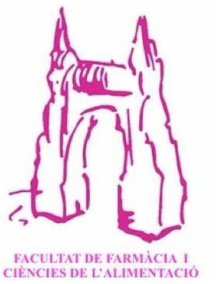
UNIVERSITAT DE BARCELONA



FINAL DEGREE PROJECT

FIBROMYALGIA THROUGH GENETICS: DEVELOPMENT OF A MICROARRAY-BASED DIAGNOSTIC ALGORITHM

José Manuel Borrego Burón

March 2020

Biochemistry and molecular biology

Mathematics and computer science

Public Health

Facultat de Farmàcia i Ciències de l'Alimentació

Universitat de Barcelona

1. Abstract

Fibromyalgia syndrome (FMS) is an incapacitating multifactorial disease characterised by widespread pain. Its pathophysiology is still unknown and its diagnosis traditionally difficult. New research on possible genetic and epigenetic factors has shed light into its possible pathways and better diagnostic methods. The goal of this study is to design and implement a diagnostic algorithm for microarray data regarding RNA and microRNA expression. To do so, on the one hand, we studied several classification methods and tested their adequacy and feasibility given the data and available computational power, and on the other hand, we analysed gene expression data in an interaction network and microRNA related pathways. The final algorithm used Support Vector Machine based Recursive Feature Elimination and holdout cross validation to assess the minimum probeset that provided the best accuracy. The results provided a set of 56 RNA probes with an accuracy of 95.72% and a set of 20 microRNA probes with 98.95%. Since it is based on a very limited dataset, the results are not meant to be conclusive but to serve as a steppingstone to future studies. The interaction network, as well as microRNA analysis provided useful insights into possible FMS-related neural system genesis and, specially, inflammatory pathways (through miR-145, miR-150 and miR-451 and TNF-α interactions). We advise future work on the subject to finally unveil fibromyalgia's aetiology and provide accurate and useful diagnosis and treatment.

Resumen

El síndrome de fibromialgia (FMS) es una enfermedad multifactorial incapacitante caracterizada por dolor generalizado. Su fisiopatología es aún desconocida y su diagnóstico tradicionalmente difícil. Nuevas investigaciones sobre posibles factores genéticos y epigenéticos han arrojado luz sobre posibles vías y sobre mejores métodos de diagnóstico. El objetivo de este estudio es diseñar e implementar un algoritmo de diagnóstico aplicado en datos de *microarray* de expresión de ARN y microARN. Para hacerlo estudiamos varios métodos de clasificación y probamos su adecuación y viabilidad dados los datos y la potencia computacional disponible. También analizamos los datos de expresión génica en una red de interacción y vías relacionadas con microARN. El algoritmo final usó Eliminación Recursiva de Atributos basada en Máquinas de Soporte Vectorial y validación cruzada aleatoria para evaluar el mínimo conjunto de sondas que proporcionara la mejor precisión. Obtuvimos un conjunto de 56 sondas de ARN con una precisión del 95,72% y otro de 20 sondas de microARN con el 98,95%. Dado que se basa en un conjunto de datos muy limitado, los resultados no pretenden ser concluyentes, sino que sirven como paso para futuros estudios. La red de interacción, así como el análisis de microARN proporcionaron información útil sobre posibles vías relacionadas con FMS: la génesis del sistema neural y, especialmente, vías inflamatorias (a través de las interacciones miR-145, miR-150 y miR-451 y TNF-α). Sugerimos la

necesidad de más estudios sobre el tema para finalmente revelar la etiología de la fibromialgia y proporcionar un diagnóstico y tratamientos precisos y efectivos.

2. Integration of academic fields

The main focus of this project is to analyse genetic and transcriptomic data on fibromyalgia syndrome. Hence, this work is mainly framed into the Biochemistry and molecular biology field. Moreover, the project also aims to design a classification algorithm to apply on such data and hopefully find an objective diagnostic tool for future FMS patients. The development of the algorithm implies analysis of machine learning methods and statistical tools, subjects from the Mathematics and computer science field. Additionally, since the goal of the algorithm implementation is to provide improve the FMS diagnosis procedure, the field of Public Health is involved.

3. Introduction

"Having fibromyalgia is like being fifty years older: daily pain, constant effort. Everything is twice or three times harder. Thanks to the medication I have some hours a day of normal activity, but I can't have a regular job. That was hard to accept. When I was 24, I started having pain, but fibromyalgia wasn't known at the time. I began going from doctor to doctor; they couldn't find anything specific in all the radiographies and analysis. Years later, after the pregnancy, someone recommended I went to an internist and he gave the right diagnosis: fibromyalgia. We also went to medical investigation centres and to visit specialists in fibromyalgia and rheumatology and we tried to participate in studies for new treatments; but the options were few. To better understand this illness and to find support, I got myself involved in a fibromyalgia patients' association. We also raised awareness of this almost unknown disease. "

As Mari Carmen Burón, my mother, states, fibromyalgia syndrome (FMS) is an incapacitating multifactorial disease. It is characterized by the presence of deep and diffuse musculoskeletal pain associated with other subjective manifestations such as fatigue, disturbed sleep and variable degrees of anxiety and depression, among others. A recent review of the state of the issue (1) reports its prevalence in the general population at 2.2 % and that various risk factors have been identified: age, gender, level of education and socio-economic status. The main concern about this disorder is that its causes have been a mystery. In recent years there has been several hypothesis and further breakthroughs about its pathophysiology but there still isn't consensus on the root of the problem. Neuprez et al. (1) state that, as of 2017, the proposed potential mechanisms include genetic predisposition, central amplification, diffuse inhibitory control failure, muscle as peripheral nociceptive afferents. Given its characteristics, it is not unusual that FMS patients have to suffer from stigma. Some moralizing attitudes, disbelief as to the reality of pain, and pain's invisibility are the main causes of this stigmatization of patients (2) and it hinders severely an effective management of the disorder.

The official diagnosis was firstly stablished by the American College of Rheumatology in 1990 (3) and it consisted of the presence of widespread pain in combination with tenderness at 11 or more of 18 specific tender point sites. However, due to its somewhat subjective nature, the diagnosis was rather difficult and usually was made by exclusion of other disorders. It was and still is sometimes confused with chronic fatigue syndrome, due to a similar symptomatology. There was a revision in 2010-2011 (4) that included a patient questionnaire in order to better assess the correct diagnosis. Later, in 2016, another revision combined physician and questionnaire criteria while minimizing misclassification of regional pain disorders and is generally considered an improvement over the 1990 criteria (5).

There have also been attempts at classifying FMS patients into subgroups

according to symptoms and other attributes as to better understand it and provide better treatment. Docampo and colleagues (6) conducted a cluster analysis of clinical data and described three FMS subgroups according to familial and personal comorbidities (such as stress disorders, family history of autoimmune disorders, etc.) and symptoms and their characteristics (such as muscle weakness, sleep disturbances, etc.). Furthermore, for years it has been reported that there exists a familial susceptibility of FMS and Buskila et al. (7) already suggested in 2007 that it may be attributable to genetic factors.

Further investigation into this side of the issue has shed light into the syndrome. In 2007 there were reports of evidence that polymorphisms of genes in the serotoninergic, catecholaminergic and dopaminergic systems were related to FMS (7). In recent years, proteomic analysis has found several possible biomarkers or FMS-related pathways such as the G-protein coupled estrogen receptor (8), kinins and their B1 and B2 receptors in mice (9) and haptoglobin and fibrinogen (10). Clos-Garcia and colleagues (11) analysed gut microbiome and metabolome of patients and controls and found that the abundance of the *Bifidobacterium* and *Eubacterium* genera in patients was significantly reduced. They also found there are altered levels of glutamate and serine that had correlation to the gut microbiome results, reflecting the effect of the microbiome on metabolic activity.

Genetic studies have recently vastly improved FMS knowledge. The Al-Ándalus project (12) identified associations of the rs841 and rs2097903 SNPs, from the guanosine triphosphate cyclohydrolase 1 and catechol-O-methyltransferase genes respectively, with higher risk of fibromyalgia susceptibility. They also confirmed that the rs1799971 SNP (opioid receptor μ1 gene) might confer genetic risk of fibromyalgia. D'Agnelli and colleagues conducted a review on genetic and epigenetic data. They found that beside a genetic predisposition, environmental factors also play a fundamental role in the onset and development of FMS, through epigenetic modulations. Particularly, FMS patients show hypomethylation especially in promoter regions of genes implicated in DNA repair, immune system, and membrane transport genes. It is also reported that there are some studies that investigated microRNA expression.

In light of these and several other new findings there has been some development of new clinical diagnostic criteria such as the use of using slowly repeated evoked pain responses in addition to clinical symptoms to enhance the diagnosis (13). Furthermore, there have been new breakthroughs on the use of diagnostic biomarkers such as succinic acid, taurine and creatine levels (14) or alpha-enolase, phosphoglycerate-mutase 1 and serotransferrin (15). Most of the genetic studies cited above also suggest using the found the genetic or transcriptomic differences in diagnosis.

Even though the studies conducted over the past twenty years have not yet completely explained the molecular mechanisms of FMS, the possibility that there are

important and targeted genetic basis is rather unlikely. Nevertheless, all found genetic and transcriptomic modifications related to FMS or to an increased risk of FMS might reveal new information and truly prove to be an objective and accurate diagnostic tool.

4. Objectives

This study follows two main objectives. Firstly, we aimed to learn to design and implement a selective discriminatory algorithm on available microarray data from FMS patients. Secondly, we wanted to further assess possible physiopathological causes and the use of such data in diagnosis and/or treatment.

5. Methods

We conducted a search in several genetic databases (NCBI PubMed, NCBI dbSNP, NCBI GEO, OMIM, GWAS Catalog, PheWAS Catalog). Docampo and colleagues (16) conducted a genotypic profiling study and found two mutated or modified genes associated with FM. However, very few databases provided other useful or relevant data for the purpose of this study. The main source of data was the GEO database in which we found three relevant studies with available microarray datasets: three studies regarding gene expression, gene methylation and microRNA expression in FMS, respectively.

On the one hand, Ciampi de Andrade and colleagues (17) characterized DNA methylome in peripheral blood, using bisulphite converted DNA hybridised to the Illumina Infinium 450k Human Methylation Beads. They found changes in genes implicated in immune system and showed relation to a dysfunctional connectivity in pain network. On the other hand, Jones et al. (18) analysed gene expression in whole blood samples of FMS patients and healthy matched controls. The RNA was isolated using the PAXgene RNA isolation kit and total RNA was quantified afterwards on a Nanodrop spectrophotometer and only samples with good quality RNA (RNA integrity number > 8) were processed to be hybridized to Affymetrix® Human Gene 1.1 ST Peg arrays. They found that genes related to inflammatory pathways were hyper-expressed while specific pathways related to hypersensitivity and allergy were hypo-expressed, as well as known pathways for pain processing and axonal development were differently expressed. They used a Support Vector Machines-based algorithm to classify samples into either patient or control classes and tested their results with corrections for optimism based on the bootstrapping method, applied to the model generated using a Logistic Regression algorithm.

