Master in Artificial Intelligence (UPC-URV-UB)

Master of Science Thesis

Brain tumor Microarray Data analyzed using clustering and novel consensus clustering techniques

José Antonio Magaña Mesa

Director: Sergio Gómez Jiménez, DEIM, URV
Co-director: Alexandre Arenas Moreno, DEIM, URV

July 8th, 2014

> “You get nothing for free. There are never lucky accidents in the computer; only hard-won victories”
> Pixar 25th anniversary. The Art Exhibition.

To my parents

To Laura, see you on July 25th

To FM, for his involuntary advice to start this adventure

I. ABSTRACT

Micro Array data contains the whole gene expression of a tissue, not a person as the gene expression of every tissue is different as different functions/genes are activated on each. Over the last years it has become a powerful tool in the diagnosis of genetic related diseases, and particularly cancer.

This study has applied heterogeneous clustering techniques and defined consensus techniques in order to extract a robust partition from a Micro Array dataset. The different algorithms generate a variety of partitions which none of them is able to extract all the information that the dataset contains. Hence, it is required to find ways to combine these partial solutions in order to reach a comprehensive understanding of the dataset information. Existing methods in the literature underestimate the complexity of such process and suggest methods that do not cope with situations that arise from the combination of different solutions.

At the same time, the available ground truth for the dataset used in the study cannot be granted full reliability what has made that the methods used are fully unsupervised. The ground truth has been considered, in the final results analysis phase, in order to have an estimate of the agreement with the oncologists diagnose.

The data used is a Microarray dataset, GSE4290, from NIH database, with 180 samples and 54613 probes (features) corresponding to a study on brain tumors. The approach used has been constructing a distance matrix from the dataset in order to reduce the dimensionality of the problem.

In order to generate an ensemble of partitions, the clustering methods used have been chosen to be heterogeneous (Hierarchical Clustering, MSTKNN and Complex Networks based). For each of the algorithms used, a number of executions have been run with subsamples of the dataset. The results obtained for each subsample have been combined using an evolution of a well-known method for consensus clustering as is the work by Monti et al. in [1], that has been adapted to incorporate available information from the clustering methods used, in order to weight differently the partitions depending on their stability.

In order to obtain a single partition the two more different results obtained from the three algorithms have been combined using robust consensus generating a unique partition that becomes the solution partition for the dataset. The experiment has been run for 6 different settings as the dataset has been normalized in several ways and two different distance functions have been applied.

Finally, the robustness of the methodology proposed has been evaluated by repeating the experiments a number of times and the results compared among them using a battery of well-known similarity measures, both for the individual algorithm results and the proposed consensus solution. In addition, the variability of the different clustering methods has been measured using the membership coefficient of the consensus matrixes. The results show that the sought consensus occurs for one of the settings configuration used while the other configurations converge much later or do not converge at all.

The obtained partition has 6 clusters, one containing the whole control group, a second one containing most of the Oligoblastoma types and two others that split the Glioblastoma group. There are two small clusters with 5 and 1 samples respectively. The Astrocytoma group on the dataset has not been separated having a much smaller number of samples.

The agreement of the result obtained with the physicians' diagnosis has been measured using a modified Purity Index in order to consider, as it is the case, that the partitions obtained identify 2 subtypes of one of the brain tumor classes.

As a direct application of the method, from the partitions obtained for the Microarray dataset GSE4290, it has been possible to identify the features (probes that correspond to genes) that better help to classify (diagnose) samples (patients) with several types of brain tumors.

In order to achieve this, the features that explain the partitions have been extracted using the CM1 indicator and the results have been compared to other studies that have used the same dataset. The function of the genes identified corresponds, in a very high percentage, to genes related to oncological processes and metabolic pathways with an incidence in the development of the disease.

The accuracy of 72% on the purity indicator is in the same level that the best results obtained in other studies using the same dataset. It is also in the same level of precision that several studies attribute to physicians when considering the varieties of tumors with higher degree of agreement in the diagnosis, Glioblastoma, while it significantly improves the diagnosis of other subtypes, for example Oligoblastoma.

The results have been validated by an expert in biomarkers in order to support the conclusions from a domain point of view also based on extense available literature on the topic.

Table of contents

I. Abstract ... 3
II. Background and Motivation ... 15
III. Introduction ... 16
IV. State of the Art ... 17
A. Microarray ... 17
B. Clustering ... 18
C. Clustering methods ... 21
1) Hierarchical Clustering(HC) ... 21
a) Bottom-up methods ... 21
b) Top Down methods ... 21
c) Disadvantages of Hierarchical Clustering ... 22
2) Multi-resolution clustering based on complex networks ... 22
3) MSTKNN ... 23
D. Consensus clustering ... 24
V. Dataset ... 28
A. Main Characteristics ... 28
B. Data Acquisition ... 29
VI. Methods ... 30
A. Pre-processing: Outliers filtering ... 31
B. Pre-processing: Sample distance matrix generation ... 33
1) Jensen-Shannon Divergence Square Root ... 34
a) Normalization (for sqrtJSD) ... 35
2) Pearson distance ... 36
C. Pre-processing: Normalization and inversion ... 37
D. Pre-processing: Resampling ... 39
E. Clustering ... 39
1) Hierarchical clustering ... 40
2) Complex Networks based, modularity optimization ... 41
3) MSTKNN ... 43
F. Intra-method consensus ... 44
1) General process ... 44

2) MSTKNN 44
3) Hierarchical clustering 45
4) Complex Networks, modularity optimization 45
G. Selection of the two more different partitions 46
H. Inter-method consensus 48
1) Discussion of existing methods 48
2) Proposed method for inter-method consensus: Robust clustering 49
I. Feature extraction 50
J. Solution analysis 50
1) External robustness 51
2) Internal robustness 53
3) Clustering vs Decision Tree 54
VII. Experiments 55
VIII. Results 57
IX. Discussion and Conclusions 61
X. Future Work 63
XI. Appendix 64
A. Bio-interpretation of the results 64
1) Qualitative evidence of relevance 64
2) Expert evaluation: a bio perspective (by Prof. Pablo Moscato) 66
a) Preliminary analysis 66
b) Annotation of the results 71
c) Conclusions 73
B. Experiments results 74
1) All methods, N Iterations=100 74
2) All settings, N Iterations=200 79
3) All settings, N Iterations=500 84
4) Raw, Column-Normalization, Jensen-Shannon, N Iterations=100 to 400 90
5) Comparison of Accuracy for all methods 95
6) Selection of Best K 96
7) Solution variability for each algorithm for RAW-JS configuration 103
8) Solution variability for each algorithm for WO Outliers-JS 103
C. Feature selection using CM1 104

1) Probeset to Gene matching from CM1 extracted features 104
2) CM1 feature extraction diagrams 111
D. Hierarchical clustering linkage functions... 116
E. Similarity and dissimilarity measures 117
1) Agreements and Disagreements 118
2) Jaccard Index 118
3) Fowlkes Mallows Index.. 118
4) Normalized Mutual Information Index 118
5) Normalized Mirkin Metric.. 119
6) Normalized Van Dongen Metric 120
7) Normalized Variation of Information Metric.. 120
8) Adjusted Rand Index 120
9) Purity.. 121
10) Homogeneity and Completeness 121
F. Data Normalization tests 122
G. Brain tumour diagnosis reliability.. 124
H. Known results obtained from dataset GSE4290 125
XII. References 129
XIII. References for bio-informatics study 133

List of figures

Figure 1 Tutorial on Genetics and Microarray Data 17

Figure 2 Classification of clustering methods... 19

Figure 3 Number of clusters as function of R showing the stable number of clusters for two datasets 22

Figure 4 MSTKNN applied to the GSE4290 dataset... 23

Figure 5 Data points, consensus matrixes and CDF from [1] 25

Figure 6 Processing pipeline 30

Figure 7 Number of features below the threshold for MAD.. 33

Figure 8 Histogram for the range of values for the 54K genes 33

Figure 9 Histogram for the range of values for the sum of probes for the 180 samples 34

Figure 10 Histogram for inversions proposed (left w=1/d, right w=1-d) 37

Figure 11 Dendrograms of the GSE4290 dataset for different HC linkages 40

Figure 12 Evolution of number of clusters for GSE4290 dataset 42

Figure 13 Distribution of k for 200 subsampled executions of MSTKNN.................. 44

Figure 14 Separation not possible in the same K than original 49

Figure 15 Consensus partition result for RAW-JS settings 59

Figure 16 Probesets and genes used in Scikit Decision Tree for the cluster separation60

Figure 17 Scikit Decision Tree applied to the cluster solution 60

Figure 18 Probesets and genes from 1-vs-all CM1.. 68

Figure 19 Venn-diagram for G1X-vs-G2X, OX, CX 69

Figure 20 G1X and G2X Venn diagram analysis 70

Figure 21 Accuracy for N Iterations=100 79

Figure 22 Accuracy for N Iterations=200 84

Figure 23 Accuracy for best partition N Iterations=500.. 89

Figure 24 Membership coefficients for the clustering methods (RAW-JS) 89

Figure 25 Accuracy for different Nr of Iterations for Raw Jensen-Shannon settings .. 93

Figure 26 Clustering of GBM-vs-AC and GBM-vs-OL, from [77] 126

Figure 27 GSE4290 Prognosis classification.. 127

List of tables

Table 1 Class distribution in GSE4290 28

Table 2 Pseudo-code for processing pipeline 31

Table 3 Histogram of distance matrixes 38

Table 4 Comparison of the different clustering algorithms 40

Table 5 How r is obtained for each sample in the scanned range 42

Table 6 Comparison of best K (central column) for each algorithm with K±1 47

Table 7 Measures to choose the two more different partitions 48

Table 8 Voting of pair differences 48

Table 9 Confusion Matrix for solution partition, including J factor used in EP 51

Table 10 Parameter Set Scenarios 55

Table 11 Qualitative analysis of the best result for each configuration 57

Table 12 Block model clustering confusion matrix 63

Table 13 Genes also found in other GSE4290 studies 65

Table 14 Genes related to Epilepsy 65

Table 15 Genes related to Glioma, Glioblastoma, Gliosarcoma, Astrocytoma, Oligodendroglioma, Limphoma, Neuroblastoma 65

Table 16 Genes related to Cancer 65

Table 17 Genes related to Cancer and Nervous System Disease 66

Table 18 Genes common to G1X and G2X, literature references 72

Table 19 Genes only in G1X 72

Table 20 Genes G1-vs-G2 and G1X/G2X 73

Table 21 Genes G1-vs-G2 and G2X/G1X 73

Table 22 Partition results for N Iterations=100, Raw Dataset 75

Table 23 Partition results for N Iterations=100, WO Outliers 76

Table 24 Indexes and Agreements for N Iterations=100 78

Table 25 Partition results for N Iterations=200, Raw Dataset 80

Table 26 Partition results for N Iterations=200, WO Outliers 82

Table 27 Indexes and Agreements for N Iterations=200 83

Table 28 Partition results for N Iterations=500, Raw Dataset 85

Table 29 Partition results for N Iterations=500, WO Outliers 87

Table 30 Indexes for N Iterations=500 88

Table 31 Agreements for N Iterations=500 88

Table 32 Partition results for N Iterations=100 to 400, Raw Dataset, column normalization, Jensen-Shannon 91

Table 33 Agreements and Indexes for Raw Jensen-Shannon data 92

Table 34 Membership coefficients for the different clustering algorithms 94

Table 35 Evolution of accuracy convergence for different N Iterations 95

Table 36 Selection of best K for CN 98

Table 37 Selection of best K for HC 100

Table 38 Selection of best K for MSTKNN 102

Table 39 Variability of the solutions for individual algorithms in RAW-JS configuration 103

Table 40 Variability of the solutions for individual algorithms in WO Outliers-Row-JS configuration 103

Table 41 Probes IDs not found on SOURCE database 104

Table 42 Unique genes 1-vs-All 105

Table 43 Unique Gene IDs in G1X-vs-G2X, G2X-vs-G1X not in 1-VS-all 105

Table 44 CM1 Underexpressed genes, 1-VS-ALL 106

Table 45 CM1 Overexpressed genes, 1-VS-ALL 107

Table 46 CM1 indicators for 1-vs-1, G1X-vs-G2X 108

Table 47 Unique Gene IDs, down-regulated 109

Table 48 Unique Gene IDs, upregulated 110

Table 49 Number of unique genes per class 110

Table 50 CM1 weights for top/bottom 50 1-vs-All 111

Table 51 CM1 weights for CX-vs-1 112

Table 52 CM1 weights for OX-vs-1 113

Table 53 CM1 weights for G1X-vs-1 114

Table 54 CM1 weights for G2X-vs-1 115

Table 55 Synthetic dataset 122

Table 56 Column normalized dataset 122

Table 57 Dataset after row normalization 123

Table 58 Dataset after subsequent row and column normalization 123

Table 59 Sample to sample factor for any gene with column normalization 123

Table 60 Sample to sample norm factor for any gene with row-column normalization 124

