



Article

Intelligent Microarray Data Analysis through Non-negative Matrix Factorization to Study Human Multiple Myeloma Cell Lines

Gabriella Casalino ^{1,*} , Mauro Coluccia ², Maria L. Pati ², Alessandra Pannunzio ² ,
Angelo Vacca ³, Antonio Scilimati ^{2,*} and Maria G. Perrone ^{2,*}

¹ Department of Informatics, University of Bari “Aldo Moro”, Via Orabona 4, 70125 Bari, Italy

² Department of Pharmacy-Pharmaceutical Sciences, University of Bari “Aldo Moro”, Via Orabona 4, 70125 Bari, Italy; mauro.coluccia@uniba.it (M.C.); marialaura.pati@uniba.it (M.L.P.); alessandra.pannunzio@uniba.it (A.P.)

³ Department of Biomedical Sciences and Human Oncology, Internal Medicine Unit G. Baccelli, University of Bari Aldo Moro Medical School, 70125, Bari, Italy; angelo.vacca@uniba.it

* Correspondence: gabriella.casalino@uniba.it (G.C.); antonio.scilimati@uniba.it (A.S.); mariagrazia.perrone@uniba.it (M.G.P.); Tel.: +39-0805442203 (G.C.); +39-0805442753 (A.S.); +39-0805442747 (M.G.P.); Fax: +39-0805442724 (A.S.); +39-0805442724 (M.G.P.)

Received: 23 October 2019; Accepted: 10 December 2019; Published: 17 December 2019



Abstract: Microarray data are a kind of numerical non-negative data used to collect gene expression profiles. Since the number of genes in DNA is huge, they are usually high dimensional, therefore they require dimensionality reduction and clustering techniques to extract useful information. In this paper we use NMF, non-negative matrix factorization, to analyze microarray data, and also develop “intelligent” results visualization with the aim to facilitate the analysis of the domain experts. For this purpose, a case study based on the analysis of the gene expression profiles (GEPs), representative of the human multiple myeloma diseases, was investigated in 40 human myeloma cell lines (HMCLs). The aim of the experiments was to study the genes involved in arachidonic acid metabolism in order to detect gene patterns that possibly could be connected to the different gene expression profiles of multiple myeloma. NMF results have been verified by western blotting analysis in six HMCLs of proteins expressed by some of the most abundantly expressed genes. The experiments showed the effectiveness of NMF in intelligently analyzing microarray data.

Keywords: nonnegative matrix factorization; intelligent data analysis; feature extraction; dimensionality reduction; unsupervised learning; human multiple myeloma cell lines; gene expression profile; arachidonic acid metabolism

1. Introduction

The massive use of technologies, in any domain, is leading to an exponential growth of available data. This abundant data and their complexity very often make them unusable. For this reason, automatic tools are more and more frequently used to facilitate their management and analysis. However, when completely automatic tools are used without proper technical knowledge, they provide results that are not completely understandable.

Intelligent data analysis (IDA) is a methodology to extract useful hidden knowledge from data by involving human expertise in the analysis process [1,2]. It includes different methodologies to analyze real-world problems, such as statistics, artificial intelligence, machine learning, data mining, and data visualization. Domain experts and computer scientists are both necessary during the data analysis

process. The former pertains to the domain expertise that is used to design the research questions that drive the analysis. The latter use automatic techniques to extract unknown patterns from data.

However, there is a language gap between these two figures. Data visualization, dimensionality reduction, and clustering techniques allow the simplification of original data, making them more understandable and easier to analyze by non-technicians [3].

In this scenario, the non-negative matrix factorization (NMF) was shown to be suitable for IDA [4]. NMF is a low-rank approximation technique that, due to its non-negativity constraint, is able to describe original data as a linear additive combination of hidden factors: i.e., pieces of images in image recognition, topics in text mining, and metagenes (MGs) in the micro-array analysis [5–8]. NMF allows dimensionality reduction and clustering of original data, both useful for data understanding. On one hand clustering methods are used to group data according to similarity criteria [9]. On the other hand, dimensionality reduction methods represent data in a low dimensional space by discarding noise and worthless information and highlighting noteworthy elements. Both methods allow simplifying the complexity of original data, to bring out hidden structures among them. Differently from other similar methods, such as singular value decomposition (SVD), principal component analysis (PCA), and k-means, NMF gives easily interpretable results due to its non-negativity constraint [3].

In the last decade, NMF was widely used to analyze biological data [9,10]. Gene expression profiles, coming from microarray chips, are represented as numerical non-negative high-dimensional data. Thus, dimensionality reduction and clustering techniques are commonly used to discard unimportant information and highlight the relevant data.

In this work, 40 human myeloma cell lines (HMCLs) were investigated. Particularly a selection of gene expression profiles (GEPs) that are representative of the human multiple myeloma diseases were analyzed through NMF. The NMF part-based representation simplifies the data and the relationships in them. Moreover, result visualizations are used with the aim to facilitate the data understanding from the domain experts.

Multiple myelomas (MM) is still an unmet clinical need, being an incurable malignant disease of plasma cells affecting approximately 25,000 new patients per year in Europe and characterized by a marked genetic heterogeneity [11,12]. The major limitation to identify a proper MM therapy is mainly caused by a wide inter individual variation in response to the clinically available drugs [11,13–15]. This heterogeneous response to the treatment is mostly due to the molecular characteristics of the tumor, i.e., the differences in the involved GEP could determine the drug-resistance [16–18]. Deciphering key changes in gene expression levels underlying personalized sensitivity to chemotherapy is, therefore, essential to predict the efficacy of anticancer drugs and to prevent delay in the selection of more effective treatment strategies [19].

Cell lines are currently preferred in vitro pre-clinical tools to screen the effects of new drugs or combinations of drugs. Often, the efficacy of a drug candidate may vary depending on the cell line based model used for testing, probably also due to the different GEP of the used cells.

The possibility to associate a specific cell line to a particular pathology condition would help to better design pharmaceutical studies, in terms of targets such as receptors, channel proteins or enzymes, and biological pathways. It is well-known that under some circumstances expressed genes do not produce the corresponding proteins necessary to exert a specific function. Thus, in continuation with our scientific interests, gene expression related to arachidonic acid (AA) metabolism was looked for in selected HMCLs and the presence of their coding proteins was verified by the western blotting technique. AA metabolism was chosen because, in its very complex network, several eicosanoids, as crucial mediators of several physio-pathological pathways, are produced.

Prostaglandin E2 (PGE2), one of the known eicosanoids, is one of the pro-inflammatory and angiogenesis network mediators and supports the growth of several solid tumors [20–23]. In addition, PGE2 stimulates gene transcription, influences the mitogenesis of normal human bone cells, and promotes tumor metastases formation. Cyclooxygenase (COX) is responsible for PGE2 bio-synthesis from AA, in turn, released from the cell membrane upon mechanical or mitogen

stimuli. Both the two known COX isoforms (COX-1 and COX-2) contribute to PGE2 production, even though COX-1 seems to be its major source in normal tissue. COX-1 is also the target of a low dose of acetylsalicylic acid (ASA), the well-known aspirin active principle ingredient that irreversibly inactivates the platelet cyclooxygenase-1 by acetylating its S530, preventing thromboxane formation. In fact, thrombotic complications in patients with newly diagnosed MM treated with lenalidomide, dexamethasone, and thalidomide chemotherapy benefit from aspirin prophylaxis [24–26]. Tens of millions of worldwide adults take aspirin to reduce their risk of heart attack or stroke. Studies over the last two decades have suggested that regular use of aspirin may have another important benefit: decreasing the risk of developing or dying from some types of cancer [27,28]. These aspects potentiate the idea that the genes and their encoded proteins involved in the COX-mediated AA cascade would of relevance and noteworthy investigations also in multiple myeloma as in other types of cancer. For this purpose, a microarray data matrix containing the GEP of 40 HMCLs was used. The experiments, here below described, were designed to study a subset of 64 genes that were selected as involved in arachidonic acid metabolism. Particularly, for this purpose, the gene taxonomy proposed by [29] was used, 64 genes were grouped according to their role in arachidonic acid metabolism. Indeed the aim of the analysis was to detect expression patterns possibly connected to different multiple myeloma histotypes, stages, and grades that could emerge from the interaction of the selected genes. The NMF part-based representation was used to reduce the problem dimensionality and to describe the HMCLs as a composition of metagenes (i.e., groups of genes that emerged automatically from data). Finally, the western blotting technique was used to verify if the relationships between genes and HMCLs that have been suggested by NMF were actually coded.

2. Materials and Methods

2.1. Materials

Cell culture reagents were purchased from EuroClone (Milan, Italy). Radioimmunoprecipitation assay buffer (RIPA) and protease inhibitor cocktail were obtained from Sigma–Aldrich (Milan, Italy). Anti- μ -ACTIN, anti-COX-1 (COX-111), and anti-rabbit secondary peroxidase antibodies were purchased from Thermo Fisher Scientific Italia (Monza, Italy). Anti-17 μ -HSD4, anti-GGT1, anti-LTA4H, anti-ALOX-15B, anti-ACAA1, anti-CBR1, anti-CYP1A2, anti-PGEs2, and anti-VINCULIN were obtained from Santa Cruz Italia (Milan, Italy). Anti-mouse secondary peroxidase antibody and all reagents for western blotting were purchased from Bio-Rad Laboratories Srl (Milan, Italy). The other chemicals were prepared in AS and MGP laboratory [30–36].

