the case of the two SVMs, the standard errors are nearly four times smaller, indicating a greater stability and confidence around their means. Note also that in all cases the distributions have a slightly negative skew, since the medians are larger than the means.

Table 2. Test set accuracy statistics using **Algorithm 1**. Lower and upper limits are computed with 99% confidence

Accuracy	LDC	SVM (C = 1)	SVM (C = 20)
Median	91.33	92.00	92.02
Mean	90.60	91.86	91.87
Std. error	0.15	0.04	0.04
C.I. (99%)	(90.20,92.00)	(91.75,91.96)	(91.76,91.98)

The corresponding results using Welch's t-test are shown in Table 3. On average, and for all three classifiers, performance is 2% worse, both in mean and median, and the standard errors are much larger, showing a more unstable result. The results using the mRmR criterion are shown in Table 4. Although the overall results are, as one would expect, better than those using a univariate test, they are still worse than those obtained by the FBGS method, in terms of both average prediction error and stability.

To better compare average performances, we perform a ROC analysis, a technique widely applied in many fields of medical research and clinical practice. A ROC curve displays the relationship between the proportion of true positives ('sensitivity') and false positives ('1-specificity') classifications, resulting from each possible decision threshold in a two-class classification task (Parodi et al., 2003). Figure 2 depicts the ROC curves for all three methods overlayed. It is seen that the curve corresponding to the four genes model shows better sensitivities (only marginally, but consistently) for all possible false positive rates. The AUC values are as follows: four genes model (AUC = 0.9583), Welch's ten genes (AUC = 0.9444), and mRmR (AUC = 0.9405).

Figure 3 depicts a dendrogram of cases and standardized gene expression levels of the top four genes (see

Table 3. Test set accuracy statistics using the top ten genes according to Welch's t-test (eqn. (1)). Lower and upper limits are computed with 99% confidence

Accuracy	LDC	SVM (C = 1)	SVM (C = 20)
Median	88.95	90.19	89.67
Mean	88.61	89.84	89.72
Std. error	0.29	0.22	0.23
C.I. (99%)	(87.86,89.36)	(89.28,90.40)	(89.13,91.31)

Table 4. Test set accuracy statistics using the mRmR criterion (eqn. (5)). Lower and upper limits are computed with 99% confidence

Accuracy	LDC	SVM (C = 1)	SVM (C = 20)
Median	88.86	91.33	91.33
Mean	89.04	90.97	91.02
Std. error	0.21	0.12	0.12
C.I. (99%)	(88.49,89.59)	(90.66,91.29)	(90.72,91.32)

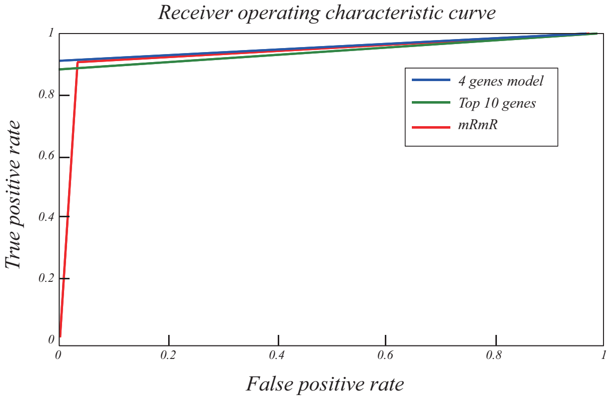

Fig. 2. ROC curves for all three methods.

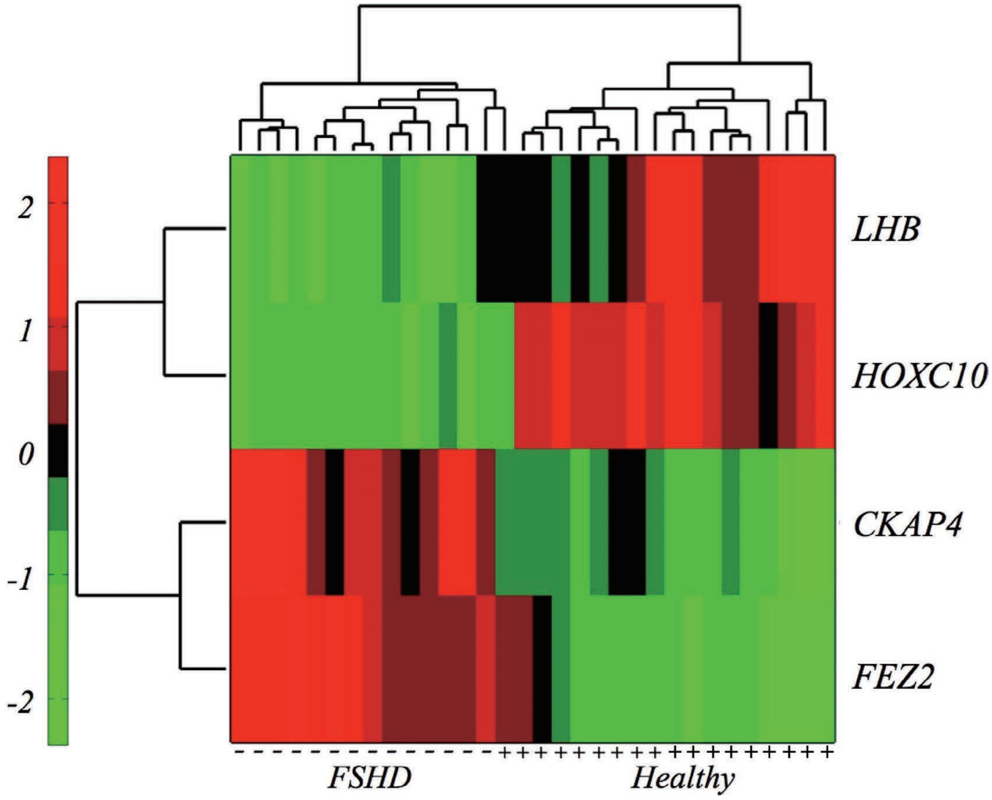

Fig. 3. Dendrogram of cases and standardized gene expression levels of the top four genes. The symbols + and indicate positive (healthy) and negative (FSHD) cases, respectively.

