

Article

# Combinatorial Discriminant Analysis Applied to RNAseq Data Reveals a Set of 10 Transcripts as Signatures of Exposure of Cattle to *Mycobacterium avium subsp. paratuberculosis*

Michela Malvisi <sup>1,2,†</sup>, Nico Curti <sup>3,†</sup> , Daniel Remondini <sup>3,\*,†</sup> , Maria Grazia De Iorio <sup>2</sup>, Fiorentina Palazzo <sup>4</sup>, Gustavo Gandini <sup>2</sup>, Silvia Vitali <sup>3</sup>, Michele Polli <sup>2</sup>, John L. Williams <sup>5</sup>   
and Giulietta Minozzi <sup>2,\*,†</sup> 

<sup>1</sup> Parco Tecnologico Padano, 26900 Lodi, Italy; michela.malvisi@yahoo.com

<sup>2</sup> Department of Veterinary Medicine DIMEVET, University of Milan, 20133 Milan, Italy; mariagrazia.deiorio1993@gmail.com (M.G.D.I.); Gustavo.gandini@unimi.it (G.G.); Michele.polli@unimi.it (M.P.)

<sup>3</sup> Department of Physics and Astronomy, University di Bologna, 40126 Bologna, Italy; nico.curti2@unibo.it (N.C.); silvia.vitali4@unibo.it (S.V.)

<sup>4</sup> Faculty of Bioscience and Technology for Food, Agriculture and Environment, University of Teramo, 64100 Teramo, Italy; fpalazzo@unite.it

<sup>5</sup> Davies Research Centre, School of Animal and Veterinary Sciences, University of Adelaide, Roseworthy, South Australia 5005, Australia; john.williams01@adelaide.edu.au

\* Correspondence: Daniel.remondini@unibo.it (D.R.); Giulietta.minozzi@unimi.it (G.M.)

† These authors contributed equally to this work.

Received: 16 January 2020; Accepted: 28 January 2020; Published: 5 February 2020



**Simple Summary:** Paratuberculosis or Johne's disease (JD) in cattle is a chronic granulomatous gastroenteritis caused by infection with *Mycobacterium avium subspecies paratuberculosis* (MAP). JD is not treatable; therefore, the early identification of infected animals is a key point in reducing its incidence worldwide. In this paper, combinatorial discriminant analysis was applied to transcriptomic data to find a small set of differentially expressed genes able to discriminate between exposed cattle from non-exposed animals. Results of the discriminant analysis identified 10 transcripts that differentiate between ELISA-negative animals belonging to paratuberculosis-positive herds that were potentially exposed to MAP and negative-unexposed animals belonging to paratuberculosis-negative herds. The 10 transcript signature was also used for the classification of 5 cattle positive-for-MAP infection samples, with the underlying hypothesis that their transcriptional profile should be closer to potentially exposed animals than to controls. The same set of 10 transcripts was able to differentiate ELISA-negative unexposed animals from positive animals based on the results of the ELISA test for bovine paratuberculosis and faecal culture. In conclusion, these findings suggest the possible use of the identified RNA expression signature as a new diagnostic test for paratuberculosis.

**Abstract:** Paratuberculosis or Johne's disease in cattle is a chronic granulomatous gastroenteritis caused by infection with *Mycobacterium avium subspecies paratuberculosis* (MAP). Paratuberculosis is not treatable; therefore, the early identification and isolation of infected animals is a key point to reduce its incidence. In this paper, we analyse RNAseq experimental data of 5 ELISA-negative cattle exposed to MAP in a positive herd, compared to 5 negative-unexposed controls. The purpose was to find a small set of differentially expressed genes able to discriminate between exposed animals in a preclinical phase from non-exposed controls. Our results identified 10 transcripts that differentiate between ELISA-negative, clinically healthy, and exposed animals belonging to paratuberculosis-positive herds and negative-unexposed animals. Of the 10 transcripts, five (*TRPV4*, *RIC8B*, *IL5RA*, *ERE*, *CDC40*) showed significant differential expression between the three groups while the remaining 5 (*RDM1*,

*EPHX1*, *STAU1*, *TLE1*, *ASB8*) did not show a significant difference in at least one of the pairwise comparisons. When tested in a larger cohort, these findings may contribute to the development of a new diagnostic test for paratuberculosis based on a gene expression signature. Such a diagnostic tool could allow early interventions to reduce the risk of the infection spreading.

**Keywords:** Johne's disease; bovine; *Mycobacterium avium subsp. paratuberculosis*; RNAseq; combinatorial discriminant analysis; biomarker discovery

---

## 1. Introduction

Paratuberculosis or Johne's disease (JD) in cattle is a chronic granulomatous gastroenteritis caused by infection with *Mycobacterium avium subspecies paratuberculosis* (MAP) [1,2]. JD is present worldwide, is a welfare issue and causes significant economic losses. Cattle are usually infected as young calves but typically do not show clinical signs before 24 months of age; however, not all infected animals progress to clinical disease [1]. During this extended period of silent infection, animals may shed MAP in their feces into the environment before infection can be diagnosed with the current methods. JD is not treatable; therefore, the early identification and isolation of infected animals before they start shedding the bacteria is a key point to reduce its incidence in cattle herds worldwide.

An association between MAP and Crohn's disease (CD) in humans has been suggested. Both JD and CD share similar symptoms and lesions, and MAP has been identified in tissue isolates of CD patients; however, the causative association is still controversial [3,4]. Given the economic losses and welfare concerns for livestock, and possible human health risk, the research interest in JD has been focused on the substantial difficulty in early diagnosis of infected animals and the exploration of potential new diagnostic techniques [5].

Several studies have analyzed transcriptomic profiles of both RNA and MicroRNA in cattle to identify indicators of MAP infection status [6–10]. One of these studies [9] used RNAseq transcriptomic profiling of bovine monocyte-derived macrophages (MDM) cultured in vitro and infected with the L1 MAP strain and identified 3212 differentially expressed transcripts at 2, 6 and 24 h post-infection. Understanding the host–pathogen interaction in the initial phases is essential to understand the biology of infection. Nevertheless, the transcriptomic profile of the later phases of the disease may guide the identification of changes occurring in the pre-clinical and clinical phase of disease.

In a previous study conducted by the authors of the current manuscript, the whole blood transcriptomic profiles of 5 animals serologically positive to the ELISA test for MAP (PP), 5 ELISA-negative potentially exposed animals (NP) and 5 serologically negative unexposed control animals (NN) were described [8]. In total, in our previous work, we identified 258 differentially expressed genes (DE) between the three groups: 162 were differentially expressed between PP vs. NN, 94 between NP vs. NN and 2 between PP and NP.