2.2. Cell Culture

The human NCI-H929, KMS-12-BM, and SK-MM-2 multiple myeloma (MM) cells were obtained from The Leibniz Institute DSMZ (Braunschweig, Germany). Human RPMI-8226, U266 B1, MM1S, HEK-293, and HepG2 cells were purchased from ATCC (Manassas, VA, USA). huCOX-1-IRES-mPGES-1 cells were a gift from William Smith [37]. NCI-H929 cells were grown in RPMI-1640 supplemented with 20% fetal bovine serum, 50 μ M μ -mercaptoethanol, 1 mM sodium pyruvate, 2 mM glutamine, 100 U/mL penicillin, and 100 μ g/mL streptomycin in a humidified incubator at 37 °C with a 5% CO₂ atmosphere. KMS-12-BM and SK-MM-2 cells were grown in RPMI-1640 supplemented with 20% fetal bovine serum, 2 mM glutamine, 100 U/mL penicillin, and 100 μ g/mL streptomycin in a humidified incubator at 37 °C with a 5% CO₂ atmosphere. MM1S cells were grown in RPMI-1640 supplemented with 10% fetal bovine serum, 1 mM sodium pyruvate, HEPES 10 mM, 1.4 M glucose, 2 mM glutamine, 100 U/mL penicillin, and 100 μ g/mL streptomycin in a humidified incubator at 37 °C with a 5% CO₂ atmosphere.

RPMI-8226 and U266 B1 cells were grown in RPMI-1640 supplemented with 10% fetal bovine serum, 2 mM glutamine, 100 U/mL penicillin, and 100 μ g/mL streptomycin in a humidified incubator at 37 °C with a 5% CO₂ atmosphere. Human HEK-293 embryonal kidney cells were grown in DMEM

(Dulbecco's Modified Eagle Medium) supplemented with 10% fetal bovine serum, 2 mM glutamine, 100 U/mL penicillin, and 100 µg/mL streptomycin in a humidified incubator at 37 °C with a 5% CO₂ atmosphere. The human HepG2 hepatocellular carcinoma cells were grown in MEM supplemented with 10% fetal bovine serum, 2 mM glutamine, 100 U/mL penicillin, and 100 µg/mL streptomycin in a humidified incubator at 37 °C with a 5% CO₂ atmosphere. HEK-293-COX-1 were grown in DMEM supplemented with 10% fetal bovine serum, 2 mM glutamine, 100 U/mL hygromycin B, and 100 µg/mL blasticidin in a humidified incubator at 37 °C with a 5% CO₂ atmosphere.

2.3. Western Blotting

The western blotting analysis was conducted by using the methods reported in Perrone et al. with minor modifications [38]. Briefly, all cells were washed twice with 10 mL phosphate-buffered saline (PBS), scraped in 1 mL PBS and centrifuged for 10 min at 1500 rpm, 4 °C. Proteins were extracted from cells by homogenization in cold RIPA buffer (Sigma–Aldrich) containing 1X protease inhibitor cocktail and centrifuged at 14,000 rpm for 10 min at 4 °C. The supernatant was recovered, and the protein concentration was measured using a DC Protein Assay Reagent Kit (BIO-RAD, Hercules, CA, USA). Protein extract (25 µg) was separated on 10% polyacrylamide gel (BIO-RAD) and then transferred onto a polyvinylidene difluoride membrane (PVDF) by a Trans-Blot Turbo Transfer System (BIO-RAD). The membrane was blocked for 1 h at room temperature with blocking buffer (1% blotting grade blocker, 0.1% Tween20 in Tris-buffered saline, TBS). The membrane was then incubated with either anti-COX-1 (1:500 mouse monoclonal, overnight at 4 °C), anti-LTA4H, anti-PGEs2, anti-CYP1A2, anti-CBR1 (1:200 mouse monoclonal, overnight at 4 °C), anti-ALOX-15-B (1:1000 rabbit monoclonal, 1 h at room temperature), or anti-17µ-HSD4, anti-GGT1, anti-ACAA1, anti-vinculin, anti-µ-actin (1:1000 mouse monoclonal, 1 h at room temperature) antibodies, diluted in blocking buffer. After the incubation time, the membrane was washed with washing buffer (0.1% Tween-20 in Tris-buffered saline, TBS) three times and incubated with a secondary peroxidase antibody (1:3000 anti-mouse or anti-rabbit) for 1 h at room temperature. After washing, the membrane was treated with enhanced chemiluminescence (ECL, BIO-RAD) according to the manufacturer's instructions and the blot was visualized by UVITEC Cambridge (LifeTechnologies, Carlsbad, CA, USA). The expression level was evaluated by densitometric analysis using UVITEC Cambridge software (LifeTechnologies) and the µ-actin or vinculin expression level was used as a loading control to normalize the sample values.

2.4. Non-Negative Matrix Factorization

NMF is a low-rank approximation technique that decomposes the original data matrix $X \in R_+^{n \times m}$ in the product of two non-negative matrices W and H ($X \approx WH$, with $W \in R_+^{n \times k}$ and $H \in R_+^{k \times m}$) called basis and encoding matrix, respectively, where k is the factorization rank that indicates the number of hidden genes (metagenes) to extract.

An explicative example is reported in Figure 1. A microarray data matrix with 30 genes (rows) and 10 HMCLs (columns) was decomposed in the matrices $W \in R_+^{30 \times 2}$ and $H \in R_+^{2 \times 10}$.

The basis matrix W , contains the latent factors that are automatically extracted by the algorithm. In the context of the microarray data analysis, these factors are called metagenes: groups of genes that emerged from the computation and used to describe the original data in a sub-dimensional space. Each column of W represents a metagene that is described by the same genes used in the original data. Each entry w_{ij} is the membership values of the i -th gene to the j -th metagene. The metagenes represent the relationships among the genes. From the mathematical point of view, these are linear relationships that emerge from data. One gene can be involved in the definition of different bases (metagenes); indeed, these are not exclusive relationships. This well represents the biological nature of the metagenes that could actually involve different genes depending on their role.

The encoding matrix H represents the original data (HMCLs) in terms of the metagene expression profile. Each entry h_{ij} denotes the importance of the i -th metagene in reconstructing the j -th HMCL.

The non-negativity constraint of NMF allows describing the HMCLs as a linear combination of these metagenes, weighted by the coefficients in H (part-based decomposition).

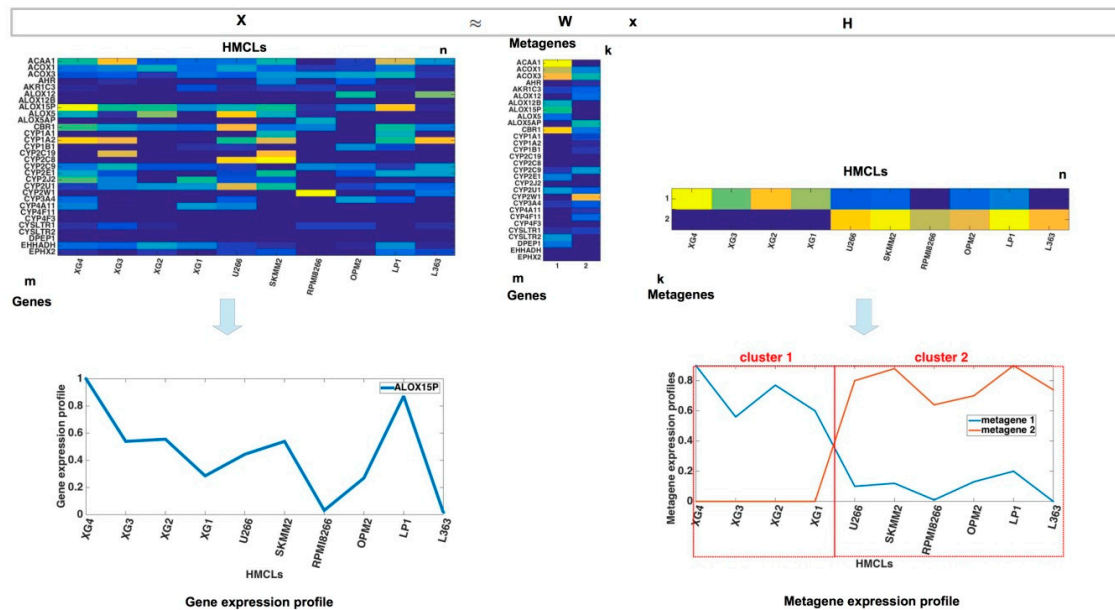


Figure 1. Non-negative matrix factorization decomposition schema. Each column of X contains a different human myeloma cell lines (HMCL) described by the expression profiles of the selected genes; each row represents the expression profile of the corresponding gene, through the HMCLs.

Graphical representations of the metagene expression profiles could reveal different metagene behaviors for the considered HMCLs, as in Figure 1. Therefore these patterns are used to automatically group the HMCLs based on their similarities.

For this study, the standard multiplicative update algorithm [39] based on the Euclidean distance was used. Moreover, since the NMF algorithms are strictly dependent on the initial starting point, the non-negative double singular value decomposition (NNDSVD) algorithm was applied since it was proven to improve the quality of clustering results [40].