Table 61 Results in [77] showing MEA 125

Table 62 11 GB serum markers 125
Table 63 42 Probesets relevant for glioma prognosis 127
Table 64 List of gliogenesis related genes as in [80] 128
Table 65 Genes overexpressed in GBM for EGFR and 12q13-15 128

List of equations

Equation 1 Modularity 20
Equation 2 Consensus Matrix 24
Equation 3 Cumulative Distribution Function for consensus 24
Equation 4 Area Under the Curve 24
Equation 5 Delta AUC 25
Equation 6 Individual cluster membership coefficient 25
Equation 7 Partition membership coefficient 25
Equation 8 Mean Absolute Deviation 32
Equation 9 MAD to Standard Deviation conversion for normal distributions 32
Equation 10 Jensen-Shannon Divergence 34
Equation 11 Shannon Entropy 34
Equation 12 Jensen-Shannon Divergence square root triangular inequality 35
Equation 13 Pearson Correlation Coefficient 36
Equation 14 Pearson distance 36
Equation 15 Weight definition from distance 37
Equation 16 Alternate weight definition 37
Equation 17 CM-1 50
Equation 18 Intra-Purity 52
Equation 19 Extra-Purity 52
Equation 20 extra-Purity per cluster 52
Equation 21 Extra-Purity general case 52
Equation 22 Numerical example of the calculation of P, IP and EP 53
Equation 23 Single linkage 116
Equation 24 Complete linkage 116
Equation 25 Average linkage 116
Equation 26 Weighted linkage 116
Equation 27 Centroid linkage 116

Equation 28 Median linkage.. 117
Equation 29 Ward linkage... 117
Equation 30 Jaccard Index 118
Equation 31 Fowlkes-Mallows Index 118
Equation 32 Normalized Mutual Information arithmetic... 119
Equation 33 Normalized Mutual Information geometric... 119
Equation 34 Mutual Information... 119
Equation 35 Mutual Information for Clustering 119
Equation 36 Mirkin Metric... 119
Equation 37 Normalized Mirkin Metric... 119
Equation 38 Van Dongen Metric 120
Equation 39 Normalized Van Dongen Metric... 120
Equation 40 Variation of Information... 120
Equation 41 Normalized Variation of Information.. 120
Equation 42 Rand Index... 120
Equation 43 Adjusted Rand Index 120
Equation 44 Purity... 121
Equation 45 F-Measure 121
Equation 46 Homogeneity.. 121
Equation 47 Completeness 121
Equation 48 V-measure 122

II. Background and Motivation

The term Bioinformatics was coined in 1970 by Paulien Hogeweg[2] with a much narrower meaning that it has nowadays. The concept has broadened while maturing supported by other, also emerging, technical and scientific disciplines such as Artificial Intelligence, Genetics Engineering and the rest of bio-x fields that constitute a solid cluster where each pillar reinforces the others while benefiting from them in a perfectly harmonious symbiosis.

Currently we can describe Bioinformatics as the discipline that using diverse computing technologies studies and provides the technologies to capture, store and analyze the information related to biological processes. Within this wide definition, genetics occupies a central space.

Bioinformatics has experienced a magnificent progress in the last decade, as can be read in [3]. The cost and time required for DNA sequencing has been reduced from 30 days and 100K$ to 1 day and 3K$. The focus is now on analysis of the data that can be obtained at such affordable price as it is estimated that the information retrieved in one month takes 5 months to be analyzed. Despite this staggering progress made on Bioinformatics, or because of this, it can be considered a young discipline, with many and ambitious challenges ahead.

Some of these challenges, that have inspired and guided this thesis are, according to [4], (1) the processing of large-scale robust genomic data, (2) the interpretation of the functional effect and the impact of genomic variation, (3) integrating systems and data to capture complexity and (4) make results clinically relevant so that they are translated into medical practice. Underlying needs to achieve this as explained in [5], are, even now (13 years after publication), (1)the understanding of the sources of noise and variation in Microarray experiments, (2) the combination of expression data with other sources of information to improve their range and quality and (3) the reconstruction of networks of genetic interactions in order to create integrated and systematic models of biological systems.

It has been my desire to take advantage of this Master's Thesis to modestly contribute to the field in a way that it produces a benefit to society (even if just tiny). This desire is the outcome of a summer stay at the Hunter Medical Research Institute (HMRI) in Newcastle (Australia) where I have had the opportunity to be exposed and learn about the research projects in place at the Center for Information Based Medicine (CIBM) under the supervision of Prof. Pablo Moscato. The focus of most of such projects is the discovery of Biomarkers that assist in the diagnosis of diseases with a genetic origin and the selection of treatments for the different illnesses subtypes.

Cancer, being one of the most fatal diseases at a global level, is the one being object of more studies currently. Brain tumors have been the selected disease to study in this thesis.

III. INTRODUCTION

Gene Expression Microarray Data based experiments typically fall into three types of problems, according to [6]: “(i) identification of new tumor classes using gene expression profile – unsupervised learning; (ii) classification of malignancies into known classes- supervised learning; (iii) identification of marker genes that characterize the different tumor classes – feature selection.”

From a bioinformatics point of view, as stated in [5]: “clustering methods are now more routinely being evaluated with respect to criteria such as robustness, computational cost, clarity of cluster definitions and reproducibility”, introducing the desirable characteristics of the algorithms.

At the same time, Kleinberg, in [7], states that no clustering algorithm exists that can satisfy three basic properties (scale-invariance, richness, consistency) that are required in order to grant clustering results major trust. In opposition to the former, Zadeh’s theory in [8], relaxes Kleinberg’s axioms, although restricted to clustering methods where the number of clusters is provided, to the identification of Hierarchical Clustering with Single Linkage as the only method satisfying the three basic axioms defined (scale-invariance, order-consistent, k-richness).

As a consequence of this, it comes that different clustering algorithms may generate very different partitions depending on their characteristics, what represents a limitation that adds to the lack of specific meaningfulness result of the unsupervised nature of clustering.

In short, there is no guarantee on the kind of separation a clustering algorithm is going to generate. The only certainty is that it will, in a certain but unknown space, maximize the difference between members of different clusters while minimizing the difference between cluster members and that to a certain resolution, as the number of clusters may be different. The reason for that being the unsupervised nature of the process, there is no constrain in the kind of separation among samples that will be extracted.

This being the case, it must be considered when analyzing a complex dataset using clustering, that the algorithm applied may not be able to extract all the relevant information implicit in the dataset. From this, it comes as a direct conclusion that it is necessary to apply a variety of algorithms and try to combine the results obtained from the different algorithms to produce a partition that represents all the groups in the data.

Bearing in mind these guidelines, the objective of this work has been to define a methodology that can be used to obtain robust and reliable partitions of datasets from very heterogeneous partitions generated by different clustering algorithms applied to the same dataset. From the partition, the genes that help to explain the separation will be extracted and assessed its relevance to explain the tumor classification.

IV. STATE OF THE ART

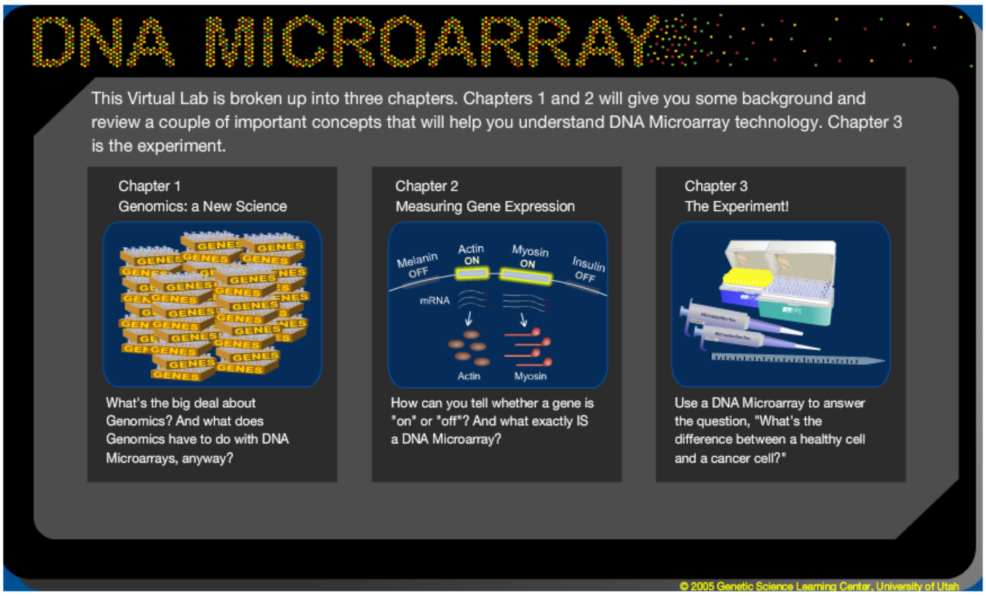
Both clustering and bioinformatics are hot topics that have generated numerous publications. A comprehensive study of the state of the art on any of either topics or the combination of them would suppose an overwhelming effort that would consume the time this thesis is supposed to take. Hence, this section is a brief review of the literature in order to have a first insight of the current status of the field. It is then a best effort task that would require a deeper research in order to be considered complete, what is beyond the objective of this thesis.

A. *Microarray*

Genetic data commonly proceeds from Microarray chipsets. A Microarray chipset can be seen as a tray full of microscopic spots, called probes, containing, each, multiple identical DNA strands that match to one of the genes the human (or other organism) genome has. The human genome contains 21K genes. Several probes correspond to the same gene. The mapping of the probes to the matrix position in the Microarray surface is precisely registered. The manufacturing process of Microarray chips is alike to that of a microprocessor.

The process of extracting a genetic signature of a certain tissue using Microarray technology requires a lot of steps involving manual processes and bio-chemical reactions making it quite sensitive. Rather than explaining the process how a microarray experiment is performed, what is not a core knowledge of this Masters, those requiring an explanation can view the following interactive and highly pedagogical tutorial available from Utah University[9].

http://learn.genetics.utah.edu/content/labs/microarray/



DNA MICROARRAY

This Virtual Lab is broken up into three chapters. Chapters 1 and 2 will give you some background and review a couple of important concepts that will help you understand DNA Microarray technology. Chapter 3 is the experiment.

Chapter 1
Genomics: a New Science

What's the big deal about Genomics? And what does Genomics have to do with DNA Microarrays, anyway?

Chapter 2
Measuring Gene Expression

How can you tell whether a gene is "on" or "off"? And what exactly IS a DNA Microarray?

Chapter 3
The Experiment!

Use a DNA Microarray to answer the question, "What's the difference between a healthy cell and a cancer cell?"

© 2005 Genetic Science Learning Center, University of Utah

Figure 1 Tutorial on Genetics and Microarray Data

From the explanations about the process it can be quickly observed that the process is far from exact and hence data proceeding from Microarray data is affected by high variability on its measures due to the many noise sources that along the whole process can affect the sample.

In this case, the Affymetrix Human Genome U133 Plus 2.0 Microarray chipset has been used. This chipset contains 54613 probes mapping to the whole human genome.

Being well established the variability of Microarray experiments they are designed with normalization controls so that the measurements are tractable with a reasonable level of confidence that guides the researchers that use it to make the adequate data pre-processing of the information.

Many techniques have been developed in order to address the main complexities inherent to Microarray data. From a sampling on the vast existing literature [10-13], a large group is related to Pre-processing (including Image Analysis), Normalization and Outlier detection. Other important problems addressed are related to the link of the biological information in the processing algorithm from the beginning in order to guide any search/combinatorial process involved. A third group tackles the different uses that the data can be given (class prediction, classification or discovery). Finally, the techniques are applied to specific problems (diseases / organisms). Clustering appears recurrently as an effective technique for class discovery and also related to feature selection as a means of validation of the extracted features.

B. *Clustering*

As aforementioned, clustering is the AI technique commonly used in Bioinformatics for class discovery. The literature provides many examples in which the use of clustering techniques have permitted to discover or re-discover, new subtypes of different diseases. In [14], Alizadeth et al. discover a new variety of lymphoma: “have conducted a systematic characterization of gene expression in B-cell malignancies” while, in [15] Golub et al., “A class discovery procedure automatically discovered the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) ".

The approaches taken to provide clustering solutions are of different nature: neural networks (SOM), spectral, genetic algorithms or probabilistic models are just some of the techniques applied to the problem.

