BMC
Genomics

RESEARCH

Open Access

A machine learning heuristic to identify biologically relevant and minimal biomarker panels from omics data

Anna L Swan[1], Dov J Stekel[1], Charlie Hodgman[1,2], David Allaway[4], Mohammed H Alqahtani[5], Ali Mobasheri[2,3,5,6], Jaume Bacardit[2,7*]

Abstract

Background: Investigations into novel biomarkers using omics techniques generate large amounts of data. Due to their size and numbers of attributes, these data are suitable for analysis with machine learning methods. A key component of typical machine learning pipelines for omics data is feature selection, which is used to reduce the raw high-dimensional data into a tractable number of features. Feature selection needs to balance the objective of using as few features as possible, while maintaining high predictive power. This balance is crucial when the goal of data analysis is the identification of highly accurate but small panels of biomarkers with potential clinical utility. In this paper we propose a heuristic for the selection of very small feature subsets, via an iterative feature elimination process that is guided by rule-based machine learning, called RGIFE (Rule-guided Iterative Feature Elimination). We use this heuristic to identify putative biomarkers of osteoarthritis (OA), articular cartilage degradation and synovial inflammation, using both proteomic and transcriptomic datasets.

Results and discussion: Our RGIFE heuristic increased the classification accuracies achieved for all datasets when no feature selection is used, and performed well in a comparison with other feature selection methods. Using this method the datasets were reduced to a smaller number of genes or proteins, including those known to be relevant to OA, cartilage degradation and joint inflammation. The results have shown the RGIFE feature reduction method to be suitable for analysing both proteomic and transcriptomics data. Methods that generate large 'omics' datasets are increasingly being used in the area of rheumatology.

Conclusions: Feature reduction methods are advantageous for the analysis of omics data in the field of rheumatology, as the applications of such techniques are likely to result in improvements in diagnosis, treatment and drug discovery.

Background

The 'omics' (genomics, epigenomics, transcriptomics, proteomics, metabolomics and lipidomics) are making significant contributions to the study of chronic diseases, especially the identification of novel biomarkers. A biomarker is defined as a characteristic that may be objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention [1]. Biomarkers are actively investigated in the areas of clinical rheumatology, orthopaedics and sports medicine. Osteoarthritis (OA) is a degenerative joint disease that affects the entire joint structure [2]. It is characterised by progressive degeneration of cartilage, menisci, ligaments and subchondral bone [3,4]. Synovial inflammation (synovitis) is a major contributor to disease progression [5-7] and is responsible for the increased production of catabolic and pro-inflammatory mediators that alter the

* Correspondence: Jaume.Bacardit@newcastle.ac.uk
[2]The D-BOARD European Consortium for Biomarker Discovery, The Universities of Surrey, Nottingham and Newcastle, United Kingdom
Full list of author information is available at the end of the article

balance of cartilage matrix degradation and repair, leading to excess production of the proteolytic enzymes responsible for cartilage breakdown [6]. OA is currently diagnosed by radiography, once clinical signs of pain and loss of mobility have already appeared, and therefore biomarkers that could identify early signs of OA would significantly aid in diagnosis [8]. Current research is aimed at identifying panels of clinically useful biochemical and imaging markers into single diagnostic algorithms that can be used for diagnostic and prognostic applications and for testing the efficacy of new drugs [9]. Applying 'omics' results in the generation of large datasets that are suitable for bioinformatic analysis using machine learning, to extract important information [10].

Bioinformatics tools play an important role in the analysis of data from omics technologies, such as microarrays, next generation sequencing and mass spectrometry (MS), and as a result a wide range of methods have been developed [11,12]. Such methods include supervised machine learning (ML) techniques, which are used to build classification models. Models are used to automatically label samples of unknown class by using a training set of known labelled samples. There are many different types of ML methods, some of which can be used to identify putative biomarkers from data by observing the attributes (genes or proteins) used to build the models. Rule-based methods are an example of this, as it is possible to read the rules generated to form the model [13].

BioHEL is a rule based machine learning method which has been used for sample classification in highly dimensional datasets because of its fine-grained embedded feature selection [14]. It has been successfully applied to accurately classify many different types of biological data [15-18]. Rule-based methods construct rule sets that contain at least one rule for each sample group, based on the values associated with the attributes, for example the expression value of the genes. An example of a rule set is shown in Figure 1. ML can also be used to identify possible biomarkers in the form of feature selection (FS), a method of data reduction. FS techniques identify a subset of attributes, for example genes or proteins, which could be used

to build a more successful model, compared to using the whole dataset.

Supervised FS methods analyse data with known class labels, with the aim to remove irrelevant or redundant features. Using FS techniques can improve model performance, increase classification accuracy, and provide a clearer understanding of attributes, which are useful for generating an effective classification model. Their key challenge is to avoid a loss of information by dropping features that contain crucial information for the data being analysed. The application of FS methods is particularly relevant for transcriptomic datasets that often have very large numbers of attributes, in relation to the number of samples. There are three main forms of feature selection: filter, wrapper and embedded methods [19]. Filter methods, such as correlation-based feature selection [20], are used to estimate the effectiveness of a reduction in attributes. Wrapper methods, including Genetic Algorithms [21], use classifiers to determine if the subset of attributes gives successful classifications. The third method, embedded, includes feature reduction within the process of classification and uses such machine learning techniques as Naïve Bayes [22] and Support Vector Machines (SVM) [23,24].

The aim of this study was to develop an FS heuristic specialized on identifying very reduced sets of variables built on top of the BioHEL rule-based ML method, in order to increase the accuracy of classification models and to identify putative biomarkers. This heuristic performs an iterative feature elimination process that is guided (in choosing which features to eliminate first) by an analysis of the rule sets generated by BioHEL. This guiding process avoids numerous iterations of blind trial-and-error attempts at removing features from the dataset, and quickly finds, in most datasets, very reduced subsets of features. We call this method RGIFE (Rule-guided Iterative Feature Elimination).

Other work in this area has involved iterative processes for feature reduction, however these have been based on different methods [23]. One example of such a process for feature reduction is Support Vector Machines Recursive

If the expression of 207211_at is greater than 100 **then** the sample belongs to the OA group
If the expression of 207201_s_at is less than 45 **then** the sample belongs to the RA group
If the expression of 208806_at is greater than 150 **and** 206269_at is less than 20 **then** the sample belongs to the RA group
If the expression of 204462_s_at is greater than 250 **and** 207591_s_at is greater than 35 **then** the sample belongs to the OA group
Any sample not assigned to a group belongs to the control group

Figure 1 Example of a rule set generated by BioHEL. Rule sets are generated by BioHEL to classify samples. The combination of rules in the rules sets are used to assign samples to their respective treatment groups. Each rule contains one or more gene and an expression value which each gene should either be above or below, depending on the rule. At the end of each line is the group to which each rule relates. For example, the 1st rule of the rule set shown classifies all samples as belonging to the OA class if the value of the gene attribute 207211_at is greater than 100.

Feature Elimination (SVM RFE), which is a form of backward feature elimination. SVM RFE has three main steps. Firstly it trains the SVM classifier, followed by computing a specific ranking criterion for all features. This criterion is based on the weights of a linear SVM. The features are ranked based on this value and the feature(s) with the smallest ranking criterion are removed. These steps are then repeated in an iterative process, resulting in a reduced number of features.

The RGIFE method is tested on several proteomics and transcriptomics OA datasets to demonstrate its suitability for multiple data types. Its performance is compared to several combinations of feature selection and classification algorithms. The results show that RGIFE improves Bio-HEL's performance in all datasets, is able to identify very reduced sets of variables in most of them and shows competitive performance to other methods from the literature.