Table 1). The symbols + and – indicate positive (healthy) and negative (FSHD) cases, respectively. The *CKAP4* and *FEZ2* genes clearly show an up-regulation in most of the FSHD cases, while *LHB* and *HOXC10* show a clear down-regulation in expression. It is also seen that, considering only these four genes, two well-defined clusters or branches of cases are identified. Remarkably, this result matches the class labels of the data set.

Scatter plots of the top four genes are given in Fig. 4. The most frequently selected genes define well-separated clusters of FSHD and healthy samples. In order to have a graphical representation of the four genes together, we use a visualization method based on the decomposition of the scatter matrix with the property of maximizing the separation between the projections of data. Being a linear method, it is easier to use in real scenarios that might require an intuitive representation of results (Lisboa et al., 2008). The result of such visu-

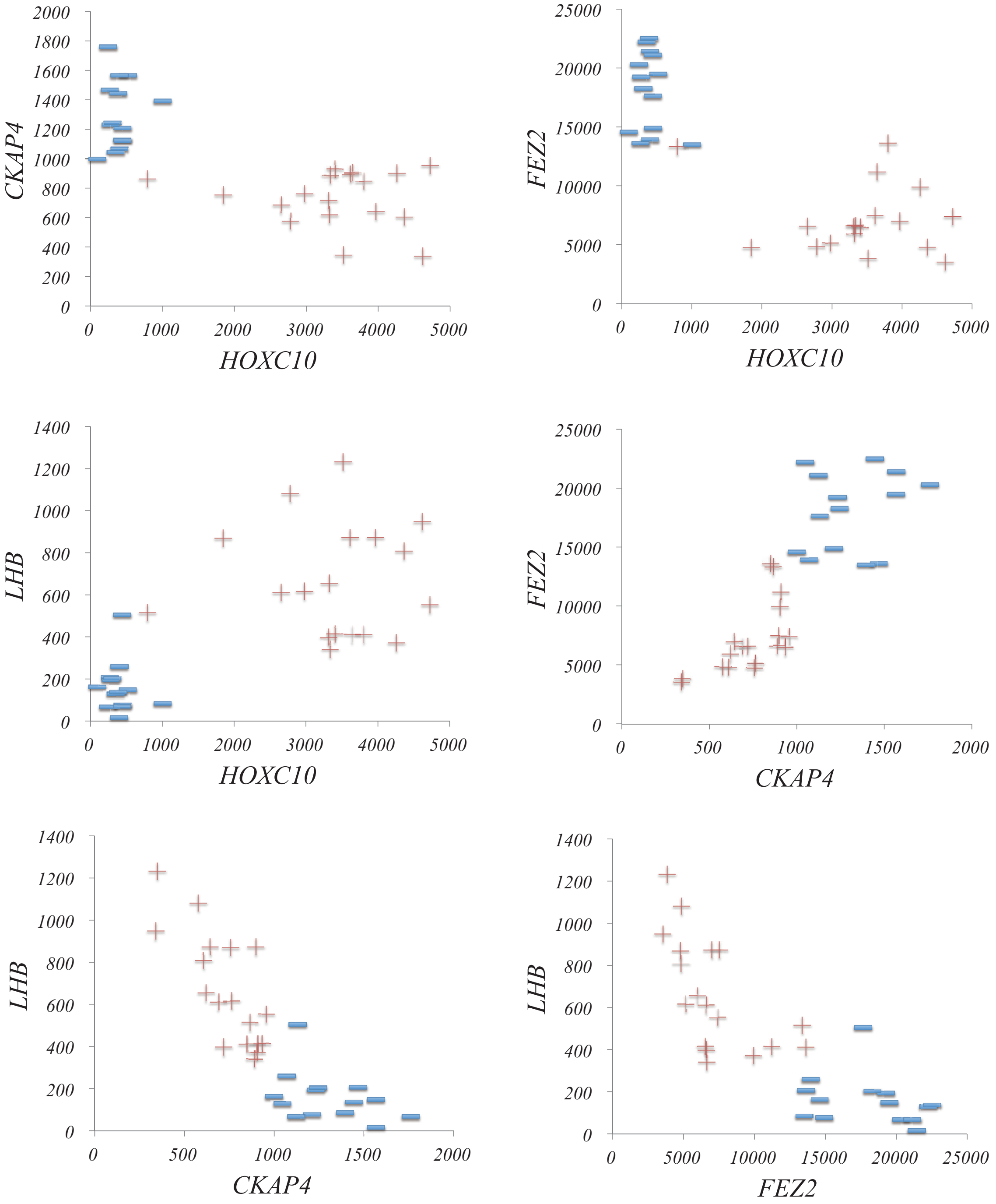

Fig. 4. Pairwise scatter plots for the most frequently selected genes: (+) indicates healthy cases and (–) indicates FSHD cases.

alization is illustrated in Fig. 5. This scatter plot is the two-dimensional projection of the two classes onto the first two eigenvectors of the scatter matrix as coordinate system, using as data the top four genes only.