Some well-known state of the art studies have been published that summarize in a very comprehensive way the work done in this field. From [16] and [17], the following classification of clustering methods can be extracted:

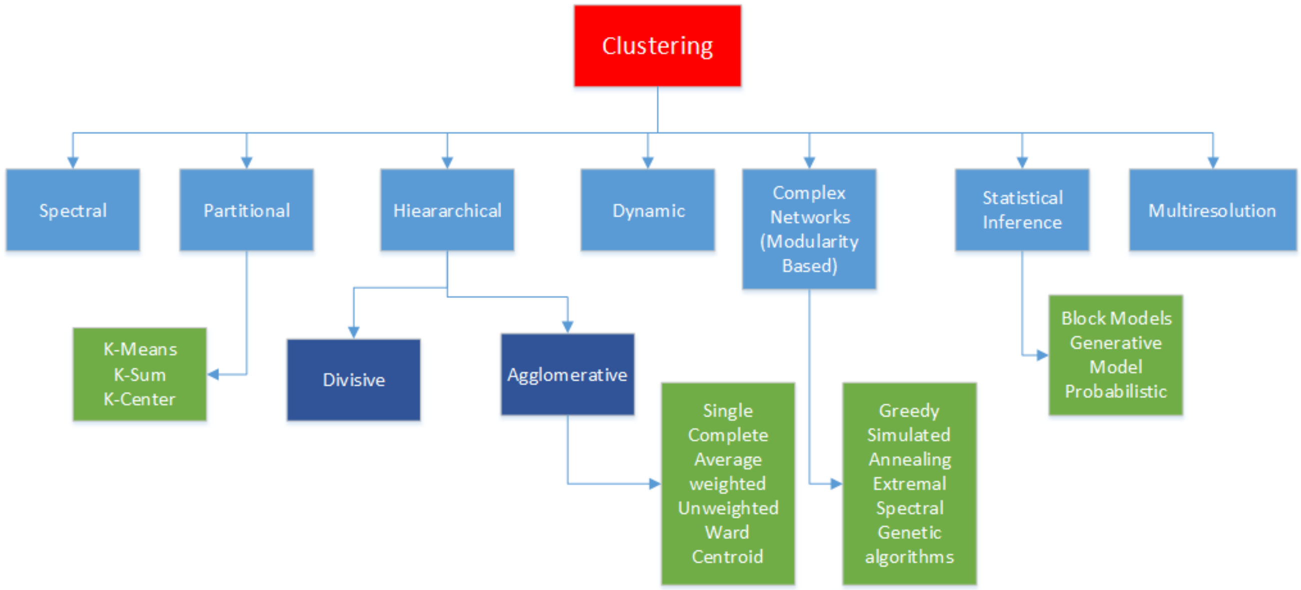


Figure 2 Classification of clustering methods

If we survey [6, 10-13, 18] the clustering algorithms used in the study of Microarray(MA) Data we find that the most widely used are very traditional non-sophisticated ones, like Hierarchical Clustering (HC) (in any of the linkage varieties) or among Partitional: the K-means family and Self Organizing Maps (SOM), as referred in[19]. A third approach is based on the application of component analysis and value decomposition (like PCA or SVD) to transform the data into a different feature space. It is very common also to apply Two-way clustering, consisting in clustering not only the samples but also the features (sometimes in an interdependent way) what helps in feature selection. Some other less traditional approaches have also been successfully applied to MA clustering like SQVT, a form of divisive HC, or kernel-PLS, a predictive model. CAST(Cluster Affinity Search Technique)[20, 21], CLICK(Cluster Identification via Connectivity Kernels)[21, 22], CURE[23] or PAM(Partition Around Medioids) [24, 25] are also examples that have appeared several times during the literature review either as methods to benchmark to or inspiration for new methods.

Some clustering methods used in Bioinformatics, and in particular the ones selected in this work, are based on measuring the pairwise distance in the feature space of the samples to generate a distance matrix. A wide range of options have been proposed as distance functions to be used: Minkowsky[26] in any of its particular cases (Euclidean, Manhattan, Maximal), Mahalanobis[26] and Pearson(see section VI.B.2) or Kullback-Leibler[27] are just some of them.

In order to measure the quality of the results obtained a variety of metrics exists. These metrics fall into different categories. Graph measures are related to some variety of Modularity. Modularity calculates the difference between inter-cluster and intra-cluster edge weights, giving a measure of the modular structure of the network defined by the partition being evaluated. A general version of the Modularity that considers weighted and un-weighted networks [28] is:

$$Q = \frac{1}{2w} \sum_{i,j} \left(w_{ij} - \frac{w_i w_j}{2w} \right) \delta(C_i, C_j)$$

Equation 1 Modularity

where w_{ij} is the weight/strength of the edge connecting samples i and j, w_x is the sum of all the edges of node x, 2w the sum of weights of all the nodes in the network (note that every edge will be added twice, what explains the 2 factor) and the $\delta(C_i, C_j)$ is the Kronecker function indicating whether samples i and j are in the same cluster. The higher the modularity value, the better is the partition.

Supervised measures, also known as externally supported, are based on some calculation performed on the confusion matrix obtained by crossing the obtained partition with the ground truth for the data set. Mutual Information, Variation of Information, Jaccard Index, Rand Index, Mirkin Metric, Wallace Index and some Normalized and Adjusted versions of them are just a few examples of this family. On Appendix XI.D an overview of the main indicators and the ones used in this study is included.

Among unsupervised methods, Silhouette [29-31] and Dunn [29, 32-34] are the most referenced ones. Both indexes are based on relationships among maximum inter-cluster distances and some variety of maximum or average intra-cluster distance.

In respect to metrics, [29] states, that any Cluster Validity Index chosen will be biased towards some of the desirable properties of a cluster result (compactness, separability, connectivity) that may be different from the criterion applied by the clustering method. Hence, it is required to carefully choose any metric used for this purpose.

The similarity with modularity methods is evident, the difference being that Modularity is based on weights among samples edges what does not require a complete connection among samples. Distances, in addition, can be seen as some form of inverse of weights, meaning that for two samples being on the same cluster, intuitively, it will be the case that there will be either small distances or large weights among the samples.

Clustering, due to its unsupervised nature, is a common and powerful tool in bioinformatics as it provides a second diagnosis isolated from that obtained from other sources. The possibility of misdiagnosis, what is common even when made by highly qualified doctors, must be taken in consideration when analyzing the data. A brief summary of some papers on the reliability of tumor diagnosis by physicians can be found in Appendix XI.G. Both studies confirm the lack of consensus in glioma diagnosis with percentages of non-agreement typically around 30% in the best case that can get to 70% for certain subtypes.

C. Clustering methods

The number of available clustering methods is endless as it is the literature about the topic. This study is based in three different Clustering algorithms that have different characteristics but that have something in common: the three can be applied to a distance or weight matrix for the dataset.

1) Hierarchical Clustering(HC)

Hierarchical Clustering (HC) is a wide family of algorithms rather than an algorithm itself. It has plenty of variants whose common ground is the fact that rather than generate a single partition, a series of nested partitions are obtained. The series of partitions are characterized for having strict borders, that is, if two elements are in separate clusters in the partition with k clusters, the same elements will always be in separate clusters for any partition with k'>k clusters.

Naturally, a first classification of HC separates the methods in two groups, bottom-up or aggregative and top-down or divisive.

a) Bottom-up methods

Bottom-up methods start from the list of samples and at each step put together the closest two elements. The two elements are removed from the list and a new element (cluster) is added with distances to the rest of elements of the list calculated according to the selected linkage criteria.

Several are the linkage methods most commonly used that can be classified in two groups. In the first linkage criteria group we find the methods Single, Complete, Average or Weighted Linkage. On the second group we find Centroids, Median and Ward linkage. See Appendix XI.D for a whole description and discussion of the methods.

The first group has a common characteristic; they are computationally more affordable than the second group as all distance generation for the generated clusters can be generated from the original distance matrix. This makes the clustering calculation independent of the dimensionality of the feature space what in the dataset is very high and would introduce a computational overhead.

b) Top Down methods

In top-down methods the best partition at each step cannot be easily identified. Hence, many methods are based on search of a local optimum. Examples of top-down methods, as described in [6], chapter 4, are the Tree Structure Vector Quantization and Macnaughton-Smith.

The former is based on K-means, as it is based on recursively applying a 2-means clustering to each obtained partition until the whole dataset has been separated in individual clusters.

The latter is based on finding the point with greatest mean dissimilarity to other points. This point will be the centroid of the new group and the points in the group moved one at a time until no member is closer to the splinted group.

c) *Disadvantages of Hierarchical Clustering*

In general, HC has been accused of important shortages compiled in *[26]* " Kaufman and Rousseeuw (1990) commented that hierarchical clustering suffers from the defect that it can never repair what was done in previous steps. Morgan and Ray (1995) showed that hierarchical clustering suffers from lack of robustness, non-uniqueness, and inversion problems that complicate interpretation of the hierarchy."

2) *Multi-resolution clustering based on complex networks*

The algorithm was first published in [28] , with subsequent enhancements in [35] and permits evaluating the system being studied at multiple scales or resolutions. To make this possible, the diagonal of the correlation matrix is modified increasing the value in every iteration. Each iteration represents a new version of the network for which a partition is generated while the modularity of the network is optimized.

The effect of modifying the diagonal of the correlation matrix is equivalent to adding a self-loop to each node whose weight is incremented along the process. The self-loop acts as a modifier of the node strength to become an individual cluster instead of being co-clustered with his neighbors. The effect of the self-loop in the whole network causes the emergence of different partitions for each self-loop value.

The weight of this self-loop, referred as r, needs to be within a range that makes possible to create partitions that have from 1 to the total number of nodes clusters. Since we are modifying the network at each step there is no relationship among the modularity obtained for every iteration.

The algorithm is non-parametric as the minimum and maximum required values for r can be automatically calculated. Both positive and negative weights can appear in the graph. R can also take negative values, even when the matrix is strictly positive, as it is the case. The higher r, the more clusters will appear in the partition obtained, although it is not monotonic and due to the stochasticity of the algorithms used there may be a decrease in the number of clusters as r grows, as can be observed in Figure 12. The increase on the number of clusters does not have a fixed slope and different regions will have different derivatives.

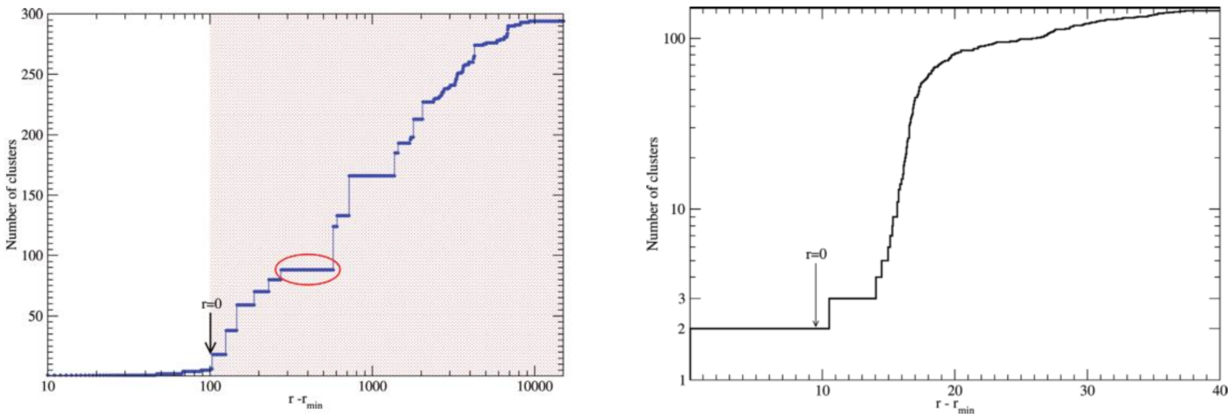


Figure 3 Number of clusters as function of R showing the stable number of clusters for two datasets

Figure 3 shows the evolution on the number of clusters for the whole r spectrum of two examples in [36]. The number of clusters covers the whole range of number of partitions. The diverse plateaus show the size of the stable partitions in the dataset.

3) *MSTKNN*

MSTKNN[21] is a technique developed at the CIBM Research Centre. The algorithm is based on the intersection of two independently generated graphs. On one side, the Minimum Spanning Tree of the graph made by the pairwise distance matrix of the samples. On the other side, the k-Nearest Neighbor for the same base graph. For the second graph, k is defined automatically as the minimum k that makes the graph fully connected. The intersection of the two connected graphs does not need to be connected as in the generation of the MST graph, having to avoid the creation of loops will cause that different edges than the ones included in the k-NN graph are used.

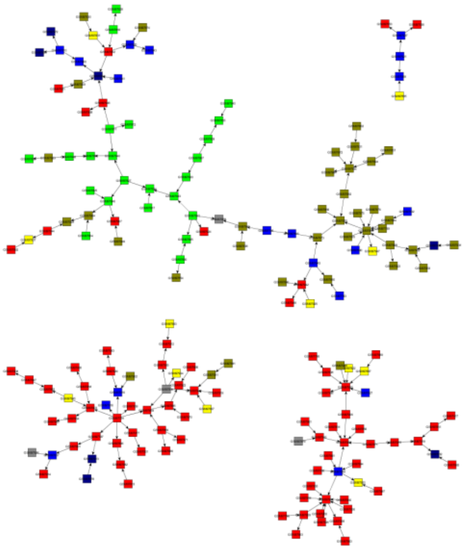


Figure 4 MSTKNN applied to the GSE4290 dataset

The method, as can be understood from its definition, is highly variant depending on the samples. The separations in a result partition will correspond to the removal of edges that being in the MST graph where not in the k-NN graph. This means that the edge was not one of the k best edges. The weakest edges in the MST graph are the ones that the intersection process causes to disappear.

In Figure 4, the result of applying the MSTKNN to the original dataset distance matrix can be seen. A total of 4 clusters are generated. Different node colors correspond to different sample diagnosis.