Methods

Proteomics and transcriptomics datasets

A proteomics dataset was selected, which had been previously analysed by BioHEL to identify putative biomarkers [25]. The proteomics dataset analysed here consisted of 23 samples analysed by mass spectrometry to identify the proteins present in each sample [26]. The dataset included six samples treated with the pro-inflammatory cytokine interleukin-1 beta (IL-1β), to simulate OA in culture, and five samples treated with IL-1β followed by carprofen, which is a non-steroidal anti-inflammatory drug (NSAID) used to treat OA. Also included in the study were six control samples and six samples treated with only carprofen. Mascot was applied, using the Uniprot database, to determine proteins present in the samples. From this two datasets were generated. The first included an emPAI quantitation value assigned to each protein present and alternatively, the second provided each protein with a ProteinProphet probability [27,28]. This probability is a measure of how likely it is each protein is present in a sample.

To assess the method developed on transcriptomics data, datasets were identified from ArrayExpress [21,23-29] and NCBI GEO [30] by searching for the term 'Osteoarthritis'. Of the datasets that this search term returned, those that had more than five samples per group (disease or treatment group) were analysed. This resulted in five datasets (Table 1) that vary in size, but which all contain many more attributes than the proteomics dataset. The sample numbers also vary between 25 samples over 5 classes to 48 samples over 3 classes. These datasets were classified using BioHEL and other machine learning methods, as the canine proteomic dataset was, reported in a recent paper by Swan et al., [25].

All datasets were partitioned into training and test sets following the leave-one-out cross-validation methodology.

The Rule-guided Iterative Feature Elimination heuristic

With the goal of identifying very reduced and highly accurate sets of variables we propose a feature selection heuristic built on top of the BioHEL rule-based machine learning method and based on the iterative feature elimination (IFE) principle. The basic IFE method would pick an attribute, remove it from the dataset and train a classifier (a rule set) with the remaining attributes. If the prediction capacity of the new model is equal or better than with the whole set of attributes (or the previous model), the attribute is eliminated. Otherwise it is inserted back into the dataset. The basic IFE process becomes extremely computationally demanding in high dimensional datasets, such as the case of omics data. Therefore, in order to make such IFE heuristic feasible it needs to be improved in several directions.

The main change is that, rather than picking attributes to be removed at random, we will pick them based on their relevance, estimated from the rule sets generated by BioHEL from the whole set. Then, the first attribute(s) to be picked for elimination are those at the bottom of BioHEL's ranking.

The second change is remove attributes in blocks, rather than one by one. At the start of the heuristic the block size is set up to a very large number, 25% of the problem's attributes. If a block cannot be removed because it would decrease the prediction capacity of the classification model then the next block (following the ranking of attributes) is tested. The block size is reduced to be 25% of the previous block size whenever (a) all attributes have been tested or (b) five consecutive unsuccessful trials have been attempted.

As a final change the acceptance criteria of a trial is relaxed in certain scenarios: whenever five consecutive unsuccessful trials have been performed, before reducing the block size the heuristic checks if one of these trials suffered an accuracy drop corresponding to just one sample. If so, this trial is accepted.

The iterative process then finishes when the attributes are being removed one at a time and either the removal of all attributes left in the dataset had been tested or five iterations in a row resulted in a drop in the percentage classification accuracy equivalent to more than one sample. The overall workflow of the heuristic is represented in Figure 2.

Experimental design

For the experiments of this paper our aim is to show that the RGIFE heuristic is both able to find feature subsets that are both (a) small and (b) highly accurate. To this aim we have designed a series of experiments that firstly compares the heuristic with other FS methods, and then compares the accuracy obtained by BioHEL with the reduced feature subsets with other machine learning methods.

Table 1 Descriptions of datasets analysed, including both proteomic and transcriptomic. Those prefixed 'GSE' were from NCBI GEO and those prefixed 'E-GEOD' were from ArrayExpress

Dataset	No. of samples	No. of genes	No. of classes	Description
Proteomics datasets:				
emPAI canine	23	178	4	Articular cartilage dataset treated with IL-1β to stimulate inflammation. Some samples were also treated with carprofen, a non-steroidal anti-inflammatory drug. Other samples were treated with carprofen only or nothing, as a control. The emPAI dataset includes emPAI label-free quantitation to compare protein quantities across samples, for the proteins with Mascot scores above 30. The ProteinProphet dataset includes a probability for each protein identified in each sample indicating how likely it is to be present in the sample.
ProteinProphet canine	23	1322	4	
Transcriptomics datasets:				
GSE36700	25	54675	5	Comparison between gene expression in synovial biopsies from patients with OA, RA, Systemic Lupus Erythematosus (SLE), seronegative arthritis (SA), and microcrystalline arthritis (MIC) [48]
GSE3698	48	17048	3	Comparison between OA, RA & Pigmented villonodular synovitis (VS), a rare group of lesions with morphological features suggesting an inflammatory as well as a neoplastic nature. All three diseases result in a progressive destruction of affected joints and remain a diagnostic difficulty because of nonspecific symptoms. Tissue samples obtained from knee surgery [49]
E-GEOD-12021	31	22284	3	Gene expression variances were tested in synovial membrane samples of RA patients, OA patients, and normal controls [50]
E-GEOD-29746	31	44397	3	Comparison of gene expression between two pathological groups of human synovial fibroblasts (SF) from RA and OA synovial tissues with normal SF from healthy individuals [51]
E-GEOD-27390	19	54675	2	Gene expression profiling of bone marrow-derived mononuclear cells from patients with RA vs. OA [52]

The feature selection techniques used were Correlation-based (CFS), SVM RFE, Random Forest, Naïve Bayes and Chi Squared, some of which are feature selection methods based on the machine learning methods used. All methods used were those implemented in the machine learning software, WEKA [31]. For those methods that ranked their selected attributes, rather than identifying a small selection of attributes, the top 10 features were used. The machine leaning methods included in the comparison are Naïve Bayes, Support Vector Machines (SVM), k-nearest neighbour (IBk), JRip (rule-based), J48 and Random Forest (RF), also using WEKA implementations [31].

For all classifications performed, the true positive rate (TPR/sensitivity) and true negative rate (TNR/specificity)
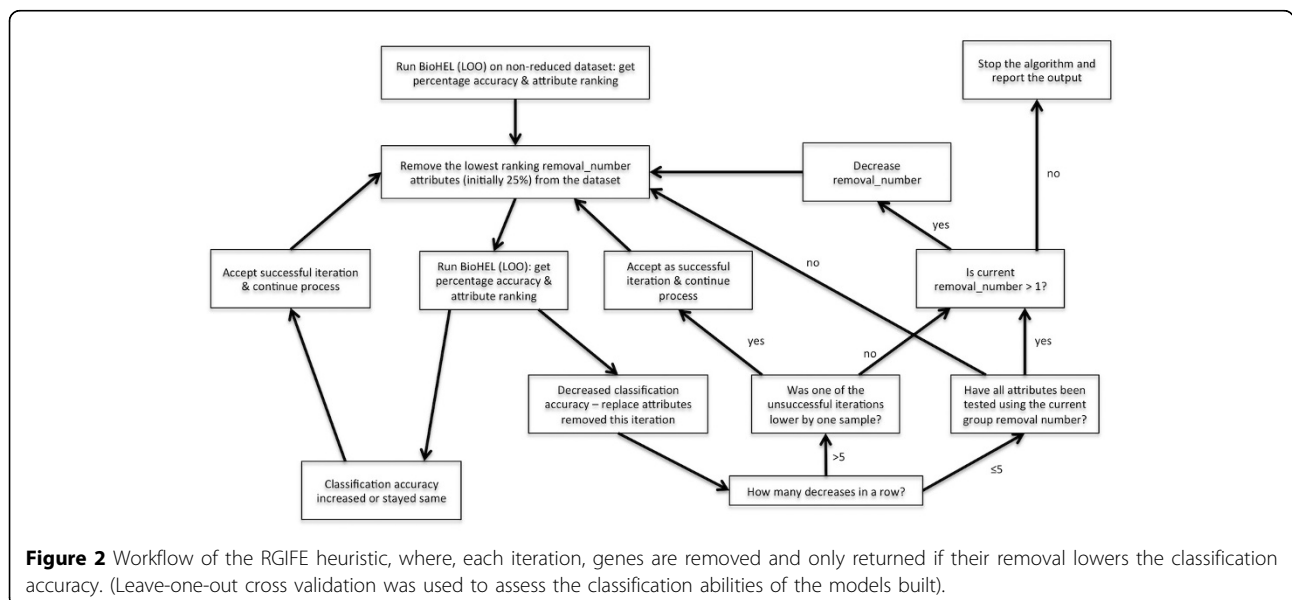