Apart from methylation and RNA expression data, another microarray dataset was obtained regarding microRNA expression in FMS patients and matched controls. MicroRNA (miRNA) are small, singlestranded, non- coding RNA molecules that regulate

several protein-coding genes (19). miRNA attaches to the target mRNA by base-pairing to downregulate its expression and, as so, negatively regulate protein synthesis. miRNA profiles (miRnome) have been analysed in several diseases to find out whether it plays an important role in their aetiology or physiopathology. miRNome profiles are actually altered in specific tumours, indicating that miRNA might be implicated in the development of cancer and other diseases (19). The importance of miRNA-mediated gene regulation indicates that the study of miRNome in FMS might unveil new information. Cerdá-Olmedo et al. (20) conducted a study to identify changes in miRNome of FMS patients to, firstly, develop a quantitative diagnostic method and, secondly, provide a deeper understanding of FM. miRNA samples were extracted from peripheral blood mononuclear cells of FMS patients and population-age-matched controls using human v16-miRbase 3D-Gene microarrays (Toray Industries, Japan). Selected miRNAs were further validated by RT-qPCR. They found a marked downregulation of 4-fold or more of hsamiR223-3p, miR451a, miR338-3p, miR143-3p, miR145-5p and miR- 21-5p. About 20% of the miRNA were hypo-expressed (2-fold or more). They concluded that this might implicate a general de-regulation of the miRNA synthetic pathway in FM. Nonetheless, they found no significant correlations between miRNA inhibition and FMS fundamental symptoms.

Onwards, these three datasets will be referred to as DM (DNA Methylation), GE (Gene Expression) and ME (miRNA Expression), respectively.

5.1. Algorithm design

The diagnosis of a patient can be modelled through a classification problem, such that, given a series of values belonging to some features, we can decide if they correspond to either a (+) or a (-) class, patient or healthy control, respectively. These features could be glycohemoglobin blood levels in diabetes diagnosis or blood pressure in arterial hypertension. The development of genetic techniques provides substantial opportunities in classification through genetic data (21). In these cases, the input data is a vector consisting of *N* features: gene expression or methylation coefficients for N genes, among others.

In order to develop such a classification model, we need a training set of vectors **X** (corresponding to patients and controls) with known class labels **Y**. For example, **X** correspond to the probe expression coefficients for all patients and controls and **Y**, to either +1 or -1 if the given sample is from the patient or control classes, respectively.

$$\mathbf{X} = \{\bar{x}_1, \bar{x}_2, ...\bar{x}_i, ...\bar{x}_l\}$$
$$\mathbf{Y} = \{\bar{y}_1, \bar{y}_2, ...\bar{y}_i, ...\bar{y}_l\} \qquad \bar{y}_k \in \{-1, +1\}$$

These training vectors are used to generate a decision function $D(\bar{x})$. Once built, it can be used to classify new patterns:

$$D(\bar{x}) > 0 \Rightarrow \bar{x} \in \text{class } (+) \qquad D(\bar{x}) < 0 \Rightarrow \bar{x} \in \text{class } (-)$$
$$D(\bar{x}) = 0, \ \bar{x} \text{ is on the decision boundary}$$

The decision function can be a simple weighted sum of the training vector plus a bias, a linear discriminant function:

$$D(\bar{x}) = \bar{w} \cdot \bar{x} + b$$

where $\bar{w}$ is the weight vector and b is a bias value.

In the case of using genetic data in a classification problem, usually the number of features is very high (thousands of genes) and the number of training patterns is smaller in comparison (the number of patients and controls). In such cases the problem of data overfitting can arise; it is easy to find a $D(\bar{x})$ that separates the training but will classify poorly test data. To avoid the issue, it's not unusual to perform algorithms that reduce the dimensionality of the vector space.

One such method is projecting on the first few principal directions of the data, obtaining new features that are linear combinations of the original features (22). However, this implies that one cannot discard any of the original features, even though some might be irrelevant to the classification. Given that the objective is to build diagnostic tests, we need to select a small subset of genes due cost effectiveness and ease of verification of the results.

The simplest way to find such subset would be to exhaustively train the decision function to all subsets of features and selecting the one with highest discriminatory power with test data. However, it is obvious that exhaustive enumeration on large numbers of features is highly impractical.

Feature selection in such cases can be achieved through various methods; feature-ranking being especially useful (21). If all features are ranked according to a ranking function, we can select a fixed number of the top features or establish a threshold on the ranking criterion. Furthermore, it is possible to use the ranking to defined nested subsets of the feature space leaving out successively the lowest ranked feature. In doing so, we can find an optimum subset of features modifying only one variable: the number of features.

In order to apply feature selection to the available data, we firstly proposed using correlation classifiers as Lukkahatai et al. performed in 2018 (23). They attempted to apply a predictive algorithm to identify a group of genes whose differential expression discriminated individuals with FMS diagnosis from healthy controls.

They followed the predictive algorithm using filter methods and recursive feature elimination established by Saligan and colleagues (24). These researchers studied the difference between nonmetastatic patients that developed fatigue and those that did not during radiation therapy, and through raw microarray transcripts from whole-blood RNA, they identified genes that discriminated from both classes. Lukkahatai et al. (25), similarly, applied raw microarray gene expression data from peripheral blood mononuclear cell RNAs of FMS subjects and age- and gender-matched healthy controls to the same predictive algorithm. It uses fold-change differential (FC) and Fisher's ratio

(FR) as feature ranking criterion. The discriminatory accuracy of the gene subset was established via leave-one-outcross-validation (LOOCV) iterating over all the samples. The decision function in the subset was based on a nearest-neighbour classifier (k-NN).

However, as Guyon et al. state, correlation classification methods select the genes that individually classify best the training data (21). The feature elimination procedure usually does not yield compact subsets because gene data is redundant and also it eliminates genes that individually do not separate the data but do when considered together. A good feature ranking criterion is not necessarily a good feature subset ranking criterion.

Taking this information into consideration, we decided to use, as the original study our GE data is from, Support Vector Machines (SVMs).

5.1.1. SVM

SVMs are machine learning models with associated learning algorithms that map the data examples of separate categories (in this case a binary separation between patients and healthy controls) so that they are divided by a clear margin that is as wide as possible. More formally, a SVM builds a hyperplane or set of hyperplanes in a high-dimensional space (22), as seen in figures Figure 1 and Figure 2.

Guyon et al. demonstrated that SVMs are very effective for discovering informative features and that they have quantitative advantages over other gene selection models. If the training dataset, **X,** is linearly separable, a linear SVM is a maximum margin classifier (26). The decision boundary (the hyperplane that separates data points from each class) is located to leave the largest possible margin on either side, so that the classification can be as robust as possible. In SVMs, the weights w_i of the decision function $D(\bar{x})$ depend only on a small subset of the training points. Those are the examples that are closest to the decision boundary and lie on the margin and they are called *support vectors*. They are one of the reasons for the competitive performance of SVMs.

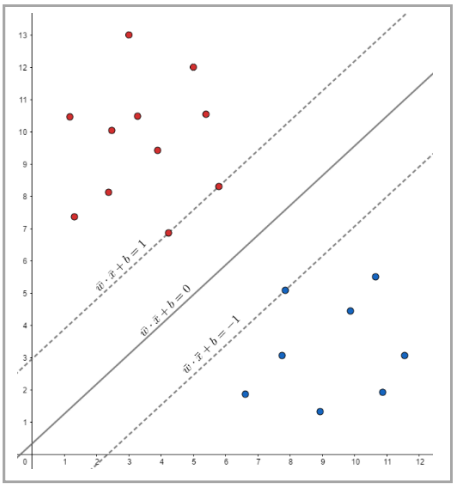


Figure 1: example of a 2D SVM classification where the central line represents the decision function and the dotted lines, the margin limits designated by the support vectors (data points on the lines)

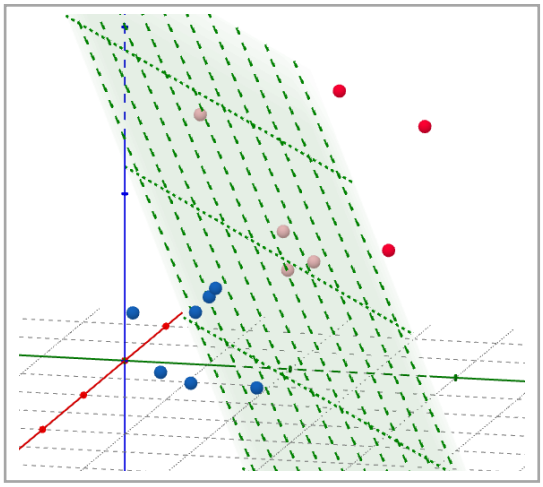


Figure 2 :example of a 3D SVM classification in which the dotted plain represents the decision function

The decision boundary is defined by the weight vector $\bar{w}$, which is perpendicular to the hyperplane. Given a new data point, noted by vector $\bar{x}$, the decision function will indicate if it falls on either side of the decision boundary, mathematically either:

$$\bar{x} \cdot \bar{w} + b \leq 0 \Rightarrow \text{class} -$$
$$\bar{x} \cdot \bar{w} + b \geq 0 \Rightarrow \text{class} +$$

The constraint of finding the hyperplane with the largest margin is crucial in the optimization of $\bar{w}$ and b. In order to do so, we define the two hyperplanes that define the margin, given by the support vectors (as seen in Figure 1 as a 2D example). These are defined as

$$\bar{w} \cdot \bar{x}_{-SV} + b = -1$$
$$\bar{w} \cdot \bar{x}_{+SV} + b = 1$$

for all (+)-class and (-)-class support vectors. Taking into consideration the class labels, y_i:

$$y_i \left(\bar{w} \cdot \bar{x}_i + b\right) - 1 = 0$$

This equation is the formal constraint for support vectors that allows us to calculate the optimal weight vector and bias. The constraint function for every featureset is, simply, that the above result has to be greater than or equal to 0, since we want that the training data do not fall within the decision margin even though unknown data might do so.

Therefore, our goal is to maximize the width between separating hyperplanes: minimizing the magnitude of the weight vector, $||\bar{w}||$, and maximizing b, with the constraint such that $y_i \left(\bar{w} \cdot \bar{x}_i + b\right) \geq 1$.