In the context of high-throughput data analysis, a challenge is defining an optimal choice of variables (a “signature”) to classify groups of samples or regress trends with optimal performance and minimum dimensionality. Usually, high-throughput omics data (e.g., transcriptomics, genomics, methylomics) provide datasets with tens to hundreds of samples, and often 1000 times larger numbers of variables [11]. The objective of dimensionality reduction through the choice of an optimal signature is twofold: (1) the identification of relevant variables that should separate the signal from the noise (i.e., variables not significantly associated to, or descriptive of the studied process); and (2) in a practical context, it is important to establish future diagnostic criteria that can be implemented in cheap and simple toolkits, such as PCR cards or dedicated microarray chips, that usually test a small number of transcripts (ranging from tens to hundreds, at most). We have recently applied dimensionality reduction methods in other high-throughput gene expression contexts, with, e.g.,  $10^4$  transcripts, and obtained an approximately 1000-fold reduction of the number of transcripts required while retaining

the discriminatory signatures [12–14]. These studies show the feasibility of such approaches, which could be translated into diagnostics for clinical practice, or to suggest different therapeutic strategies based on the patient's genomic profile (in the scope of so-called personalized medicine).

The purpose of the present study was to apply machine learning approaches, specifically a custom algorithm involving Combinatorial Discriminant Analysis, to RNAseq data on cattle belonging to three groups based on different MAP infection status [8], to find an optimal low-dimensional signature able to discriminate between ELISA-negative potentially exposed animals compared to negative-unexposed controls. Such a signature, when validated in a larger cohort of animals, could aid early interventions to reduce sanitary and economic burdens and to reduce the risk of JD infection spreading.

## 2. Materials and Methods

### 2.1. Animal Resource

Holstein dairy cows (15 animals in total) were selected from two herds, one MAP-free (verified by ELISA and the absence of clinical cases) and the other positive for MAP (verified by ELISA, faecal culture and the presence of clinical cases). From the positive herd, 5 positive (PP) cows based on ELISA test results and 5 ELISA-negative, potentially exposed (NP) animals were chosen. All animals were 4 to 5 years old, had a body condition score (BCS) of 3 points and were at 170 to 190 days in milk (DIM) when sampled. The five ELISA positive subjects (PP) were confirmed by faecal culture. Five negative non-exposed control animals (NN) were selected from the negative herd, matched for age, stage of lactation, BCS and DIM with the positive and exposed animals [8].

### 2.2. Sample Preparation, RNA Extraction and Quality Control

Samples preparation, DNA extraction, library preparation and quality control are detailed in Malvisi et al. (2016) [8]. In brief, samples used were taken during obligatory routine animal sanitary controls by an authorised veterinarian. No ethical approval was required, in compliance with the European Directive 2010/63/UE and the Italian Regulation D. Lgs n. 26/2014. Whole blood was collected into PAXgene® Blood RNA tubes (PreAnalytiX GmbH). Total RNA was extracted using the PAXgene® Blood miRNA kit (PreAnalytiX GmbH). RNA concentration was measured by a NanoDrop™ 1000 spectrophotometer (Thermo Scientific) and RNA integrity was assessed with an Agilent 2200 TapeStation system (Agilent Technologies).

### 2.3. RNA-Seq Library Preparation and Sequencing

Libraries were prepared with the Illumina Truseq RNA sample prep kit (Illumina Inc., USA), size and yield were evaluated using an Agilent TapeStation 2200. Fifteen libraries were prepared and equimolar amounts of 5 samples were mixed before NaOH denaturation. Each of the pools was run in a lane of a HiSeq Flowcell. The Illumina Truseq PE cluster kit v3 was used to generate clusters on the grafted Illumina Flowcell and the hybridised libraries were sequenced on a HiSeq2000 with a 100 cycles of paired-end sequencing module using the Truseq SBS kit v3 (Illumina Inc., San Diego, CA, USA).

### 2.4. Availability of Data

All data generated or analysed during this study are available online (GEO reference number GSE130124). Furthermore, all data are available on request. In addition, all transcript counts per sample are given as supplementary information files GSE130124.

### 2.5. RNA-Seq Data Analysis

Preliminary quality control of raw reads was carried out with FastQC software v0.11.2 [15]. Illumina raw sequences were trimmed using Trimmomatic [16] and PCR primers and Illumina adapter sequences were removed. Minimum base quality 15 over a 4-base sliding window was required, then only sequences longer than 36 nucleotides were included in the downstream analysis. Reads

which successfully passed trimming were mapped against the *Bos taurus* UMD3.1.68 reference genome sequence, using STAR [17] aligner to obtain BAM alignment files. The BAM files were sorted and indexed using Samtools [18]. In order to quantify counts for each sample, a list of genes and relative abundance of mapping reads were extracted using htseq-count [19]. These count files were used for downstream statistical analysis. Data from Malvisi et al. [8] was used (GEO reference number GSE130124).

## 2.6. Signature Identification

A combinatorial discriminant analysis was conducted on the RNAseq data of the 15 samples derived from the study published by the authors [8], which has been previously analysed using a traditional pairwise comparison between groups at a single gene level. In detail, the dataset comprised 15,036 transcripts from 15 samples, classified as “serologically negative non-exposed cows/healthy” (5 samples, labelled as NN), “serologically negative exposed cows” (5 samples, NP) and “serologically positive/clinical cows” (5 samples, PP). Only transcripts with non-zero counts for all samples were considered, reducing the dataset to 13,529 transcripts.

A custom feature selection algorithm based on Quadratic Discriminant Analysis was used to identify an optimal low-dimensional signature able to discriminate between NN and NP samples based on a method developed by the authors, described in [20–22], which had been used previously to analyse biological and social data [13,14,21]. The algorithm evaluates all possible pairs of transcripts which it groups onto a transcript–transcript network weighted by classification performance. The putative signatures are obtained through thresholding based on classification performance for the disconnected components resulting from the initial weighted network. However, optimal performance of the algorithm requires a larger sample size than that available in this study [14,20–22], to allow more performance threshold values to be tested. The smaller sample size used here tends to generate a single giant component with a large number of transcripts. Thus, to reduce the dimensionality of the classification signature, (1) the most central nodes of the giant component were considered, with a combination of two centrality measures (betweenness centrality [23] and clustering coefficient [24]) that reduced the signature size to 123 transcripts; (2) the pendant nodes (i.e., with only one link) were recursively removed, until no further removals were allowed, leading to a final signature of 10 transcripts. In order to validate the classification performance of the signature, as the low number of samples did not allow the application of procedures such as robust k-fold cross-validation with a subset of samples of the same classes, we used the signature to classify the PP data, which was not used for signature identification, with the hypothesis that these samples should be classified more closely to the NP samples than the NN ELISA-negative samples.