Figure 2 Workflow of the RGIFE heuristic, where, each iteration, genes are removed and only returned if their removal lowers the classification accuracy. (Leave-one-out cross validation was used to assess the classification abilities of the models built).

were calculated. The TPR is the measure of the proportion of correctly classified samples and multiplying it by 100 gives the percentage classification accuracy. TPR is calculated by:

$$TPR = \frac{True\ positives}{True\ positives\ +\ False\ negatives} \quad (1)$$

In contrast, the TNR is a measure of the proportion of negatives that are correctly classified. TNR is calculated by:

$$TNR = \frac{True\ negatives}{True\ negatives + False\ positives} \quad (2)$$

The combinations of methods were compared in two ways. Firstly, the classification accuracies were compared, to determine which ones gave the highest accuracy. Then, the attributes selected by the FS techniques, were analysed using DAVID bioinformatics resource to identify the genes or proteins in each reduced dataset, from their microarray identifiers [32]. The genes and proteins included in the BioHEL reduced datasets were assessed by literature searches with PubMed, to determine any known relevance of the attributes to OA and any other disease or treatment classes included in the datasets.

Results

The RGIFE heuristic increased the classification accuracy achieved for both proteomics and transcriptomics data

By applying the RGIFE FS heuristic to the canine proteomics dataset, the classification accuracies achieved were considerably higher compared to the accuracy achieved without any feature selection (Table 2). This result is robust to choice of protein quantification (emPAI or ProteinProphet). In previous work [25] we showed that BioHEL performed favourably as compared with other classification techniques, best performing with ProteinProphet probabilities (classification accuracy of 73.9%), but out-performed by JRip with emPAI values (classification accuracy of 78.3%). Thus our new feature reduction method, with TPR and TNR of 96% and 99% respectively, has considerably outperformed all other methods for classification accuracy on this dataset.

All five transcriptomics datasets were analysed using the seven ML methods and BioHEL+RGIFE (Table 3). For three of the five transcriptomics datasets, BioHEL without feature reduction gave a classification that was higher

or equal to the other methods. For the other two datasets, GSE36700 and E-GEOD-29746, the best methods were IBk and SVM. The application of RGIFE increased the classification of all five datasets, except for E-GEOD-27290, which was already at 100%. After the feature reduction, only one dataset, E-GEOD-27946, was classified better using an alternative method, SVM.

RGIFE compares favourably with other feature reduction methods

The comparison shown above could be construed as unfair, as it is comparing a classification using the attributes identified by a feature selection method with classifications using the whole dataset. Therefore, it is plausible that application of other feature selection methods could provide improvements comparable to the RGIFE technique. To test this, we have analysed both the proteomics and the transcriptomics datasets using all combinations of the FS and classification method methods. Detailed results are given in Supplementary Tables 1-7 (addition file 1).

For the proteomics dataset, RGIFE is the best feature reduction method, both with emPAI and ProteinProphet values. Using emPAI values (Supplementary Table 1, addition file 1), the highest accuracy was achieved by BioHEL for both FS and classification (TPR 96%; TNR 99%). This is compared to the highest achieved prior to FS, of 78.3%, using JRip, a rule-based classifier. Using ProteinProphet values (Supplementary Table 2, addition file 1), the accuracy rose from the highest accuracy of 73.9% using BioHEL to 95.7% after application of FS. The two FS and classification method combinations that gave this highest classification are: RGIFE with IBk (TPR 96%; TNR 99%) and Naïve Bayes for both FS and classification (TPR 96%; TNR 98%). The combination of RGIFE+BioHEL compares reasonably favourably with these scores (TPR 91%; TNR 97%).

The five transcriptomics datasets were also analysed using the different FS and classification method combinations, with mixed results. For GSE3698 (Supplementary Table 3, addition file 1), classification accuracies of 100% were achieved using RGIFE+BioHEL, as well as with a combination of SVM RFE FS with either SVM or IBk. For GSE36700 (Supplementary Table 4, addition file 1), SVM RFE FS again gave the highest accuracy of 100%, when used with NB, SVM or IBk. For this dataset, RGIFE misclassified only one of the 25 samples. Only one sample was also misclassified when RGIFE was applied to E-GEOD-12021 (Supplementary Table 5, addition file 1), and again SVM RFE combined with NB, SVM or IBk gave the highest accuracy of 100%. For dataset E-GEOD-27390 (Supplementary Table 6, addition file 1), a number of FS and classification methods resulted in classifications of 100% accuracy. This included the combination of RGIFE with either BioHEL

Table 2 Classification accuracies achieved using BioHEL with and without also using the RGIFE heuristic

Classifier	BioHEL, no feature reduction		RGIFE+BioHEL	
	TPR	TNR	TPR	TNR
ProteinProphet	0.74	0.90	0.91	0.97
emPAI	0.57	0.81	0.96	0.99

Table 3 TPR and TNR achieved by BioHEL compared to the other best methods for the five transcriptomics datasets, using leave-one-out cross validation

	NaiveBayes		SVM		IBk		Jrip		J48		RandomForest		BioHEL		RGIFE+ BioHEL	
	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR
GSE36700	0.24	0.67	0.28	0.72	0.84	0.97	0.48	0.66	0.24	0.62	0.44	0.83	0.76	0.91	0.96	0.98
GSE3698	0.58	0.73	0.39	0.60	0.67	0.79	0.67	0.78	0.75	0.84	0.54	0.72	0.73	0.83	1.00	1.00
E-GEOD-12021	0.84	0.91	0.39	0.61	0.71	0.83	0.58	0.73	0.58	0.74	0.77	0.87	0.87	0.93	0.97	0.98
E-GEOD-29746	0.48	0.63	0.48	0.68	0.65	0.79	0.61	0.76	0.42	0.59	0.55	0.71	0.77	0.87	0.84	0.90
E-GEOD-27390	1.00	1.00	1.00	1.00	0.95	0.95	0.90	0.89	0.90	0.91	0.84	0.86	1.00	1.00	1.00	1.00

or RF. The FS methods CFS and RF also resulted in classifications of 100% when used with various classification methods. Dataset E-GEOD-29746 (Supplementary Table 7, addition file 1) was generally less well classified by the various methods, with only two FS and classification method combinations resulting in classifications of 100% accuracy. These were SVM RFE with SVM and RF for FS with RF. Thus for these datasets, no single method stands out as being best.

For each combination of classifier and dataset, the best and worse methods were recorded, considering both the TPRs and TNRs. Tables 6 and 7 compare the six FS methods; they show that the method that most frequently resulted in the highest accuracy was SVM RFE, however it also frequently resulted in the lowest TPR and TNR as well, showing it to be very unstable. The comparison showed RGIFE to be the second best for the highest accuracy, and the second worst for the lowest accuracy, showing it to be the most stable across the datasets tested.