This problem is quadratic with linear inequality constraints, so in order to solve it we can use Lagrange multipliers:

$$L(\bar{w}, b) = \frac{1}{2} \|\bar{w}\|^2 - \sum_i \alpha_i \left[y_i \left(\bar{w} \cdot \bar{x}_i + b \right) - 1 \right]$$

Then we substitute into the Lagrange primal function the corresponding derivatives. Finally, the training of the SVM given a training dataset **X** and **Y** consists of minimizing the following equation over α_i:

$$L = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \left(\bar{x}_i \cdot \bar{x}_j \right)$$

Subject to $0 \leq \alpha_i \leq C$

The result is the list of parameters α_i which is used to calculate both $\bar{w}$ and b, most α_i being equal to 0. *C* is a positive soft margin parameter that ensures convergence even when the problem is non-linearly separable. As stated by Guyon et al. in the case of study, a value of *C*=100 is adequate. The decision function $D(\bar{x})$ can be calculated through:

$$\bar{w} = \sum_i \alpha_i y_i \bar{x}_i$$
$$b = \langle y_i - \bar{w} \cdot \bar{x}_i \rangle$$

As we stated before, in cases where the dimensionality of the feature space is very high, it is useful to perform feature elimination. We can, therefore use the SVM to assess the importance of each feature for the classification of the data and successively eliminate the least important features; such a procedure is SVM-based recursive feature elimination (SVM-RFE) (27). It is a specific application of recursive feature elimination that uses the weight magnitude as ranking criterion of the features. An outline of the algorithm is as follows:

- Given training datasets **X**,**Y**, initialize the subset of surviving features *s* and the ranked feature list *r:*
  $$s = [1, 2, \ldots N]$$
  $$r = [\,]$$
- Train the classifier α_i
- Compute the weight vector of dimension length |*s*|
  $$\bar{w} = \sum_i \alpha_i y_i \bar{x}_i$$
- Compute the ranking criteria for all i
  $$c_i = (w_i)^2$$
- Find the feature f with smallest ranking criterion
  $$f = s_k \quad \{k \,|c_k = min(\mathbf{c})\}$$

- Update r by adding f
- Eliminate f from s
- Repeat until $s = [\,]$

5.1.2. Internal Validation

The resulting SMV model, after applying RFE to a certain number of features, should, therefore, classify new datasets into either class and, as such, act as a diagnostic criterion. However, it is essential that its performance is tested and validated. Given the limited amount of data available and that repeating the sample collection and processing is highly impractical, we resort to internal validation methods (28). In general, they split the available data into two subsets, one to train the model and the other to test it.

Jones et al. used a bootsrapping-based method to assess the discrimination of the optimal and minimal probeset. According to Efron and colleagues (29), the most efficient validation is achieved by computer-intensive resampling techniques such as the bootstrap. Bootstrapping imitates the process of sampling from an underlying population by generating samples with replacement from the original data set (28).

However, due to computational constraints, we decided to apply classical cross-validation methods following a similar procedure to Saligan et al.

5.1.3. Final Algorithm

In summary, after assessing several algorithms and methods and the feasibility with our available assets, we designed the final algorithm to be applied to both the GE and ME datasets. The first dataset, since Jones et al. already applied a more powerful SVM-based algorithm on it, will serve as a reference to assess the external validity of the results. All procedures were implemented in Python 3.7.6 using the *numpy*, *pandas* and *scikit-learn* libraries.

Firstly, data from all features was normalized according to the N samples and then defined into **X**,**Y** sets, **X** containing all $\bar{x}_i$ vectors with the expression profile and **Y**, all y_i class labels for each sample i. These sets were then divided into random paired training-testing subsets such that the testing subset contained $0.2N$ elements. Both cohorts had equal representation of FMS.

Next, SVM-RFE was performed through training the SVM-classifier with the training dataset, calculating the weight and ranking of each feature and recursively eliminating the lowest ranking one until a variable number of features was achieved. The accuracy of the reduced featureset was tested through 2000 iterations of k cross-validation, all accuracy values were stored in a data matrix.

The whole process was repeated 10 times for each number of features, k. The selected k values for each microarray dataset were:

GE:	ME:
k=[2,3,4,5,6,7,8,10,12,14,16,20,24,28,32,40, 48,56,64,80,96,112,128,160,192,224,256]	k=[2,3,4,5,6,7,8,10,12,14,16,20,24,28,32,40, 48,56,64,80,88,96,104,112,120,128,144,160, 176,192,208,224,240,256]

They were selected by successive addition of powers of 2 so as to cover a greater range of k without computing all natural numbers in-between.

The **X,Y** sets that were used in the algorithm were subsets and did not contain all features from the microarray datasets. Since GE dataset contains 33297 probes (features) and ME only 1213, the former was used to assess the computational capabilities and the limits on the algorithm. The algorithm was tested by including increasing number of features into the **X**,**Y** sets. Including more than 8000 features exceeded the available computational power.

Therefore, less than 8000 could be included from the GE and ME datasets. In order to select the features to be included we performed a correlation ranking as proposed at the beginning of the algorithm design. All features were ranked according to their Fischer's Ratio (FR) and the top 8000 features were selected to perform the algorithm.

5.2. Microarray data general analysis

Apart from performing the classification algorithm on GE and ME data, and taking into consideration that the results discussed in section 6.1 did not provide truly useful insights into the role of epigenetic data in FMS, we decided to analyse the three datasets in parallel to the algorithm by selecting the 250 top differentiated probes for all datasets (by ranking through |FC|).

The data obtained for DM and ME can be observed in Table 1 and Table 2. It is important to note that the fold change is calculated by

$$FC_i = \log_2 \frac{\mu_{i\mathrm{FMS}}}{\mu_{i\mathrm{Control}}}$$

Where $\mu_{i\mathrm{FMS}}$ and $\mu_{i\mathrm{Control}}$ are the mean coefficients for patient and control classes for probe i. The fold change values represent different magnitudes in each dataset: in DM they represent methylation levels and in GE, RNA levels. Even though a gene with high methylation is usually underexpressed and, as such, these magnitudes affect in opposite directions, the changes in both parameters are indicative of gene expression changes.

We firstly selected 80 of the most differently expressed genes for each subset to be input in the STRING database so as to create a protein interaction network. However, no satisfactory networks were obtained. Therefore, we input all 250 genes from each subset and obtained two complex networks seen in Figure 3 and Figure 4.

ID	P value	FC	2^{FC}	Probe SNPs	Gene Symbol	Relation to UCSC CpG Island
cg26044428	4.75e-07	0.415	1.333		ANGEL1	Island
cg10140678	2.67e-10	-4.013	0.062			N_Shore
cg03517506	4.76e-08	0.537	1.451		ANKRD36	N_Shore
cg09413645	4.71e-07	-3.039	0.122			
cg02458875	2.96e-09	-2.402	0.189	rs72843885		N_Shore
cg12817782	5.17e-08	0.666	1.587		ANO1	Island
cg11094953	5.41e-07	0.311	1.241		ARCN1	N_Shore
cg14930904	1.25e-07	0.475	1.390		ARHGAP12	N_Shore
cg01019484	1.77e-07	-1.659	0.317			
cg03794530	9.03e-08	-1.654	0.315			N_Shore

Table 1: first 10 items from DM data with calculated FC (in alphabetical order of Gene Symbol)

ID	P value	FC	2^{FC}	Gene Symbol	Gene name
8128795	1.24e-03	0.178	0.884	AK9	adenylate kinase 9
7892779	2.67e-04	0.56	0.678		
7893526	2.67e-04	0.56	0.678		
8122807	3.32e-05	0.200	0.871	AKAP12	A-kinase anchoring protein 12
8150439	3.21e-03	-0.294	1.226	ANK1	ankyrin 1
7895647	7.71e-04	0.506	0.704		
8069511	2.76e-03	0.249	0.842	ANKRD20A11P	ankyrin repeat domain 20 family member A11
8069508	3.44e-04	0.479	0.718		
7919139	1.12e-03	0.179	0.883	ANKRD20A12P	ankyrin repeat domain 20 family member A12
7919146	1.71e-03	0.181	0.882	ANKRD20A8P	ankyrin repeat domain 20 family member A8

Table 2: first 10 items from GE data with calculated FC (in alphabetical order of Gene Symbol)

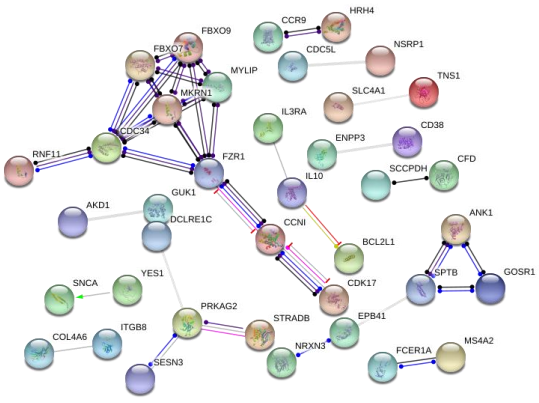


Figure 3: protein interaction network from GE data in STRING database (high confidence)

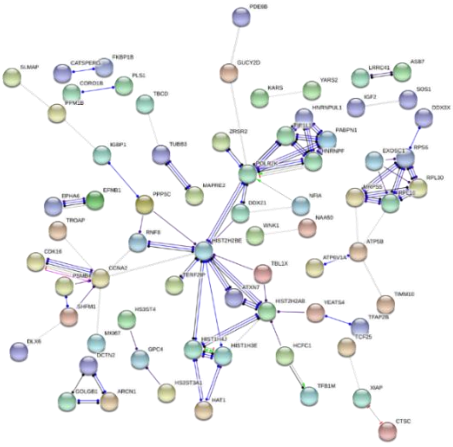


Figure 4: protein interaction network from DM data in STRING database (high confidence)

Nonetheless, as we could only find one gene that appeared in both subsets, the comparison of both networks would probably be unfruitful, thus, we combined all genes and (their respective FC values) in a single dataset to be input in the STRING database. Some formatting work was needed to achieve better results (namely, eliminating duplicates, commas and other punctuation symbols, etc.). The network we obtained (Figure 5) was difficult to analyse due to the high number of nodes and edges and due to the lack of the expression data. We did have a succesful protein interaction network, but no readily available information about the difference in expression in FMS.

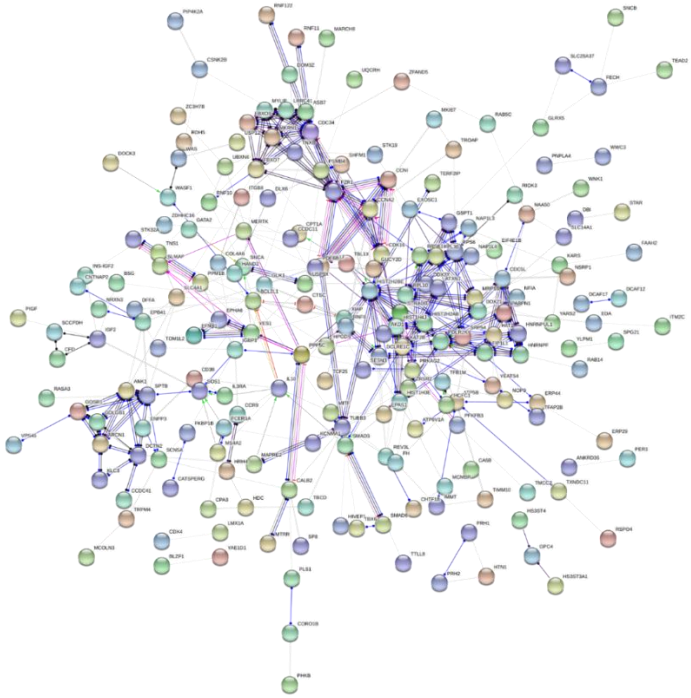


Figure 5: protein interaction network from GE and DM conjoined data in STRING database (medium confidence)

In order to create a network which included the expression data, we downloaded the graph information from the STRING database. We obtained an incidence list with scores for each edge (weight) and additional information on each node. The edge scores were calculated by STRING according to the confidence of the given interaction. We put together the incidence list and weight with the FC for each node and the dataset node 1 was part of (to facilitate further identification). Table 3*: random 10 item incidence list from conjoined STRING and expression data* represents a random 10-row sample from the conjoined graph information.