Biological evidence In this section, two kinds of knowledge about the selected group of four genes are compiled and reported. First, scientific findings in the literature about the genes and their primary functions in cellular processes are summarized; due to space reasons, this is an abbreviated compendium of information found in major specialized sites. Next, protein-protein association networks (PPANs) are rendered by means of the STRING (search tool for the retrieval of interacting genes/proteins) database and software (Mering et al., 2005; Szklarczyk et al., 2011).

HOXC10. *HOXC10* (homeobox C10) is one of several homeobox *HOXC* genes located in a cluster on chromosome 12, encoding a highly conserved family of transcription factors playing an important role in morphogenesis in all multicellular organisms. The protein level is controlled during cell differentiation and proliferation, which may indicate that this protein has a role in origin activation (Gene, HOXC10, 2015). Isolating HOXC10 using a yeast one-hybrid system, Gabellini et al. (2003) concluded that HOXC10 is a homeoprotein with the potential to influence mitotic progression, which might provide a link between developmental regulation and cell cycle control.

CKAP4. CKAP4 is cytoskeleton-associated protein 4. The role of p63 (CKAP4) in binding of surfactant protein-A (SP-A) to type II pneumocytes has been assessed (Bates et al., 2008). Also, Rufini et al. (2011) propose a biological role of p63-GM1 interaction in regulation of p63 during epidermal differentiation.

FEZ2. The function of FEZ2 (fasciculation and elongation protein zeta 2 (zygin II)) remains unknown; however, using the yeast two-hybrid system, Alborghetti et al. (2011) found that FEZ2 interacts with up to 59 other proteins. Most importantly, Maturana et al. (2010) found that FEZ1, a FEZ2 homolog, is associated with neuronal development, neuropathologies and viral infection.

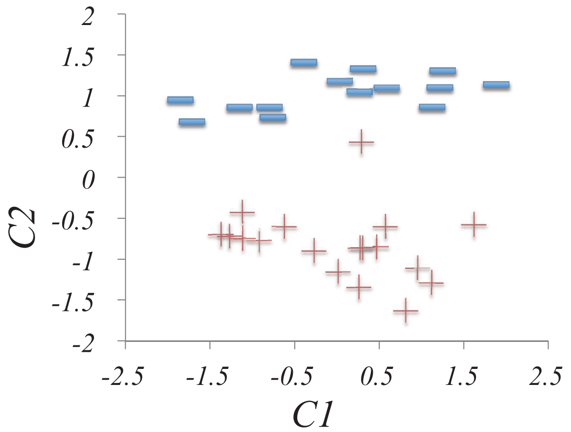

Fig. 5. Projection of selected genes onto the first two eigenvectors C_1, C_2 of the scatter matrices as coordinate system: (+) indicates healthy cases and (–) indicates FSHD cases.

LHB. The *LHB* (luteinizing hormone beta polypeptide) gene is a member of the glycoprotein hormone beta chain family and encodes the beta subunit of luteinizing hormone (LH) (Gene, LHB, 2015). Luteinizing hormones play an essential role in normal pubertal development and reproductive function in humans (Themmen and Huhtaniemi, 2000).

Protein-protein association networks The STRING platform integrates known functional associations between proteins, either by direct physical binding or by participation in the same metabolic pathway or cellular process. The associations are obtained from statistical analysis of co-occurrence in documents, physical interaction databases and curated biological pathway databases. These are then calibrated against previous knowledge, using the manually curated Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway maps (Kanehisa et al., 2014).

To this end, STRING uses two strategies: the first (‘COG–mode’) is based on externally provided orthology assignments and transfers interactions in an all-or-nothing fashion, while the second (‘protein mode’) uses quantitative sequence similarity searches and often distributes a given interaction partly among various pairs of target organism proteins (Mering et al., 2005). A final combined score S between proteins pairs is then computed between any two proteins. This score is often higher than the individual sub-scores, expressing larger confidence when an association is supported by several types of evidence. It is calculated under the (tenable) assumption of independence for the different sources (Mering et al., 2005):

$$S = 1 - \prod_i (1 - S_i) \tag{6}$$

There exist several examples about the use of PPANs generated by STRING software for the exploration of specific biological conditions. Bhutani et al. (2015) investigated the structural and dynamic features of two enzymes encoded by *Mycobacterium tuberculosis*, using molecular 3D modeling to simulate the interactions. The STRING software was used as a tool to confirm the results. Other statistical tools were used in the analysis, such as Clustering and PCA, the latter to reveal atomic motion fluctuations in both enzymes.

Olsen et al. (2014) examined tumor antigens as potential biomarkers for breast cancer using genomics and proteomics data. Normal vs. invasive ductal carcinomas tissue samples were analyzed via the Spearman’s rank correlation measure in order to extract expression levels of genes. A group of 30 genes (potential biomarkers), atypically expressed at the protein level, were found, 28 of them being analyzed *a posteriori* using the STRING software, a methodology reminiscent of the one proposed in the present paper. Díaz-Beltran et al. (2013) conducted a comparative analysis of several neurological diseases to describe multi-disorder subcomponents of autism spectrum disorders. The STRING software was used to generate gene networks of each member of autism sibling groups in order to find genetic overlaps. The next four paragraphs summarize the main findings about the biological role of the principal protein partners of the selected genes' products according to the scored partners as suggested by the STRING software.