This method is the one generating a wider variety of partitions when subsampling is applied as it depends a lot on the samples being clustered.

D. *Consensus clustering*

With all this information, and recalling the introduction, it becomes necessary to find methods to combine the different partitions that are result of the different clustering algorithms that may be applied in order to obtain a 360º view of the dataset.

A disruptive method, and widely adopted, for combination of multiple partitions is [1], that defines a method to select automatically the number of clusters that best represents the dataset. As this algorithm is the basis for the implemented in this study a more complete description follows.

The algorithm takes a series of partitions of the same dataset and combines the partitions with the same number of partitions (k) into a single partition. To achieve this, a consensus matrix is computed, Equation 2, where each (i, j) position of the matrix corresponds to the ratio number of times in the same cluster for samples (i, j), $M^{(h)}(i,j)$, divided by the number of times in one of the dataset sub-samplings, $I^{(h)}(i,j)$ used to generate the range of partitions. The matrix is in the [0, 1] range, meaning for 0 that the two samples are never in the same cluster, and for 1 that the two samples are always in the same cluster.

$$\mathcal{M}(i, j) = \frac{\sum_h M^{(h)}(i, j)}{\sum_h I^{(h)}(i, j)}$$

Equation 2 Consensus Matrix

From the consensus matrix for each k a Cumulative Distribution Function (CDF) is obtained. Each point on the curve indicates the number of elements in the matrix whose value is less than the abscissa value.

$$\mathrm{CDF}(c) = \frac{\sum_{i<j} 1\{\mathcal{M}(i, j) \leq c\}}{N(N-1)/2}$$

Equation 3 Cumulative Distribution Function for consensus

In a second step, the Area Under the Curve (AUC) of this CDF is calculated and finally the relative increment respect to the previous (or largest so far) k is computed. The method states that the best k to cluster the dataset corresponds to the one with greatest relative increase of the AUC. The relative increase is measured respect to the largest AUC for smaller k's than the one being measured.

$$A(K) = \sum_{i=2}^{m} [x_i - x_{i-1}]\, \mathrm{CDF}(x_i)$$

Equation 4 Area Under the Curve

$$\Delta(K) = \begin{cases} A(K) & \text{if } K = 2 \\ \dfrac{A(K+1) - A(K)}{A(K)} & \text{if } K > 2, \end{cases}$$

Equation 5 Delta AUC

In order to measure the stability of the clustering, the membership coefficient can be defined for each cluster according to Equation 6 and from the average of clusters for the whole partition.

$$m(k) = \frac{1}{N_k(N_k - 1)/2} \sum_{\substack{i, j \in I_k \\ i < j}} \mathcal{M}(i, j)$$

$$I_k = set(\, j{:}\, e_j \in k)$$

Equation 6 Individual cluster membership coefficient

$$m_c = \frac{1}{K} \sum_{k=1}^{K} m(k)$$

Equation 7 Partition membership coefficient

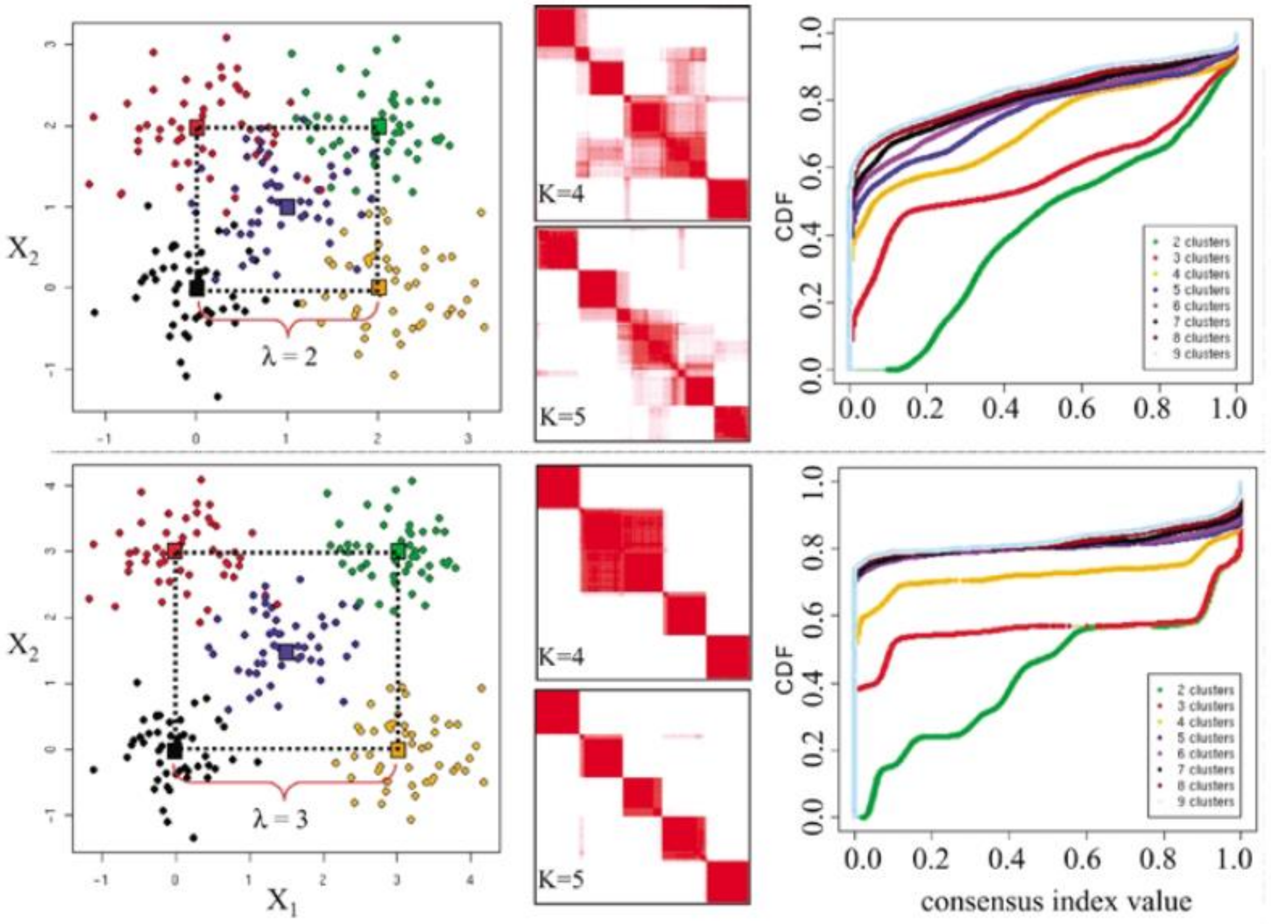


Figure 5 Data points, consensus matrixes and CDF from [1] .

Figure 5 shows for two synthetic datasets, the data points, the consensus matrixes obtained for the true k (k=5) and (k=4) and the AUC for all K's, showing in yellow and blue the curves for the mentioned k's. In the second row, the separation among clusters is clearer and so are the consensus matrixes and the differentiation between the consecutive AUC. For k>5 there is virtually no increase in the AUC.

The algorithm has several drawbacks, as it is the need of visually inspecting the curves to assess if the first k (k=2 and ΔAUC= ∞ as it compares to 0) is a better partition than the best k obtained from the algorithm definition. This is not a problem in our case, where we want the process to be automatic, as our dataset has 5 different classes so any partition with only 2 clusters will not be relevant for our purposes as we expect to be able to differentiate at least 3 different large clusters in our dataset. If the best partition corresponds to k=2 we would not take it and consider instead the second best option, without this fact being relevant for our purposes.

In order to resample the dataset, the authors suggest sub-sampling the feature set and taking subsets of the probes in the microarray matrix. Given that for this study the aim is to obtain the most representative biomarkers a different alternative will be used and the subsampling will be done for samples (patients). This, as will be shown, has some important benefits.

The method assigns the same weight to all the original partitions what is far from realistic as some algorithms like K-means require K to be defined blindly and others like hierarchical clustering produce a whole range of partitions for any K but not all of them have the same stability, as will be discussed later.

In [24], the authors differentiate between consensus clustering and robust consensus clustering and contribute a method who can produce both. The latter corresponds to the cluster that results from considering the strict assignment agreement of samples to the same cluster in several original partitions. The method permits combination of partitions from different algorithms. The weighted-kappa indicator is used, that is in its conception a supervised metric.

In [37], a consensus method is proposed based on a probabilistic (and generative) model where the different partitions obtained from a selected clustering algorithm are averaged. The concept of refined consensus clustering is used to solve the complexity introduced when the consensus clustering presents heterogeneity, a concept that will keep some attention in this study. For the same topic, a solution is suggested in [38] that measures when the consensus matrix obtained can be merged. The main remaining problem for the approach is that no alternative is proposed other than discarding them.

[25] is built directly on the work in [1], the method taken as basis in this study, to extend the method in order to make possible the combination of partitions from multiple algorithms by assigning arbitrary weights to the different methods. The method is still naïve in considering each of the individual partitions as equally weighted.

In [29], the consensus algorithm incorporates a variable weight for each k, that is calculated from a Cluster Validity Index (i.e. Dunn, Silhouette,…).

The stability and accuracy of consensus clusters is largely improved respect to the average of the individual partitions generated by multiple runs of the same algorithms, as is stated in [39].

With all this information, it can be expected that the application of consensus clustering techniques helps to improve the result obtained by individual algorithms. Furthermore, key elements to be considered in the algorithm have been identified such as differentiated weighted partitions or the choice of appropriate Cluster Validity Indexes (CVI).

Last but not least, the techniques found in the literature, [1, 24, 25, 37, 38] , have been successfully used with microarray datasets creating a solid ground to build the algorithm required for this study.

V. Dataset

A. *Main Characteristics*

Being restricted by ethical and regulatory constraints, a public dataset on brain tumors has been used for the study keeping in mind that the developed methodology should be able to be used in the future with other samples part of ongoing research for which a ground truth may not be available.

The dataset is made of 180 samples (patients), with a control group of 23 individuals that have a diagnosis of epilepsy this implying that their brain is not a healthy one and some gene signature divergence may not be related to the disease being studied but to epilepsy disorders.

The study has patients with 3 different types of tumors and, for two of the types, subtypes have also been identified. The classes and distribution can be seen in the following table:

GSE4290 Class Distribution			Sub-Class		Class	
			Number	Percentage	Number	Percentage
CLASS	Control	C	23	13%	23	13%
	Glioblastoma Grade IV	G4	77	43%	77	43%
	Astrocitoma Grade II	A2	7	4%	26	14%
	Astrocitoma Grade III	A3	19	11%		
	Oligoblastoma Grade II	O2	38	21%	50	28%
	Oligoblastoma Grade III	O3	12	7%		
	Unclassified	U	4	2%	4	2%
Total			180	100%	180	100%

Table 1 Class distribution in GSE4290

In [40], the authors define that clustering has some limits on the size of the partitions that can be identified. This resolution limit is defined as:

$$l < \sqrt{n}$$

being n the number of samples. In this study we have n=180 then l=13.41

Given that for some of the subtypes, namely A2 and O3, the number of samples is below the resolution limit aforementioned, 7 and 12 respectively, for the study we will work with the Class Partition, with 4 main classes (C, G4, A, O) and a small group of Unclassified samples.

The microarray platform used is Affymetrix Human Genome U133 Plus 2.0, that contains 54613 probes, providing complete coverage of the human genome, 21K genes. The complete technical description can be found here:

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL570

While the number of samples in the dataset may seem small (for typical datasets in the AI data repositories) it is indeed quite large in the context of Microarray data and genetic studies. In the NCBI database, there are only 1210 that have 180 samples or more out of 681.887 datasets for the species Homo Sapiens. This places the dataset in the 99.8 percentile in terms of number of samples.

The existence of unclassified samples in the dataset is not negative by itself and may produce some benefits. While it is true that no category can be assigned to the samples, given that our dataset is relatively small, and considering that clustering is an unsupervised method, it has been decided to keep the samples in the dataset for our study. The reason is that these samples may help to create the solution partition either because they help to connect nodes, acting as bridges among them, i.e. two samples that are close to an unclassified sample may get clustered together when otherwise they would be separated and consequently they may be separated from other nodes. At the same time, if we didn’t use them, we should better go for a supervised technique and make full use of the available ground truth, with the required caution given the accuracy that can be expected from it. Chances are that these samples are particularly difficult to classify.

B. *Data Acquisition*

Microarray Data studies produce a CEL file, that contains the result of scanning the different samples, without any processing. For this study, the already pre-processed information has been used as without having access to the manufacturer tools and processing software making a good normalization is difficult although can be done and in some cases may be necessary, for example, if samples from different platforms must be compared as in [41].

According to [42], the original study that made the dataset available, the dataset was processed following manufacturer defined protocol and using the manufacturer provided software in order to analyze very specific genes related to brain tumors:

“the CEL files were normalized to a median-intensity array, and model-based expression values were calculated using PM/MM difference model. Based on the latest annotation from Affymetrix NETAFFX service, four probe sets for SCF(KITLG)gene were present in each chip. The signal intensities of each probe set were used for analyzing SCF (KITLG) expressions. The significance in differential SCF (KITLG) expressions in different grades of gliomas versus human non tumor brains was determined using log2-transformed expression values by standard unpaired two-tailed Student’s t test of two groups without assuming equal variance between groups.”