Node 1	Node 2	Score	FC node 1	FC node 2	Dataset
TERF2IP	HIST2H2BE	0.900	1.515	1.303	DM
TFAP2B	YEATS4	0.964	1.724	2.241	DM
TFB1M	HCFC1	0.902	1.285	0.531	DM
TFB1M	NOP9	0.575	1.285	1.477	DM
TFB1M	PPP5C	0.436	1.285	1.313	DM
TFB1M	RPL10	0.405	1.285	1.213	DM
TIMM10	ATP5B	0.902	1.274	1.227	DM
TIMM10	IMMT	0.558	1.274	1.453	DM
TNS1	SLC4A1	0.708	1.210	1.229	GE
TNS1	BCL2L1	0.498	1.210	1.144	GE

Table 3: random 10 item incidence list from conjoined STRING and expression data

The incidence list was converted into an adjacency matrix which included the weight for each edge. Table 4 is the 7x7 top-left extract from the original 210x210 matrix.

	AKD1	ANK1	ANKRD36	ARCN1	ASB7	ATP5B	ATP6V1A
AKD1	0	0	0	0	0	0.645	0.445
ANK1	0	0	0	0.901	0	0	0
ANKRD36	0	0	0	0	0	0	0
ARCN1	0	0.901	0	0	0	0	0
ASB7	0	0	0	0	0	0	0
ATP5B	0.645	0	0	0	0	0	0.928
ATP6V1A	0.445	0	0	0	0	0.928	0

Table 4: 7x7 extract from adjacency matrix of table 5 data

The adjacency matrix along with node attribute list (the FC value of each node) was then processed in the Gephi network analysis software. It resulted in a graph (Figure 6) in which proteins and their interactions as well as their expression value (indistinctly as methylation and RNA expression FC) are represented. The thickness of the edges indicates the confidence according to the STRING database and the colour of the nodes indicate the FC in the following scale: red for hypo-differentiated genes (FC<0), green for non-differentiated genes (FC=0) and blue for hyper-differentiated genes (FC>0). The nodes where distributed using the ForceAtlas 2 layout algorithm to better identify clusters.

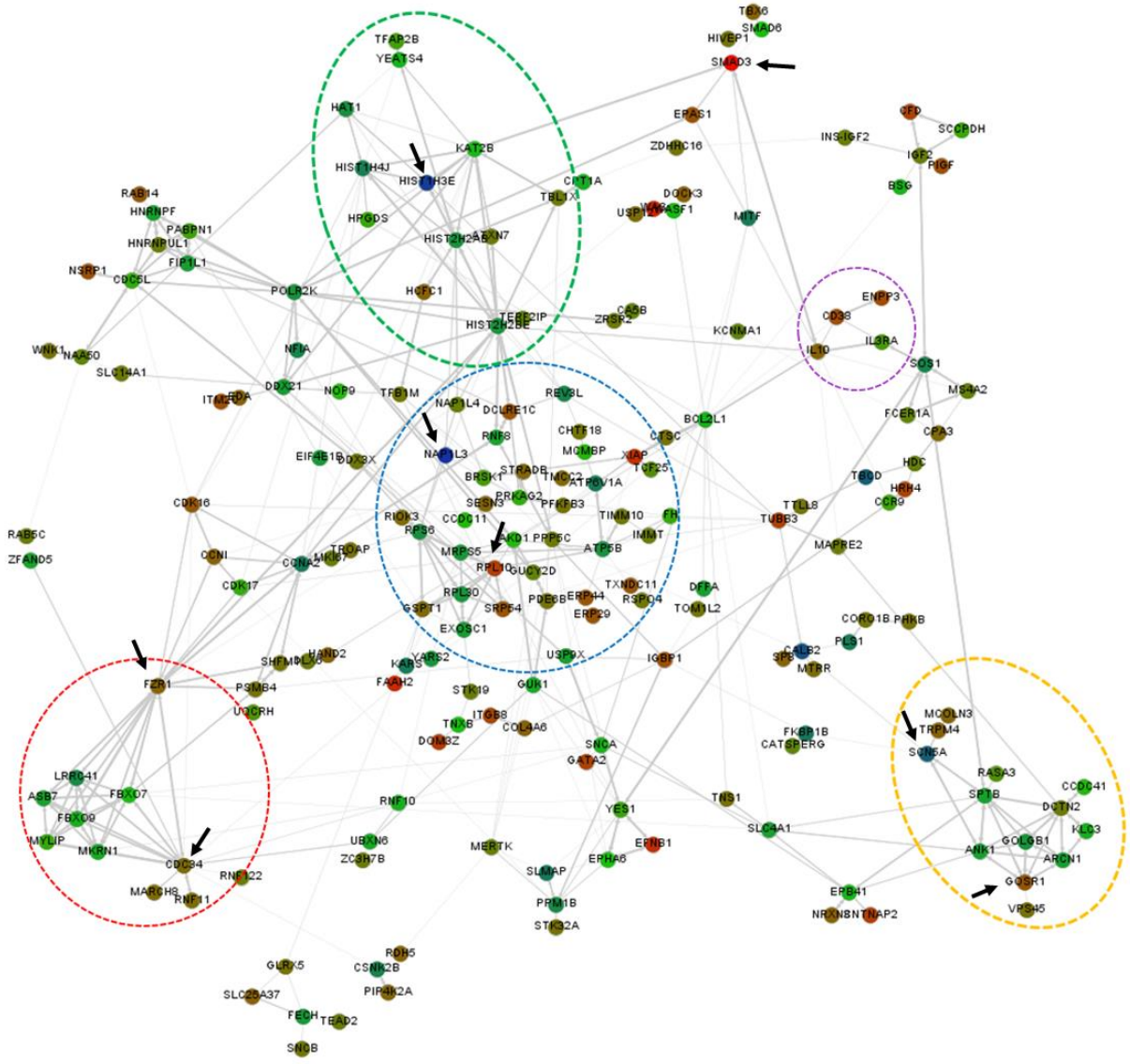


Figure 6: protein interaction network from DM, GE and STRING data. The thickness of the edges indicates the confidence according to the STRING database and the colour of the nodes indicate the FC (red for FC<0, green for FC=0, and blue for FC>0). Dotted lines represent clusters, defined imprecisely, of highly differentiated gene.

6. Discussion

6.1. Algorithm results

After computing the algorithm on any set of data, the results consist of the set of selected probes for each of the 10 iterations of the SVM-RFE, each containing k probes for all k in the corresponding array (27 and 34 different values of k for the GE and ME datasets, respectively). In addition to all probesets, we stored the accuracy values of the

2000 iterations on the 5-fold cross-validation, such that for any k we obtained 20.000 values of accuracy in percentage. The algorithm, although yielding good accuracy results, shown and discussed in subsequent sections, is deemed in need of improvement.

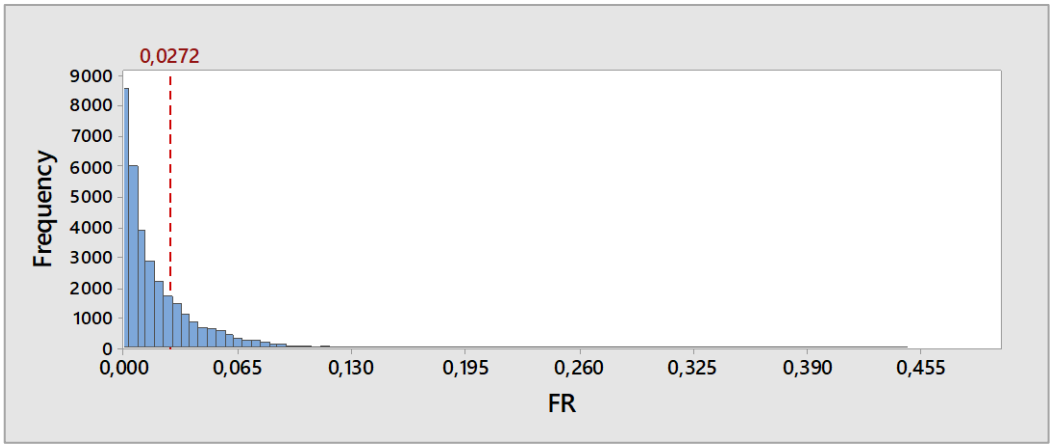


Figure 7: histogram of FR in GE dataset, the dotted line represents the 8000th FR value, data points below this are not included on the algorithm

We included only data from the 8000 top differentiated features through FR ranking, forcibly leaving out 76% of the available features in the GE dataset (Figure 7). Even though it might be logical that the mostly differentiated features between FMS patients and controls provide the better discriminatory power, genes, as stated, are redundant and this initial feature selection might leave important features out of the algorithm. In consequence and to fully analyse the data, the algorithm needs better implementation and optimization tools and higher computational power.

6.1.1. GE dataset

From the original 33297 features, the data input to the algorithm consisted of 8000 features with 141 samples (67 FMS and 75 healthy controls). The mean accuracy obtained in this dataset was 85.63% (CI 81.72- 89.54 %). However, the relationship between accuracy and the number of selected features is clear in Figure 8. The data points show a good linear correlation ($p<0,05$) from 64 to 256 features, reaching accuracy values between 92 and 96% (Figure 9). The actual highest accuracy, however, corresponds to the probesets with 56 features: 95.72% (CI 95.55-95.90%). Even though the regression yields better accuracy at longer probesets, we decided to further analyse the 56 features-probeset to favour minimal length over a small increase in accuracy.

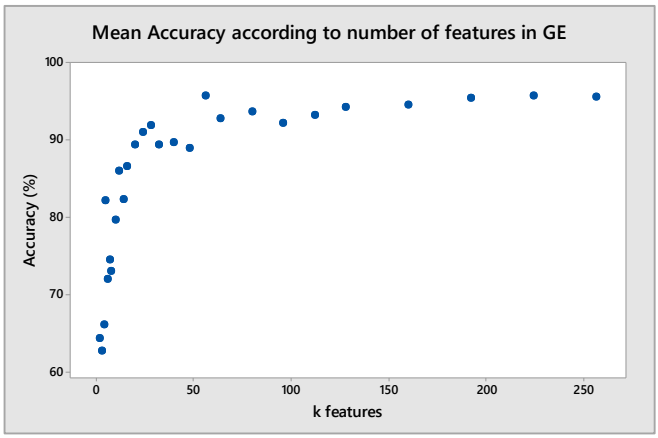


Figure 8: mean accuracy versus number of features achieved in SVM-RFE for the GE data

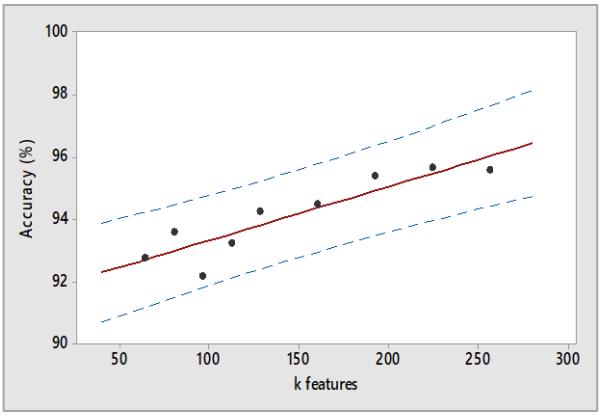


Figure 9: linear regression of mean accuracy versus k and CI region (dotted lines) for k≥64

The specific accuracy results on the 10 iterations of the 56 features-probeset are shown in Figure 10. These sets did not fully coincide, they all included some features but most of them were not shared between all 10 sets. As a consequence, we decided to list all appearing features and rank them according to their frequency between all probesets, such that probes that were included in all 10 sets were ranked first. We then selected the top 56 features to create a new probeset (TOP) that might, as a hypothesis, have better accuracy. Its results are also shown in Figure 10.