The rank of the factorization is an input parameter of the problem. It is crucial for the results since it indicates the number of metagenes and the clusters that should be extracted. The correlation matrix gives a qualitative measurement of the stability of the algorithm. Quantitative measurement of this stability is given by the cophenetic correlation coefficient that in gene expression analysis it is used to quantify the stability of the clustering results of microarray gene expression [10].

3. Results and Discussion

In this section, we report and discuss the results obtained by NMF on the selected dataset (Section 3.1). Clustering results and the HMCLs part-based representation, in terms of the extracted metagenes, are firstly described and analyzed in Section 3.2. Section 3.3 describes the hidden metagenes and their role in arachidonic acid metabolism; finally, Section 3.4 reports western blotting analysis to validate NMF results.

3.1. Microarray Data

A microarray data matrix containing the gene expression profile of 40 HMCLs was used to study the genes involved in arachidonic acid metabolism, to detect genes expression profile patterns likely connected to the different multiple myeloma types. All the numerical results were obtained by implementing the algorithms in Matlab 8.4 codes and running them on a machine equipped with an Intel® core 2 Duo CPU with RAM 8.00 GB.

Microarray data of multiple myeloma cell lines available on the public database ArrayExpress [41] were analyzed.

The microarray data were collected using the Affymetrix GeneChip Human Genome U133 Plus 2.0 [HG-U133-Plus-2]. The analysis was conducted on a subset of the original data starting from 40 human myeloma cell lines (HMCLs) that provided a signature for stratification of patient risk [41] and 64 genes involved in arachidonic acid metabolism as depicted in Figure 2.

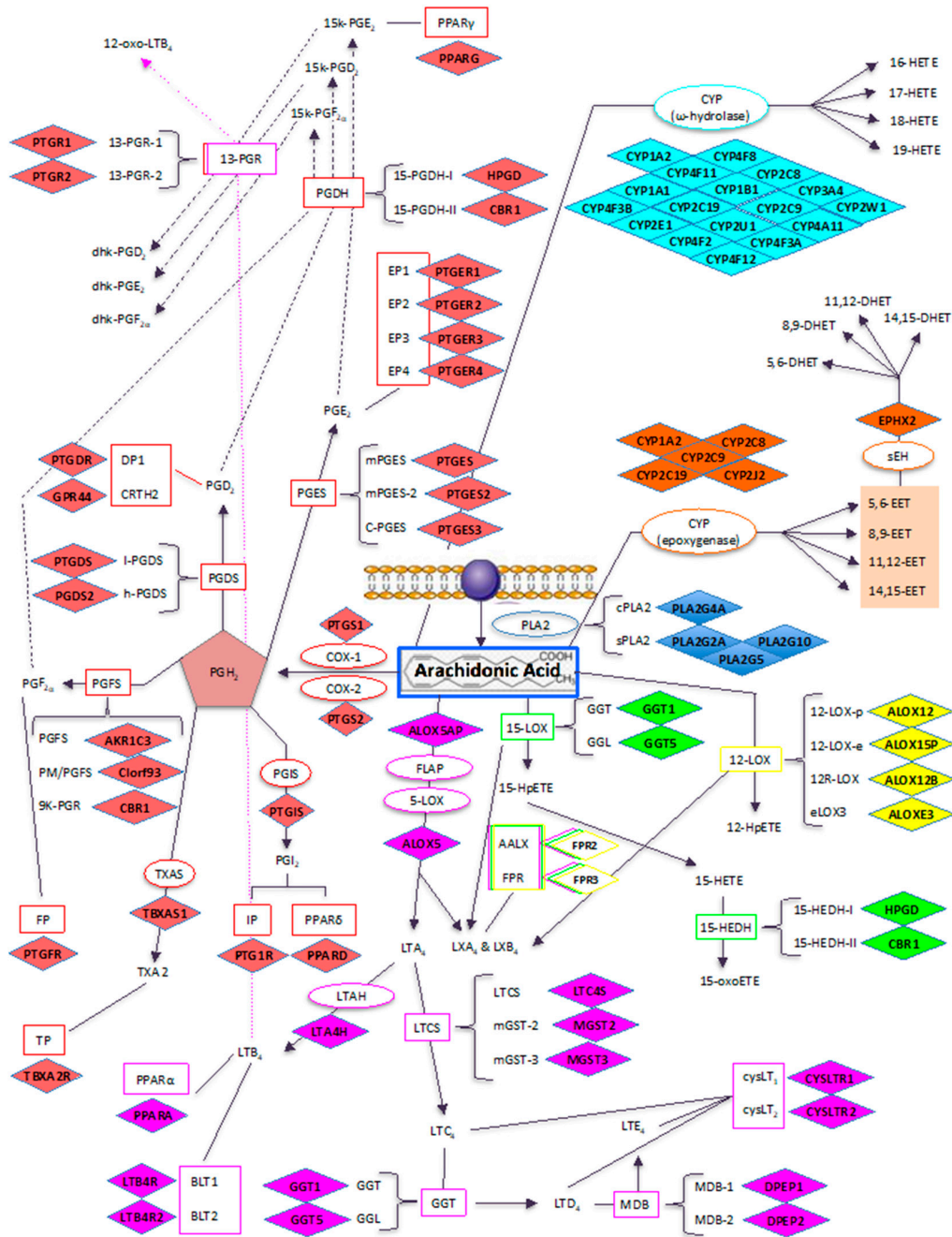


Figure 2. Arachidonic acid (AA) metabolism, released from the membrane upon mechanical mitogen stimuli. Eicosanoids are the major products of AA metabolism. In this figure, the biosynthetic pathways of the main eicosanoids together with genes coding for proteins are differently colored: cyclooxygenase (COX) (red), 5- lipoxygenase (5-LOX) (purple), 15-LOX (green), 12-LOX (yellow), cytochrome P (CYP) epoxygenase (orange), CYP w-hydroxylase (cyan). Modified image from [35].

A data matrix was built by using 64 rows (genes) and 40 columns (HMCLs) (Figure 3). It was evident that the dimensionality of the data was too big to directly identify any GEP pattern. Hence, a dimensionality reduction was needed to bring out useful information that could be easily visualized through the part-based representation.

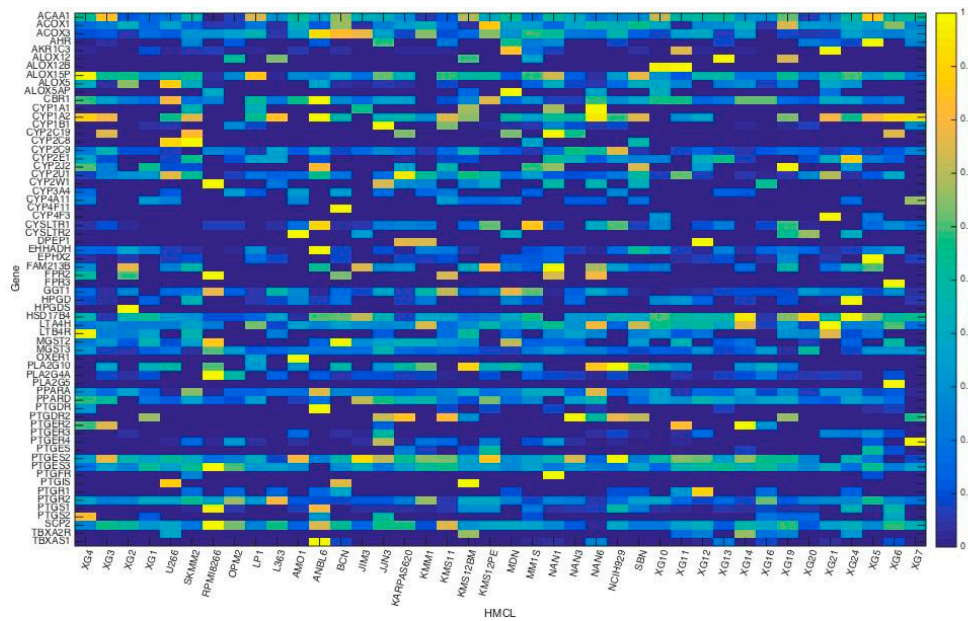


Figure 3. Visual representation of the data matrix. Different colors are used to represent the gene expression level: yellow grades for high expression, green grades for medium expression, blue grades for low or no expression.

3.2. HMCL Clustering Results

Fifty runs of the NMF algorithm varying the rank value in the interval 2–10 were executed [42,43], and both the consensus matrices and the cophenetic correlation coefficient were collected from the analysis of the encoding matrix *H*. Whilst the consensus matrix gives a qualitative measurement of the cluster quality, the cophenetic correlation coefficient is a quantitative measure, and its trend is used to choose the optimal rank (i.e., the number of clusters).

Figure 4 shows the cophenetic correlation coefficient trend in the interval 2–10 (Each point represents the mean value of the cophenetic correlation coefficient on the fifty runs). From value five we observed a significant reduction of this trend, so it has been chosen as factorization rank.

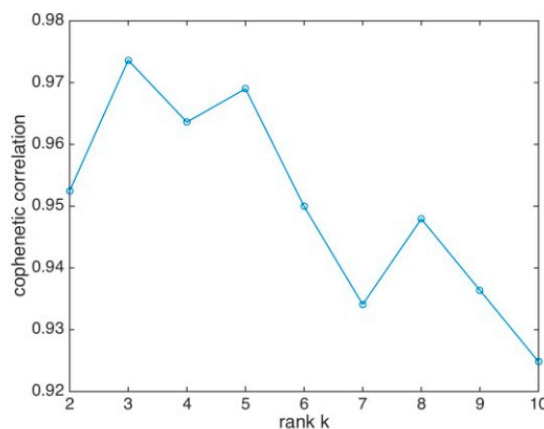


Figure 4. Cophenetic index values in 2–10, as a function of the factorization rank.