## 3. Results

### 3.1. Results of the NGS Pipeline Analysis

The results of the RNAseq analysis pipeline and list of differentially expressed genes between the three animal groups (PP, NP and NN) are described by Malvisi et al. [8]. In summary, RNAseq data for whole blood of the 5 PP, 5 NP and 5 NN animals identified 12,366 genes, of which 258 were differentially expressed (DE) in the three comparisons: 162 genes were DE comparing PP vs. NN, 94 genes for NP vs. NN and 2 genes for PP vs. NP. Differential expression was defined as  $\log_2$  fold change value above 1 or below -1 and a FDR threshold of 0.05. The complete list of DE genes for each comparison, derived from Malvisi et al. [8], is given in Supplementary tables (Tables S1–S3).

### 3.2. Results of the Signature Analysis

Starting from the top-performing pairs of transcripts, with the first thresholding step of the weighted transcript network (see the Methods section) we obtained an initial signature of 123 transcripts capable of correctly classifying 4 out of the 5 NN samples (80%) and all 5 NP samples (100% performance). The average performance was therefore 90%, with a Matthews correlation coefficient  $MCC = 0.82$  (a complete list of the 123 transcripts is given in Supplementary Table S4).

Processing the 123-transcript network by removing all pendant nodes, we obtained a final signature with 10 transcripts (Table 1) with a 100% performance classifying all NN and NP samples (Figure 1). Five transcripts of the signature show significant difference in the pairwise comparisons between the NN and NP groups (TRPV4, RIC8B, IL5RA, ERF and CDC40, see Table 1). The minimum  $\log_2$  fold change difference was  $-0.016$  (*ASB8* gene transcript) while maximum was  $-1.556$  for the *IL5RA* transcript.

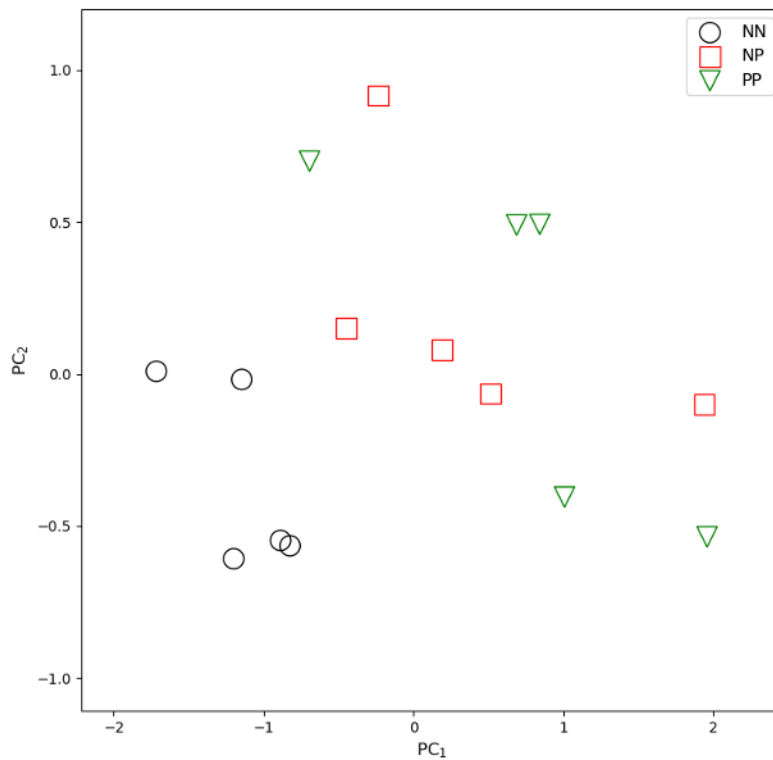
**Table 1.** The set of 10 transcripts that best discriminate between 5 animals serologically positive to the ELISA test for MAP (PP), 5 ELISA-negative potentially exposed (NP) and 5 serologically negative unexposed control animals (NN), and their corresponding pairwise  $\log_2$  fold change in the PP vs. NN, NP vs. NN and PP vs. NP comparisons.

| Gent Name    | PP vs. NN | NP vs. NN | PP vs. NP |
|--------------|-----------|-----------|-----------|
| <i>RDM1</i>  | 0.153     | 0.492     | -0.338    |
| <i>TRPV4</i> | 1.535 *   | 1.457 *   | 0.948 *   |
| <i>EPHX1</i> | 0.132     | 0.104     | 0.028     |
| <i>RIC8B</i> | -0.538 *  | -0.415    | -0.122    |
| <i>STAU1</i> | 0.081     | 0.155     | -0.074    |
| <i>TLE1</i>  | 0.238     | 0.445     | -0.207    |
| <i>IL5RA</i> | -1.556 *  | -1.191    | -0.366    |
| <i>ASB8</i>  | -0.095    | -0.078    | -0.016    |
| <i>ERF</i>   | 0.681 *   | 0.554     | 0.126     |
| <i>CDC40</i> | -0.429 *  | -0.387    | -0.042    |

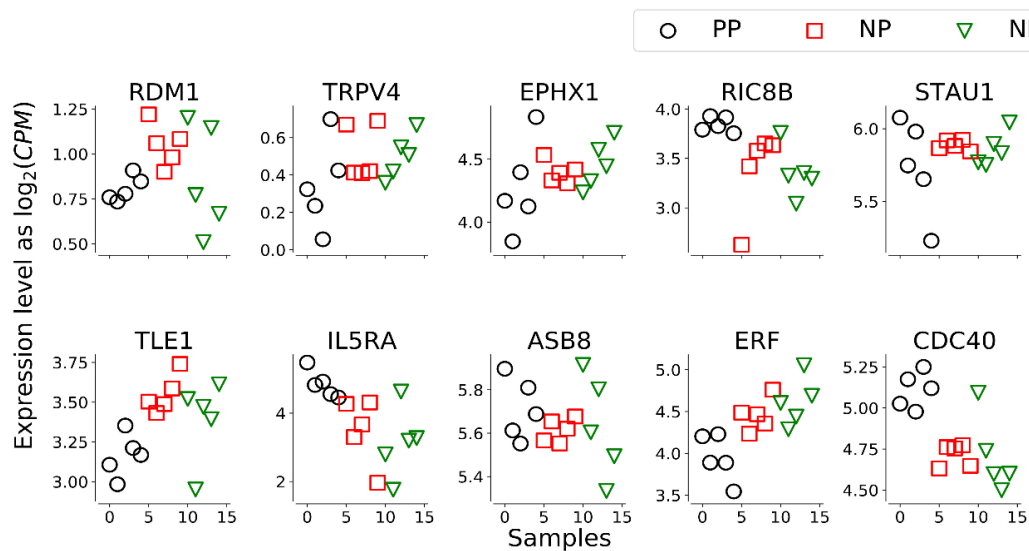
\*  $\log_2$  fold changes that show significant difference ( $p < 0.05$ ) in the comparison tests (Student's *t*-test).