HOXC10. Figures 6 (top left) and 7 show the scored functional partners and the PPAN for HOXC10, respectively. The top four are TBX4, TBX5, CDC27 and TBX6. It is believed that these *TBX* (T-box) gene products are transcription factors involved in the regulation of key developmental processes; moreover, CDC27 may be involved in controlling the timing of mitosis (Gene, CDC27, 2015). Among the other partners, DUX4L6 may be involved with transcriptional regulation for mesoderm differentiation and TBX2 very likely plays a role in extremity pattern formation. It is notable that the DUX4L6 protein appears in this interaction network, because it is already known to be related to FSHD (Gene, DUX4L6, 2015).

CKAP4. Two CKAP4 partners, the PLAT and ZDHHC2 proteins, are strongly related according to STRING (Figs. 6 (top right) and 8). The most prominent is PLAT, a plasminogen activator of tissue that converts plasminogen to plasmin zymogen by hydrolysis, playing an important role in remodeling and tissue degradation in cell migration (Gene, PLAT, 2015). ZDHHC2 has been proposed as a putative tumor/metastasis suppressor and its expression level is decreased in human cancers (Yan et al., 2013).

FEZ2. Figures 6 (bottom left) and 9 show the FEZ2 protein-protein partners, CRIM1 and LZTS1 being the top

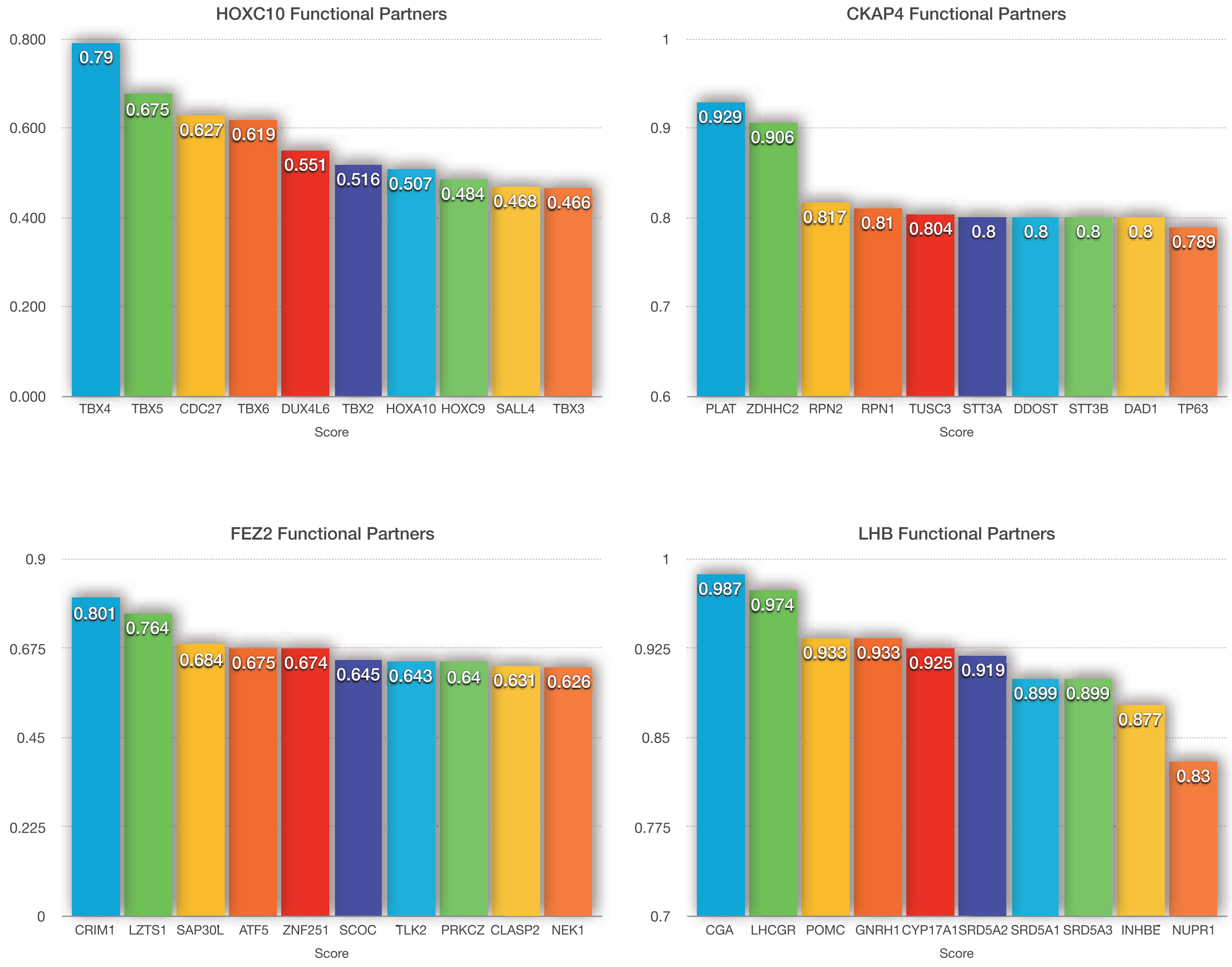

Fig. 6. Computed functional partners and scores for the top four genes' products (see text for a description).