VI. METHODS

The method proposed has the following process pipeline, which will be explained in detail in this chapter. The first row, acquisition, with green background, corresponds to the process performed by the authors of the dataset. The rest of the sections have been developed as part of this study.

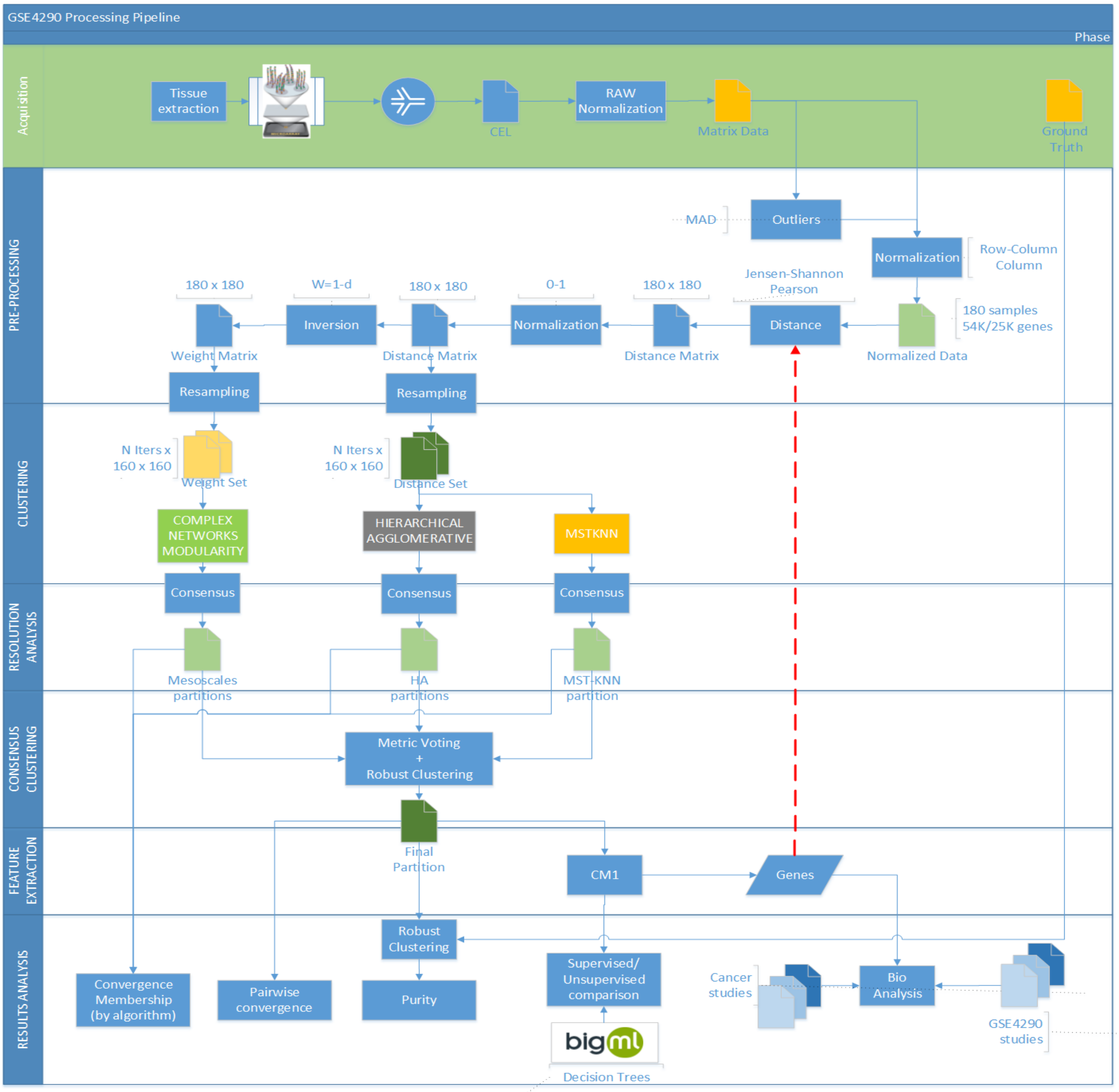


Figure 6 Processing pipeline

The pseudo-code of the data would be as follows:

```
multiClustering_consensus( no_outliers, Jensen-Shannon,[row_normalization]: bool;
                          N_times, N_iterations: integer
                          ground_truth: partition)

if no_outliers then
        data=RemoveOutliers(MAD>5, data)
if Jensen-Shannon then
        if row_normalization then
                data=RowNormalize( data, sum=1)
        data=ColumnNormalize( data, sum=1)
        distanceMatrix=Jensen-Shannon( data )
else  // Pearson as distance
        distanceMatrix=Pearson(data)
distanceMatrix =ApplyNormalization01(distanceMatrix)
weightMatrix=Invert(distanceMatrix)

// as CN based method is so heavy, we do the subsampling once
weightSubMatrixArray=subsample(100, weightMatrix)
modularityClusterArray=optimizeModularityClustering(weightSubMatrixArray)

repeat N_times:  // for statistical significance
      // hierarchical clustering
      hcSubDistanceMatrixArray = subsample(N_iterations, distanceMatrix)
      hcCluster = consensuate(hcSubDistanceMatrixArray)
      // mstknn
       mstknnSubDistanceMatrixArray = subsample(N_iterations, distanceMatrix)
      mstknnCluster = consensuate(mstknnSubDistanceMatrixArray)
      // cn
      modularityCluster = consensuate(N_selections, modularityClusterArray)
      cl1, cl2 = two_more_differents ( hcCluster, mstknnCluster, modularityCluster)
      cluster = robust_clustering(cl1, cl2)
      update_variability_internal (cluster)
      update_variability_external(cluster, ground_truth)

cm1=feature_extraction(cluster)
domain_analysis(cluster, cm1, ground_truth)
```

Table 2 Pseudo-code for processing pipeline

A. *Pre-processing: Outliers filtering*

Even after the pre-processing in the acquisition phase, microarray data has a very high variability due to the process required for its generation. It is then necessary to consider as a first step in the process to filter the features used in the analysis.

Because the number of samples is reduced, and more if we consider the number of samples for each class, this process is very sensitive and its convenience needs to be carefully considered. The reasons are that different subclasses may have different distributions particularly in those probes (genes) that explain the classification as quite often what causes a disease is an over-expression or under-expression of a set of genes and this miss-expression can be considered an outlier. Adding that the class distribution is unbalanced makes, intuitively, that what from one perspective can be seen as an outlier, from other perspective is an essential feature in our analysis.

Despite this, we have wanted to see the influence of adding an outlier removal step in our processing pipeline and being able to compare the results obtained for both configurations. Nevertheless, being outlier detection a complex topic that could inspire a whole study, the strategy has been that of defining a single aggressive criteria and eliminating all those features that have a sample that classifies as an outlier under such criteria.

By applying this, we will have two different datasets, one with the full set of features and a second one with a subset of the features after applying the outlier detection criteria.

The criteria defined to consider a data point a feature is based on the Relative Mean Absolute Deviation (RMAD) that is calculated for each of the features independently. A data point will be considered an outlier if its RMAD is greater than 5.

MAD is defined as:

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - m(X)|.$$

Equation 8 Mean Absolute Deviation

where m(X) corresponds, for our case, to the mean of the distribution. RMAD will be obtained dividing MAD by m(X).

MAD is related to standard deviation for normally distributed data:

$$\sigma = 1.2533\, MAD = \sqrt{\frac{\pi}{2}}\, MAD$$

Equation 9 MAD to Standard Deviation conversion for normal distributions

Our factor RMAD=5 would translate, if the distribution was normal (what it’s not claimed) to σ=6.2666. A high σ is not synonym of outlier more when the distribution cannot be considered Gaussian but gives an idea of the criteria being applied.

Calculating the number of features that would remain in the dataset if the MeanAD and MedianAD were used for different thresholds, we obtain the distribution in Figure 7.

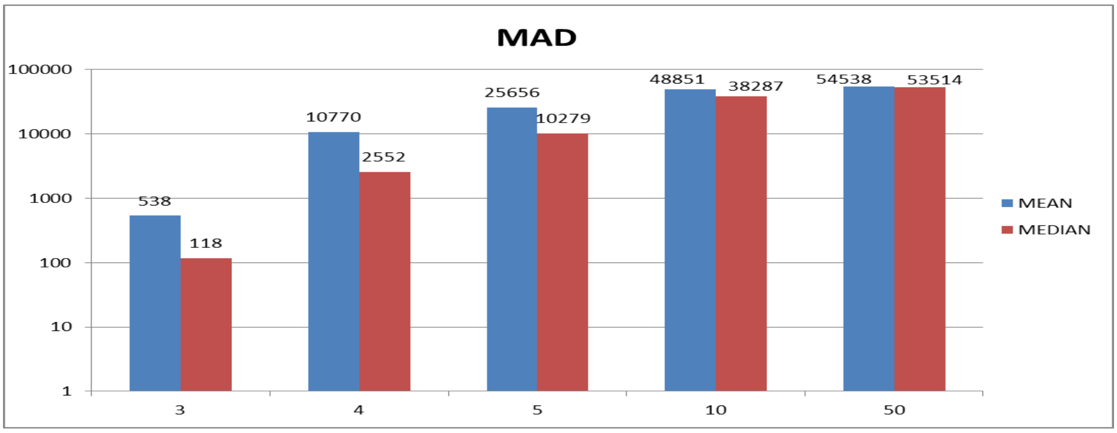


Figure 7 Number of features below the threshold for MAD

Applying this criterion, a total of 56.455 data points are considered outliers. These data points belong to 28.957 features, keeping then 25.656 features free of outliers. On average each feature removed will have less than 2 outliers, and based on this, the whole feature will be removed from our dataset as a first scenario definition variable. This method, that is quite aggressive given the low number of outliers per feature we have and the high number of features filtered, should guarantee that any outlier is removed and possibly other features are removed too. Despite the high number of features removed we expect to be able to cluster with reasonable results the resulting dataset.

B. *Pre-processing: Sample distance matrix generation*

The different features (probes) have a very diverse value range with differences of 4 orders of magnitudes from 6 to 60K when measured across genes, Figure 8. If the range is measured for the gene expression of an individual the difference is in one order of magnitude, Figure 9.

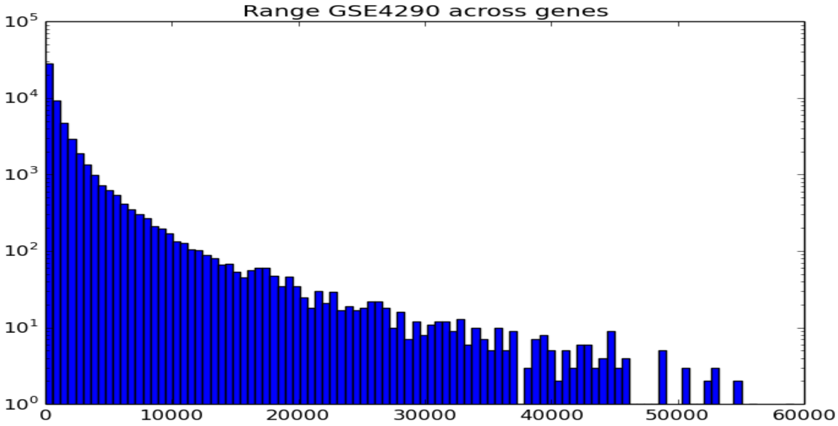


Figure 8 Histogram for the range of values for the 54K genes

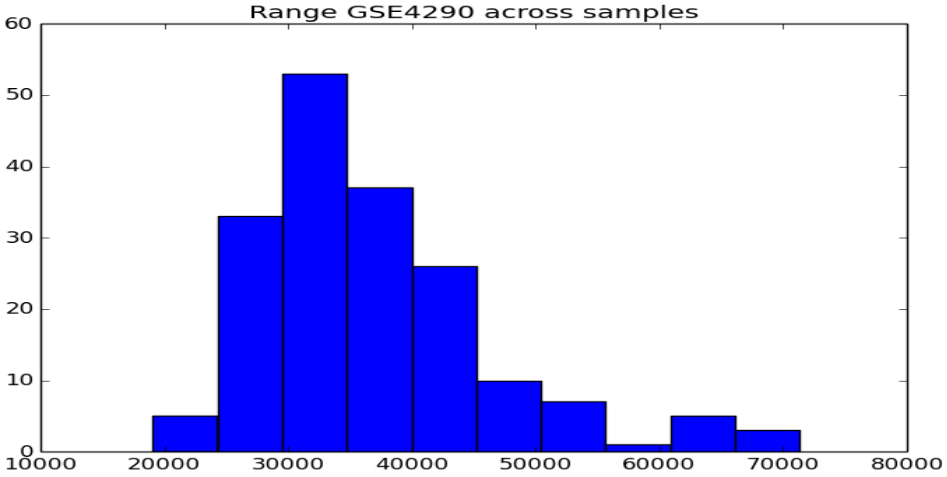


Figure 9 Histogram for the range of values for the sum of probes for the 180 samples

Because of the high-dimensionality of our data, a common strategy in order to reduce the computing required for the processing of the dataset is to obtain a pairwise distance among the samples. As explained in the State of the Art section, the available options are endless.