The first two principal components in a PCA of the 10-transcript signature (Figure 1) highlight a clear separation between NN and NP + PP samples. Moreover, the expression levels of the single transcripts of the signature show a progressive increase or decrease from a healthy (NN) samples to a positive (PP) samples, passing through the exposed (NP) sample class, specifically for *EPHX1*, *IL5RA*, *ERF* and *CDC40* (Figure 2). In the 123-transcript network, all nodes are directly connected to a central 10-transcript core, while only these core 10 transcripts, which form the signature, are connected to each other (see Figure 3).

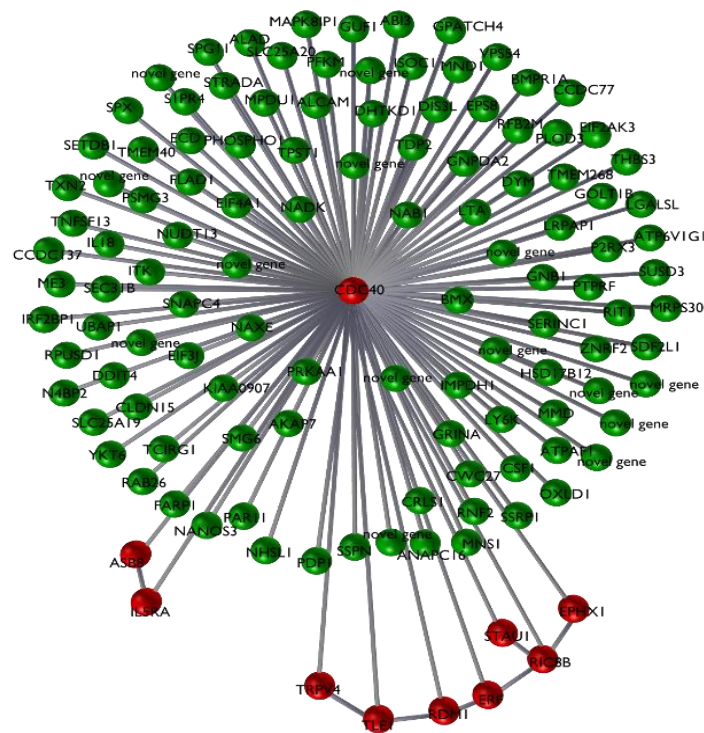
To further validate the goodness of the signature, we generated 10,000 different signatures with 10 randomly chosen transcripts, and then applied a Leave-One-Out cross validation procedure to quantify the classification performance of these signatures. Only 50 of these signatures (i.e., the 0.5% of the total random signatures) produced a better classification performance than that achieved using the core 10 transcripts. Thus, it is unlikely to obtain a signature of similar performance with the same number of transcripts by pure chance. The statistical significance of the signature identified using our feature selection algorithm would be improved by extending the data set.



**Figure 1.** Principal Component Analysis (first 2 components) applied to the three groups of animals analyzed by RNAseq analysis (5 animals serologically positive to the ELISA test for MAP (PP), 5 ELISA-negative potentially exposed animals (NP) and 5 serologically negative unexposed control animals (NN), based on the RNA-seq profile of the 10 signature transcripts.



**Figure 2.** Plot of transcript levels for the 10 genes belonging to the signature. Some transcripts (EPHX1, IL5RA, ERF, CDC40) show a clear trend when moving from 5 animals serologically positive to the ELISA test for MAP (PP), to 5 ELISA-negative potentially exposed animals (NP) and 5 serologically negative unexposed control animals (NN).



**Figure 3.** Plot of the 123-transcript network (green nodes), with a detail of the 10-probe signature (red nodes).

#### 4. Discussion

Earlier detection of subclinical disease is urgently required to enhance control of MAP. However, the mechanisms involved in progressive evolution of JD phases are still unknown, which limits the development, identification and sensitivity of diagnostic tools. Gene expression profiling from peripheral blood has been shown to discriminate patients affected by Crohn’s Disease, ulcerative colitis and non-inflammatory diarrheal disorders [25]. Recently, a smaller panel of only 6 genes that are differentially expressed during inflammatory bowel disease, identified by supervised learning methods, were shown to discriminate between the different stages of the disease such as mild severity, moderate-to-severe, inactive forms of Crohn’s disease and ulcerative colitis with very high sensitivity [26].

Further examples involve the development of a small panel, composed of seven differentially expressed genes (*ANXA3*, *CLEC4D*, *LMNB1*, *PRRG4*, *TNEAIP6*, *VNN1* and *IL2RB*), able to discriminate between subjects according to their relative risk for colorectal cancer in an average-risk population based on 642 samples [27]. Therefore, mRNA expression profiling is seen as a potential source of biomarkers for early diagnosis of JD [7,28,29]. Several DE transcripts have been identified in whole blood of Holstein–Friesian calves challenged with a high or low dose of MAP three months after inoculation and control calves [28]. Furthermore, the same authors identified that calves receiving a high dose of MAP showed more up-regulated transcripts. In a study of a mixed population of Holstein and Holstein Red calves aged between 2 to 4 months inoculated experimentally with MAP, 63 DE genes were identified early in infection [29]. Of the genes identified in these studies, some have a role in immune function (*IL7*, *TRIM13*, *ICOS*, *CD46*), but none are in common with the 10 genes forming the core signature identified in the present study. However, experimental designs differed, and responses of young calves exposed to experimental MAP infection may not coincide with those of naturally infected (PP) adult cattle examined here.