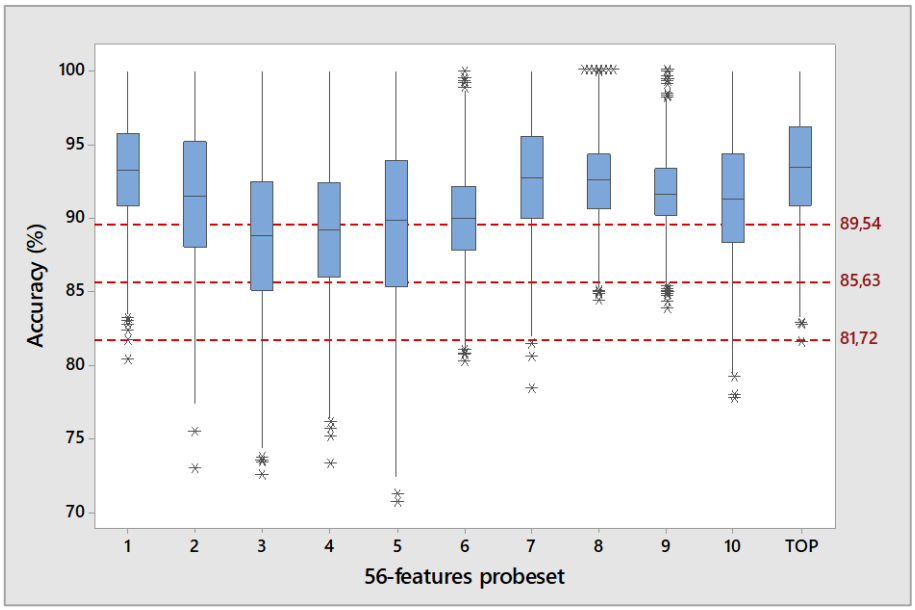


Figure 10: accuracy boxplots for all 10 iterations of SVM-RFE at k=56 and TOP probeset (56 most frequent probes in the previous sets); the dotted lines are the mean accuracy and CI for all k

The difference in accuracy was tested through a one-way analysis of variance with post-hoc Dunnett's test comparing each probeset to the TOP set. Figure 11 shows that the new set yields better results than most of the previous sets except for probeset 1. When contrasting them, turns out that 38 out of the 56 features in both sets are shared and this level of coincidence is highest among all comparisons between sets 1 and TOP.

This might provide an explanation to 1 and TOP not being statistically different in accuracy.

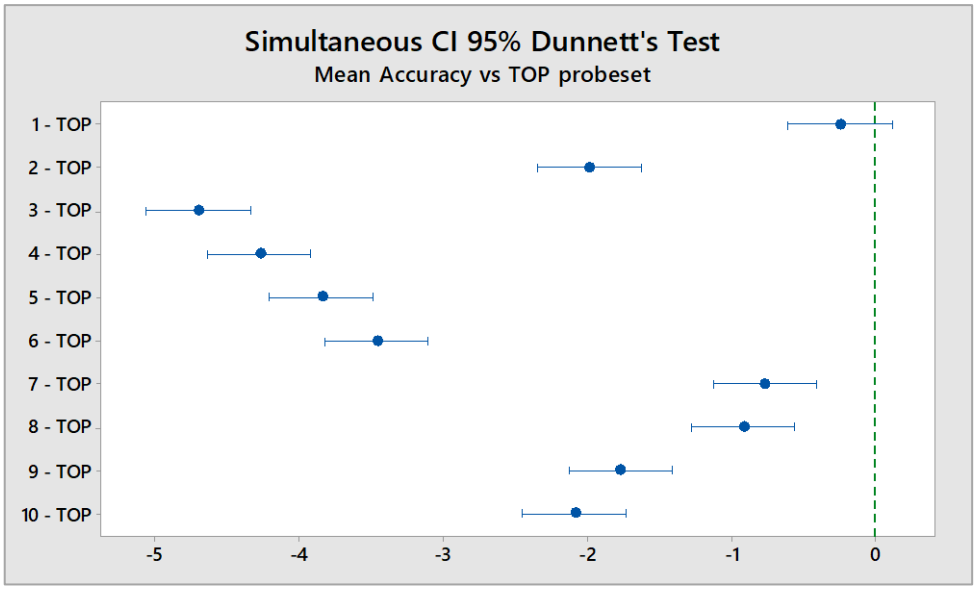


Figure 11: Dunnett's test results, if an interval does not contain zero, the mean accuracy is significantly different to the TOP probeset

Finally, we input all probe IDs from the TOP probeset into DAVID database so as to get the gene the probe corresponds to (Table 5).

Probe ID	Gene Name	Gene Symbol	FC	LogFC	FR
8083260	carboxypeptidase A3	CPA3	0.008	1.006	0.029
7927548	translocase of inner mitochondrial membrane 23	TIMM23	0.005	1.004	0.030
8179559	prefoldin subunit 6	PFDN6	0.016	1.011	0.030
7895158			0.021	1.015	0.030
8098342	Sin3A associated protein 30	SAP30	0.011	1.007	0.030
7990815	suppressor of tumorigenicity 20	ST20	0.027	1.019	0.030
8073842	tetratricopeptide repeat domain 38	TTC38	0.008	1.006	0.030
8046527	homeobox D12	HOXD12	-0.007	0.995	0.030
8048114			-0.016	0.989	0.031
8015349	keratin 19	KRT19	-0.012	0.991	0.031
8037315	pleckstrin homology like domain family B member 3	PHLDB3	0.007	1.005	0.031
8149289	SRY-box 7	SOX7	0.015	1.011	0.031
8008868			0.014	1.010	0.031
8139031			0.013	1.009	0.031
8171026	H2A histone family member B2	H2AFB2	-0.022	0.985	0.031
8058591	acyl-CoA dehydrogenase, long chain	ACADL	-0.012	0.992	0.031
7894806			0.015	1.010	0.031
8032804	SH3 domain containing GRB2 like 1, endophilin A2	SH3GL1	0.006	1.004	0.031
8169598	zinc finger CCHC-type containing 12	ZCCHC12	0.019	1.014	0.031
7895793			0.030	1.021	0.031
8154670	intraflagellar transport 74	IFT74	-0.006	0.996	0.031

8109350	solute carrier family 36 member 1	SLC36A1	-0.011	0.993	0.031
7994541	linker for activation of T-cells	LAT	-0.008	0.994	0.031
7893149			0.009	1.007	0.031
7986755	MAGE family member L2	MAGEL2	0.009	1.006	0.031
7913705	cannabinoid receptor 2	CNR2	-0.012	0.992	0.031
7894857			-0.024	0.984	0.031
8000676	nuclear pore complex interacting protein family member B5	NPIPB5	0.019	1.013	0.032
7909441	G0/G1 switch 2	G0S2	0.014	1.010	0.032
8120194	transcription factor AP-2 beta	TFAP2B	-0.010	0.993	0.032
8122705	protein-L-isoaspartate , D-aspartate O-methyltransferase	PCMT1	0.012	1.009	0.033
7965627	leukotriene A4 hydrolase	LTA4H	0.014	1.010	0.033
7971134	proline and serine rich 1	PROSER1	0.013	1.009	0.033
8124394	histone cluster 1 H2B family member b	HIST1H2BB	-0.008	0.994	0.033
7994308	KIAA0556	KIAA0556	-0.015	0.990	0.037
7970388	ankyrin repeat domain 20 family member A9, pseudogene	ANKRD20A9P	-0.066	0.955	0.037
7894365			0.018	1.013	0.039
7892545			0.017	1.012	0.041
7896535			0.017	1.012	0.044
8083463	chromosome 3 open reading frame 79	C3orf79	0.016	1.011	0.045
8149629	GDNF family receptor alpha 2	GFRA2	0.017	1.012	0.047
8158059	syntaxin binding protein 1	STXBP1	-0.021	0.986	0.050
8125447	major histocompatibility complex, class II, DQ beta 1	HLA-DQB1	-0.033	0.977	0.051
8069517	ankyrin repeat domain 20 family member A9, pseudogene	ANKRD20A9P	0.018	1.013	0.054
7895033			0.013	1.009	0.057
8090469	GATA binding protein 2	GATA2	0.017	1.012	0.058
8113664			0.020	1.014	0.059
8165752	interleukin 3 receptor subunit alpha	IL3RA	0.029	1.021	0.065
7925448			0.009	1.007	0.076
8044124	G protein-coupled receptor 45	GPR45	0.024	1.017	0.079
7972936	transmembrane protein 255B	TMEM255B	-0.024	0.983	0.082
7896649			0.021	1.014	0.099
7893252			-0.017	0.988	0.103
7949971	carnitine palmitoyltransferase 1A	CPT1A	-0.023	0.984	0.107
7894782			-0.030	0.979	0.109
8155248			0.038	1.027	0.139

Table 5: set of 56 gene expression probes that yields best accuracy results

A characteristic to note from these probes is that they do not correspond to the most differentiated probes in the whole GE dataset. They actually mostly fall on the low end of the FR distribution as seen in Figure 12. This result further confirms Guyon and colleagues' idea that the most differentiated genes might not provide the best discriminatory accuracy.

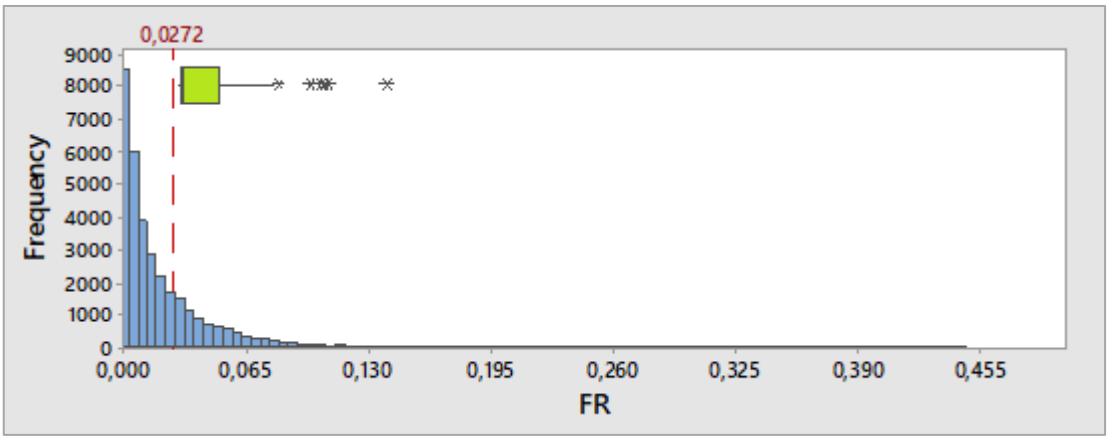


Figure 12: histogram of FR with superposed FR boxplot for the 56 probes included in the selected set

On the one hand, from the most selected features, only CPA3 corresponds to one of the most discriminatory genes according to Jones et al. They found that the most accurate probesets included only 10 features in contrast to the 56 we found. However, their 10-long probeset yielded an accuracy of about 95%, while our 56-long probeset yielded 95,72%. On the other hand, Lukkahatai and colleagues (23) found a 57 features-long set to be more accurate but none of their selected features matched our set. These results confirmed our initial suspicions that, given the stochastic nature of the algorithm (due to the 5-fold data split for validation and training and leaving 76% of probes out), the algorithm wouldn't yield the same probeset.