Two options have been chosen for this study. The first is the Jensen-Shannon Divergence (JSD), successfully used in microarray data studies [43, 44] and other domains [45], the second being Pearson Distance, that according to [46], happens to be used in 95% of the studies for them considered where cluster analysis is based on similarity.

1) *Jensen-Shannon Divergence Square Root*

Although Jensen-Shannon Divergence (JSD) was defined in [47], it is studied more in depth in [48] that has been the reference used in this study. Its formula is given by:

$$JSD(X,Y) = H(w * X + (1-w) * Y\ ) - w * H(X) - (1-w) * H(Y)$$

Equation 10 Jensen-Shannon Divergence

Where w belongs to [0,1], in our case w=1/2, and H is the Shannon's entropy:

$$H(X) = -\sum_{i=1}^{k} p_i \log p_i$$

Equation 11 Shannon Entropy

The measure requires then X and Y to be probability distribution functions, that is, their components to be positive and sum 1.The formula can be generalized to consider any number of distributions X, Y with different weights for each element in the distribution.

The JSD has important and useful characteristics for the type of data involved: (i) JSD is symmetric; (ii) it is non-negative; (iii) JSD is only 0 if the parameters are identical; (iv) JSD is well defined even if the distributions are not perfectly continuous, that is if

X_i vanishes without Y_i also vanishing. In addition, its square root is a metric since it satisfies the triangular inequality:

$$\mathrm{sqrtJSD}(X,Y) \leq \mathrm{sqrtJSD}(X,Z) + \mathrm{sqrtJSD}(Z,Y)$$

Equation 12 Jensen-Shannon Divergence square root triangular inequality

With these properties sqrtJSD becomes a very promising distance function for this study.

a) *Normalization (for sqrtJSD)*

As aforementioned, in order to apply sqrtJSD, it is required that the data meets certain requirements. The requirement is that the data columns (samples) must correspond to Probability Distribution Functions, that is, they must add to 1.

In addition to this, in [1], it is stated that: “The data used [in their experiments] were row- and column-normalized (so that both rows and columns sum to 0 and have a standard deviation of 1). This is necessary when using consensus clustering with HC, because it yields well-balanced hierarchical trees, which can in turn be split into non-trivial (i.e., non-singleton) clusters”. This normalization can be achieved following the method described in [49]. While it may be true that data so normalized, may generate non-trivial clusters, no evidence is provided in the mentioned paper that the results are more relevant. This type of normalization has then not been considered in our study. Also the requirements are different for the distance function chosen, sqrtJSD.

The requirement of sqrtJSD being that columns add to 1, we could have just applied a column normalization dividing each data point by the sum of its column. While this would have allowed us to meet the criteria, the high different range of values that data has, causes that by normalizing to sum to unity we may be introducing variations in our data.

A very simplistic test has been performed with synthetic data in order to confirm this hypothesis and measure the effect of it and an alternative normalization. The alternative normalization consists on row normalizing to sum to unity (same criteria) prior to applying the column normalization. By doing this additional step the high variability among features in the dataset is eliminated and the variability introduced by the column normalization is much less. The results measured with the synthetic dataset can be seen in Appendix XI.F Data Normalization tests.

Despite this significant difference observed, because the row normalization is not usually considered in similar studies performed with Microarray data, it has been decided to apply both normalization cases (with and without row normalization prior to column normalization) in the study.

2) *Pearson distance*

Pearson correlation coefficient is among the most common measurements used to compare data distributions. First mentioned in [50], its definition is given, as shown in [51], by:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Equation 13 Pearson Correlation Coefficient

Where cov(X, Y) is the covariance, σ_X is the standard deviation of X, μ_X is the mean of X, and E is the expectation. Pearson is bounded [-1, 1] where 1 indicates that both distributions are perfectly correlated and -1 indicates perfect anti-correlation, with 0 meaning both distributions being completely independent.

The existence of ρ just requires the distributions to be bounded, what is our case, despite the high range difference among different samples. ρ is sensitive to outliers so it could be expected to have quite different results for the experiment settings with and without outliers in the study.

From the Pearson Correlation factor, a distance can be calculated as:

$$d_{X,Y} = 1 - \rho_{X,Y}$$

Equation 14 Pearson distance

The obtained variable will be [0, 2] bounded, with value 0 meaning the two distributions are perfectly correlated (not necessarily equal) and 2 meaning the distributions are perfectly uncorrelated.

C. *Pre-processing: Normalization and inversion*

In all cases, the obtained distance matrix will be [0, 1] normalized. The reason for this normalization is driven by several aspects. The normalization is a linear transformation and will not affect relative magnitudes of the distances, the order is preserved. This makes that the change has no effect for the clustering algorithms chosen: MSTKNN and Hierarchical Clustering with Complete Linkage. In the case of Complex Networks, the change does not affect the Modularity calculation, since the pairwise distance is 0 when a sample is compared to itself, matrix distance diagonal, this being the case the r factor will compensate for the normalization as no offset is applied in the normalization.

In addition, as the complex networks method requires weights instead of distances, a weight will be calculated from each distance by just inverting the distance. In this case, by applying the subtraction:

$$w_{X,Y} = 1 - d_{X,Y}$$

Equation 15 Weight definition from distance

Note that in the case of the Pearson distance we are obtaining the initial Pearson Coefficient with a change in scale because of the [0, 1] normalization.

This has been chosen respect to:

$$w_{X,Y} = \frac{1}{d_{X,Y}}$$

Equation 16 Alternate weight definition

because the former maintains the [0, 1] normalization and it doesn’t introduce singularities/outliers when dividing by values close to 0 and maintains a distribution that it’s symmetric.

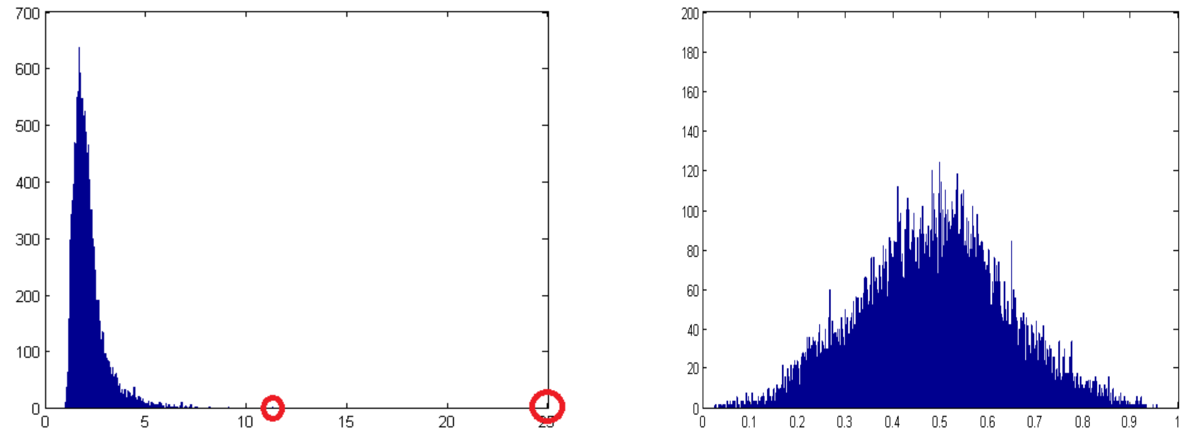


Figure 10 Histogram for inversions proposed (left w=1/d, right w=1-d)

The left figure in Figure 10 shows the histogram obtained for the latter option. Note the two red circled outliers. On the right, the histogram corresponding to the former option, that shows a (more) symmetric distribution and a Gaussian-like shape.

The data distribution for the distance matrixes obtained for all the settings used in the study can be seen in Table 3. The bar at 0.0 corresponds to the diagonal of the matrix. The Pearson cases show a less symmetric distribution.

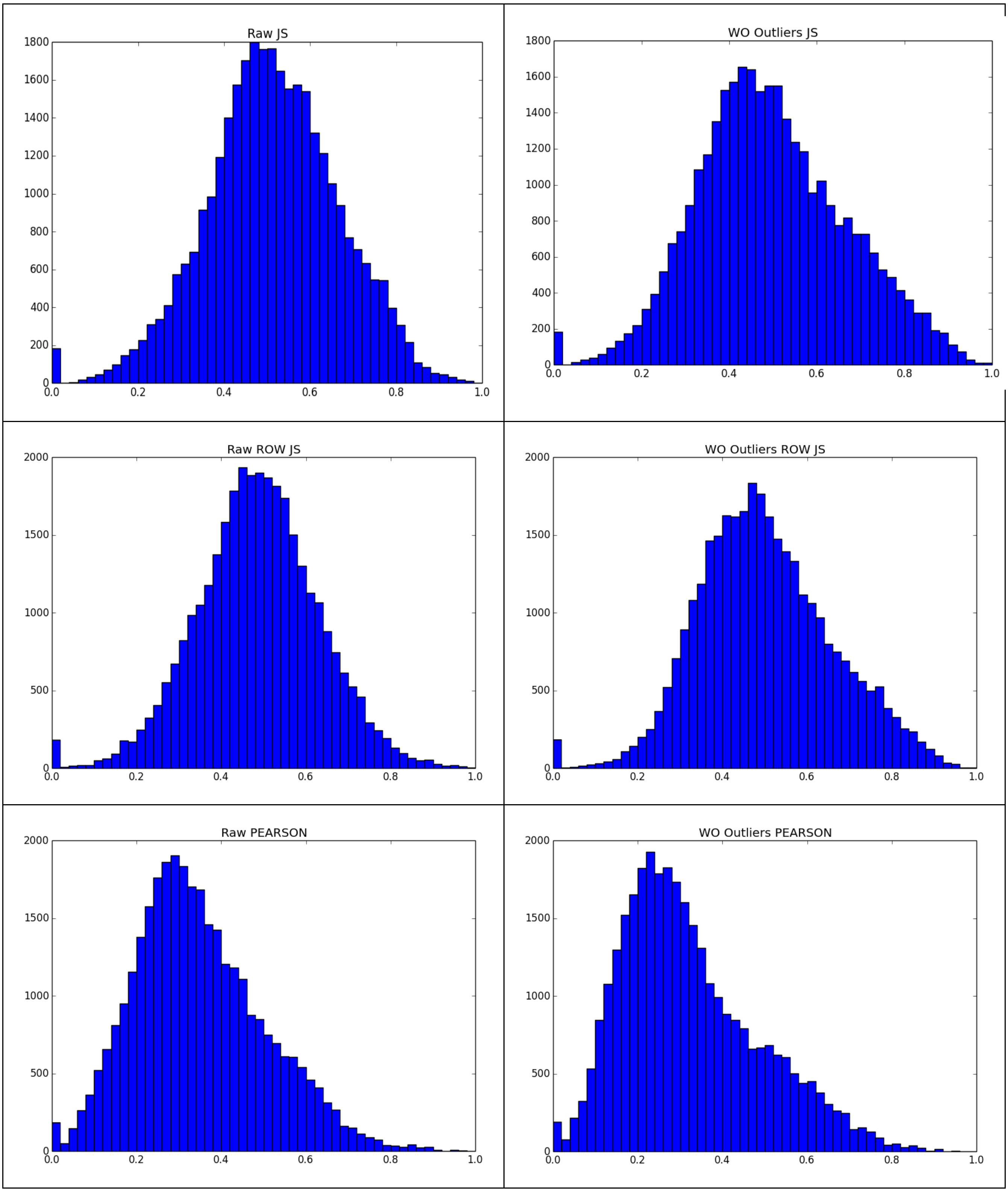


Table 3 Histogram of distance matrixes

D. *Pre-processing: Resampling*

Because we need our results to be relevant it is required to be able to perform multiple executions of our algorithms. Some papers, including [1, 25], perform subsampling based on selection of features but since our objective is the detection of the relevant features for disease diagnosis that would become a source of variability and potential loss of interpretability since we are aiming for a very reduced number of genes to explain the different classes in the dataset. In general, though, the decision of subsampling features is arguable since it modifies the description of the samples and hence the problem being solved.

Typically, when the number of samples is small, Bootstrapping can be used in order to generate a larger dataset. As the methods used are based on distances, adding a copy of the same element does not increase the dataset as the added elements would be identical and the distances are equal to 0 respect to the cloned samples. Then, it is not an alternative in this case.

We need then to subsample our dataset in the traditional way, taking a subset of the samples. The size of our subsampling instances has been decided to be set to 160, leaving out 20 samples on every iteration, roughly 10% of the total dataset.

The subsampling is intuitively the more effective way of having different nodes networks since it removes samples from the complete original network. This modifies/moves the hubs present in the network and really permits the clustering to generate partitions that can be very different. In particular, for MSTKNN the resampling generates partitions with very different number of clusters. The effect is also significant in the case of Hierarchical Clustering, as the Complete Linkage is used. In both methods, as part of the calculation, the most distant elements become the seeds/hubs that generate the clusters, based on the minimum or maximum distance respectively.