The aims of high-throughput discovery studies are to gain deeper biological insight into the underlying phenomena (as in Malvisi et al. [8]), or to find an optimal subset of variables able to describe

the phenomenon (e.g., in terms of sample classification or stratification) by means of robust statistical analyses. In this paper, we applied a custom dimensionality reduction method to the initial set of variables to obtain a multivariate signature, with optimal performance in terms of sample classification using a small subset of variables. The authors have successfully applied this approach in other contexts involving high-throughput transcriptomics data. These analyses allowed samples to be stratified and different therapeutic strategies based on individual genomic profiling to be proposed, and in some cases, specific pathways associated to the clinical outcome to be identified [12–14]. The methodology identifies DE genes that individually may not be significant or biologically relevant, but that taken together as a group, gain discriminative value.

Results of supervised classification analysis of the RNAseq data reported here identified a signature of 10 transcripts that is able to differentiate between ELISA-negative, likely exposed, but clinically healthy animals belonging to JD positive herds (NP), and negative-unexposed animals (NN). Furthermore, the same set of 10 transcripts differentiate negative-unexposed (NN) animals from positive animals defined by the ELISA test for bovine paratuberculosis and faecal culture (PP).

Within the 10 transcripts of the signature, 5 genes (*TRPV4*, *RIC8B*, *IL5RA*, *ERF* and *CDC40*) showed significant differential expression between the NN, NP and/or PP groups, while the remaining 5 transcripts (*RDM1*, *EPHX1*, *STAU1*, *TLE1*, *ASB8*) did not show a significant difference in least in one of the pairwise comparisons (Supplementary Table S4).

Among the significantly differentially expressed genes in the signature, the cell division cycle 40 gene (*CDC40*) showed the smallest fold change between classes, this gene is the most central node of the signature network associated with JD health status (Figure 3). *CDC40* is located on BTA9, and was under-expressed in the NP and PP groups, compared with the NN group. *CDC40* has been shown to be involved in clathrin-mediated endocytosis [30]. Clathrin is the best characterised coat protein involved in the endocytosis process, specifically in receptor-mediated internalization [31]. *Mycobacterium paratuberculosis* enters the host macrophages, its primary target cell, and manages to survive within the phagosome [32]. It is possible that the under-expression of *CDC40* in infected and sick animals, compared to unexposed animals, may facilitate the survival of the pathogen within the host target cell. In addition to the *CDC40* gene, other transcripts in the core set are involved with immune response mechanisms; these include *IL5RA*, *ERF* and *TRPV4*. These genes potentially have functions related to the biology of the progression of JD. Interleukin 5 receptor subunit alpha (*IL5RA*) is under-expressed in infected NP and serologically positive PP animals compared with unexposed NN animals. IL5 is a cytokine that acts as a growth and differentiation factor for both B cells and eosinophils. Recently *IL5RA* was found to be under-expressed in human tuberculosis (TB) patients compared with healthy controls [33]. Further, after 2 and 6 months of treatment for TB, *IL5RA* was found overexpressed in the treated patient group compared to the healthy controls, indicating a reactivation of the immune cells [33]. This is in line with the present study, where this gene is significantly under-expressed in the para-tuberculosis infected and exposed samples compared with controls. *IL5RA* may possibly be incorporated into a diagnostic tool to monitor disease progression. It is worth noting that *IL5RA* belongs to the most affected gene network associated with PP vs. NN [8]. This network is enriched in genes associated with the inflammatory response. *IL5RA* shows several connections with other DE genes in the study, indicating that *IL5RA* acts as a central node [8]. The ETS2-repressor factor gene (*ERF*) encodes a transcriptional repressor factor that is ubiquitously expressed and known to have tumor suppressor activity [32]. *ERF* is one of the five genes that was over-expressed in the PP and NP group compared with the NN group (with  $\log_2$  fold change of 0.68 and 0.55  $\log_2$  fold change, respectively, between the PP and NP groups vs. the NN group). However, *ERF* was not differentially expressed between the PP and NP groups (0.126  $\log_2$  fold change and 0.48 p-value), although it was one of the five genes out of 10 that showed statistical significance in the pairwise comparisons [8]. Several studies suggest a role of *ERF* in the immune response and macrophage biology and is a candidate for ParaTB diagnostics [34]. It has been shown that the human Erf protein decreases the expression of mouse c-Myc mRNA in RAS-3T3 cells [34]. In normal conditions, c-Myc is a



transcription factor that binds to more than 10% of human promoters and regulates several crucial cell functions [34]. ERF binding to the *c-Myc* promoter has been shown to reduce *c-Myc* transcription [34], and overexpression of *ERF* causes down-regulation of *c-Myc*. Interestingly, *c-Myc* is a key factor in macrophage differentiation [35]. Inhibition of *c-Myc* expression may prevent macrophage activation induced by infecting bacteria. Furthermore, the overexpression of *ERF* and repression of *c-Myc* is consistent with the under-expression we observed of *IL-15RA* [8]. The protein encoded by the *c-Myc* has been shown to affect *IL-15RA* mRNA levels in murine studies [36]. The Transient Receptor Potential Vanilloid 4 gene (*TRPV4* located on chromosome 17 in cattle), was overexpressed in the NP and PP animals compared with the unexposed animals. *TRPV4* encodes a non-selective cation channel (for, e.g., sodium, calcium, potassium) that is expressed widely and is involved in numerous metabolic processes in humans, including inflammation, where it contributes to disease progression [37]. Interestingly, *TRPV4* has a role in gastrointestinal inflammatory diseases such as ulcerative colitis (UC) and Crohn's disease (CD), which are part of a group of inflammatory bowel diseases (IBD) [38]. *TRPV4* is expressed widely in the intestinal tract and activation of *TRPV4* causes the release of pro-inflammatory cytokines and an inflammatory response a few hours after administration [38]. Furthermore, over-expression of *TRPV4* has been observed in murine inflamed colon tissues compared with healthy controls [38]. This suggests that as the inflammatory environment increases, both *TRPV4* expression and the activation of the non-selective cation channel help to maintain the inflammatory state. This is consistent with the increased level of expression of *TRPV4* in NP and PP animals compared to the unexposed animals (NN) seen in the present study, and suggests that *TRPV4* should be included in the set of diagnostic markers for paratuberculosis.

The other transcripts identified by the discriminant analysis do not show an obvious role in JD or bacterial infections in general. The *RDM1* (RAD52 motif containing 1) gene codes for a protein that is involved in a DNA damage repair pathway, and in the cellular response to cisplatin, a drug commonly used in chemotherapy [39,40]. *STAU1* is a member of the family of double-stranded RNA (dsRNA)-binding proteins involved in the transport of mRNAs to different organelles or cell compartments. In cattle, the protein has been shown to have a role during mammalian oocyte maturation and early embryogenesis [41].