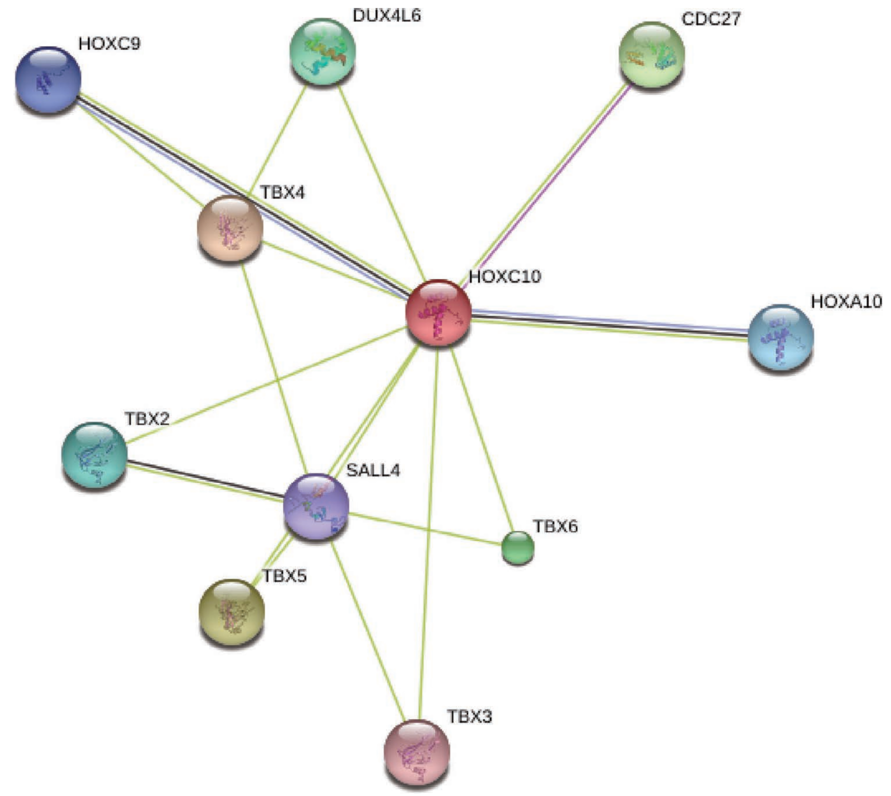

Fig. 7. Evidence in the protein association network around HOXC10. The different line colors represent evidence types for the association and the spheres represent the associated proteins.

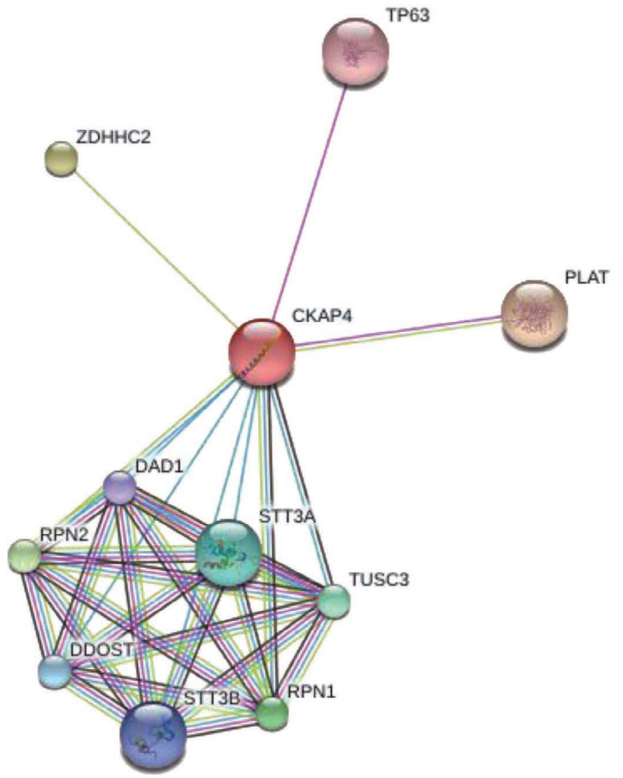

Fig. 8. Evidence in the protein association network around CKAP4 (see Fig. 7 for notation).

two. The former may play a role in CNS development through interaction with growth factors involved in motor neuron differentiation and survival, and possibly in capillary formation and maintenance during angiogenesis; it also modulates BMP activity by affecting its processing and delivery to the cell surface (Gene, CRIM1, 2015). The latter (LZTS1) may play a tumor-suppressing role in some types of cancer (Onken et al., 2008; Lovat et al., 2014; Wei et al., 2015).

LHB. This protein shows a strong PPAN association with at least six partners (Figs. 6 (bottom right) and 10). CGA belongs to the four human glycoprotein group, together with the luteinizing, follicle-stimulating and thyroid-stimulating hormones (GeneCards, 2012). LHCGR acts as a receptor for both luteinizing hormone and choriogonadotropin. Disorders of male secondary sexual character development, including familiar male precocious puberty, hypogonadotropic hypogonadism, Leydig cell adenoma with precocious puberty, and male pseudohermaphroditism with Leydig cell hypoplasia are disorders associated with this gene (Gene, LHCGR, 2015). POMC stimulates suprarenal glands to liberate cortisol; CYP17A1 participates in sexual development during fetal life and puberty, and SRD5A2 performs a decisive role at sexual differentiation. SRD5A3 transforms testosterone to 5–α–dihydrotestosterone; INHBE is an inhibine involved in the regulation of a variety of functions such