Feature reduction identifies inflammation-associated proteins, which have been associated previously with cartilage matrix degradation

RGIFE identified ten proteins: six using the ProteinProphet quantification and five using emPAI values, with one protein (MMP-3) in common (Table 4). Six of these

Table 4 The proteins included in the reduced datasets identified by RGIFE for the canine proteomics emPAI and ProteinProphet data

Protein ID	Protein name	Identified from emPAI or ProteinProphet	Protein description	Known link to cartilage inflammation or OA
MMP-3	matrix-metalloproteinase 3	ProteinProphet and emPAI	MMP-3 is a proteolytic enzyme known to degrade components of the ECM, including collagens and cartilage proteoglycans [53].	Found to be down-regulated in late OA [33].
IL-8	interleukin-8	ProteinProphet	IL-8 is a chemotactic factor that attracts neutrophils, basophils, and T-cells, but not monocytes. It is also involved in neutrophil activation. It is released from several cell types in response to an inflammatory stimulus [54].	IL-8 is the major chemotactic factor released in response to proinflammatory cytokines in synovial tissues from RA and OA affected joints [34].
TSP1	thrombospondin-1	ProteinProphet	Adhesive glycoprotein that mediates cell-to-cell and cell-to-matrix interactions [55].	Levels of TSP1 are increased after the onset of OA [35]
APOE	apolipoprotein E	ProteinProphet	APOE mediates the binding, internalization, and catabolism of lipoprotein particles [56].	No known link.
HPLN1	hyaluronan and proteoglycan link protein 1	ProteinProphet	Stabilizes the aggregates of proteoglycan monomers with hyaluronic acid in the extracellular cartilage matrix [57].	HPLN1 has been associated with OA and osteophyte formation [36].
TPIS	triosephosphate isomerase	ProteinProphet	Catalyses the reaction D-glyceraldehyde 3-phosphate = glycerone phosphate [58].	No known link.
CLUS	clusterin	emPAI	A glycoprotein that functions as extracellular chaperone that prevents aggregation of non-native proteins, which is involved in many diverse biological functions [59].	Higher levels of clusterin have been observed in synovial fluid of advanced primary knee and hip OA patients [37].
FETUA	alpha-2-HS-glycoprotein/ fetuin-A	emPAI	Promotes endocytosis, possesses opsonic properties and influences the mineral phase of bone [60].	FETUA levels have been found to decrease as the severity of knee OA increases [38].
POLG	Genome polyprotein	emPAI	Bacterial protein.	No known link.
ATPX	ATP synthase subunit b'	emPAI	Bacterial protein.	No known link.

proteins found have been previously associated with cartilage inflammation and OA: MMP-3 [33], IL-8 [34], thrombospondin-1 [35], hyaluronan and proteoglycan link protein 1 [36], clusterin [37] and fetuin-A [38].

The reduced datasets generated from the transcriptomics datasets vary in size and utility

All the transcriptomics datasets were vastly reduced in the number of genes using BioHEL reduction. Table 5 shows the number of genes in each dataset both before and after application of RGIFE. The numbers of genes the datasets were reduced to varied: four out of five were between five and twenty-four genes (a reasonable number). However, E-GEOD-29746 gave 669 genes (out of the original 44397 genes), a much larger number.

The genes present in all the reduced subsets identified by RGIFE are listed in Supplementary Tables 8 – 11 (addition file 1), except for E-GEOD-29746, due to the large number of genes included. Of the genes included in the GSE36700 reduced dataset, two genes, RSAD2 and CXCL9, were found to be associated with RA, one of the disease groups included in the datasets [39,40]. In the GSE3698 reduced dataset, four genes are clearly relevant to the diseases analysed with this dataset. These genes are FN1, DDR2, MMP-9 and NOTCH3, all of which have been associated with either OA or RA [41-44]. Dataset E-GEOD-12021 was reduced to only five genes. Included in this small subset of genes was CXCL13, which has previously been suggested as a possible biomarker for RA [45]. None of the genes in the E-GEOD-27390 reduced dataset were found to be specifically related to OA or RA, the two disease classes included in the dataset. However, there were a number of genes whose functions are currently unknown.

Discussion

Comparison of classification accuracies achieved by feature selection methods

From the results shown, it can be seen that there is not one FS and classification method combination that is best for all datasets. For the proteomics data, the RGIFE+ BioHEL combination compared very well with other methods. For all five of the transcriptomics datasets, at least one combination gave a classification accuracy of 100%. However, there was a lot of variation both in the accuracies of the classifications for each dataset, using the different methods, and the classifications performed by each method across the different datasets.

RGIFE gave the equal highest accuracies for two of the five transcriptomics datasets and was close to the highest accuracies for another two datasets. However, for dataset E-GEOD-29746 (the dataset where the heuristic was not very effective at identifying a very reduced set of features), other methods were better.

Proteins identified from the proteomics dataset by RGIFE

Using the two different values for the proteins, emPAI and ProteinProphet, did result in different proteins being selected, however MMP-3 was included in both. This is because emPAI is a measure of relative protein quantitation and ProteinProphet gives a probability based on how likely it is a protein is present in a sample. Therefore, whilst these measures are related, they are not the same.

Included in the proteomics emPAI-reduced dataset are two bacterial proteins, Genome polyprotein (POLG) and ATP synthase subunit b' (ATPX). These proteins have fairly low Mascot scores, indicating a lack of confidence in this identification and so, based on this, it may be suitable to increase the Mascot score threshold to reduce the likelihood of false positives. Also, POLG was found in only three samples and ATPX was present in only one sample. Through analysis of the rules generated by BioHEL it is also clear that these proteins were used to distinguish between the control and carprofen-only treated samples, indicating that carprofen treatment has no detectable direct physiological effect.

Supplementary tables 12 and 13 (addition file 1) show the proteins selected by the other feature selection methods tested. All the proteins identified by RGIFE for the MS dataset, using ProteinProphet probabilities, were also identified by at least one of the other FS techniques tested. For the emPAI values dataset, two of the five proteins, CLUS & MMP-3, were identified by other FS methods. These proteins have previously been identified in OA [33,46]. For the dataset using emPAI values, other methods did also select bacterial proteins.

BioHEL reduction method applied to transcriptomics datasets

The genes identified by RGIFE were also compared to the genes selected by the other FS techniques used. Supplemental tables 14-17 (addition file 1) list the genes

Table 5 The number of genes present in each dataset before and after feature reduction with RGIFE

Dataset	Whole dataset No. of genes identifiers	Reduced datasets No. of genes identifiers
GSE36700	54675	24
GSE3698	17048	19
E-GEOD-12021	22284	5
E-GEOD-29746	44397	669
E-GEOD-27390	54675	14

identified by each method, for comparison with the genes included by RGIFE. For datasets E-GEOD-12021 and E-GEOD-27390 no genes selected by RGIFE were included in the lists of genes identified by the other FS methods. There was some overlap between the genes selected by the other methods, however only very few. Dataset GSE36700 had one gene, RSAD2, which was included in the reduced dataset of another FS method. RSAD2 has been found to be up-regulated in RA [39], which was a disease assessed by the GSE36700 investigation. Five of the genes selected by RGIFE for this dataset were also identified by other FS methods. When faced with such high dimensional datasets, the results show that the natural bias of each FS method produces very divergent results. Perhaps this is not surprising since all of them are heuristic.

The transcriptomics reduced-datasets were generally larger than those of the proteomics datasets. Quite likely this is due to proteomics datasets having far fewer features than transcriptomics ones, and the features present are less noisy [19].

Conclusions

The RGIFE feature-reduction method has been shown to be suitable for the analysis of both transcriptomics and proteomics data. The classification accuracies achieved by this method in combination with the BioHEL rule-based machine learning method were better than other machine learning methods used (without FS) for all datasets and better or equal for the majority of datasets with feature selection also applied.

The feature reduction method resulted in the selection of a subset of genes for all datasets, some of which had clear links to the diseases related to the datasets. A number of genes were identified that may be suitable as possible biomarkers, however they require further individual analysis to determine their relevance and suitability.

Given that this is a generic supervised machine-learning technique, this method should also be suitable for analysis of other forms of complex data, including data from other omics areas, such as metabolomics and lipidomics, however testing in this area is still required.