*EPHX1* (Epoxide Hydrolase 1) codes for an enzyme involved in the degradation of aromatic compounds to less reactive and more water-soluble compounds [42]. Mutations of *EPHX1* may alter its enzymatic function and lead to Familiar Hypercholanemia, a genetic disorder characterized by high serum bile acid concentrations, itching, and fat malabsorption [43]. *ASB8* (Ankyrin Repeat and SOCS Box Containing 8) has a putative role in the growth and proliferation of lung cancer, possibly as a positive regulator [44]. The *TLE1* (TLE Family Member 1, Transcriptional Corepressor) gene product is involved in hemopoiesis, epithelial and neuronal differentiation and has a role in repressing cell differentiation [45]. *TLE1* has been used as a immunohistochemical marker for Synovial Sarcoma, an adult soft tissue sarcoma [46]. Lastly, the *RIC8B* (RIC8 Guanine Nucleotide Exchange Factor B) gene product is known to interact with G-alpha proteins [47] and potentiates odorant receptor signal transduction by increasing G (olf)- alpha-dependent cAMP accumulation [48].

In summary, the results of the discriminant analysis identified a 10-gene signature, the expression levels of which could discriminate between the exposed and sero-converted animals. Not all of the 10 genes were DE or relate to immune function. To further validate these findings, a larger group of animals should be tested. The study was limited by the unknown exposure status of the NP group, which gives uncertainty in the interpretation of the signature identified, and by the lack of a longitudinal information on the disease outcome of the NP group.

## 5. Conclusions

In conclusion, the discriminant analysis described here identified a 10-gene signature, the expression levels of which could discriminate between exposed and sero-converted animals and suggest the possible use of RNA expression analysis as a new diagnostic test for JD. In particular,

the approach may be able to identify infected or/exposed animals prior to sero-conversion and a positive ELISA test result. However, further tests for specificity and validation in a larger cohort are required.

**Supplementary Materials:** The following are available online at <http://www.mdpi.com/2076-2615/10/2/253/s1>, Table S1. Differentially expressed genes in the positive subject when compared with the control group. Genes with an FDR < 0.05 and a Log<sub>2</sub> fold change (Log<sub>2</sub>FC) <−1 or >1 were considered as differentially expressed. Table S2. Differentially expressed genes in the exposed subject when compared with the control group. Genes with an FDR < 0.05 and a Log<sub>2</sub> fold change (Log<sub>2</sub>FC) <−1 or >1 were considered as differentially expressed. Table S3. Differentially expressed genes in the positive subject when compared with the exposed group. Genes with an FDR < 0.05 and a Log<sub>2</sub> fold change (Log<sub>2</sub>FC) <−1 or >1 were considered as differentially expressed. Table S4. List of 123 transcripts identified by Combinatorial Discriminant Analysis (CDA) that discriminates between cattle infected and non-infected by *Mycobacterium avium* ssp. *Paratuberculosis*.

**Author Contributions:** Conceptualization: D.R., J.L.W., G.M. Formal analysis: N.C. S.V. Funding acquisition: J.L.W. Investigation: M.M., G.M., F.P., M.G.D.I., M.P. Methodology: D.R., J.L.W. Project administration: J.L.W. Resources: F.P., M.L. Supervision: J.L.W., G.M. Validation: D.R. Visualization: M.M., G.M. Writing—original draft: G.M., D.R. Writing—review & editing: M.M., J.L.W., G.M., G.G., D.R. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by The Italian Ministry of Education and Research ([www.istruzione.it](http://www.istruzione.it)), grant number PON01\_01841 PON EPISUD.

**Acknowledgments:** This work was carried out with the support of grant PON01\_01841 PON EPISUD from the Italian Ministry of Education University and Research ([www.istruzione.it](http://www.istruzione.it)) awarded to J.L.W. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Olsen, I.; Sigurðardóttir, G.; Djønne, B. Paratuberculosis with special reference to cattle. *Vet. Q.* **2002**, *24*, 12–28. [[CrossRef](#)] [[PubMed](#)]
- Whittington, R.J.; Sergeant, E.S.G. Progress towards understanding the spread, detection and control of *Mycobacterium avium* subsp. *paratuberculosis* in animal populations. *Aust. Vet. J.* **2001**, *79*, 267–278. [[CrossRef](#)] [[PubMed](#)]
- Bach, H. What role does *mycobacterium avium* subsp. *paratuberculosis* play in Crohn’s disease? *Curr. Infect. Dis. Rep.* **2015**, *17*, 463. [[CrossRef](#)] [[PubMed](#)]
- Liverani, E.; Scaioli, E.; Cardamone, C.; Dal Monte, P.; Belluzzi, A. *Mycobacterium avium* subspecies *paratuberculosis* in the etiology of Crohn’s disease, cause or epi-phenomenon? *World J. Gastroenterol.* **2014**, *20*, 13060–13070. [[CrossRef](#)]
- Britton, L.E.; Cassidy, J.P.; O’Donovan, J.; Gordon, S.V.; Markey, B. Potential application of emerging diagnostic techniques to the diagnosis of bovine Johne’s disease (*paratuberculosis*). *Vet. J.* **2016**, *209*, 32–39. [[CrossRef](#)]
- Casey, M.E.; Meade, K.G.; Nalpas, N.C.; Taraktsoglou, M.; Browne, J.A.; Killick, K.E.; Park, S.D.; Gormley, E.; Hokamp, K.; Magee, D.A.; et al. Analysis of the Bovine Monocyte-Derived Macrophage Response to *Mycobacterium avium* Subspecies *Paratuberculosis* Infection Using RNA-seq. *Front. Immunol.* **2015**, *6*, 23. [[CrossRef](#)]
- Farrell, D.; Shaughnessy, R.G.; Britton, L.; MacHugh, D.E.; Markey, B.; Gordon, S.V. The Identification of Circulating miRNA in Bovine Serum and Their Potential as Novel Biomarkers of Early *Mycobacterium avium* subsp. *paratuberculosis* Infection. *PLoS ONE* **2015**, *10*, e0134310. [[CrossRef](#)]
- Malvisi, M.; Palazzo, F.; Morandi, N.; Lazzari, B.; Williams, J.L.; Pagnacco, G.; Minozzi, G. Responses of Bovine Innate Immunity to *Mycobacterium avium* subsp. *paratuberculosis* Infection Revealed by Changes in Gene Expression and Levels of MicroRNA. *PLoS ONE* **2016**, *11*, e0164461. [[CrossRef](#)]
- Marino, R.; Capoferri, R.; Panelli, S.; Minozzi, G.; Strozzi, F.; Trevisi, E.; Snel, G.G.M.; Ajmone-Marsan, P.; Williams, J.L. Johne’s disease in cattle: An in vitro model to study early response to infection of *Mycobacterium avium* subsp. *paratuberculosis* using RNA-seq. *Mol. Immunol.* **2017**, *91*, 259–271. [[CrossRef](#)]
- Shaughnessy, R.G.; Farrell, D.; Riepema, K.; Bakker, D.; Gordon, S.V. Analysis of Biobanked Serum from a *Mycobacterium avium* subsp. *paratuberculosis* Bovine Infection Model Confirms the Remarkable Stability of Circulating miRNA Profiles and Defines a Bovine Serum miRNA Repertoire. *PLoS ONE* **2015**, *10*, e0145089. [[CrossRef](#)]