E. *Clustering*

For the purpose of the study, three clustering methods have been chosen. The three methods are based on dimensionality reduction and from the feature space the pairwise distance among all samples is used instead. All the methods chosen are non-parametric. Other than that they have different characteristics. Table 4 summarizes the characteristics of the different algorithms. “Stochastic” refers to the algorithm generating different results for the same data. “Edges” refers to the meaning of the connections between samples that could be distances or weights. “K” indicates if the algorithm generates one partition or a series of them.

These algorithms have been chosen because having different characteristics they were candidates to generate discrepant partitions. Any algorithm could be included in the study if the proper adjustments are included in order to weight the different partitions obtained when the algorithm is executed several times for different subsets of the datasets. In this case, all the algorithms implemented are based on pairwise comparisons but algorithms working directly in the feature space could also be included.

Algorithm	**Multi-resolution Complex Networks**	**Hierarchical Clustering Complete Linkage**	**MST-KNN**
Stochastic	Yes	No	No
Edges	Weight	Distance	Distance
K	Multiple, non-nested	Multiple, nested	Unique, automatic

Table 4 Comparison of the different clustering algorithms

1) *Hierarchical clustering*

Based on the linkage methods described, the preliminary exploration of linkage methods has been restricted to the first group, more computationally affordable. An execution of the methods for the complete GSE4290 dataset produces the outcome in Figure 11.

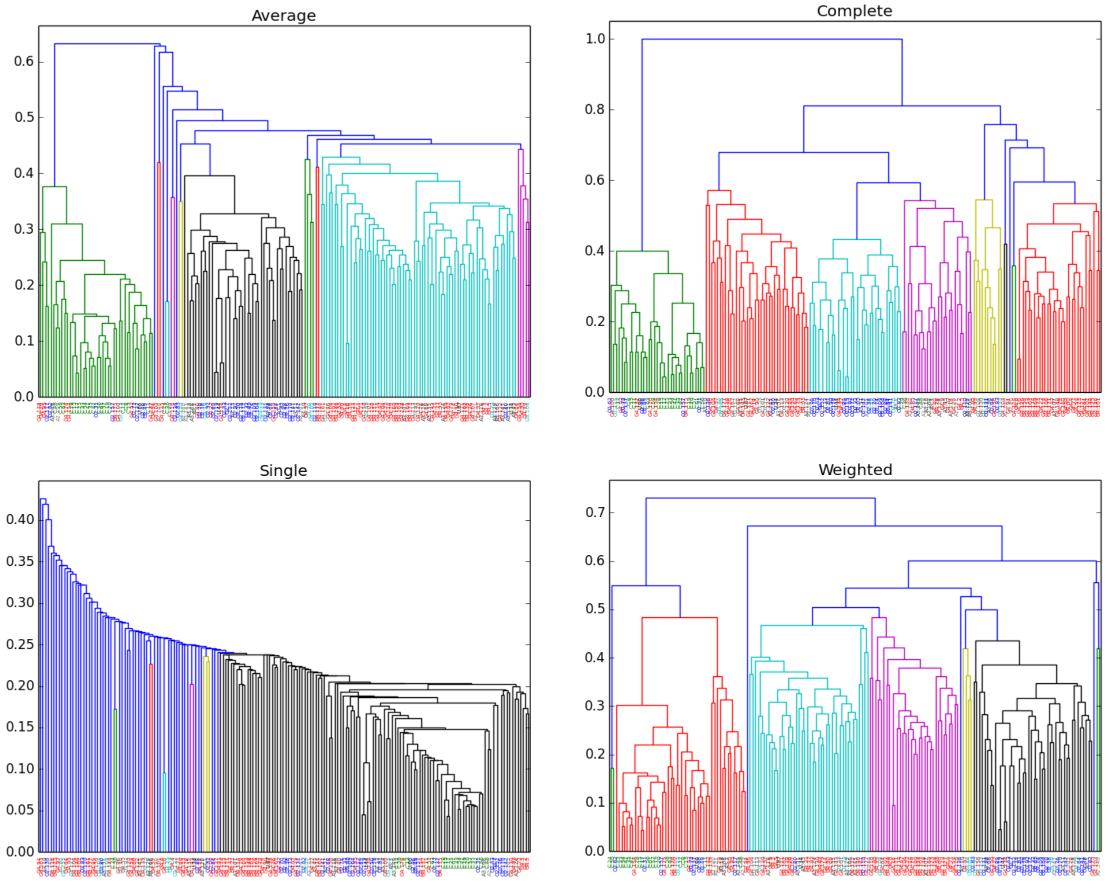


Figure 11 Dendrograms of the GSE4290 dataset for different HC linkages

The differences among the different methods are significant. Single linkage produces very small clusters that are aggregated almost in a 1 by 1 basis. Although the colors in the x-axis show that the method can identify some meaningful structures, see the green and red labels concentrations, it is difficult to find a threshold cut that generates a partition not dominated by clusters with 1 sample only. Complete linkage is the one with less clusters of small size and that even with a cut threshold lower than the other methods, implying the partition is more relevant as differences among clusters are larger.

From the graphs and based on its definition, it is immediate the agreement with chapter 15 in [26], where the author states that both Single and Complete linkage are invariant under monotonic transformations of the distance matrix. This is not the case for Average as the same author defines neither it is for Weighted Average. The Complete Linkage has another advantage, the height of the dendrogram is fixed and equal to the maximum distance in the dataset, that is 1 for the whole dataset. Making an analogy with supervised learning, the fact that Complete Linkage is based on the maximum function, makes more difficult the existence of false positives, that is, samples that are very different to be put together.

The Complete Linkage result used to decide shows three GB groups, what doesn’t confirm the hypothesis that the experts have about the dataset. If a higher threshold cut is specified, less clusters, then two of the GB groups are mixed with a group combining A and O samples.

While this early design decision may, at a first glance, seem opposed to one of the methodological guidelines stated in [19]: “Don’t select the clustering method that gives the best result; class discovery should not be result driven.”, it is not our driver. To this respect, the result obtained by Weighted Average would be a better bias as it has the same four main groups that the final solution obtained shows. The Complete Linkage seems is a feasible option given that other options are not useful for the study based on characteristics not related to the results.

Also the decision is based on observation of the dendrograms for one of the settings in the experiment while 6 different scenarios are considered in the study.

2) *Complex Networks based, modularity optimization*

The clustering method based on Complex Networks, requires significant computing resources even for small or medium size datasets like GSE4290. An exact result based on exhaustive search is only feasible for small and/or very sparse networks. Therefore, the algorithms used must be based on stochastic search. The resolution method chosen will be a sequence of algorithms all of them devoted to the optimization of the network modularity.

The high cost is due to the fact that the algorithm is run once for each value of r in the interval selected. The interval is initially set to cover the whole scale of resolutions, meaning that the dataset will be split in partitions with a number of clusters from 1 to the total number of samples. The resolution of this scanning is a parameter of the algorithm as can be the r minimum and maximum. If the r range is not specified the algorithm will find them. The r_{min} (the dataset clusters in 1 partition) is approximate while the r_{max} (the dataset clusters in partitions with 1 sample) is exact. The algorithm then finds the best clustering for each of the networks resulting from adding a self-loop of weight r_i to the network where r_i is given by:

$$r_i = r_{min} + i * \frac{r_{max} - r_{min}}{N\ samples} \qquad i = 0 \ldots N\ samples$$

Table 5 How r is obtained for each sample in the scanned range

The result of each iteration will be a unique partition made up of a certain number of clusters. As previously mentioned, while in theory, the number of clusters should be increasing, the stochasticity of the algorithm may cause that there are points in which it decreases as r is increased. Figure 12 shows the evolution of the number of clusters for an execution of the algorithm. As an example, note that around r=45 the number of clusters decreases.

Partitions for different values of r may be different even if they have the same number of clusters. Plateaus in the graph indicate the dataset stably partitions in that amount of clusters but not necessarily in the same way.

As obtaining a high resolution on the whole range may be an expensive endeavor, it has been decided to limit the exploration to partitions with k<30. For the interval, 101 equally spaced executions have been requested for each of the scenarios considered. As for the other methods, the initial dataset has been subsampled to create 100 subsamples of 160 samples each.

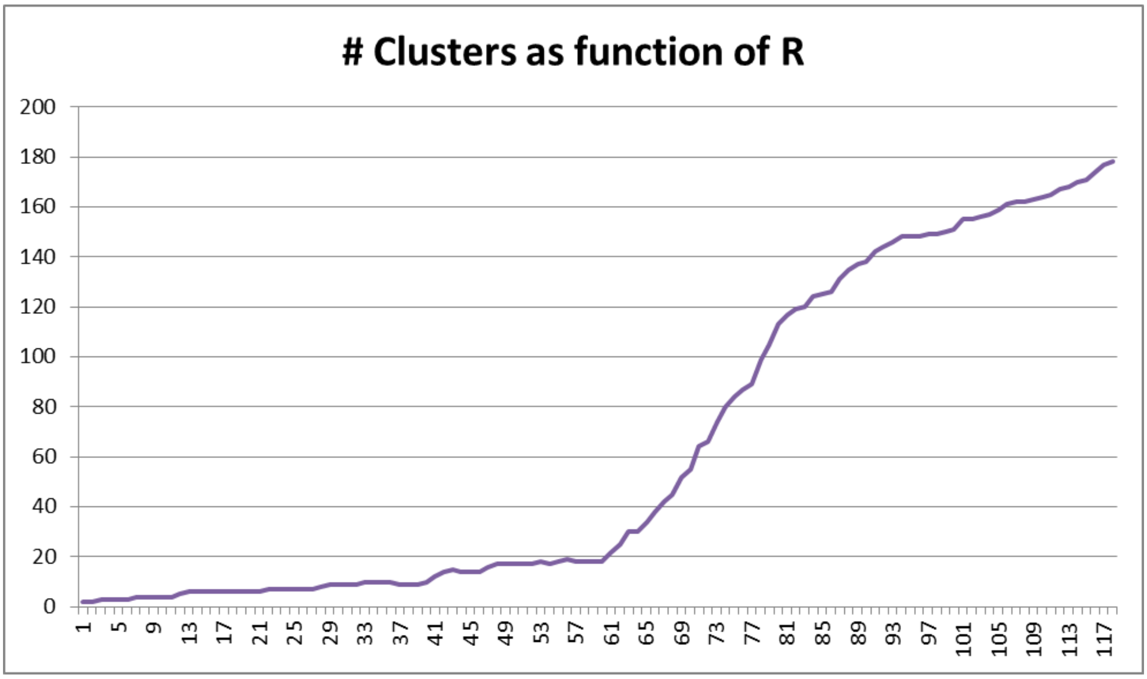


Figure 12 Evolution of number of clusters for GSE4290 dataset

The problem then remains how to specify the $r_{max.}$ to be used in each case without paying a tremendous computing effort. Since the r range is dependent on the sub-dataset considered, ideally, it would be required to study the r-k relation for each case.

Since this is unaffordable, for each of the experiment scenarios, 5 sub-datasets have been randomly chosen and for each of them an execution of the algorithm has been run for the whole r range. From it, the minimum r such that $k(r) \geq 30$ for the 5 cases has been chosen as r_{max} for all the executions.

The choice of k=30 is driven by the analysis of Figure 12, note that around r=61 we have k=20 and this is the "tipping point" for a much faster increase of k. This indicates that any important partition will happen to the left of the graph. A safety margin has been added and then, instead of k=20, k=30 has been set for the executions performed. In addition, for k=20 or 30 the average size of the clusters in the partition will be 9 or correspondingly 6, what is also below the resolution limit of the problem and unlikely to provide any meaningful clusters.

In summary, the sub-datasets generated for each experiment have been screened in the r interval expected to extract partitions with k in the range 1 to 30.

The software package Radatools[52] has been used to perform the required executions for the study. The algorithm permits many different implementations that must be chosen based on the size and characteristics of the dataset. For this study, the number of nodes, 160, is too high for exact algorithms and the method used is a combination of 3 different algorithms combined in a sequence of 4 steps:

- *Tabu search[53]*, consists on the move among existing clusters or segregation to a new one preventing, in order to constrain the number of options, the same nodes are moved repeatedly or reversed for a number of moves. The tabu constraint is not strict and tabu moves are allowed if they generate a solution better than the best generated so far.
- *Reposition[54]*, Kernighan-Lin algorithm, as described in [55], is based on the swapping of pairs of nodes to improve the modularity. As in the previous step, moves are locked for a period of time.
- *Newman fast algorithm[56]*, is in its conception an agglomerative Hierarchical algorithm but in the pipeline it takes the partition generated on the previous step and try to improve the modularity by merging pairs of the existing communities.
- *Reposition*, same than step 2.

3) *MSTKNN*

This method generates a unique partition for each execution and hence is the simpler one to incorporate. For each execution on one subsample a partition with an unpredictable number of clusters will be generated. As an example, the distribution k for a round of 200 executions of the algorithm is shown in Figure 13. No assumption is made about the distribution of k or the similarity of the different partitions with the same k.