The consensus matrix in Figure 5 represents the HMCLs (rows and columns) grouped according to the metagene expression profile similarities. Five clusters (yellow squares) are depicted showing some regularities within the groups. The algorithm proved to be stable and it converged to the same solution in all 50 runs. This result is shown in Figure 5 where the squares are clearly separated and only yellow and blue values are present.

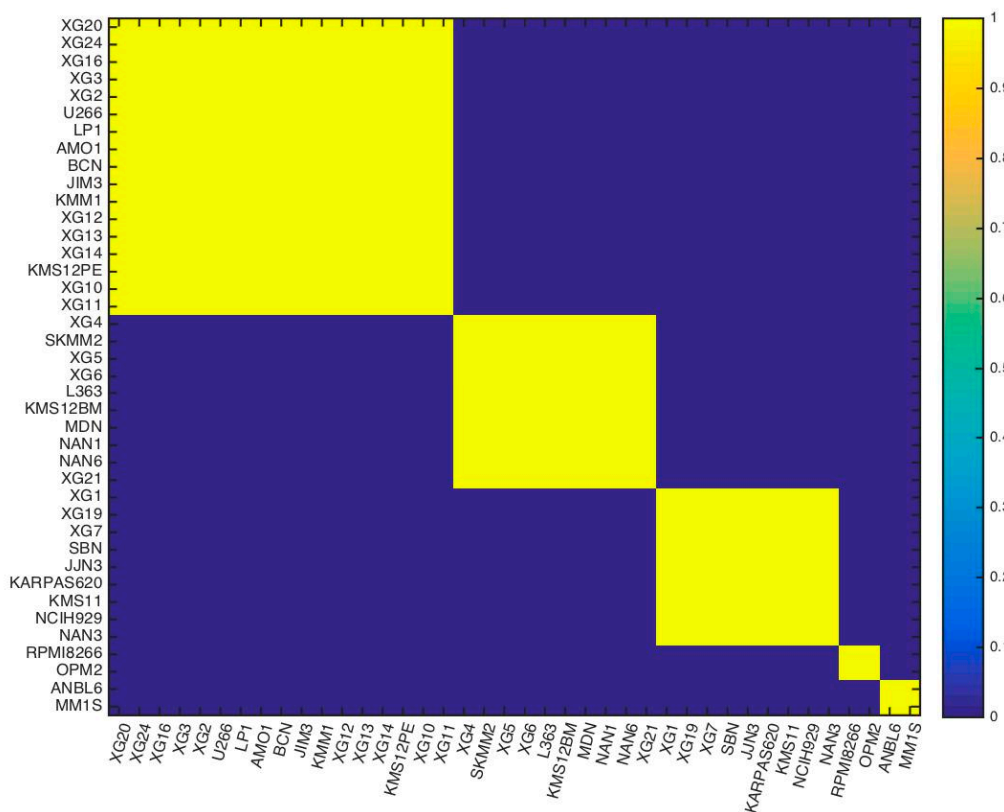


Figure 5. Consensus matrix obtained with 50 runs of non-negative matrix factorization (NMF) and $k = 5$.

It is known that NMF allows a part-based representation of the original data by a linear additive composition of parts that can be interpreted as building blocks. In this scenario, the ‘blocks’ are the metagenes. Figure 6 shows the HMCLs composition in terms of metagenes in each of the five groups that have been returned from the consensus matrix. The whole bar is given by the sum of the single metagene bars, expressed as a percentage.

The emerging patterns are now clearer. The dimensionality reduction and the clustering have made it possible. They allow the study of both the metagene expression profiles through the HMCLs and the different behaviors of the metagenes among the clusters. This simplified representation helps the domain experts in better understanding the relationships hidden in data. It is a starting point for conducting the biological analysis.

Figure 6 shows that the HMCLs in the first cluster are characterized by a high expression of the first metagene (dark blue bars), and the HMCLs in the second cluster from high expression of the third metagene (green). The HMCLs in the third cluster are characterized by a high expression of the fourth metagene (yellow) bars, and the bars in clusters four and five have very high expression of the metagenes two (light blue) and five (dark yellow), respectively.

Furthermore, it is noteworthy that, even if the metagenes two and four are highly expressed in clusters four and three, they are always expressed in all the HMCLs. Note that even if this representation has allowed the identification of the most expressed metagenes per HMCLs, the analysis has to take into account the less expressed metagenes. Indeed, all the extracted metagenes contribute to defining each HMCL, thus they cannot be removed as noise. In fact, the dimensionality reduction has already

removed the noise from data, whilst it has highlighted the relationships among genes and HMCLs. The emerged patterns are then used by the domain experts, involved in the data analysis process (biologists and physicians in this case), to better understand their role in arachidonic acid metabolism, as it is described in Section 3.4.

The expression profiles of the single metagenes are depicted (same colors code used before) in Appendix A, reporting the same behavior summarized in Figure 6.

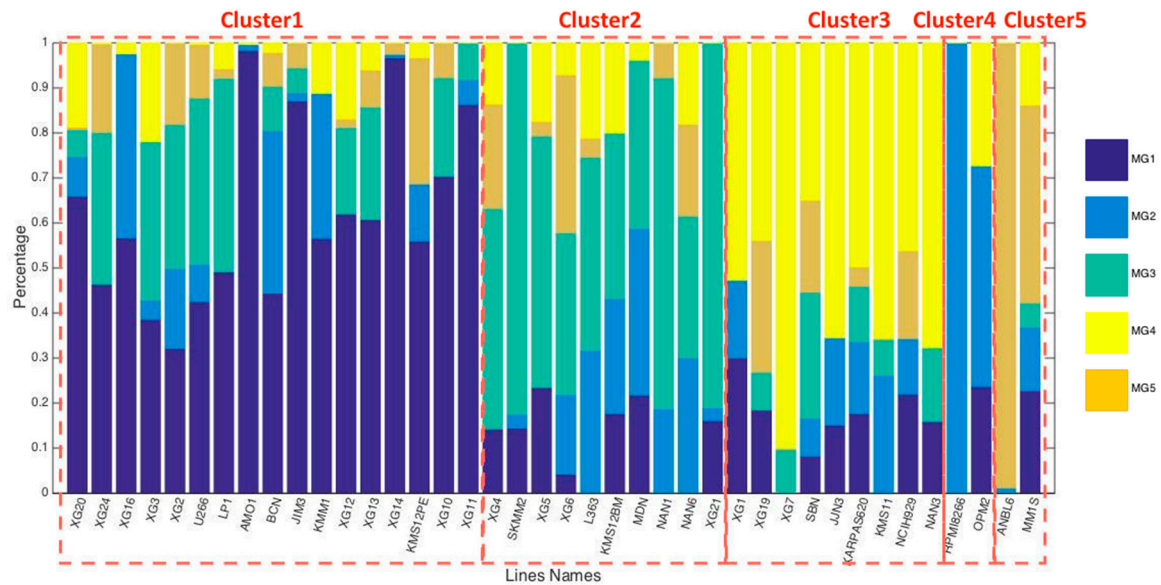


Figure 6. Metagene expression profile through the HMCLs.

Moreover, visual analysis allows us to detect some subgroups within the groups. It is the case that some HMCLs such as XG24, XG3, XG2, U266, LP1, XG12, XG13, and XG10 of the first cluster have a medium expression of the metagene 3 (the green one), different than the other HMCLs of the same groups; the HMCLs XG4, XG6, and NAN6 of the second cluster, that use metagene 4 (the yellow bars); in the same clusters we recognize another subgroup composed of the HMCLs XG6, L363, KMS12BM, MDN, NAN1, and NAN6 that have a medium expression of metagene 2 (light blue). Differences among HMCLs in the same group are the object of study for the domain experts.

At this point, it appeared interesting to analyze the genes forming each metagene. Since our analysis started using Moreaux’s classification of HMCLs, [41] it seemed important to compare his results to our obtained clusters. According to Moreaux’s study, the 40 HMCLs could be clustered into 6 groups using unsupervised hierarchical clustering, whereas our clustering led to five clusters.

This difference was predictable. Moreaux’s analysis was conducted on a large set of genes. Our analysis is focused on a smaller subset of genes, the arachidonic acid metabolism involved genes.

3.3. Metagene Analysis

The encoding matrix H was used to cluster the HMCLs and to extract the metagene expression profiles, whereas the basis matrix W contains the metagenes, represented as sets of genes with different weights. We extracted five different metagenes. Figure 7 depicts the top most 20 genes involved in each metagene. The genes have been ordered and visualized, according to their influence (expressed in percentage) on each metagene. For a broader analysis, twenty genes are reported even though we observed the highest weights for the first five/six genes. These graphs give more insights to the domain experts that are helpful to drive the biological analysis of the HMCLs.