11. Johnstone, I.M.; Titterton, D.M. Statistical challenges of high-dimensional data. *Philos. Trans. A Math. Phys. Eng. Sci.* **2009**, *367*, 4237–4253. [[CrossRef](#)] [[PubMed](#)]
12. Scotlandi, K.; Remondini, D.; Castellani, G.; Manara, M.C.; Nardi, F.; Cantiani, L.; Francesconi, M.; Mercuri, M.; Caccuri, A.M.; Serra, M.; et al. Overcoming resistance to conventional drugs in Ewing sarcoma and identification of molecular predictors of outcome. *J. Clin. Oncol.* **2009**, *27*, 2209–2216. [[CrossRef](#)] [[PubMed](#)]
13. Terragna, C.; Renzulli, M.; Remondini, D.; Tagliafico, E.; Di Raimondo, F.; Patriarca, F.; Martinelli, G.; Roncaglia, E.; Masini, L.; Tosi, P.; et al. Correlation between eight-gene expression profiling and response to therapy of newly diagnosed multiple myeloma patients treated with thalidomide-dexamethasone incorporated into double autologous transplantation. *Ann. Hematol.* **2013**, *92*, 1271–1280. [[CrossRef](#)] [[PubMed](#)]
14. Terragna, C.; Remondini, D.; Martello, M.; Zamagni, E.; Pantani, L.; Patriarca, F.; Pezzi, A.; Levi, G.; Offidani, M.; Proserpio, I.; et al. The genetic and genomic background of multiple myeloma patients achieving complete response after induction therapy with bortezomib, thalidomide and dexamethasone (VTD). *Oncotarget* **2016**, *7*, 9666–9679. [[CrossRef](#)] [[PubMed](#)]
15. FastQC Software. Babraham Institute—Babraham Bioinformatics. Available online: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (accessed on 3 May 2019).
16. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics* **2014**, *30*, 2114–2120. [[CrossRef](#)] [[PubMed](#)]
17. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15–21. [[CrossRef](#)]
18. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [[CrossRef](#)]
19. Anders, S.; Pyl, P.T.; Huber, W. HTSeq—A Python framework to work with high-throughput sequencing data. *Bioinformatics* **2015**, *31*, 166–169. [[CrossRef](#)]
20. Curti, N.; Giampieri, E.; Levi, G.; Castellani, G.; Remondini, D. DNetPRO: A network approach for low-dimensional signatures from high-throughput data. *bioRxiv* **2019**, 773622. [[CrossRef](#)]
21. Mizzi, C.; Fabbri, A.; Rambaldi, S.; Bertini, F.; Curti, N.; Sinigardi, S.; Luzi, R.; Venturi, G.; Micheli, D.; Muratore, G.; et al. Unraveling pedestrian mobility on a road network using ICTs data during great tourist events. *EPJ Data Sci.* **2018**, *7*, 44. [[CrossRef](#)]
22. Curti, N.; Giampieri, E.; Mizzi, C.; Fabbri, A.; Bazzani, A.; Castellani, G.; Remondini, D. A network approach for dimensionality reduction from high-throughput data, SDPS18. 2018; in press.
23. Brandes, U. A Faster Algorithm for Betweenness Centrality. *J. Math. Sociol.* **2001**, *25*, 163–177. [[CrossRef](#)]
24. Fagiolo, G. Clustering in complex directed networks. *Phys. Rev. E* **2007**, *76*, 026107. [[CrossRef](#)] [[PubMed](#)]
25. Burakoff, R.; Chao, S.; Perencevich, M.; Ying, J.; Friedman, S.; Makrauer, F.; Odze, R.; Khurana, H.; Liew, C.C. Blood-based biomarkers can differentiate ulcerative colitis from Crohn’s disease and noninflammatory diarrhea. *Inflamm. Bowel. Dis.* **2011**, *17*, 1719–1725. [[CrossRef](#)] [[PubMed](#)]
26. Burakoff, R.; Pabby, V.; Onyewadume, L.; Odze, R.; Adackapara, C.; Wang, W.; Friedman, S.; Hamilton, M.; Korzenik, J.; Levine, J.; et al. Blood-based biomarkers used to predict disease activity in Crohn’s disease and ulcerative colitis. *Inflamm. Bowel. Dis.* **2015**, *21*, 1132–1140. [[CrossRef](#)]
27. Marshall, K.W.; Mohr, S.; Khattabi, F.E.; Nossova, N.; Chao, S.; Bao, W.; Ma, J.; Li, X.J.; Liew, C.C. A blood-based biomarker panel for stratifying current risk for colorectal cancer. *Int. J. Cancer.* **2010**, *126*, 1177–1186. [[CrossRef](#)]
28. David, J.; Barkema, H.W.; Mortier, R.; Ghosh, S.; Guan, L.L.; De Buck, J. Gene expression profiling and putative biomarkers of calves 3 months after infection with *Mycobacterium avium* subspecies paratuberculosis. *Vet. Immunol. Immunopathol.* **2014**, *160*, 107–117. [[CrossRef](#)]
29. Purdie, A.C.; Plain, K.M.; Begg, D.J.; De Silva, K.; Whittington, R.J. Expression of genes associated with the antigen presentation and processing pathway are consistently regulated in early *Mycobacterium avium* subsp. paratuberculosis infection. *Comp. Immunol. Microbiol. Infect. Dis.* **2012**, *35*, 151–162. [[CrossRef](#)]
30. Kozik, P.; Hodson, N.A.; Sahlender, D.A.; Simecek, N.; Soromani, C.; Wu, J.; Collinson, L.M.; Robinson, M.S. A human genome-wide screen for regulators of clathrin-coated vesicle formation reveals an unexpected role for the V-ATPase. *Nat. Cell Biol.* **2013**, *15*, 50–60. [[CrossRef](#)]
31. Pieters, J. Entry and survival of pathogenic mycobacteria in macrophages. *Microbes Infect.* **2001**, *3*, 249–255. [[CrossRef](#)] [[PubMed](#)]