6.1.2. ME dataset

The original dataset consisted of 1212 features with expression coefficients for 21 samples (from 11 FMS patients and from 10 and healthy controls). Several values were registered in the GEO database as NULL. Even though the algorithm can compute all features (less than 8000), we decided to drop any feature with less than 10 samples or less than or equal to 5 or 6 in the control and the FMS classes, respectively. The reason is to only use the features whose data we deemed trustworthy as having less than 10 sample could prove statistically deficient. Those features were eliminated, and the resulting dataset was updated by substituting all remaining NULL values with -9999 such that the algorithm recognised them as outliers. In doing so, we did not get many data points, since only 259 features did not have missing attributes. The final dataset contained 454 features.

The mean accuracy obtained through all probesets was 98.25% (CI 97.96- 98.54 %) and there seemed to be no readily identifiable correlation between accuracy and the number of selected features (Figure 13). This, as well as the mean accuracy, is clearly different from the previous results as seen in Figure 14. The microRNA dataset used shows statistically higher ($p<0.0001$) accuracy results in classifying FMS patients from

controls than the RNA expression dataset. The highest accuracy corresponds to the probesets with 20 features: 98.95% (CI 98.88-99.01%).

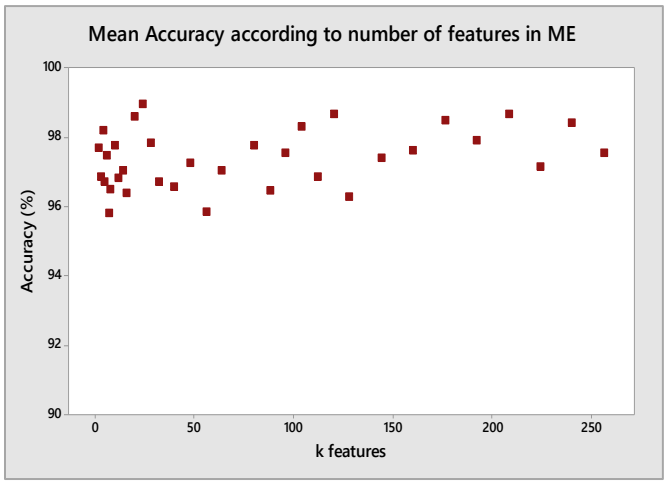


Figure 13: mean accuracy versus number of features achieved in SVM-RFE for the ME data

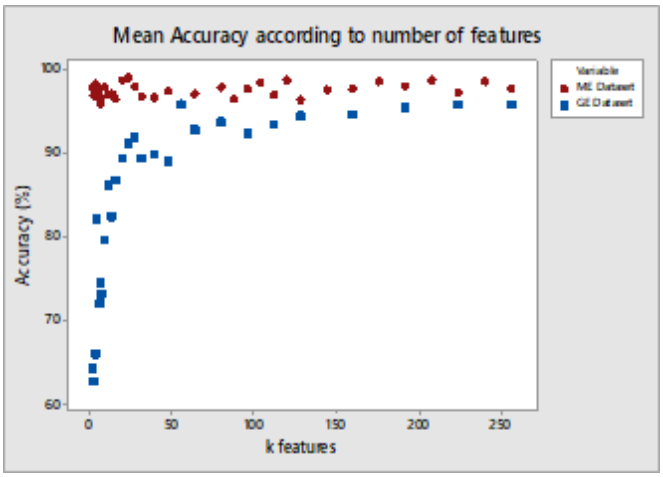


Figure 14: mean accuracy versus number of features achieved in SVM-RFE for ME (red) and GE (blue) datasets

Following the previous procedure, Figure 15 shows the specific accuracy results on the 10 iterations of the 20 features-probeset as well as the new TOP probeset (constructed with the most frequent probes). The results of the comparison between each iteration and the TOP set are also shown in Figure 16.

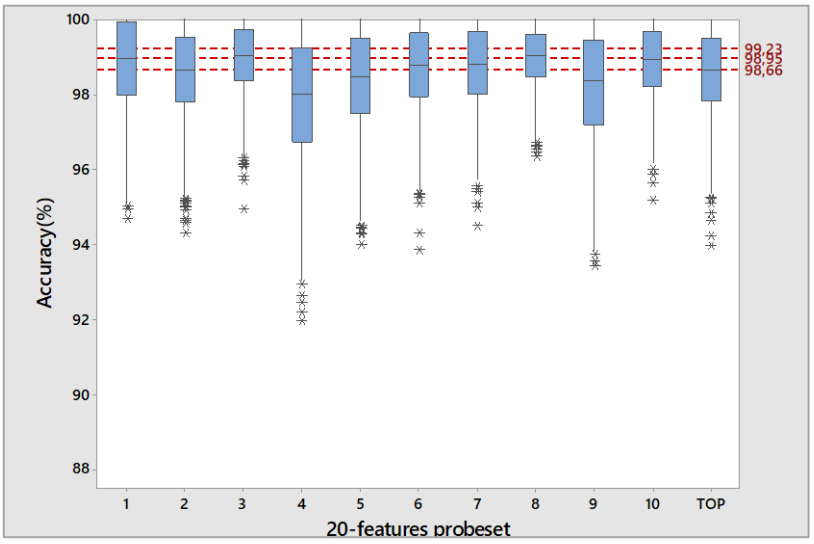


Figure 15: accuracy boxplots for all 10 iterations of SVM-RFE at k=20 and TOP probeset (56 most frequent probes in the previous sets); the dotted lines are the mean accuracy and CI for all k

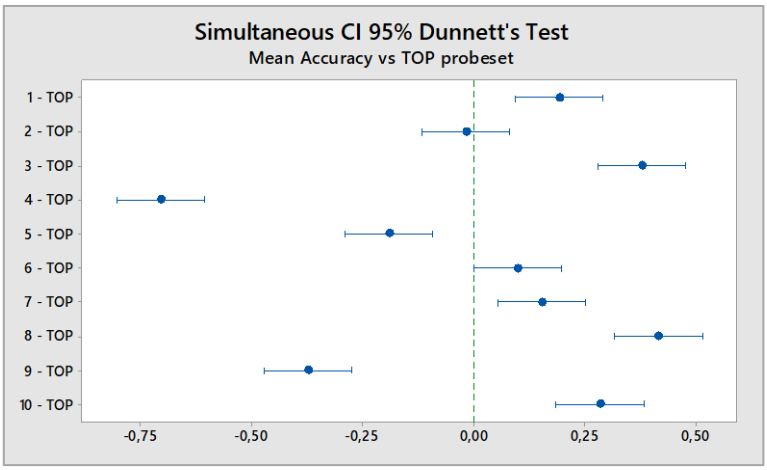


Figure 16: Dunnett's test results, if an interval does not contain zero, the mean accuracy is significantly different to the TOP probeset

The results state that the new probeset was not more accurate than most of the previous sets. However, since it is composed by the most appearing features in the other sets, we show its elements in Table 6, as well as the microRNA descriptor, FC.

Probe ID	microRNA	FC	2^{FC}
MIMAT0001631	hsa-miR-451	-3,679	0,078
MIMAT0000280	hsa-miR-223	-2,735	0,150
MIMAT0000078	hsa-miR-23a	-2,235	0,212
MIMAT0000433	hsa-miR-142-5p	-2,080	0,237
MIMAT0000067	hsa-let-7f	-1,999	0,250
MIMAT0000076	hsa-miR-21	-1,994	0,251
MIMAT0000101	hsa-miR-103	-1,989	0,252
MIMAT0000074	hsa-miR-19b	-1,985	0,253
MIMAT0000065	hsa-let-7d	-1,941	0,260
MIMAT0000418	hsa-miR-23b	-1,920	0,264
MIMAT0000062	hsa-let-7a	-1,905	0,267
MIMAT0000417	hsa-miR-15b	-1,883	0,271
MIMAT0000069	hsa-miR-16	-1,877	0,272
MIMAT0000100	hsa-miR-29b	-1,756	0,296
MIMAT0000082	hsa-miR-26a	-1,569	0,337
MIMAT0000086	hsa-miR-29a	-1,306	0,404
MIMAT0000414	hsa-let-7g	-1,057	0,481
MIMAT0016916	hsa-miR-4286	-0,991	0,503
MIMAT0000451	hsa-miR-150	-0,609	0,656
MIMAT0015041	hsa-miR-1260b	-0,369	0,774

Table 6: set of 20 miRNA expression probes that yields best accuracy results

From this set of miRNA probes we can see all features present down-regulation from healthy controls and in all but 3 cases it is more than 50% decreased (FC<-1).

6.2. General analysis

DNA methylation is a quite relevant method of gene expression regulation, therefore, important in embryogenesis, genetic imprinting, and X chromosome inactivation. Methylation profiles change across the genome due to epigenetic factors and the aging process (30). DNA methylome analysis as a mean to better understand causes and/or possible treatments for FMS is, therefore, a good course of action. Gene expression changes might be determinant to some symptoms or, maybe, acquired epigenetic changes are the underlying cause for the onset of this disorder.

According to Menzies et al. (31) significant differences in DNA methylation patterns were found between healthy controls and FMS patients. Mostly, those changes were due to an increased methylation in women with FMS, located in relevant biological clusters involved in chromatin compaction, nervous system development and skeletal/organ system development. They also compared the frequency of spontaneously occurring micronuclei, which are small nuclei that occur when a chromosome or a fragment is not integrated in one of the daughter nuclei during mitosis. They usually indicate genotoxic episodes and chromosomal instability, commonly seen in cancerous cells (32). They may also increase the risk of developmental or degenerative diseases. The mean micronuclei incidence of women with FMS was significantly higher than that of healthy controls.

The DM data (Table 1) includes the column "Relation to UCSC CpG Island" which indicates the position of the differently methylated nucleotide with respect to a CpG island (included in the USCS Genome Browser database). One of the basis of DNA methylation is the transfer of a methyl group from S-adenosyl-L-methionine to the cytosine of a CpG dinucleotide (cytosine and guanine nucleotides adjacent in the same strand) (33). This methylation in cytosine mostly occurs when being adjacent and five prime to guanine (hence, the nomenclature of CpG). DNA methylation is a major mechanism to modulate chromatin access of transcription factors and the basal transcriptional machinery. About 15% of the CpG sites are inside CpG islands in the promoters of some 70% of protein-coding genes (33). CpG islands are generally between 300 and 3000 bp long and with a GC content of greater than 50%. It is clear that CpG site methylation plays a major role in gene expression modulation (33). Furthermore, CpG islands that previously seemed to not be associated with any known genes, have been associated with long non-coding RNA (lncRNA), miRNAs and other non-coding genes, and these orphan CpG islands may be important in the control of non-coding RNA expression (34). In DM, the great majority of most differently methylated sites were in N-shores in relation to a CpG island, giving rise to the idea that FMS might be significantly related to gene dysregulations.