Some cytokines and chemokines in the joint, such as IL8 (see table 6), have increased activity during OA, [47] which drive the increased production and secretion of matrix degrading enzymes such as matrix metalloproteinases (MMPs) (i.e. MMP-3, see table 7) that mediate the destruction of articular cartilage. Therefore, the feature-reduction methods used in this study identified biologically relevant proteins.

In summary, to our knowledge bioinformatics feature-reduction tools have never been applied to 'omics' data in the area of rheumatology. However, as more and more investigators are applying 'omics' techniques to tissues and cells from arthritic joints, feature-reduction methods such as this are likely to make a significant contribution to basic and clinical research in this area, especially the stratification of patients based on data from molecular markers of joint inflammation. Despite the growing burden of arthritic diseases, many pharmaceutical companies have abandoned the development of disease modifying OA drugs (DMOADs) because OA is a heterogeneous disease with a variety of phenotypes and pathophysiological drivers. Identification of novel biomarkers and further validation of existing "experimental" markers are likely to facilitate OA drug development. Bioinformatic studies of currently available data from *in vitro* cartilage models, animal models and OA patients will consolidate existing knowledge of markers of disease progression and reveal new targets that may be invaluable for DMOAD development. The inevitable growth of public and private datasets derived from large cohort studies of joint inflammation will provide further opportunities for applying feature reduction techniques for biomarker identification and validation. This should provide a paradigm shift in the diagnosis and treatment of arthritis, facilitating new drug discovery and improving the range of effective therapy options for chronic joint diseases.

Disclosures

Table 6 Comparison of FS methods applied to both transcriptomics and proteomics, for each combination of classifier and dataset. The number of times each FS method resulted in the highest TPR and the lowest TPR are shown

Method	No. of times the method results in the highest TPR	No. of times the method results in the lowest TPR
CFS	3	15
Chisquared	2	4
NaiveBayes - FS	6	6
Random Forest - FS	10	6
RGIFE	13	5
SVM RFE	16	13

Table 7 Comparison of FS methods applied to both transcriptomics and proteomics, for each combination of classifier and dataset. The number of times each FS method resulted in the highest TNR and the lowest TNR are shown

Method	No. of times the method results in the highest TNR	No. of times the method results in the lowest TNR
CFS	3	14
Chisquared	3	4
NaiveBayes - FS	6	7
Random Forest - FS	10	7
RGIFE	12	5
SVM RFE	15	12

Additional material

Additional file 1:

List of abbreviations
ATPX: ATP synthase subunit b'; CFS: Correlation-based Feature Selection; CLUS: Clusterin; CXCL9: Chemokine (C-X-C motif) ligand 9; DDR2: Discoidin domain receptor family, member 2; emPAI: Exponentially modified protein abundance index; FN1: Fibronectin 1; FS: Feature Selection; IBk: k-nearest neighbour; IFE: Iterative Feature Elimination; IL-1β: Interleukin 1 beta; ML: Machine Learning; MMP-3: Matrix metalloproteinase-3 (stromelysin-1); MMP-9: Matrix metalloproteinase-9; MS: Mass spectrometry; NB: Naïve Bayes; NOTCH3: Notch homolog 3; NSAID: non-steroidal anti-inflammatory drug; OA: Osteoarthritis; POLG: Genome polyprotein; RA: Rheumatoid Arthritis; RF: Random Forest; RGIFE: Rule-guided Iterative Feature Elimination; RSAD2: Radical S-adenosyl methionine domain containing 2; SVM: Support Vector Machines; SVM RFE: SVM-Recursive Feature Elimination; TNR: True Negative Rate; TPR: True Positive Rate.

Authors information
Jaume Bacardit and Ali Mobasheri are member of the following European Commission Consortia:
The D-BOARD European Consortium for Biomarker Discovery, The Universities of Surrey, Nottingham and Newcastle, United Kingdom; http://cordis.europa.eu/projects/rcn/105314_en.html; http://ec.europa.eu/research/health/medical-research/severe-chronic-diseases/projects/d-board_en.html
The APPROACH Consortium (Applied Public-Private Research enabling OsteoArthritis Clinical Headway) (Website under construction).

Authors' details
[1]School of Biosciences, Faculty of Science, University of Nottingham, Sutton Bonington Campus, Leicestershire, LE12 5RD, United Kingdom. [2]The D-BOARD European Consortium for Biomarker Discovery, The Universities of Surrey, Nottingham and Newcastle, United Kingdom. [3]School of Veterinary Medicine, Faculty of Health and Medical Sciences, University of Surrey, Duke of Kent Building, Guildford, Surrey, GU2 7XH, United Kingdom. [4]WALTHAM® Centre for Pet Nutrition, Waltham-on-the-Wolds, Melton Mowbray, Leicestershire, LE14 4RT, United Kingdom. [5]Center of Excellence in Genomic Medicine Research (CEGMR), King AbdulAziz University, Jeddah, 21589, Kingdom of Saudi Arabia. [6]Arthritis Research UK Centre for Sport, Exercise, and Osteoarthritis, Arthritis Research UK Pain Centre, Medical Research Council-Arthritis Research UK Centre for Musculoskeletal Ageing Research, Faculty of Medicine and Health Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom. [7]The Interdisciplinary Computing and Complex BioSystems (ICOS) research group, School of Computing Science, Newcastle University, Claremont Tower, Newcastle-upon-Tyne, NE1 7RU, United Kingdom.

References
1. Atkinson AJ, Colburn WA, DeGruttola VG, DeMets DL, Downing GJ, Hoth DF, Oates JA, Peck CC, Schooley RT, Spilker BA, *et al*: **Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework***. *Clin Pharmacol Ther* 2001, **69(3)**:89-95.
2. Buckwalter JA, Saltzman C, Brown T: **The Impact of Osteoarthritis: Implications for Research**. *Clinical Orthopaedics and Related Research* 2004, **427**:S6-S15.
3. Goldring MB, Goldring SR: **Osteoarthritis**. *Journal of Cellular Physiology* 2007, **213(3)**:626-634.
4. Goldring MB, Goldring SR: **Articular cartilage and subchondral bone in the pathogenesis of osteoarthritis**. *Annals of the New York Academy of Sciences* 2010, **1192(1)**:230-237.
5. Goldring MB, Otero M: **Inflammation in osteoarthritis**. *Current Opinion in Rheumatology* 2011, **23(5)**:471-478.
6. Sellam J, Berenbaum F: **The role of synovitis in pathophysiology and clinical symptoms of osteoarthritis**. *Nat Rev Rheumatol* 2010, **6(11)**:625-635.
7. Berenbaum F: **Osteoarthritis as an inflammatory disease (osteoarthritis is not osteoarthrosis!)**. *Osteoarthritis and Cartilage* 2013, **21(1)**:16-21.
8. Mobasheri A: **Osteoarthritis year 2012 in review: biomarkers**. *Osteoarthritis and Cartilage* 2012, **20(12)**:1451-1464.
9. Lotz M, Martel-Pelletier J, Christiansen C, Brandi M-L, Bruyère O, Chapurlat R, Collette J, Cooper C, Giacovelli G, Kanis JA, *et al*: **Value of biomarkers in osteoarthritis: current status and perspectives**. *Annals of the Rheumatic Diseases* 2013, **72(11)**:1756-1763.
10. Sun CS, Markey MK: **Recent advances in computational analysis of mass spectrometry for proteomic profiling**. *Journal of Mass Spectrometry* 2011, **46(5)**:443-456.
11. Deutsch EW, Lam H, Aebersold R: **Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics**. *Physiological Genomics* 2008, **33(1)**:18-25.

12. Kanehisa M, Bork P: **Bioinformatics in the post-sequence era.** *Nat Genet* 2003, **33**:305-310.

13. Swan AL, Mobasheri A, Allaway D, Liddell S, Bacardit J: **Application of Machine Learning to Proteomics Data: Classification and Biomarker Identification in Postgenomics Biology.** *Omics: a journal of integrative biology* 2013, **17**(12):595-610.