32. Armstrong, J.A.; Hart, P.D.A. Response of cultured macrophages to *Mycobacterium tuberculosis*, with observations on fusion of lysosomes with phagosomes. *J. Exp. Med.* **1971**, *134*, 713–740. [[CrossRef](#)]
33. Van Rensburg, I.C.; Wagman, C.; Stanley, K.; Beltran, C.; Ronacher, K.; Walzl, G.; Loxton, A.G. Successful TB treatment induces B-cells expressing FASL and IL5RA mRNA. *Oncotarget* **2017**, *8*, 2037–2043. [[CrossRef](#)] [[PubMed](#)]
34. Verykokakis, M.; Papadaki, C.; Vorgia, E.; Le Gallic, L.; Mavrothalassitis, G. The RAS-dependent ERF control of cell proliferation and differentiation is mediated by c-Myc repression. *J. Biol. Chem.* **2007**, *282*, 30285–30294. [[CrossRef](#)]
35. Pello, O.M.; De Pizzol, M.; Mirolo, M.; Soucek, L.; Zammataro, L.; Amabile, A.; Doni, A.; Nebuloni, M.; Swigart, L.B.; Evan, G.I.; et al. Role of c-MYC in alternative activation of human macrophages and tumor-associated macrophage biology. *Blood* **2012**, *119*, 411–421. [[CrossRef](#)] [[PubMed](#)]
36. Thomas-Tikhonenko, A.; Viard-Leveugle, I.; Dews, M.; Wehrli, P.; Seignani, C.; Yu, D.; Ricci, S.; el-Deiry, W.; Aronow, B.; Kaya, G.; et al. Myc-transformed epithelial cells down-regulate clusterin, which inhibits their growth in vitro and carcinogenesis in vivo. *Cancer Res.* **2004**, *64*, 3126–3136. [[CrossRef](#)] [[PubMed](#)]
37. Grace, M.S.; Bonvini, S.J.; Belvisi, M.G.; McIntyre, P. Modulation of the TRPV4 ion channel as a therapeutic target for disease. *Pharmacol. Ther.* **2017**, *177*, 9–22. [[CrossRef](#)] [[PubMed](#)]
38. D'Aldebert, E.; Cenac, N.; Rousset, P.; Martin, L.; Rolland, C.; Chapman, K.; Selves, J.; Alric, L.; Vinel, J.P.; Vergnolle, N. Transient receptor potential vanilloid activated inflammatory signals by intestinal epithelial cells and colitis in mice. *Gastroenterology* **2011**, *140*, 275–285. [[CrossRef](#)] [[PubMed](#)]
39. Tong, L.; Liu, J.; Yan, W.; Cao, W.; Shen, S.; Li, K.; Li, L.; Niu, G. RDM1 plays an oncogenic role in human lung adenocarcinoma cells. *Sci. Rep.* **2018**, *8*, 11525. [[CrossRef](#)]
40. Hamimes, S.; Arakawa, H.; Stasiak, A.Z.; Kierzek, A.M.; Hirano, S.; Yang, Y.G.; Takata, M.; Stasiak, A.; Buerstedde, J.M.; Van Dyck, E. RDM1, a novel RNA recognition motif (RRM)-containing protein involved in the cell response to cisplatin in vertebrates. *J. Biol. Chem.* **2005**, *280*, 9225–9235. [[CrossRef](#)] [[PubMed](#)]
41. Calder, M.D.; Madan, P.; Watson, A.J. Bovine oocytes and early embryos express Staufen and ELAVL RNA-binding proteins. *Zygote* **2008**, *16*, 161–168. [[CrossRef](#)]
42. Sivonova, M.; Dobrota, D.; Matakova, T.; Grobarcikova, S.; Habala, V.; Salagovic, J.; Tajtakova, M.; Pidanicova, A.; Valansky, L.; Lachvac, L.; et al. Microsomal epoxide hydrolase polymorphisms, cigarette smoking and prostate cancer risk in the Slovak population. *Neoplasma* **2012**, *59*, 79–84. [[CrossRef](#)]
43. Zhu, Q.; Xing, W.; Qian, B.; Von Dippe, P.; Shneider, B.L.; Fox, V.L.; Levy, D. Inhibition of human m-epoxide hydrolase gene expression in a case of hypercholesterolemia. *Biochim. Biophys.* **2003**, *1638*, 208–216. [[CrossRef](#)]
44. Liu, Y.Z.; Li, J.J.; Zhang, F.R.; Shen, Y.; Yao, M.; Wan, D.F.; Gu, J.R. Exogenous expression of SOCS box-deficient mutant ASB-8 suppresses the growth of lung adenocarcinoma SPC-A1 cells. *Sheng Wu Hua Xue Yu Sheng Wu Wu Li Xue Bao (Shanghai)* **2003**, *35*, 548–553. [[PubMed](#)]
45. Ali, Z.; Haroon Khan, A.; Rehman, U.; Faisal, M.; Ahmad, I.N.; Mamoon, N.; Nasir, H.; Hameed, Z. Is TLE1 Expression Limited to Synovial Sarcoma? Our Experience at Shifa International Hospital, Pakistan. *Cureus* **2019**, *11*, e6259. [[CrossRef](#)] [[PubMed](#)]
46. Jagdis, A.; Rubin, B.P.; Tubbs, R.R.; Pacheco, M.; Nielsen, T.O. Prospective evaluation of TLE1 as a diagnostic immunohistochemical marker in synovial sarcoma. *Am. J. Surg. Pathol.* **2009**, *33*, 1743–1751. [[CrossRef](#)]
47. Von Dannecker, L.E.; Mercadante, A.F.; Malnic, B. Ric-8B, an olfactory putative GTP exchange factor, amplifies signal transduction through the olfactory-specific G-protein G $\alpha$ olf. *J. Neurosci.* **2005**, *25*, 3793–3800. [[CrossRef](#)]
48. Von Dannecker, L.E.; Mercadante, A.F.; Malnic, B. Ric-8B promotes functional expression of odorant receptors. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 9310–9314. [[CrossRef](#)]