In order to further analyse the interaction network, we obtained in section 5.2, it would be best to have the mean values of either methylation or RNA expression for each gene. However, and since we lack any further data we deemed best not to alter the data

so as to try to unfoundedly convert gene methylation into RNA expression values. Hence, FC values are taken into consideration as a general gene dysregulation. We located nodes that had either red or blue colour and preferably high degrees and heavy edges (indicated with arrows in Figure 6). Data from the GeneCards database provided information on the function and/or properties of every relevant node (Table 7).

CDC34	Ubiquitin-conjugating enzyme; catalyzes the covalent attachment of ubiquitin to other proteins. This protein is a part of a large multiprotein complex required for ubiquitin-mediated degradation of cell cycle G1 regulators, and for the initiation of DNA replication.
FZR1	Related to pathways such as CDK-mediated phosphorylation and removal of Cdc6 and Development of TGF-beta receptor signaling. Substrate-specific adapter for the anaphase promoting complex/cyclosome E3 ubiquitin-protein ligase complex, degrading substrates to ensure that positive regulators of the cell cycle do not accumulate prematurely.
GOSR1	Trafficking membrane protein which transports proteins among the endoplasmic reticulum and the Golgi and between Golgi compartments, considered an essential component of the Golgi SNAP receptor complex.
HIST1H3E	H3 histone, basic nuclear protein responsible for the nucleosome structure. Transcripts from this gene lack polyA tails but instead contain a palindromic termination element. This gene is found in the large histone gene cluster on chromosome 6.
NAP1L3	Member of the nucleosome assembly protein family. Linked closely to a region of genes responsible for several X-linked cognitive disability syndromes.
RPL10	Ribosomal protein that is a component of the 60S ribosome subunit.
SCN5A	Integral membrane protein and tetrodotoxin-resistant voltage-gated sodium channel subunit. Found primarily in cardiac muscle and responsible for the initial upstroke of the action potential in an electrocardiogram. Defects in this gene are a cause of long QT syndrome type 3
SMAD	Signal transducer and transcriptional modulator that mediate multiple signaling pathways. Transcriptional modulator activated by transforming growth factor-beta and is thought to play a role in the regulation of carcinogenesis.

Table 7: GeneCards database information on fig. 4 relevant nodes (available from: http://www.genecards.org/ [23/12/19])

Afterwards we analysed their neighbours to identify possible dysregulation pathways. Finally, we identified five non-specific gene clusters that represent groups of highly related and relevant genes. The green gene cluster includes proteins related to DNA transcription regulation. The blue gene cluster is characterised by ribosomal and mRNA splicing proteins, thus, generally it is involved in translation. Ubitiquin and cytoskeleton-related proteins are included in the red gene cluster. Cytoskeleton and membrane-related proteins, crucial to cell proliferation, activation and even

axonogenesis, form the yellow protein cluster. Finally, several inflammation-related proteins conform the purple gene cluster.

Apart from applying the classification algorithm, from all ME data, we computed FC for all features that were not dropped due to lack of attributes. Table 8 presents the expression data for the top 10 most hypo-expressed miRNA. Two miRNA presented about 2-fold upregulation: miR-302e and miR-488*.

P value	FC	miRNA
1.80e-09	-3.998	miR-143
2.80e-09	-3.889	miR-145
3.98e-08	-3.679	miR-451
8.98e-12	-3.498	miR-338-3p
8.26e-11	-3.030	miR-148a
9.95e-09	-2.783	miR-376c
2.25e-08	-2.750	miR-126*
1.27e-11	-2.735	miR-223
5.58e-09	-2.725	miR-199a-3p, miR-199b-3p
7.49e-09	-2.697	miR-424

Table 8: top 10 items from ME dataset (in increasing order of FC)

From the most de-regulated miRNA, we found interesting documented interactions and biological functions for three: miR-145, miR-451 and miR-223. On the one hand, Sun et al. (35) identified miR-142 and miR-223 to be haematopoietic miRNA and miR-223 to have crucial functions in myeloid lineage development. Even though the function of miR-142 wasn't fully discovered, both microRNAs presented an attenuating function on haematopoietic cells and miR-223 upregulated miR-142 expression, thus discovering a new regulating pathway between these microRNAs that is very relevant to haematopoiesis.

On the other hand, Hu and colleagues (36) demonstrated the crucial role of miR-145 in the regulation of TNF-α-mediated signalling and cartilage matrix degradation. miRNA expression profiles of TNF-α-stimulated chondrocytes showed that miR-145 expression was quickly downregulated by TNF-α. They even found that miR-145 directly targeted MKK4 (mitogen-activated protein kinase kinase 4) and largely restrained the synthesis of several TNF-α-triggered matrix-degrading enzymes. MKK4 hyper-expression increased TNF-α-mediated signalling activation, and therefore worsened cartilage degradation. Furthermore, they found that intra-articular injection of miR-145 agonist to rats with surgery-induced osteoarthritis prevented or decreased cartilage destruction. Interestingly, Ohgidani et al. (37) studied TNF-α expression levels in the central nervous system in FMS patients by transforming blood cells into microglia-like cells. They found that TNF-α was hyper-expressed in FMS microglia-like cells. Furthermore, they discovered that there was a moderate correlation between TNF-α expression upregulation and clinical parameters of subjective pain and other mental

manifestations. Therefore, the hypo-expression of miR-145 matches the hyper-expression of TNF-α. Likewise, according to Sun and colleagues (38), miR-451 may relieve chronic inflammatory pain by inhibiting microglia activation-mediated inflammation via targeting TLR4. They used complete Freund's adjuvant (CFA)-induced inflammatory pain mice model and their results show the expression of miR-451 was decreased in spinal microglia. Microglia-mediated neuroinflammation in spinal cord is key in the pathogenesis of chronic inflammatory pain. They further confirmed the anti-inflammatory effects of miR-451, specifically: miR-451 overexpression antagonized microglial activation-induced proinflammatory cytokine releases, including IL-6, IL-1β, and TNF-α. The down-regulation of both miR-145 and miR-451 shown in FMS could be synergic in their regulation of inflammatory pathways.

From miRNA selected as the classification probeset, it is interesting to note information on: miR-16, miR-21, miR-103, miR-26a and miR-150, apart from the already explained implications of miR-223 and miR-451. The first three microRNAs have been related to growth cell regulation in some form of cancer cell. On the one hand, Cutrona et al. (39) state that chronic lymphocytic leukaemia (CLL) clones lack a critical region involving miR-15a and miR-16-1. When those CLL cells were transfected with the miRNA mimics they showed a decrease in cell viability *in vitro* and substantial tumour regression in NSG mice previously engrafted with CLL clones. On the other hand, Masoudi and colleagues (40) analysed miRNA expression in glioblastoma multiform (GBM) cells. The miRNA most involved in GBM pathogenesis was miR-21. Other studies have also reported that de-regulation of this miRNA could alter a variety of molecular pathways such as insulin-like growth factor-binding protein-3 (IGFBP3), RECK and TIMP3. Finally, another study regarding miRNA interaction in cancer cell growth was conducted by Chen et al. (41). They found that miR-103, miR-195 or miR-15b were downregulated in glioma tissues and cell lines, the common highly malignant primary brain tumour. These studies show that miRNAs can indeed play a part in several key regulatory pathways. Since, to our knowledge, there is no evidence of correlation between FMS and cancer, this miRNA down-regulation might not have pivotal implication in FMS pathogenesis. However, taking into consideration that miRNAs are still subject of study, miR-16, miR-21, miR-103 might provide further alterations in cell activity in FMS.

Regarding miR-26a, a study carried out by chinese researchers (42) concluded that forced miR-26a/26b expression was able to affect chondrocytes proliferation and apoptosis. They analysed the effect of these miRNAs on chondrocytes to evaluate their impact on osteoarthritis. In osteoarthritic mice, the overexpression of miR-26a/26b by intra-articular injection significantly attenuated the disease's progression. Actually, both osteoarthritis and FMS typically present chronic pain (43) and comparing the involvement of these miRNAs in both cases might provide new drug targets and or treatment. Precisely, some studies reported that FMS patients with comorbid osteoarthritis or myofascial pain had improvement in their overall FMS pain and tenderness when treated with local therapies (44). Finally, Ji et al. (45) explored the

potential role of miR-150 in regulating the process of neuropathic pain in a rat model established by chronic sciatic nerve injury (CCI). They showed that overexpression of miR-150 greatly alleviated neuropathic pain development and reduced inflammatory cytokine expression, including COX-2, IL-6 and TNF-α in CCI rats. They further proved that miRNAs are key participators in the pathophysiological course of neuropathic pain.

7. Conclusions

- Across most of the studies reviewed, it was recurrent to find that in FMS patients there is some dysregulation in neural system genesis and in inflammatory pathways.

- miRNAs are key participators in the pathophysiological course of neuropathic pain. We found that TNF-α/miR-145 regulatory pathway might be an important factor in FMS aetiology. In addition, miR-150 over-expression alleviates neuropathic pain development and reduces inflammatory cytokine expression, including TNF-α, and might also contribute to the previous relationship. The hypo-expression of miR-145 and miR-451 could be synergic in their regulation of inflammatory pathways.

- Given that some of epigenetic alterations seen in FMS could cause significant dysregulation in important proteins and non-coding RNA, they could be potential treatment targets. Overexpressed genes could be downregulated with antisense RNA molecules, for instance. Epigenetic drugs might also potentially reverse abnormal gene expression profiles associated with FMS.

- A SVM-RFE algorithm tested through holdout crossvalidation provided a set of 56 RNA probes with an accuracy of 95.72% and a set of 20 microRNA probes with 98.95%. According to our results, miRNA may be the best genetic biomarker to diagnose FMS. There should be further clinical trials and studies to assess diagnostic sensitivity and specificity.

- The algorithm, however, needs further improvements like the implementation of better internal validation techniques and optimization strategies. It could not include as many data points as hoped for and the results of miRNA probes is based on a very limited dataset. The results of the algorithm are not meant to be conclusive but to serve as a steppingstone to future studies.

- There is a lack of big sample size genetic studies and with precise exclusion criteria (to account for FMS comorbities). FMS genetic investigation is still in its infancy and the incoming results look promising. We advise future work on the subject to finally unveil fibromyalgia's aetiology and provide accurate and useful diagnosis and treatment.