14. Bacardit J, Burke E, Krasnogor N: **Improving the scalability of rule-based evolutionary learning.** *Memetic Computing* 2009, **1**:55-67.

15. Bassel GW, Glaab E, Marquez J, Holdsworth MJ, Bacardit J: **Functional Network Construction in Arabidopsis Using Rule-Based Machine Learning on Large-Scale Data Sets.** *The Plant Cell* 2011, **23**(9):3101-3116.

16. Glaab E, Bacardit J, Garibaldi JM, Krasnogor N: **Using Rule-Based Machine Learning for Candidate Disease Gene Prioritization and Sample Classification of Cancer Gene Expression Data.** *PLoS ONE* 2012, **7**(7): e39932.

17. Fainberg HP, Bodley K, Bacardit J, Li D, Wessely F, Mongan NP, Symonds ME, Clarke L, Mostyn A: **Reduced Neonatal Mortality in Meishan Piglets: A Role for Hepatic Fatty Acids?** *PLoS ONE* 2012, **7**(11):e49101.

18. Bacardit J, Widera P, Mrquez-Chamorro A, Divina F, Aguilar-Ruiz JS, Krasnogor N: **Contact map prediction using a large-scale ensemble of rule sets and the fusion of multiple predicted structural features.** *Bioinformatics* 2012, **28**(19):2441-2448.

19. Saeys Y, Inza I, Larrañaga P: **A review of feature selection techniques in bioinformatics.** *Bioinformatics* 2007, **23**(19):2507-2517.

20. Hall M: **Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning.** *Proc 17th International Conf on Machine Learning, Morgan Kaufmann, San Francisco, CA* 2000, 359-366.

21. Holland JH: **Adaptation in Natural and Artificial Systems: The University of Michigan Press.** Ann Arbor; 1975.

22. Duda RO, Hart PE, Stork DG: **Pattern Classification.** New York: Wiley; 2001.

23. Guyon I, Elisseeff A: **An introduction to variable and feature selection.** *Journal of Machine Learning Research* 2003, **3**:1157-1182.

24. Weston J, Elisseeff A, Scholkopf B, Tipping M: **Use of the zero norm with linear models and kernel methods.** *J Mach Learn Res* 2003, **3**:1439-1461.

25. Swan AL, Hillier KL, Smith JR, Allaway D, Liddell S, Bacardit J, Mobasheri A: **Analysis of mass spectrometry data from the secretome of an explant model of articular cartilage exposed to pro-inflammatory and anti-inflammatory stimuli using machine learning.** *BMC Musculoskelet Disord* 2013, **14**(1):349.

26. Williams A, Smith J, Allaway D, Harris P, Liddell S, Mobasheri A: **Carprofen inhibits the release of matrix metalloproteinases 1, 3 and 13 in the secretome of an explant model of articular cartilage stimulated with interleukin 1beta.** *Arthritis Research & Therapy* 2013, **15**(6):R223.

27. Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J, Mann M: **Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein.** *Molecular Cell Proteomics* 2005, **4**(9):1265-1272.

28. Keller A, Nesvizhskii AI, Kolker E, Aebersold R: **Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search.** *Analytical Chemistry* 2002, **74**(20):5383-5392.

29. Array Express [http://www.ebi.ac.uk/arrayexpress/].

30. NCBI GEO [http://www.ncbi.nlm.nih.gov/geo/].

31. Witten I, Frank E, Hall M: **Data Mining: Practical Machine Learning Tools and Techniques.** Morgan Kaufmann;, 3 2011.

32. Dennis G, Sherman B, Hosack D, Yang J, Gao W, Lane H, Lempicki R: **DAVID: database for annotation, visualization, and integrated discovery.** *Genome Biol* 2003, **4**(9):R60.

33. Aigner T, Zien A, Hanisch D, Zimmer R: **Gene expression in chondrocytes assessed with use of microarrays.** *The Journal of bone and joint surgery American volume* 2003, **85-A**(Suppl 2):117-123.

34. Goldring MB: **The Role of Cytokines as Inflammatory Mediators in Osteoarthritis: Lessons from Animal Models.** *Connective Tissue Research* 1999, **40**(1):1-11.

35. Hsieh J-L, Shen P-C, Shiau A-L, Jou IM, Lee C-H, Wang C-R, Teo M-L, Wu C-L: **Intraarticular gene transfer of thrombospondin-1 suppresses the disease progression of experimental osteoarthritis.** *Journal of Orthopaedic Research* 2010, **28**(10):1300-1306.

36. Urano T, Narusawa Ki, Shiraki M, Sasaki N, Hosoi T, Ouchi Y, Nakamura T, Inoue S: **Single-nucleotide polymorphism in the hyaluronan and proteoglycan link protein 1 (HAPLN1) gene is associated with spinal**

osteophyte formation and disc degeneration in Japanese women. *Eur Spine J* 2011, **20**(4):572-577.

37. Fandridis E, Apergis G, Korres DS, Nikolopoulos K, Zoubos AB, Papassideri I, Trougakos IP: **Increased Expression Levels of Apolipoprotein J/Clusterin during Primary Osteoarthritis.** *In Vivo* 2011, **25**(5):745-749.

38. Xiao J, Wang X-R, Hu K-Z, Li M-Q, Chen J-W, Ma T, Li Z-C: **Serum fetuin-A levels are inversely associated with clinical severity in patients with primary knee osteoarthritis.** *Biomarkers* 2013, **18**(1):51-54.

39. van der Pouw Kraan TC, Wijbrandts CA, van Baarsen LG, Voskuyl AE, Rustenburg F, Baggen JM, Ibrahim SM, Fero M, Dijkmans BA, Tak PP, *et al*: **Rheumatoid arthritis subtypes identified by genomic profiling of peripheral blood cells: assignment of a type I interferon signature in a subpopulation of patients.** *Annals of the Rheumatic Diseases* 2007, **66**(8):1008-1014.

40. Ruschpler P, Lorenz P, Eichler W, Koczan D, Hanel C, Scholz R, Melzer C, Thiesen H-J, Stiehl P: **High CXCR3 expression in synovial mast cells associated with CXCL9 and CXCL10 expression in inflammatory synovial tissues of patients with rheumatoid arthritis.** *Arthritis Res Ther* 2003, **5**(5): R241-R252.

41. MacLeod JN, Burton-Wurster N, Gu DN, Lust G: **Fibronectin mRNA Splice Variant in Articular Cartilage Lacks Bases Encoding the V, III-15, and I-10 Protein Segments.** *Journal of Biological Chemistry* 1996, **271**(31):18954-18960.

42. Holt DW, Henderson ML, Stockdale CE, Farrell JT, Kooyman DL, Bridgewater LC, Seegmiller RE: **Osteoarthritis-like changes in the heterozygous sedc mouse associated with the HtrA1–Ddr2–Mmp-13 degradative pathway: a new model of osteoarthritis.** *Osteoarthritis and Cartilage* 2012, **20**(5):430-439.

43. Dreier R, Grässel S, Fuchs S, Schaumburger J, Bruckner P: **Pro-MMP-9 is a specific macrophage product and is activated by osteoarthritic chondrocytes via MMP-3 or a MT1-MMP/MMP-13 cascade.** *Experimental Cell Research* 2004, **297**(2):303-312.

44. Yabe Y, Matsumoto T, Tsurumoto T, Shindo H: **Immunohistological localization of Notch receptors and their ligands Delta and Jagged in synovial tissues of rheumatoid arthritis.** *J Orthop Sci* 2005, **10**(6):589-594.

45. Meeuwisse CM, van der Linden MP, Rullmann TA, Allaart CF, Nelissen R, Huizinga TW, Garritsen A, Toes RE, van Schaik R, van der Helm-van Mil AH: **Identification of CXCL13 as a marker for rheumatoid arthritis outcome using an in silico model of the rheumatic joint.** *Arthritis & Rheumatism* 2011, **63**(5):1265-1273.