As an example, analyzing Figure 5 we know that the HMCL KMM1 is mostly composed of the expression of metagene 1, partly by metagene 2, and minimally by metagene 4. Figure 7 specifies the genes that contribute to each metagene definition. The genes PTGES2, HSD17B4, LTA4H, ACAA, CBRI,

and PPARD are the heaviest in the definition of metagene 1, so the expression of these genes highly affects the HMCL KMM1; the genes MGST2, GGT1, SCP2, FPR2, PTGES3, PLAG10, and CYP2W1 from metagene 2 only partly influence KMM1, and finally, the genes CYSLTR1, CBR1, CYP2J2, EHHADH, CYP1A2, and ACOX3 from metagene 4 slightly condition KMM1. It is possible to conduct a similar analysis on the other cell lines, looking for a relationship among the genes and the cellular functions they are involved in.

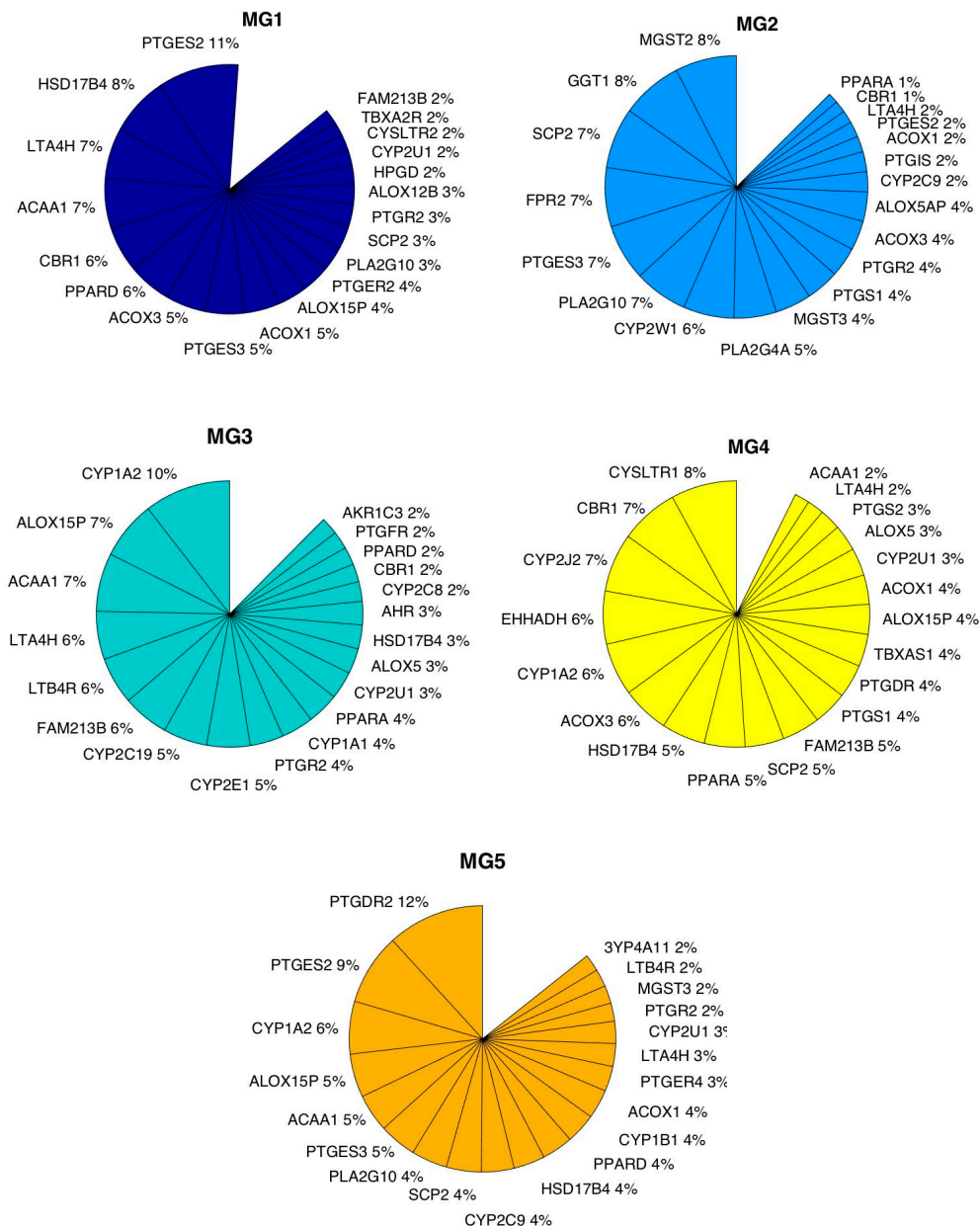


Figure 7. Composition of the metagenes MG1, MG2, MG3, MG4, and MG5 in terms of the original genes.

The taxonomy in [24] can give more insight into the role of the genes in each metagene. For example, the genes in the first metagene mostly belong to the cyclooxygenases and related enzymes involved in eicosanoid biosynthesis in humans, enzymes involved in eicosanoid catabolism in human, eicosanoid receptors, and transcription factors in humans. Similar analyses could be conducted on the other metagenes.

3.4. Proofs of Concepts: Western Blotting Analysis

With the aim to verify the correctness of the NMF results, the expression of some proteins codified by the topmost twenty genes that are part of the identified metagene were found in six representative HMCLs (U266, SKMM2, KMS12BM, NCIH929, RPMI8226, MM1S) by using the western blotting technique [44–47]. One or two HMCLs were selected from each cluster, based on their different responses to the clinically used drugs. In Figure 5, the expression of some specific proteins, analyzed by western blotting analysis and involved in the eicosanoid biosynthesis related to the arachidonic acid pathway of various human myeloma cell lines, is depicted.

The expression extent values of each protein were assigned as listed in Figure 8.

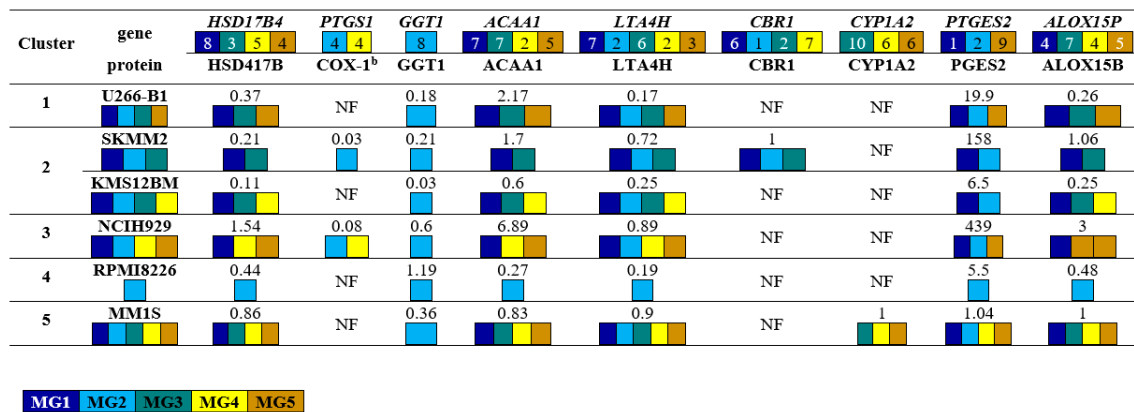


Figure 8. Protein expression in all studied cell lines (Expression values are reported with respect to HEK293 wild type or HepG2 cell line were used as a positive control, as suggested by the antibody data sheet, to which the value 1 has been arbitrarily assigned. Color codes are the same used for metagenes 1–5 depicted in Appendix A. All data are related to HEK293-COX-1, the value = 1 was assigned to the expression of COX-1. NF = not found).

U266-B1 is a B lymphocyte plasmacytoma cell line belonging to cluster 1 characterized by the expression of MG1-2-3-5 (depicted from the colored boxes under cell line name, Figure 8). *HSD17B4*, the gene coding for the protein HSD417B, was expressed at 8% in MG1, at 3% in MG3, at 5% in MG4, and 4% in MG5 (gene percentages are indicated in the colored boxes, Figures 7 and 8) but its expression (0.37) in U266-B1 was due only to MG1, 3, and 5. In fact, U266-B1 (cluster 1, Figure 5) was not represented by MG4 and HSD17B4 was not present in the top most 20 genes of MG2.

In U266-B1, COX-1 was not detected by western blot analysis (Figure 6) thus it was supposed that PTGS1 is downregulated in this cell line.

GGT1 expressed at 8% in MG2 codifies the corresponding protein which was found with a value of 0.18 in U266-B1. Comparable expression values 0.17 and 0.26, respectively, were found for LTA4H and ALOX15B; LTA4H was represented by all 5 MG with a preponderant contribution of MG1 in which LTA4H was expressed at 7% of all genes. ALOX15P was found in MG1, 3, 4, and 5 but its expression was higher in MG3 (7%).

Among the analyzed proteins, PGES2 and ACAA1 were highly expressed in U266-B1 with a value of 19.9 and 2.17, respectively. PGES2 was found in MG1, 2, and 5 with no involvement of MG3 and 4 which do not compose the U266-B1 cell line. ACAA1 derive from MG1, 3, and 5, with no contribution of MG 2 and 4. CBR1 and CYP1A2 were not detected in U266-B1, even though their corresponding genes were represented in MGs 1–5. Their absence could be justified by the downregulation of CBR1 and CYP1A2.

SK-MM-2 and KMS-12-BM, both belonging to cluster 2 (Figure 5), had a different metagene composition as depicted by the colored boxes under cell line name (Figure 8).