8. Bibliography

1. Neuprez A, Crielaard JM. Fibromyalgie: État de la question en 2017. Rev Med Liege. 2017;72(6):288–94.
2. Quintner J. Why Are Women with Fibromyalgia so Stigmatized? Pain Med. 2020;0(0):1–5.
3. Wolfe F, Smythe HA, Yunus MB, Bennett RM, Bombardier C, Goldenberg DL, et al. The American College of Rheumatology 1990 Criteria for the Classification of Fibromyalgia. Report of the Multicenter Criteria Committee. Arthritis Rheum. 1990;33(2):160–72.
4. Wolfe F, Clauw DJ, Fitzcharles MA, Goldenberg DL, Katz RS, Mease P, et al. The American College of Rheumatology preliminary diagnostic criteria for fibromyalgia and measurement of symptom severity. Arthritis Care Res. 2010;62(5):600–10.
5. Wolfe F, Clauw DJ, Fitzcharles MA, Goldenberg DL, Häuser W, Katz RL, et al. 2016 Revisions to the 2010/2011 fibromyalgia diagnostic criteria. Semin Arthritis Rheum [Internet]. 2016;46(3):319–29. Available from: http://dx.doi.org/10.1016/j.semarthrit.2016.08.012
6. Docampo E, Collado A, Escaramís G, Carbonell J, Rivera J, Vidal J, et al. Cluster Analysis of Clinical Data Identifies Fibromyalgia Subgroups. PLoS One. 2013;8(9):1–7.
7. Buskila D, Sarzi-Puttini P, Ablin JN. The genetics of fibromyalgia syndrome. Pharmacogenomics. 2007;8(1):67–74.
8. Koca T, Koçyiğit B, Seyithanoğlu M, Berk E. The importance of G-protein coupled estrogen receptor in patients with fibromyalgia. Arch Rheumatol. 2019;34(4):419–25.
9. Brusco I, Justino AB, Silva CR, Fischer S, Cunha TM, Scussel R, et al. Kinins and their B1 and B2 receptors are involved in fibromyalgia-like pain symptoms in mice. Biochem Pharmacol [Internet]. 2019;168(June):119–32. Available from: https://doi.org/10.1016/j.bcp.2019.06.023
10. Ramírez-Tejero JA, Martínez-Lara E, Rus A, Camacho MV, Del Moral ML, Siles E. Insight into the biological pathways underlying fibromyalgia by a proteomic approach. J Proteomics [Internet]. 2018;186(July):47–55. Available from: https://doi.org/10.1016/j.jprot.2018.07.009
11. Clos-Garcia M, Andrés-Marin N, Fernández-Eulate G, Abecia L, Lavín JL, van Liempd S, et al. Gut microbiome and serum metabolome analyses identify molecular biomarkers and altered glutamate metabolism in fibromyalgia. EBioMedicine. 2019;46:499–511.
12. Estévez-López F, Camiletti-Moirón D, Aparicio VA, Segura-Jiménez V, Álvarez-Gallardo IC, Soriano-Maldonado A, et al. Identification of candidate genes associated with fibromyalgia susceptibility in southern Spanish women: The al-Ándalus project. J Transl Med [Internet]. 2018;16(1):1–6. Available from: https://doi.org/10.1186/s12967-018-1416-8
13. de la Coba P, Bruehl S, del Paso GAR. Addition of Slowly Repeated Evoked Pain Responses to Clinical Symptoms Enhances Fibromyalgia Diagnostic Accuracy. Pain Med. 2019;0(0):1–9.
14. Malatji BG, Meyer H, Mason S, Engelke UFH, Wevers RA, Reenen M, et al. A diagnostic biomarker profile for fibromyalgia syndrome based on an NMR metabolomics study of selected patients and controls. BMC Neurol. 2017;17(1):1–15.
15. Ciregia F, Giacomelli C, Giusti L, Boldrini C, Piga I, Pepe P, et al. Putative salivary biomarkers useful to differentiate patients with fibromyalgia. J Proteomics [Internet]. 2019;190:44–54. Available from: https://doi.org/10.1016/j.jprot.2018.04.012
16. Docampo E, Escaramís G, Gratacòs M, Villatoro S, Puig A, Kogevinas M, et al. Genome-

wide analysis of single nucleotide polymorphisms and copy number variants in fibromyalgia suggest a role for the central nervous system. Pain. 2014;155(6):1102–9.

17. Ciampi De Andrade D, Maschietto M, Galhardoni R, Gouveia G, Chile T, Victorino Krepischi AC, et al. Epigenetics insights into chronic pain: DNA hypomethylation in fibromyalgia - A controlled pilot-study. Pain. 2017;158(8):1473–80.

18. Jones KD, Gelbart T, Whisenant TC, Waalen J, Mondala TS, Iklé DN, et al. Genome-wide expression profiling in the peripheral blood of patients with fibromyalgia HHS Public Access. Clin Exp Rheumatol. 2016;34(2):89–98.

19. Wahid F, Shehzad A, Khan T, Kim YY. MicroRNAs: Synthesis, mechanism, function, and recent clinical trials. Biochim Biophys Acta - Mol Cell Res [Internet]. 2010;1803(11):1231–43. Available from: http://dx.doi.org/10.1016/j.bbamcr.2010.06.013

20. Cerdá-Olmedo G, Mena-Durán AV, Monsalve V, Oltra E. Identification of a MicroRNA signature for the diagnosis of fibromyalgia. PLoS One. 2015;10(3):1–14.

21. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using Support Vector Machines. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2008;5139 LNAI:62–72.

22. Hastie T, Tibshirani R, Friedman J. 12. Support Vector Machines and Flexible Discriminants. In: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer Series in Statistics Trevor; 2008. p. 417–55.

23. Lukkahatai N, Walitt B, Deandrés-Galiana EJ, Fernández-Martínez JL, Saligan LN. A predictive algorithm to identify genes that discriminate individuals with fibromyalgia syndrome diagnosis from healthy controls. J Pain Res. 2018;11:2981–90.

24. Saligan LN, Fernández-Martínez JL, de Andrés-Galiana EJ, Sonis S. Supervised classification by filter methods and recursive feature elimination predicts risk of radiotherapy-related fatigue in patients with prostate cancer. Cancer Inform. 2014;13:141–52.

25. Lukkahatai N, Walitt B, Espina A, Wang D, Saligan LN. Comparing Genomic Profiles of Women With and Without Fibromyalgia. Biol Res Nurs. 2015;17(4):373–83.

26. Burges C. A Tutorial on Support Vector Machines for Pattern Recognition. Data Min Knowl Discov. 1998;2:121–67.

27. Adorada A, Permatasari R, Wirawan PW, Wibowo A, Sujiwo A. Support Vector Machine - Recursive Feature Elimination (SVM - RFE) for Selection of MicroRNA Expression Features of Breast Cancer. 2018 2nd Int Conf Informatics Comput Sci ICICoS 2018. 2019;(1):165–8.

28. Steyerberg EW, Harrell F, Borsboom G, Eijkemans M, Vergouwe Y, Habbema J. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;(54):774–81.

29. Efron B, Tibshirani R. Improvements on cross-validation: The .632+ bootstrap method. J Am Stat Assoc. 1997;92(438):548–60.

30. Petronis A. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. Nature. 2010;465(7299):721–7.

31. Menzies V, Lyon DE, Archer KJ, Zhou Q, Brumelle J, Jones KH, et al. Epigenetic Alterations and an Increased Frequency of Micronuclei in Women with Fibromyalgia. Nurs Res Pract [Internet]. 2013;2013:1–12. Available from: http://www.hindawi.com/journals/nrp/2013/795784/

32. Spektor A, Umbreit NT, Pellman D. Cell Biology: When Your Own Chromosomes Act like

Foreign DNA. Curr Biol [Internet]. 2017;27(22):R1228–31. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0960982217312411

33. Long MD, Smiraglia DJ, Campbell MJ. The genomic impact of DNA CpG methylation on gene expression; relationships in prostate cancer. Biomolecules. 2017;7(1):1–20.

34. Aune TM, Crooke PS, Patrick AE, Tossberg JT, Olsen NJ, Spurlock CF. Expression of long non-coding RNAs in autoimmunity and linkage to enhancer function and autoimmune disease risk genetic variants. J Autoimmun [Internet]. 2017;81:99–109. Available from: http://dx.doi.org/10.1016/j.jaut.2017.03.014

35. Sun W, Shen W, Yang S, Hu F, Li H, Zhu TH. MiR-223 and miR-142 attenuate hematopoietic cell proliferation, and miR-223 positively regulates miR-142 through LMO2 isoforms and CEBP-B. Cell Res [Internet]. 2010;20(10):1158–69. Available from: http://dx.doi.org/10.1038/cr.2010.134

36. Hu G, Zhao X, Wang C, Geng Y, Zhao J, Xu J, et al. MicroRNA-145 attenuates TNF-α-driven cartilage matrix degradation in osteoarthritis via direct suppression of MKK4. Cell Death Dis [Internet]. 2017;8(10):e3140. Available from: http://www.nature.com/doifinder/10.1038/cddis.2017.522

37. Ohgidani M, Kato TA, Hosoi M, Tsuda M, Hayakawa K, Hayaki C, et al. Fibromyalgia and microglial TNF-α: Translational research using human blood induced microglia-like cells. Sci Rep [Internet]. 2017;7(1):1–6. Available from: http://dx.doi.org/10.1038/s41598-017-11506-4

38. Sun X, Zhang H. miR-451 elevation relieves inflammatory pain by suppressing microglial activation-evoked inflammatory response via targeting TLR4. Cell Tissue Res. 2018;374(3):487–95.

39. Cutrona G, Matis S, Colombo M, Massucco C, Baio G, Valdora F, et al. Effects of miRNA-15 and miRNA-16 expression replacement in chronic lymphocytic leukemia: Implication for therapy. Leukemia [Internet]. 2017;31(9):1894–904. Available from: http://dx.doi.org/10.1038/leu.2016.394

40. Masoudi MS, Mehrabian E, Mirzaei H. MiR-21: A key player in glioblastoma pathogenesis. J Cell Biochem. 2018;119(2):1285–90.

41. Chen LP, Zhang NN, Ren XQ, He J, Li Y. miR-103/miR-195/miR-15b regulate SALL4 and inhibit proliferation and migration in glioma. Molecules. 2018;23(11).

42. Hu J, Wang Z, Pan Y, Ma J, Miao X, Qi X, et al. MiR-26a and miR-26b mediate osteoarthritis progression by targeting FUT4 via NF-κB signaling pathway. Int J Biochem Cell Biol [Internet]. 2018;94:79–88. Available from: http://dx.doi.org/10.1016/j.biocel.2017.12.003

43. López-Ruiz M, Losilla JM, Monfort J, Portell M, Gutiérrez T, Poca V, et al. Central sensitization in knee osteoarthritis and fibromyalgia: Beyond depression and anxiety. PLoS One. 2019;14(12):1–17.

44. Affaitati G, Costantini R, Fabrizio A, Lapenna D, Tafuri E, Giamberardino MA. Effects of treatment of peripheral pain generators in fibromyalgia patients. Eur J Pain [Internet]. 2011;15(1):61–9. Available from: http://dx.doi.org/10.1016/j.ejpain.2010.09.002

45. Ji LJ, Shi J, Lu JM, Huang QM. MiR-150 alleviates neuropathic pain via inhibiting toll-like receptor 5. J Cell Biochem. 2018;119(1):1017–26.