46. Clutterbuck AL, Smith JR, Allaway D, Harris P, Liddell S, Mobasheri A: **High throughput proteomic analysis of the secretome in an explant model of articular cartilage inflammation.** *Journal of Proteomics* 2011, **74**(5):704-715.

47. Loeser RF, Goldring SR, Scanzello CR, Goldring MB: **Osteoarthritis: a disease of the joint as an organ.** *Arthritis Rheum* 2012, **64**(6):1697-1707.

48. Toukap AN, Galant C, Theate I, Maudoux AL, Lories RJU, Houssiau FA, Lauwerys BR: **Identification of distinct gene expression profiles in the synovium of patients with systemic lupus erythematosus.** *Arthritis & Rheumatism* 2007, **56**(5):1579-1588.

49. Finis K, Sültmann H, Ruschhaupt M, Buness A, Helmchen B, Kuner R, Gross M-L, Fink B, Schirmacher P, Poustka A, *et al*: **Analysis of pigmented villonodular synovitis with genome-wide complementary DNA microarray and tissue array technology reveals insight into potential novel therapeutic approaches.** *Arthritis & Rheumatism* 2006, **54**(3):1009-1019.

50. Huber R, Hummert C, Gausmann U, Pohlers D, Koczan D, Guthke R, Kinne RW: **Identification of intra-group, inter-individual, and gene-specific variances in mRNA expression profiles in the rheumatoid arthritis synovial membrane.** *Arthritis Research & Therapy* 2008, **10**(4):R98.

51. Del Rey MJ, Usategui A, Izquierdo E, Cañete JD, Blanco FJ, Criado G, Pablos JL: **Transcriptome analysis reveals specific changes in osteoarthritis synovial fibroblasts.** *Annals of the Rheumatic Diseases* 2012, **71**(2):275-280.

52. Lee HM, Sugino H, Aoki C, Shimaoka Y, Suzuki R, Ochi K, Ochi T, Nishimoto N: **Abnormal networks of immune response-related molecules in bone marrow cells from patients with rheumatoid arthritis as revealed by DNA microarray analysis.** *Arthritis Research & Therapy* 2011, **13**(3):R89.

53. Keyszer G, Lambiri I, Nagel R, Keysser C, Keysser M, Gromnica-Ihle E, Franz J, Burmester GR, Jung K: **Circulating levels of matrix metalloproteinases MMP-3 and MMP-1, tissue inhibitor of metalloproteinases 1 (TIMP-1),**

and MMP-1/TIMP-1 complex in rheumatic disease. *Correlation with clinical activity of rheumatoid arthritis versus other surrogate markers* 1999, , **2**: 251-258.

54. van Damme J, Rampart M, Conings R, Decock B, van Osselaer N, Willems J, Billiau A: **The neutrophil-activating proteins interleukin 8 and β-thromboglobulin: in vitro and in vivo comparison of NH2-terminally processed forms.** *European Journal of Immunology* 1990, **20**(9):2113-2118.

55. Bornstein P, Alfi D, Devarayalu S, Framson P, Li P: **Characterization of the mouse thrombospondin gene and evaluation of the role of the first intron in human gene expression.** *Journal of Biological Chemistry* 1990, **265**(27):16691-16698.

56. Luo CC, Li WH, Chan L: **Structure and expression of dog apolipoprotein A-I, E, and C-I mRNAs: implications for the evolution and functional constraints of apolipoprotein structure.** *Journal of Lipid Research* 1989, **30**(11):1735-1746.

57. Hardingham TE, Fosang AJ: **Proteoglycans: many forms and many functions.** *The FASEB Journal* 1992, **6**(3):861-870.

58. Chang M, Wu X, Maquat LE, Artymiuk PJ, Hollan S, Lammi A: **Human triosephosphate isomerase deficiency resulting from mutation of Phe-240.** *Journal Name: American Journal of Human Genetics* 1993, **52**(6):1260-1269.

59. Jones SE, Jomary C: **Clusterin.** *The International Journal of Biochemistry & Cell Biology* 2002, **34**(5):427-431.

60. Lewis JG, André CM: **Effect of human alpha 2HS glycoprotein on mouse macrophage function.** *Immunology* 1980, **39**(3):317-322.

61. Tahara E Jr, Tahara H, Kanno M, Naka K, Takeda Y, Matsuzaki T, Yamazaki R, Ishihara H, Yasui W, Barrett J, *et al*: **G1P3, an interferon inducible gene 6-16, is expressed in gastric cancers and inhibits mitochondrial-mediated apoptosis in gastric cancer cell line TMK-1 cell.** *Cancer Immunol Immunother* 2005, **54**(8):729-740.

62. Conner JR, Beisswenger PJ, Szwergold BS: **Some Clues as to the Regulation, Expression, Function, and Distribution of Fructosamine-3-Kinase and Fructosamine-3-Kinase-Related Protein.** *Annals of the New York Academy of Sciences* 2005, **1043**(1):824-836.

63. Chuang SS, Helvig C, Taimi M, Ramshaw HA, Collop AH, Amad Ma, White JA, Petkovich M, Jones G, Korczak B: **CYP2U1, a Novel Human Thymus- and Brain-specific Cytochrome P450, Catalyzes ω- and (ω-1)-Hydroxylation of Fatty Acids.** *Journal of Biological Chemistry* 2004, **279**(8):6305-6314.

64. Li S, Qiao Y, Di Q, Le X, Zhang L, Zhang X, Zhang C, Cheng J, Zong S, Koide SS, *et al*: **Interaction of SH3P13 and DYDC1 protein: a germ cell component that regulates acrosome biogenesis during spermiogenesis.** *European Journal of Cell Biology* 2009, **88**(9):509-520.

65. Radaev S, Sun P: **Recognition of immunoglobulins by Fcγ receptors.** *Molecular Immunology* 2002, **38**(14):1073-1083.

66. Joos H, Albrecht W, Laufer S, Reichel H, Brenner RE: **IL-1B Regulates FHL2 and Other Cytoskeleton-Related Genes in Human Chondrocytes.** Baltimore, MD, USA: Johns Hopkins University Press; 2008**14**.

67. Samuels-Lev Y, O'Connor DJ, Bergamaschi D, Trigiante G, Hsieh J-K, Zhong S, Campargue I, Naumovski L, Crook T, Lu X: **ASPP Proteins Specifically Stimulate the Apoptotic Function of p53.** *Molecular Cell* 2001, **8**(4):781-794.

68. Mori S, Kou I, Sato H, Emi M, Ito H, Hosoi T, Ikegawa S: **Association of genetic variations of genes encoding thrombospondin, type 1, domain-containing 4 and 7A with low bone mineral density in Japanese women with osteoporosis.** *J Hum Genet* 2008, **53**(8):694-697.

69. Chin K-C, Cresswell P: **Viperin (cig5), an IFN-inducible antiviral protein directly induced by human cytomegalovirus.** *Proceedings of the National Academy of Sciences* 2001, **98**(26):15125-15130.

70. Lindbom J, Ljungman AG, Lindahl M, Tagesson C: **Increased gene expression of novel cytosolic and secretory phospholipase A[2] types in human airway epithelial cells induced by tumor necrosis factor-alpha and IFN-gamma.** Larchmont, NY, USA: Liebert; 2002**22**.

71. Pasvolsky R, Grabovsky V, Giagulli C, Shulman Z, Shamri R, Feigelson SW, Laudanna C, Alon R: **RhoA Is Involved in LFA-1 Extension Triggered by CXCL12 but Not in a Novel Outside-In LFA-1 Activation Facilitated by CXCL9.** *The Journal of Immunology* 2008, **180**(5):2815-2823.

72. Drenth JPH, GÖErtz J, Daha MR, Van Der Meer JWM: **Immunoglobulin D enhances the release of tumour necrosis factor-α, and interleukin-1β as well as interleukin-1 receptor antagonist from human mononuclear cells.** *Immunology* 1996, **88**(3):355-362.