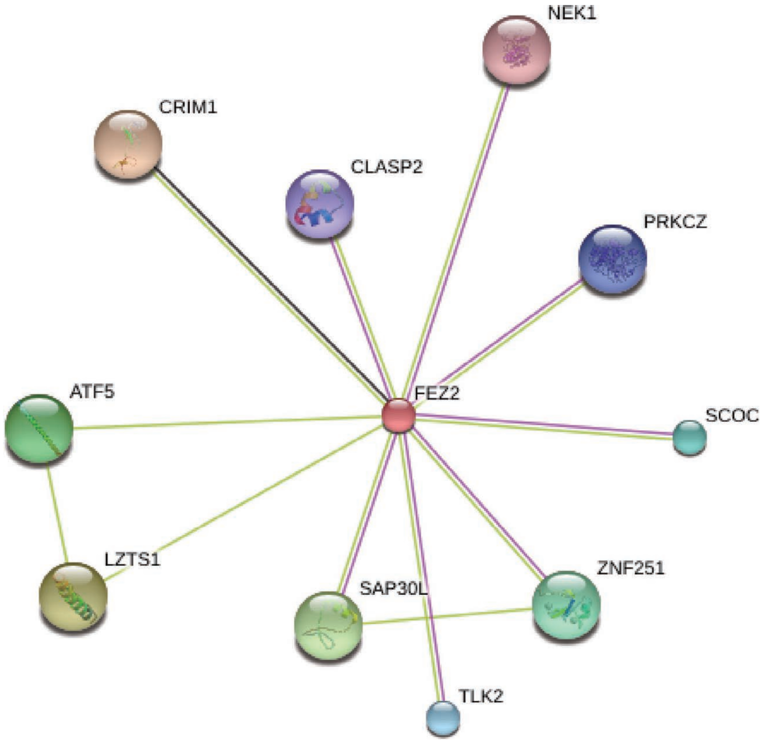

Fig. 9. Evidence in the protein association network around FEZ2 (see Fig. 7 for notation).

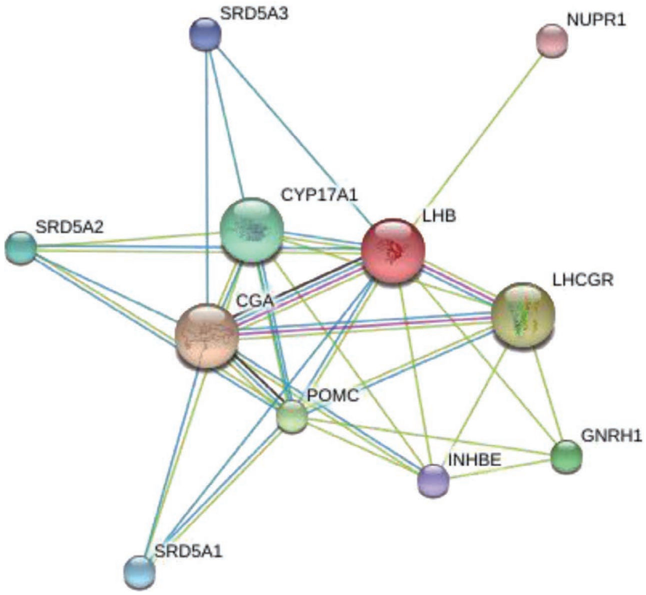

Fig. 10. Evidence in the protein association network around LHB (see Fig. 7 for notation).

as gonadal hormone secretion, germinal cell development and maturity, erythroid differentiation, insulin secretion, nervous cell survival, embryonic axial development or bone growth, depending on its subunit composition. Finally, NUPR1 is a nuclear regulator protein for transcription, promoting cellular growth in a way that helps tissues to counter a variety of injuries (Vesper et al., 2006; Winters and Moore, 2011).

CONCLUSIONS

Facioscapulohumeral muscular dystrophy is a very rare genetic muscle disorder affecting the muscles of the face, shoulder blades and upper arms. The genes responsible for the disease have not yet been identified; hence, no effective therapeutic strategies to suppress or even diminish their activity are currently possible. It is very likely that aberrant gene expression underlies FSHD, and therefore the pool of genes that should be looked at ought to be widened (Wahl, 2007). To identify possible causal gene expression for FSHD, this work uses machine learning (ML), which is not commonly addressed in health informatics.

A specific FSHD data set comprising samples of both healthy and FSHD patients has been analyzed with standard ML methods. There is no precedent in the literature for addressing this disease with these techniques. The fact that the FSHD data analyzed in this study are scarce and of high dimensionality makes computer-based automated classification a difficult undertaking. Most importantly, this high dimensionality precludes a straightforward interpretation of the obtained results, limiting their usability in a practical medical setting. In this vein, computational methods should represent low-complexity and interpretable solutions amenable to further analysis by experts. We have thus selected LDC and two linear SVMs as the target classifiers, embedded into a repeated resampling process. A simple feature selection strategy coupled with solid criteria for tie-breaking among competing genes has yielded gene subsets that constitute promising potential biomarkers. Classification models trained and tested using these genes offer acceptable recognition rates, well above 90% accuracy. The analysis has revealed the existence of a group of four genes that yield a neat discrimination between healthy and FSHD samples. This conclusion is supported both numerically (mean prediction errors) and graphically (dendrogram and low-dimensional projections). The reported results should be given careful medical evaluation, because they point in specific directions, but do not by themselves entail a medical solution to the disease, a situation common to all statistical and ML solutions. However, when predictive models use very low numbers of relevant genes, these genes are likely to be associated with the disease, and can be used as a starting point for further dedicated study from a biological point of view.

A major goal of exploratory studies of this kind should be to gain some understanding of how the variables selected by the model fit in relation to prior knowledge from the medical domain. The unveiling of triggering mechanisms of rare diseases entails the exploration of proteins, derived from specific genes, and their partners. From a functional perspective, this partnership means a direct physical binding association or an indirect interaction such as participation in the same metabolic pathway or cellular process. Information about the construction of these associations is widespread through several resources and information technologies. In this paper, STRING software was used to generate potentially useful information about genes highlighted as promising for FSHD by machine learning techniques. The supplied PPAN networks of association may then constitute a basis for exploring interactions between the FSHD region on chromosome 4q and others, as suggested elsewhere (Wahl, 2007).