SK-MM-2, a B lymphocyte plasmacytoma cell line, was mostly composed of metagene 3 and 1, with a little contribute given by metagene 2. From western blotting analysis (Figure 9) of this cell line, it was evident that all the proteins codified by the genes of metagenes 2 and 3 were expressed, except for CYP1A2.



Figure 9. Expression of some representative proteins involved in the eicosanoid biosynthetic pathways of six human myeloma cell lines analyzed by western blotting. HEK-293 and HepG2 were used as positive controls for 17- β -HSD4, LTA4H, PGES2, ALOX-15B, GGT1 and ACAA1 expressions, respectively. HEK293-COX-1 cells treated with 10 μ g/mL tetracycline for 24 h were used as a positive control for COX-1 expression. β -actin and vinculin were used as loading controls for 17- β -HSD4, COX-1, GGT1, LTA4H, ALOX-15B, and ACAA1, CBR1, CYP1A2, PGES2 respectively, except for the MM1S cell line where vinculin was used as the loading control for all the examined proteins. The image is the result of different western blotting experiments that were lined up in one single image.

Concerning KMS-12-BM, a multiple myeloma cell line, it had an almost similar contribution from all metagenes with the exception of metagene 5. In this case, PGES2 was the major expressed protein while for COX-1, CBR1, and CYP1A2, no expression was detected (Figure 9).

NCI-H929, another multiple myeloma cell line, belongs to cluster 3 with a marked expression of genes of metagene 4, and in part of metagenes 1, 3, and 5. Also, in this cell line, the expression of PGES2 was high (439) with respect to the positive control. CBR1 and CYP1A2 were not found in U266-B1, while a good expression of ALOX15B, HSD417B, ACCA1, and COX-1 was observed. In this cell line,

a good COX-1 expression was observed if considering that the reference cell line was stimulated to express COX-1.

RPMI-8266 belonging to the cluster 4 was composed of the expression of genes belonging to metagene 2. Concerning HSD17B4, ACAA1, and ALOX15B, since their encoding genes were not present in the top most 20 representing MG2, we expected not to find them. Surprisingly, western blotting experiments (Figure 9) revealed the expression of HSD17B4, ACAA1, and ALOX15B with values of 0.44, 0.27, and 0.48, respectively.

Their presence is justified by the fact that their codified genes are actually present but with a percentage of expression less than 1% and therefore not represented in Figure 8.

The last but not least important cluster was represented from the cell line MM1S with a marked component of metagene 5 but a low contribution of metagenes 1, 2, and 5. This is the only HMCL used in which CYP1A2 was expressed and with a value comparable to the positive control. Similar expression values were found for the other proteins with the exception of GGT1 and ACAA1.

4. Conclusions

Herein, a case study showing the use of NMF as a tool to intelligently analyze microarray data was proposed. Particularly the aim of the experiments was to study the genes involved in arachidonic acid metabolism, in order to detect gene patterns and relationships among the HMCLs that could be related to the different gene expression profiles of multiple myeloma. For this purpose, some numerical experiments on microarray gene expression real data belonging to 40 human multiple myeloma cell lines have been accomplished, selecting a subset of genes related to arachidonic acid metabolism, and applying the NMF algorithm to the so obtained genes–HMCLs matrix. The experiments showed the effectiveness of the NMF in intelligently analyzing microarray data to support the domain experts' activities. Through NMF we were able to select, using the genes involved in the metabolism of arachidonic acid, a certain number of multiple myeloma cell lines. This characterization has a high value from a biological point of view because it allows performing experiments using the most correct cell line. For example, the information obtained from this study led to the selection of NCIH929 and RPMI-8226 as the two most appropriate cell lines to investigate structure–inhibitory activity relationships of a set of novel compounds and the cyclooxygenase (COX) catalytic activity. Moreover, NMF results have been verified by western blotting analysis in six HMCLs of proteins expressed by some of the most abundant expressed genes. Our data confirm the correctness of NMF data outcomes for the six HMCLs. Further experiments are ongoing to increase the number of HMCLs and to be correlated to patients' disease profiles, in order to verify the presence of patterns that link different stages and grades of multiple myeloma, and the responses to the provided drug therapies.

Author Contributions: G.C. performed computational studies, M.L.P. performed biological studies, M.C. and A.P. designed and supervised G.C. work, A.V. provided MMS1 cell line and contributed to the writing of the article and A.S. and M.G.P. conceived the idea and supervised the entire work.

Funding: This work was supported by First AIRC Grant-MFAG2015 (Project Id. 17566).

Acknowledgments: Gabriella Casalino is a member of the INdAM research group GNCS (Gruppo Nazionale per il Calcolo Scientifico) of Istituto Nazionale di Alta Matematica Francesco Severi, P.le Aldo Moro, Roma, Italy.

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

Appendix A

Metagene expression profiles through the HMCLs are reported in Figures [A1–A5](#).

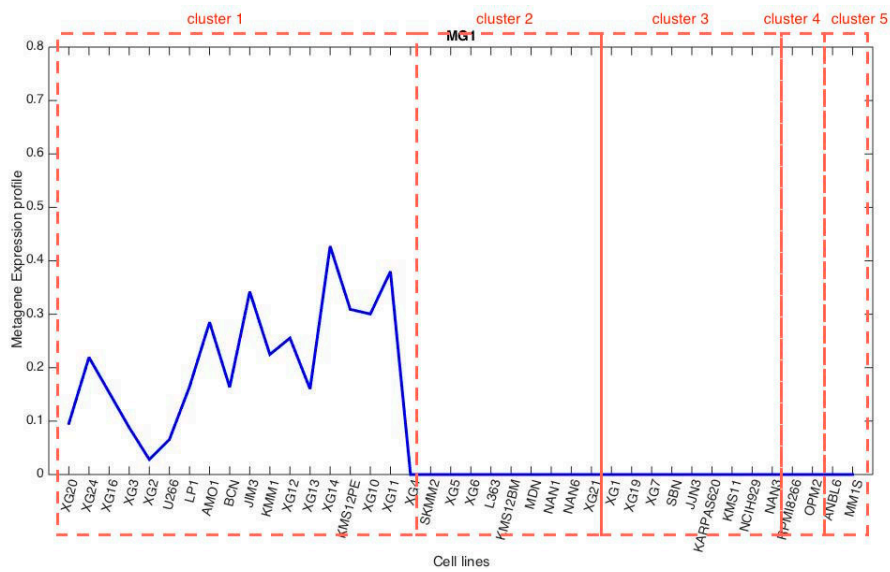


Figure A1. Metagene 1 expression profile through the HMCLs.

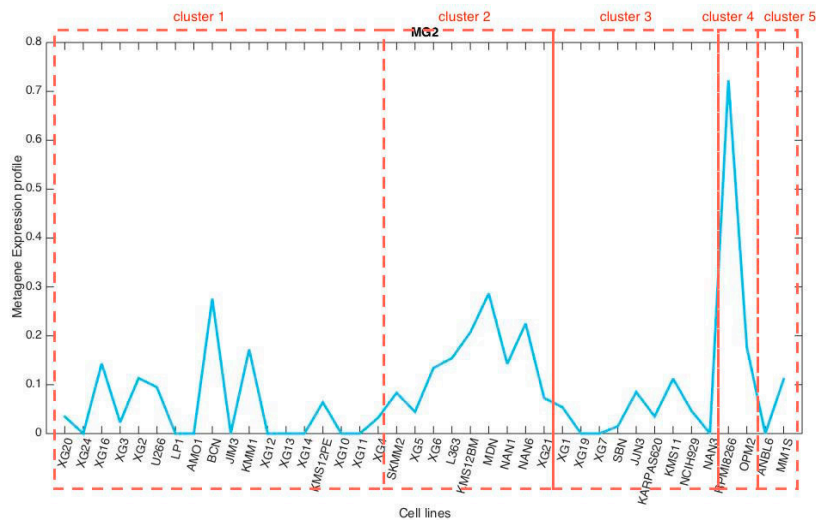


Figure A2. Metagene 2 expression profile through the HMCLs.

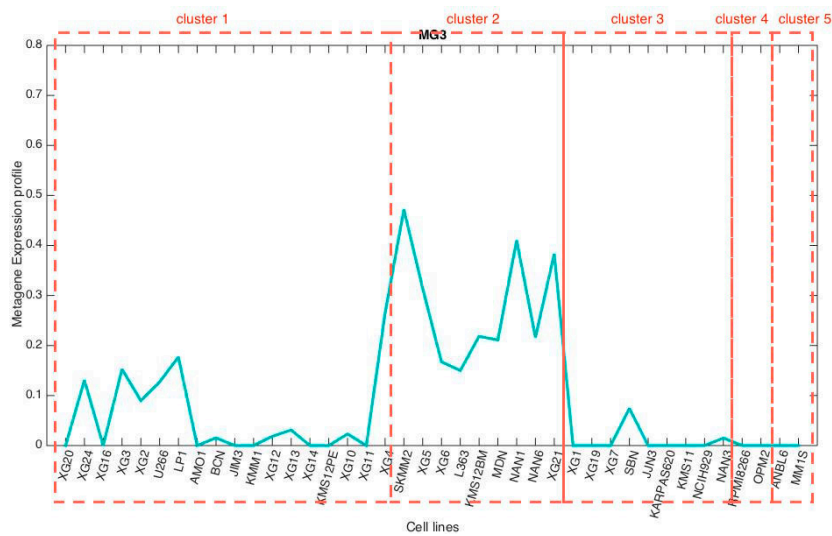


Figure A3. Metagene 3 expression profile through the HMCLs.