73. Nagel S, Burek C, Venturini L, Scherr M, Quentmeier H, Meyer C, Rosenwald A, Drexler HG, MacLeod RAF: **Comprehensive analysis of homeobox genes in Hodgkin lymphoma cell lines identifies dysregulated expression of HOXB9 mediated via ERK5 signaling and BMI1.** *Blood* 2007, **109**(7):3015-3023.

74. Anastasaki C, Longman D, Capper A, Patton EE, Cáceres JF: **Dhx34 and Nbas function in the NMD pathway and are required for embryonic development in zebrafish.** *Nucleic Acids Research* 2011, **39**(9):3686-3694.

75. Heiner I, Eisfeld RG, Halaszovich CR, Wehage E, Ngling E, Zitt C, Ckhoff A: **Expression profile of the transient receptor potential (TRP) family in neutrophil granulocytes: evidence for currents through long TRP channel 2 induced by ADP-ribose and NAD.** Colchester, UK: Portland Press; 2003**371**.

76. Brdička T, Imrich M, Angelisová P, Brdičková N, Horváth O, Špička J, Hilgert I, Lusková P, Dráber P, Novák P, *et al*: **Non–T Cell Activation Linker (NTAL): A Transmembrane Adaptor Protein Involved in Immunoreceptor Signaling.** *The Journal of Experimental Medicine* 2002, **196**(12):1617-1626.

77. Economopoulou M-al, Fragoulis EG, Sideris DC: **Molecular cloning and characterization of the human RNase κ, an ortholog of Cc RNase.** *Nucleic Acids Research* 2007, **35**(19):6389-6398.

78. Lagali PS, Kakuk LE, Griesinger IB, Wong PW, Ayyagari R: **Identification and characterization of C6orf37, a novel candidate human retinal disease gene on chromosome 6q14.** *Biochemical and Biophysical Research Communications* 2002, **293**(1):356-365.

79. McNamee CJ, Reed JE, Howard MR, Lodge AP, Moss DJ: **Promotion of neuronal cell adhesion by members of the IgLON family occurs in the absence of either support or modification of neurite outgrowth.** *Journal of Neurochemistry* 2002, **80**(6):941-948.

80. Reiss U, Oskouian B, Zhou J, Gupta V, Sooriyakumaran P, Kelly S, Wang E, Merrill AH, Saba JD: **Sphingosine-phosphate Lyase Enhances Stress-induced Ceramide Generation and Apoptosis.** *Journal of Biological Chemistry* 2004, **279**(2):1281-1290.

81. Stegert MR, Hergovich A, Tamaskovic R, Bichsel SJ, Hemmings BA: **Regulation of NDR Protein Kinase by Hydrophobic Motif Phosphorylation Mediated by the Mammalian Ste20-Like Kinase MST3.** *Molecular and Cellular Biology* 2005, **25**(24):11019-11029.

82. Koch M, Schulze J, Hansen U, Ashwodt T, Keene DR, Brunken WJ, Burgeson RE, Bruckner P, Bruckner-Tuderman L: **A Novel Marker of Tissue Junctions, Collagen XXII.** *Journal of Biological Chemistry* 2004, **279**(21):22514-22521.

83. Dhallan R, Macke J, Eddy R, Shows T, Reed R, Yau K, Nathans J: **Human rod photoreceptor cGMP-gated channel: amino acid sequence, gene structure, and functional expression.** *The Journal of Neuroscience* 1992, **12**(8):3248-3256.

84. Hjelmervik TOR, Petersen K, Jonassen I, Jonsson R, Bolstad AI: **Gene expression profiling of minor salivary glands clearly distinguishes primary Sjögren's syndrome patients from healthy control subjects.** *Arthritis & Rheumatism* 2005, **52**(5):1534-1544.

85. Juszczynski P, Kutok JL, Li C, Mitra J, Aguiar RCT, Shipp MA: **BAL1 and BBAP Are Regulated by a Gamma Interferon-Responsive Bidirectional Promoter and Are Overexpressed in Diffuse Large B-Cell Lymphomas with a Prominent Inflammatory Infiltrate.** *Molecular and Cellular Biology* 2006, **26**(14):5348-5359.

86. Bates EEM, Fournier N, Garcia E, Valladeau J, Durand I, Pin J-J, Zurawski SM, Patel S, Abrams JS, Lebecque S, *et al*: **APCs Express DCIR, a Novel C-Type Lectin Surface Receptor Containing an Immunoreceptor Tyrosine-Based Inhibitory Motif.** *The Journal of Immunology* 1999, **163**(4):1973-1983.

87. Gray GE, Mann RS, Mitsiadis E, Henrique D, Carcangiu M-L, Banks A, Leiman J, Ward D, Ish-Horowitz D: **Artavanis-Tsakonas S: Human Ligands of the Notch Receptor.** *The American Journal of Pathology* 1999, **154**(3):785-794.

88. Nakagawa O, McFadden DG, Nakagawa M, Yanagisawa H, Hu T, Srivastava D, Olson EN: **Members of the HRT family of basic helix-loop-helix proteins act as transcriptional repressors downstream of Notch signaling.** *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**(25):13655-13660.

89. Ruse M, Broome A-M, Eckert RL: **S100A7 (Psoriasin) Interacts with Epidermal Fatty Acid Binding Protein and Localizes in Focal Adhesion-Like Structures in Cultured Keratinocytes.** *J Investig Dermatol* 2003, **121**(1):132-141.

90. Cisternas FA, Vincent JB, Scherer SW, Ray PN: **Cloning and characterization of human CADPS and CADPS2, new members of the Ca2+-dependent activator for secretion protein family.** *Genomics* 2003, **81(3)**:279-291.

91. Hardy R, Juarez M, Naylor A, Tu J, Rabbitt E, Filer A, Stewart P, Buckley C, Raza K, Cooper M: **Synovial DKK1 expression is regulated by local glucocorticoid metabolism in inflammatory arthritis.** *Arthritis Research & Therapy* 2012, **14(5)**:R226.

92. Imoto I, Sonoda I, Yuki Y, Inazawa J: **Identification and Characterization of Human PKNOX2, a Novel Homeobox-Containing Gene.** *Biochemical and Biophysical Research Communications* 2001, **287(1)**:270-276.

93. Lee A, Westenbroek RE, Haeseleer F, Palczewski K, Scheuer T, Catterall WA: **Differential modulation of Cav2.1 channels by calmodulin and Ca2 +-binding protein 1.** *Nat Neurosci* 2002, **5(3)**:210-217.

94. Wu BX, Chen Y, Chen Y, Fan J, Rohrer B, Crouch RK, Ma J-x: **Cloning and Characterization of a Novel all-trans Retinol Short-Chain Dehydrogenase/ Reductase from the RPE.** *Investigative Ophthalmology & Visual Science* 2002, **43(11)**:3365-3372.

95. Zou JX, Revenko AS, Li LB, Gemo AT, Chen H-W: **ANCCA, an estrogen-regulated AAA+ ATPase coactivator for ERα, is required for coregulator occupancy and chromatin modification.** *Proceedings of the National Academy of Sciences* 2007, **104(46)**:18067-18072.

96. Horowitz DS, Krainer AR: **A human protein required for the second step of pre-mRNA splicing is functionally related to a yeast splicing factor.** *Genes & Development* 1997, **11(1)**:139-151.

97. Pandit R, Svasti S, Sripichai O, Munkongdee T, Triwitayakorn K, Winichagoon P, Fucharoen S, Peerapittayamongkol C: **Association of SNP in exon 1 of HBS1L with hemoglobin F level in β0-thalassemia/ hemoglobin E.** *Int J Hematol* 2008, **88(4)**:357-361.

98. Pörzgen P, Bönisch H, Hammermann R, Brüss M: **The human noradrenaline transporter gene contains multiple polyadenylation sites and two alternatively spliced C-terminal exons.** *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* 1998, **1398(3)**:365-370.