Authors gratefully acknowledge funding from the Universidad Autónoma de Baja California and Universitat Politècnica de Catalunya - Barcelona Tech.

REFERENCES

Alborghetti, M. R., Furlan, A. S., and Kobarg, J. (2011) FEZ2 has acquired additional protein interaction partners relative to FEZ1: Functional and evolutionary implications. PLoS ONE **6**, e17426.

Bates, S. R., Kazi, A. S., Tao, J.-Q., Yu, K. J., Gonder, D. S., Feinstein, S. I., et al. (2008) Role of P63 (CKAP4) in binding of surfactant protein-A to type II pneumocytes. Am. J. Physiol. - Lung Cell. Mol. Physiol. **295**, L658–L669.

Bell, D. A., and Wang, H. (2000) A formalism for relevance and its application in feature subset selection. Mach. Learn. **41**, 175–195.

Bhutani, I., Loharch, S., Gupta, P., Madathil, R., and Parkesh, R. (2015) Structure, dynamics, and interaction of *Mycobacterium tuberculosis (Mtb)* DprE1 and DprE2 examined by molecular modeling, simulation, and electrostatic studies. PLoS ONE **10**, e0119771.

Boulesteix, A.-L. (2007) WilcoxCV: an R package for fast variable selection in cross-validation. Bioinformatics **23**, 1702–1704.

Díaz-Beltran, L., Cano, C., Wall, D. P., and Esteban, F. J. (2013) Systems biology as a comparative approach to understand complex gene expression in neurological diseases. Behav. Sci. **3**, 253–272.

Ding, C. H. Q., and Peng, H. (2005) Minimum redundancy feature selection from microarray gene expression data. J. Bioinform. Comput. Biol. **3**, 185–206.

Duda, R., Hart, P., and Stork, D. (2001) Pattern Classification. John Wiley and Sons, Hoboken, NJ.

EMBL-EBI (2014) The European Bioinformatics Institute (http://www.ebi.ac.uk).

Flanigan, K. M. (2004) Facioscapulohumeral muscular dystrophy and scapuloperoneal disorders. In: Myology, 3rd edn., (eds: A. G. Engel and C. Franzini-Armstrong), p. 1123–1133. McGraw–Hill, New York.

Fukunaga, K. (1990) Introduction to Statistical Pattern Recognition (2nd ed.). Academic Press, San Diego.

Furlanello, C., Serafini, M., Merler, S., and Jurman, G. (2003) Entropy-based gene ranking without selection bias for the predictive classification of microarray data. BMC Bioinformatics **4**, 54.

Gabellini, D., Colaluca, I. N., Vodermaier, H. C., Biamonti, G., Giacca, M., Falaschi, A., et al. (2003) Early mitotic degradation of the homeoprotein HOXC10 is potentially linked to cell cycle progression. EMBO J. **22**, 3715–3724.

Gene, CDC27 (2015) NCBI Gene Database (http://www.ncbi.nlm.nih.gov/gene/996).

Gene, CRIM1 (2015) UniProt (http://www.uniprot.org/uniprot/Q9NZV1).

Gene, DUX4L6 (2015) NCBI Gene Database (http://www.ncbi.nlm.nih.gov/gene/653544).

Gene, HOXC10 (2015) NCBI Gene Database (http://www.ncbi.nlm.nih.gov/gene/3226).

Gene, LHB (2015) NCBI Gene Database (http://www.ncbi.nlm.nih.gov/gene/3972).

Gene, LHCGR (2015) NCBI Gene Database (http://www.ncbi.nlm.nih.gov/gene/3973).

Gene, PLAT (2015) UniProt (http://www.uniprot.org/uniprot/P00750).

GeneCards (2012) Weizmann Institute of Science (http://www.genecards.org).

Guyon, I., and Elisseeff, A. (2003) An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182.

Guyon, I., Weston, J., Barhill, S., and Vapnik, V. (2002) Gene selection for cancer classification using support vector machines. Mach. Learn. **46**, 389–422.

Hamel, A. L. (2009) Knowledge Discovery with Support Vector Machines. John Wiley & Sons, Hoboken, NJ.

John, G., Kohavi, R., and Pfleger, K. (1994) Irrelevant features and the subset selection problem. In: Proceedings of the 11th International Conference on Machine Learning, (eds: W. W. Cohen and H. Hirsh), pp. 121–129. Morgan Kaufmann Publishers, San Francisco.

Kalousis, A., Prados, J., and Hilario, M. (2007) Stability of feature selection algorithms: a study on high dimensional spaces. Knowl. Inf. Syst. **12**, 95–116.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. **42 (Database issue)**, D199–D205.

Lisboa, P., Ellis, I., Green, A., Ambrogi, F., and Dias, M. (2008) Cluster based visualisation with scatter matrices. Pattern Recogn. Lett. **2**, 1814–1823.

Liu, H., and Motoda, H. (1998) Feature Extraction, Construction and Selection. A Data Mining Perspective. Kluwer Academic Publishers, Dordrecht.

Liu, H., Li, J., and Wong, L. (2002) A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. Genome Inf. **13**, 51–60.

Lovat, F., Ishii, H., Schiappacassi, M., Fassan, M., Barbareschi, M., Galligioni, E., et al. (2014) LZTS1 downregulation confers paclitaxel resistance and is associated with worse prognosis in breast cancer. Oncotarget **5**, 970–977.