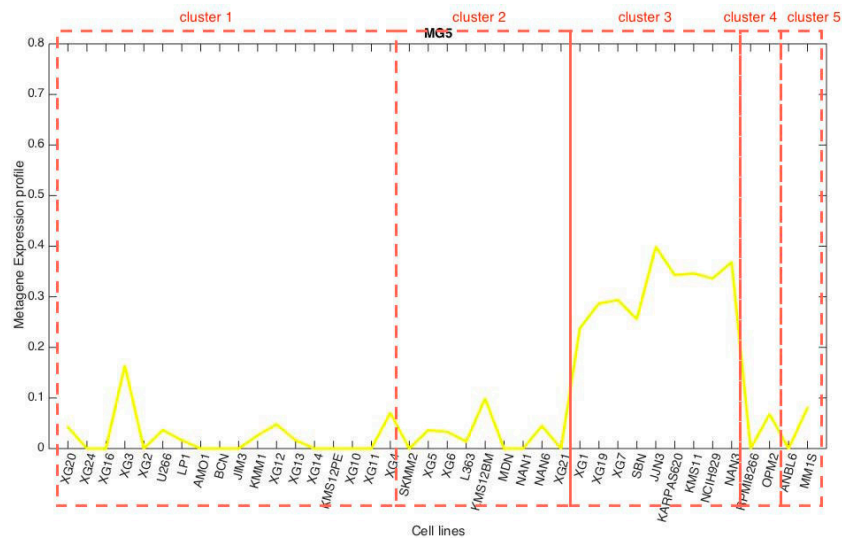


Figure A4. Metagene 4 expression profile through the HMCLs.

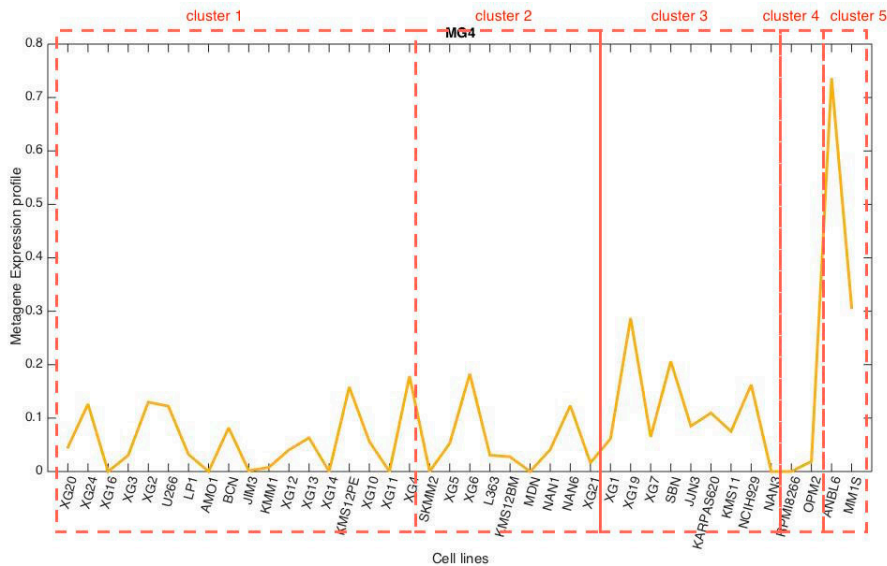


Figure A5. Metagene 5 expression profile through the HMCLs.

References

- Berthold, M.; Hand, D.J. *Intelligent Data Analysis: An Introduction*; Springer: New York, NY, USA; Secaucus, NJ, USA, 2007.
- Berthold, M.R.; Borgelt, C.; Hoppner, F.; Klawonn, F. *Guide to Intelligent Data Analysis: How to Intelligently Make Sense of Real Data*, 1st ed.; Springer: Berlin/Heidelberg, Germany, 2010.
- Nieto, J.J.; Torres, A.; Georgiou, D.N.; Karakasidis, T.E. Fuzzy polynucleotide spaces and metrics. *Bull. Math. Biol.* **2006**, *68*, 703–725. [[CrossRef](#)] [[PubMed](#)]
- Casalino, G.; Del Buono, N.; Mencar, C. Non-Negative Matrix Factorizations for Intelligent Data Analysis. In *Non-Negative Matrix Factorization Techniques, Signals and Communication Technology*; Naik, G.R., Ed.; Springer: Berlin/Heidelberg, Germany, 2016; ISBN 978-3-662-48330-5.
- Casalino, G.; Gillis, N. Sequential dimensionality reduction for extracting localized features. *Pattern Recognit.* **2017**, *63*, 15–29. [[CrossRef](#)]
- Boccarelli, A.; Esposito, F.; Coluccia, M.; Frassanito, M.A.; Vacca, A.; Del Buono, N. Improving knowledge on the activation of bone marrow fibroblasts in MGUS and MM disease through the automatic extraction of genes via a nonnegative matrix factorization approach on gene expression profiles. *J. Transl. Med.* **2018**, *16*, 217–233. [[CrossRef](#)] [[PubMed](#)]

7. Esposito, F.; Gillis, N.; Del Buono, N. Orthogonal joint sparse NMF for microarray data analysis. *J. Math. Biol.* **2019**, *79*, 223. [[CrossRef](#)] [[PubMed](#)]
8. Casalino, G.; Castiello, C.; Del Buono, N.; Mencar, C. A framework for intelligent Twitter data analysis with non-negative matrix factorization. *Int. J. Web Inf. Syst.* **2018**, *14*, 334–356. [[CrossRef](#)]
9. Filippone, M.; Camastra, F.; Masulli, F.; Rovetta, S. A survey of kernel and spectral methods for clustering. *Pattern Recognit.* **2008**, *41*, 176–190. [[CrossRef](#)]
10. Brunet, J.P.; Tamayo, P.; Golub, T.R.; Mesirov, J.P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 4164–4169. [[CrossRef](#)]
11. Marin, J.J.; Briz, O.; Monte, M.J.; Blazquez, A.G.; Macias, R.I. Genetic variants in genes involved in mechanisms of chemoresistance to anticancer drugs. *Curr. Cancer Drug Targets* **2012**, *12*, 402–438. [[CrossRef](#)]
12. Lombardi, L.; Poretti, G.; Mattioli, M.; Fabris, S.; Agnelli, L.; Bicciato, S.; Lambertenghi-Deliliers, G. Molecular characterization of human multiple myeloma cell lines by integrative genomics: Insights into the biology of the disease. *Genes Chromosomes Cancer* **2007**, *46*, 226–238. [[CrossRef](#)]
13. Richardson, P.G.; Sonneveld, P.; Schuster, M.W.; Irwin, D.; Stadtmauer, E.A.; Facon, T.; Harousseau, J.L.; Ben-Yehuda, D.; Lonial, S.; San Miguel, J.F.; et al. Safety and efficacy of bortezomib in high-risk and elderly patients with relapsed multiple myeloma. *Br. J. Haematol.* **2007**, *137*, 429–435. [[CrossRef](#)]
14. Fonseca, R.; Bergsagel, P.L.; Drach, J.; Shaughnessy, J.; Gutierrez, N.; Stewart, A.K.; Morgan, G.; Van Ness, B.; Chesi, M.; Minvielle, S.; et al. International Myeloma Working Group molecular classification of multiple myeloma: Spotlight review. *Leukemia* **2009**, *23*, 2210–2221. [[CrossRef](#)] [[PubMed](#)]
15. Vangsted, A.; Klausen, T.W.; Vogel, U. Genetic variations in multiple myeloma II: Association with effect of treatment. *Eur. J. Haematol.* **2012**, *88*, 93–117. [[CrossRef](#)] [[PubMed](#)]
16. Kumar, S.; Rajkumar, S.V. Many facets of bortezomib resistance/susceptibility. *Blood* **2008**, *112*, 2177–2178. [[CrossRef](#)] [[PubMed](#)]
17. Rajkumar, S.V.; Dimopoulos, M.A.; Palumbo, A.; Blade, J.; Merlini, G.; Mateos, M.V.; Kumar, S.; Hillengass, J.; Kastritis, E.; Richardson, P.; et al. International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma. *Lancet Oncol.* **2014**, *15*, 538–548. [[CrossRef](#)]
18. Rajkumar, S.V. Myeloma today: Disease definitions and treatment advances. *Am. J. Hematol.* **2016**, *91*, 90–100. [[CrossRef](#)]
19. Mitra, A.K.; Harding, T.; Mukherjee, U.K.; Jang, J.S.; Li, Y.; HongZheng, R.; Jen, J.; Sonneveld, P.; Kumar, S.; Kuehl, W.M.; et al. A gene expression signature distinguishes innate response and resistance to proteasome inhibitors in multiple myeloma. *Blood Cancer J.* **2017**, *7*, e581. [[CrossRef](#)]
20. Williams, T.J.; Peck, M.J. Role of prostaglandin-mediated vasodilatation in inflammation. *Nature* **1977**, *270*, 530–532. [[CrossRef](#)]
21. Pai, R.; Soreghan, B.; Szabo, I.L.; Pavelca, M.; Baatar, D.; Tarnawski, A.S. Prostaglandin E2 transactivates EGF receptor: A novel mechanism for promoting colon cancer growth and gastrointestinal hypertrophy. *Nat. Med.* **2002**, *8*, 289–293. [[CrossRef](#)]
22. Salcedo, R.; Zhang, X.; Young, H.A.; Michael, N.; Wasserman, K.; Ma, W.H. Angiogenic effects of prostaglandin E2 are mediated by up-regulation of CXCR4 on human microvascular endothelial cells. *Blood* **2003**, *102*, 1966–1977. [[CrossRef](#)]
23. Perrone, M.G.; Scilimati, A.; Simone, L.; Vitale, P. Selective COX-1 inhibition: A therapeutic target to be reconsidered. *Curr. Med. Chem.* **2010**, *17*, 3769–3805. [[CrossRef](#)]
24. Palumbo, A.; Cavo, M.; Bringhen, S.; Zamagni, E.; Romano, A.; Patriarca, F. Aspirin, Warfarin, or Enoxaparin thromboprophylaxis in patients with multiple myeloma treated with Thalidomide: A Phase III, Open-Label, Randomized Trial. *J. Clin. Oncol.* **2011**, *29*, 986–993. [[CrossRef](#)] [[PubMed](#)]
25. Baz, R.; Li, L.; Kottke-Marchant, K.; Srkalovic, G.; McGowan, B.; Yiannaki, E. The Role of Aspirin in the Prevention of Thrombotic Complications of Thalidomide and Anthracycline-Based Chemotherapy for Multiple Myeloma. *Mayo Clin. Proc.* **2005**, *80*, 1568–1574. [[CrossRef](#)] [[PubMed](#)]
26. Zonder, J.A.; Barlogie, B.; Durie, B.G.; McCoy, J.; Crowley, J.; Hussein, M.A. Thrombotic complications in patients with newly diagnosed multiple myeloma treated with lenalidomide and dexamethasone: Benefit of aspirin prophylaxis. *Blood* **2006**, *108*, 403–404. [[CrossRef](#)] [[PubMed](#)]
27. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2015. *CA Cancer J. Clin.* **2015**, *65*, 5–29. [[CrossRef](#)] [[PubMed](#)]

28. Cuzick, J.; Thorat, M.A.; Bosetti, C.; Brown, P.H.; Burn, J.; Cook, N.R.; Ford, L.G.; Jacobs, E.J.; Jankowski, J.A.; La Vecchia, C.; et al. Estimates of benefits and harms of prophylactic use of aspirin in the general population. *Ann. Oncol.* **2015**, *26*, 47–57. [[CrossRef](#)] [[PubMed](#)]
29. Perrone, M.G.; Malerba, P.; Uddin, M.J.; Vitale, P.; Panella, A.; Crews, B.C.; Daniel, C.K.; Ghebreselasie, K.; Nickels, M.; Tantawy, M.N.; et al. PET radiotracer [¹⁸F]-P6 selectively targeting COX-1 as a novel biomarker in ovarian cancer: Preliminary investigation. *Eur. J. Med. Chem.* **2014**, *80*, 562–568.
30. Vitale, P.; Perna, F.M.; Perrone, M.G.; Scilimati, A. Screening on the use of *Kluyveromyces marxianus* CBS 6556 growing cells as enantioselective biocatalyst for ketones reduction. *Tetrahedron Asymmetry* **2011**, *22*, 1985–1993. [[CrossRef](#)]
31. Catalano, A.; Carocci, A.; Corbo, F.; Franchini, C.; Muraglia, M.; Scilimati, A.; De Bellis, M.; De Luca, A.; Camerino Conte, D.; Sinicropi, M.S.; et al. Constrained analogues of tocainide as potent skeletal muscle sodium channel blockers towards the development of antitumor agents. *Eur. J. Med. Chem.* **2008**, *43*, 2535–2540. [[CrossRef](#)]
32. Di Nunno, L.; Vitale, P.; Scilimati, A.; Simone, L.; Capitelli, F. Stereoselective dimerization of 3-arylisoaxazoles to cage-shaped bis-beta-lactams syn 2,6-diaryl-3,7-diazatricyclo[4.2.0.0^{2,5}]-octan-4,8-diones induced by hindered lithium amides. *Tetrahedron* **2007**, *63*, 12388–12395. [[CrossRef](#)]
33. Vitale, P.; Scilimati, A. Functional 3-Arylisoaxazoles and 3-Aryl-2-isoxazolines from reaction of aryl nitrile oxides and enolates: Synthesis and reactivity. *Synthesis* **2013**, *45*, 2940–2948. [[CrossRef](#)]
34. Perrone, M.G.; Santandrea, E.; Bleve, L.; Vitale, P.; Colabufo, N.A.; Jockers, R.; Milazzo, F.M.; Sciarroni, A.F.; Scilimati, A. Stereospecific synthesis and bio-activity of novel beta3-adrenoceptor agonists and inverse agonists. *Bioorg. Med. Chem.* **2008**, *16*, 473–2488. [[CrossRef](#)] [[PubMed](#)]
35. Perrone, M.G.; Santandrea, E.; Scilimati, A.; Tortorella, V.; Capitelli, F.; Bertolasi, V. Baker's yeast-mediated reduction of ethyl 2-(4-chlorophenoxy)-3-oxoalkanoates suitable intermediates for potential PPARalpha ligands. *Tetrahedron Asymmetry* **2004**, *15*, 3501–3510. [[CrossRef](#)]
36. Perrone, M.G.; Santandrea, E.; Dell'Uomo, N.; Giannessi, F.; Milazzo, F.M.; Sciarroni, A.F.; Scilimati, A.; Tortorella, V. Synthesis and Biological Evaluation of New Clofibrate Analogues as Potential PPARalpha Agonists. *Eur. J. Med. Chem.* **2005**, *40*, 143–154. [[CrossRef](#)] [[PubMed](#)]
37. Yuan, C.; Smith, W.L. A Cyclooxygenase-2-dependent Prostaglandin E2 Biosynthetic System in the Golgi Apparatus. *J. Biol. Chem.* **2015**, *290*, 5606–5620. [[CrossRef](#)]
38. Pati, M.L.; Vitale, P.; Ferorelli, S.; Iaselli, M.; Miciaccia, M.; Boccarelli, A.; Di Mauro, G.D.; Fortuna, C.G.; Souza Domingos, T.F.; Rodrigues Pereira da Silva, L.C.; et al. Translational impact of novel widely pharmacological characterized mofezolac-derived COX-1 inhibitors combined with bortezomib on human multiple myeloma cell lines viability. *Eur. J. Med. Chem.* **2019**, *164*, 59–76. [[CrossRef](#)]
39. Lee, D.D.; Seung, H. Learning the parts of objects by nonnegative matrix factorization. *Nature* **1999**, *401*, 788–791. [[CrossRef](#)]
40. Boutsidis, C.; Gallopoulos, E. SVD-based initialization: A head start for nonnegative matrix factorization. *Pattern Recognit.* **2008**, *41*, 1350–1362. [[CrossRef](#)]
41. Moreaux, J.; Klein, B.; Bataille, R.; Descamps, G.; Maïga, S.; Hose, D. A high-risk signature for patients with multiple myeloma established from the molecular classification of human myeloma cell lines. *Haematologica* **2011**, *96*, 574–582. [[CrossRef](#)]
42. Buczynski, M.W.; Dumlaio, D.S.; Dennis, E.A. An integrated omics analysis of eicosanoid biology. *J. Lipid Res.* **2009**, *50*, 1505. [[CrossRef](#)]
43. Liu, W.; Yuan, K.; Ye, D. Reducing microarray data via nonnegative matrix factorization for visualization and clustering analysis. *J. Biomed. Inform.* **2008**, *41*, 602–606. [[CrossRef](#)]
44. Park, J.; Bae, E.K.; Lee, C.; Choi, J.H.; Jung, W.J.; Ahn, K.S.; Yoon, S.S. Establishment and characterization of bortezomib-resistant U266 cell line: Constitutive activation of NF-κB-mediated cell signals and/or alterations of ubiquitylation-related genes reduce bortezomib-induced apoptosis. *BMB Rep.* **2014**, *47*, 274–279. [[CrossRef](#)] [[PubMed](#)]
45. Chauhan, D.; Singh, A.V.; Ciccarelli, B.; Richardson, P.G.; Palladino, M.A.; Anderson, K.C. Combination of proteasome inhibitors bortezomib and NPI-0052 trigger in vivo synergistic cytotoxicity in multiple myeloma. *Blood* **2008**, *111*, 1654–1664. [[CrossRef](#)] [[PubMed](#)]

46. Ling, S.C.; Lau, E.K.; Al-Shabeeb, A.; Nikolic, A.; Catalano, A.; Iland, H.; Horvath, N.; Ho, P.J.; Harrison, S.; Fleming, S.; et al. Response of myeloma to the proteasome inhibitor bortezomib is correlated with the unfolded protein response regulator XBP-1. *Haematologica* **2012**, *97*, 64–72. [[CrossRef](#)] [[PubMed](#)]
47. Maïga, S.; Gomez-Bougie, P.; Bonnaud, S.; Gratas, C.; Moreau, P.; Le Gouill, S.; Pellat-Deceunynck, C.; Amiot, M. Paradoxical effect of lenalidomide on cytokine/growth factor profiles in multiple myeloma. *Br. J. Cancer* **2013**, *108*, 1801–1806. [[CrossRef](#)] [[PubMed](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).