Lukas, L., Devos, A., Suykens, J., Vanhamme, L., Howe, F., Majós, C., et al. (2004) Brain tumor classification based on long echo proton MRS signals. Artif. Intell. Med. **31**, 73–89.

Maarel, S. M. van der, Frants, R. R., and Padberg, G. W. (2007) Facioscapulohumeral muscular dystrophy. Biochim. Biophys. Acta **1772**, 186–194.

Maarel, S., Tawil, R., and Tapscott, S. (2011) Facioscapulohumeral muscular dystrophy and DUX4: breaking the silence. Trends Mol. Med. **17**, 252–258.

Maturana, A. D., Fujita, T., and Kuroda, S. (2010) Functions of fasciculation and elongation protein zeta-1 (FEZ1) in the brain. The Scientific World Journal **10**, 1646–1654.

MDC (2012) Muscular Dystrophy Campaign (http://www.muscular-dystrophy.org/).

Mering, C. von, Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., et al. (2005) STRING: known and predicted protein–protein associations, integrated and transferred across organisms. Nucleic Acids Res. **33 (suppl 1)**, D433–D437.

Onken, M., Worley, L., and Harbour, J. (2008) A metastasis modifier locus on human chromosome 8p in uveal melanoma identified by integrative genomic analysis. Clin. Cancer Res. **14**, 3737–3745.

Olsen, L., Campos, B., Winther, O., Sgroi, D., Karger, B., and Brusic, V. (2014) Tumor antigens as proteogenomic biomarkers in invasive ductal carcinomas. BMC Med. Genomics **7 (Suppl 3)**, S2.

Pan, W. (2002) A comparative review of statistical methods for discovering differentially expressed genesin replicated microarray experiments. Bioinformatics **18**, 546–554.

Parodi, S., Muselli, M., Fontana, V., and Bonassi, S. (2003) ROC curves are a suitable and flexible tool for the analysis of gene expression profiles. Cytogenet. Genome Res. **101**, 90–91.

Pudil, P., Ferri, F., Novovicova, J., and Kittler, J. (1994) Floating search methods for feature selection. Pattern Recogn. Lett. **15**, 1119–1125.

Reunanen, J. (2003) Overfitting in making comparisons between variable selection methods. J. Mach. Learning Res. **3**, 1371–1382.

Ripley, B. D. (1996) Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge.

Rose, M., and Tawil, R. (2004) Drug treatment for facioscapulohumeral muscular dystrophy. Cochrane Database Syst. Rev. **2**, DOI: 10.1002/14651858.CD002276.

Rufini, S., Lena, A. M., Cadot, B., Mele, S., Amelio, I., Terrinoni, A., et al. (2011) The sterile alpha-motif (SAM) domain of P63 binds in vitro monoasialoganglioside (GM1) micelles. Biochem. Pharmacol. **82**, 1262–1268.

Schölkopf, B., Tsuda, K., and Vert, J.-P. (2004) Kernel Methods in Computational Biology. MIT Press, Cambridge, MA.

Sotoca, J. M., Sánchez, J. S., and Mollineda, R. A. (2005) A review of data complexity measures and their applicability to pattern classification problems. In: Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA2005, p. 77–83. Thomson.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., et al. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. **39 (suppl 1)**, D561–D568.

Tawil, R. (2008) Facioscapulohumeral muscular dystrophy. Neurotherapeutics **5**, 601–606.

Tawil, R., and Maarel, S. (2006) Facioscapulohumeral muscular dystrophy. Muscle Nerve **34**, 1–15.

Tawil, R., Figlewicz, D., Griggs, R., and Weiffenbach, B. (1998) Facioscapulohumeral dystrophy: A distinct regional myopathy with a novel molecular pathogenesis. Ann. Neurol. **43**, 279–282.

Themmen, A. P. N., and Huhtaniemi, I. T. (2000) Mutations of gonadotropins and gonadotropin receptors: Elucidating the physiology and pathophysiology of pituitary-gonadal function. Endocr. Rev. **21**, 551–583.

Vapnik, V. (1998) Statistical Learning Theory. John Wiley & Sons, Hoboken, NJ.

Vesper, A., Raetzman, L., and Camper, S. (2006) Role of prophet of Pit1 (PROP1) in gonadotrope differentiation and puberty. Endocrinology **147**, 1654–1663.

Wahl, M. (2007) Impossible things: Through the looking glass with FSH dystrophy researchers. Quest Magazine, 14 (2). Available from http://quest.mda.org/sites/default/files/Quest142.pdf

Wei, Z., Mei-Rong, H., Hong-Li, J., Liu-Qing, H., Dan-Ling, D., Juan-Juan, C., et al. (2015) The tumor-suppressor gene LZTS1 suppresses colorectal cancer proliferation through inhibition of the AKT–mTOR signaling pathway. Cancer Lett. **360**, 68–75.

Winters, S. J., and Moore, J. P. (2011) PACAP, an autocrine/paracrine regulator of gonadotrophs. Biol. Reprod. **84**, 844–850.

Yan, S.-M., Tang, J.-J., Huang, C.-Y., Xi, S.-Y., Huang, M.-Y., Liang, J.-Z., et al. (2013) Reduced expression of ZDHHC2 is associated with lymph node metastasis and poor prognosis in gastric adenocarcinoma. PLoS ONE **8**, e56366.

Zhou, X., and Mao, K. Z. (2006) The ties problem resulting from counting-based error estimators and its impact on gene selection algorithms. Bioinformatics **22**, 2507–2515.