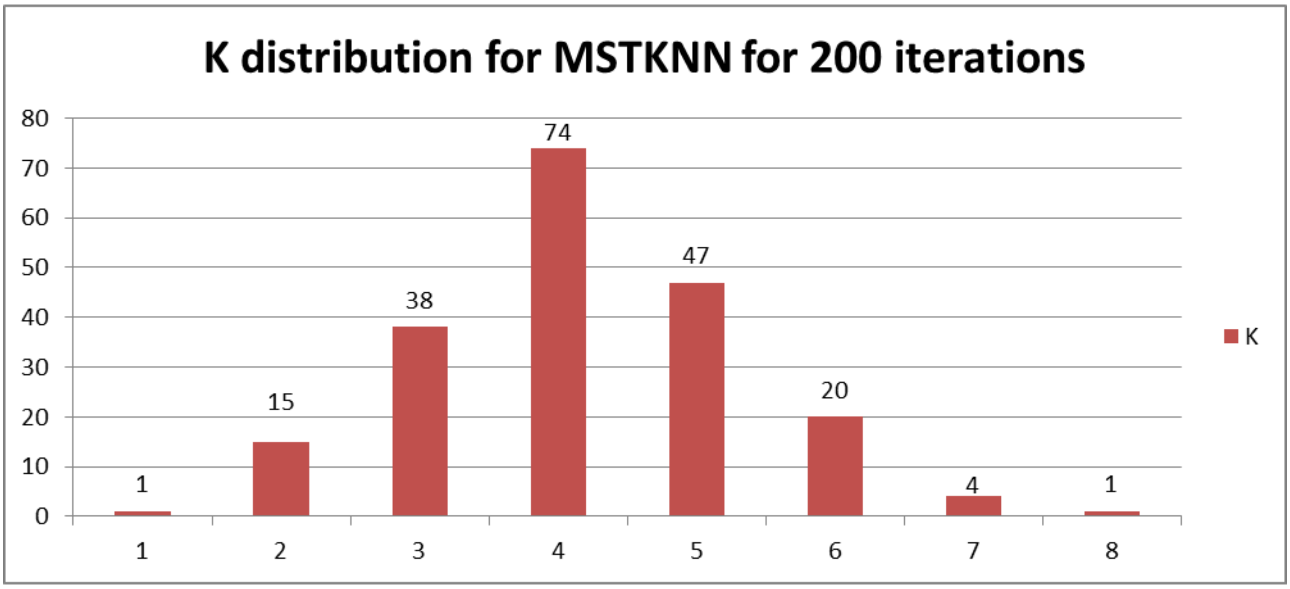


Figure 13 Distribution of k for 200 subsampled executions of MSTKNN

F. *Intra-method consensus*

1) *General process*

The idea behind consensus clustering, as explained in IV.D, is obtaining an average partition for each possible number of classes, k. For this, the algorithm chosen is executed a number of times and the partitions obtained merged. In [1], all the partitions receive the same credit what is not realistic since some clusters appear in a more natural way. The authors used k-means and HC. For the former, k must be provided as input to the algorithm and this is done blindly so the authors executed the same number of executions for all k's. For the latter, the dendrogram was cut in the different partitions without retaining information about the cophenetic distance differences among consecutive partitions or any other information. The whole process relies then in the repeatability of the partitions at a certain k. There is no influence of the representability of the partition. The method proposed enhances the original method by assigning to each partition a weight obtained from the own method and that considers then the importance of the partition in the set of solutions obtained. The following subsections will clarify how the weights are obtained and what they represent for each of the algorithms considered.

2) *MSTKNN*

In order to obtain a consensus cluster for the MSTKNN a number of executions have been performed over a subsample of the dataset. As explained, the MSTKNN will generate a unique partition with a given number of clusters, K.

Based on the consensus method explained in the State of the Art section, all the partitions with the same K will be added in a consensus matrix. The consensus matrix will have for each position i, j in the matrix the ratio number of times in the same cluster-number of times in the subsample.

As the algorithm generates partitions with a range of K's, each of the consensus matrixes is created from a different number of generated partitions.

Each of the consensus matrixes obtained will then be weighted based on the ratio of number of partitions in the set divided by the maximum number of partitions for the same k. Taking as example the k distribution in Figure 13, k=4 will have w=1 and k=5 will have w=47/74.

Despite the unbalance of the weights, it is not the case that the k with more results is the one that generates the larger ΔAUC. The reason is that the partitions generated by the algorithm are very heterogeneous and having more of them does not guarantee a better consensus.

3) *Hierarchical clustering*

In the case of hierarchical clustering, the result obtained from an execution is a hierarchy of clusters. As in the case of MSTKNN a consensus matrix will be generated for each K. As in the case of the Complex Networks algorithm, there is no interest on partitions with K>30, so the range will be restricted to this range.

In this case, all the consensus matrixes will have the same number of components as every execution will contribute one partition to each consensus matrix. While in the case of ties in distances that condition wouldn't have been true[57], the case has not occurred in the experiments, but it wouldn't have any impact if it occurred.

As an enhancement to the original algorithm, each of the partitions when added is weighted in order to consider the stability of the cluster in the hierarchy. The weight assigned is the difference among the dendrogram height (known as cophenetic distance) for the current partition, k, and partition k-1. This modulates the partition space giving more significance to partitions that keep apart clusters whose points are more separated.

As the linkage method used is Complete Linkage and the data is [0, 1] normalized, the total height of the dendrogram will be close to 1. It is not 1 because the subsampling of the dataset may not preserve the [0, 1] normalization.

By weighting this way, each of the consensus matrixes will have a different total weight (the sum of weights of the partitions included on it).The maximum total weight of the consensus matrixes will be used to normalize the weight of each in an analogous way to how it has been done for MSTKNN but in this case each partition instead of contributing a fix weight of 1, contributes by its cophenetic distance. The idea is like considering the cophenetic distance as weight for a partition and normalizing by the sum of all the dendrograms and then making sure that the consensus matrix with maximum value is multiplied so that the maximum value is 1. Then apply the same factor to all the consensus matrixes.

4) *Complex Networks, modularity optimization*

Every execution of the algorithm generates a series of partitions with different number of clusters and often more than one partition will have the same number of clusters. Each of the generated partitions is the one that has been found to have a better modularity for an r interval. As in the case of HC, the partition has a weight, but differently there is more than one partition on each solution. Also, in this case, there is no guarantee that the solution generates partitions for all the number of clusters as the resolution used may not be enough to extract them. In summary every solution will be then a sequence of partitions where no assumption can be made about the k of each of them and where each partition is the best partition for an r interval. The sum of all the r intervals is equal to the whole r range scanned. Each partition will be then assigned a percentage of the whole r.

The process defined for HC is then valid also for this other algorithm by assimilating the cophenetic distance to the r interval of the partition.

G. *Selection of the two more different partitions*

As a result of the previous steps, and calculating independently for each, the delta of AUC for the different k's, we will obtain for each clustering algorithm (and settings) a k that is the consensus partition. The result obtained when applying the consensus, is shown in XI.B.6) for a run of 10 executions and 200 iterations. With this, a partition will be generated for each of the algorithms, as can be seen, for different number of iterations in Appendix XI.B.6).

As an example in Table 6 the best partitions for an execution (central column) are shown together side by side with the consensus partitions for k-1 and k+1. Blue lines in the central column indicate the partitions. The order of the samples has been preserved in the other cases so that the difference in the cleanliness of the partitions can be observed in addition to the Membership coefficient. Note that a better membership coefficient does not imply the AUC criteria will select that k, also the limitation of k>2 plays a role.

The current step will receive as input the partitions in the central column and the two more different will be chosen to be combined in the following step. Since the similarity of the partitions can be measured in many different ways and the way it is measured may be related to how the clustering algorithm works, biasing the result, it is appropriate to use a battery of indicators. The measurements chosen are in Table 7 and its description in Appendix XI.E.

The system then votes how many times a certain pair of partitions is the most different one. The pair that receives more votes is selected. Despite having a number of indicators that is multiple of the number of elements being compared during the experiments no ties have been produced. This is a circumstance that should be paid attention if the method wants to be generalized.

The voting results for each of the settings with N_Iterations=500 can be seen in Table 8. Different settings produce a different "most distant" pair of partitions.

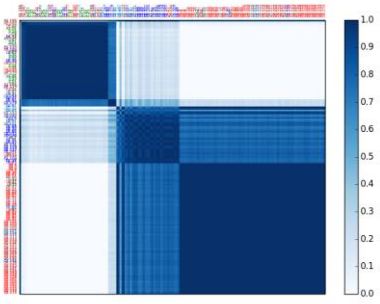
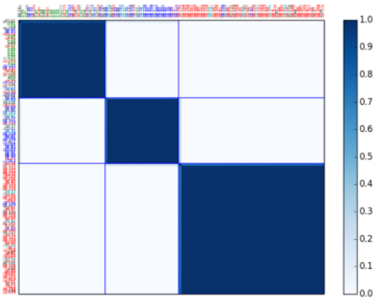
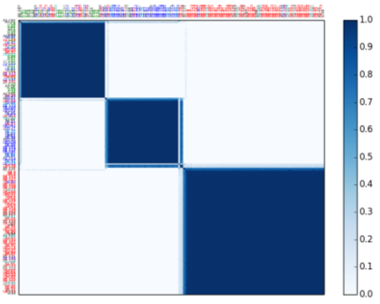
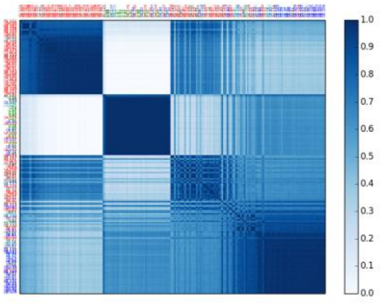
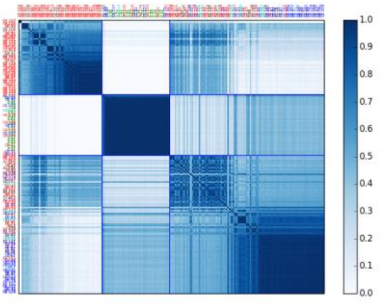
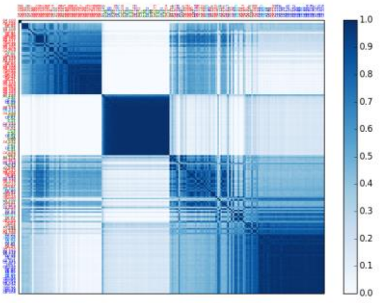
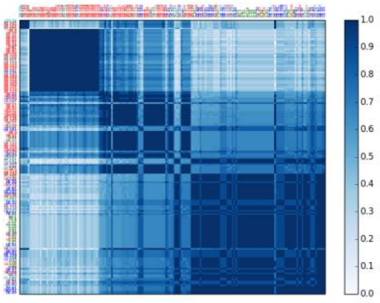
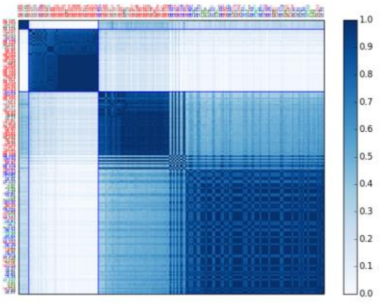
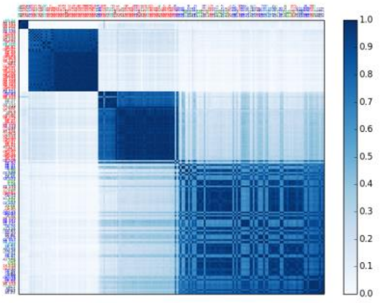
	Best K - 1	Best K	Best K + 1
CN	 M_c=0.92	 M_c=0.98	 M_c=0.95
HC	 M_c=0.90	 M_c=0.81	 M_c=0.77
MSTKNN	 M_c=0.16	 M_c=0.48	 M_c=0.83!!!

Table 6 Comparison of best K (central column) for each algorithm with K±1

Measure	S/D	Type
Same Class Agreements	Similarity	Pairwise distance
Disagreements	Dissimilarity	Pairwise distance
Jaccard Index	Similarity	Pairwise distance
Adjusted Rand Index	Similarity	Pairwise distance
Fowlkes Mallows Index	Similarity	Pairwise distance
Normalized Mutual Information Index (arithmetic)	Similarity	Entropy
Normalized Mirkin Metric	Dissimilarity	Confusion matrix
Normalized Van Dongen Metric	Dissimilarity	Confusion matrix
Normalized Variation of Information Metric	Dissimilarity	Entropy

Table 7 Measures to choose the two more different partitions

		HC-CN	CN-MSTKNN	MSTKNN-HC
Raw	**Column Norm + Jensen-Shannon**	0	94	6
	Row + Column Norm + Jensen-Shannon	3	11	86
	Pearson	0	78	22
WO outliers	**Column Norm + Jensen-Shannon**	1	29	70
	Row + Column Norm + Jensen-Shannon	5	82	13
	Pearson	0	51	49

Table 8 Voting of pair differences

H. *Inter-method consensus*

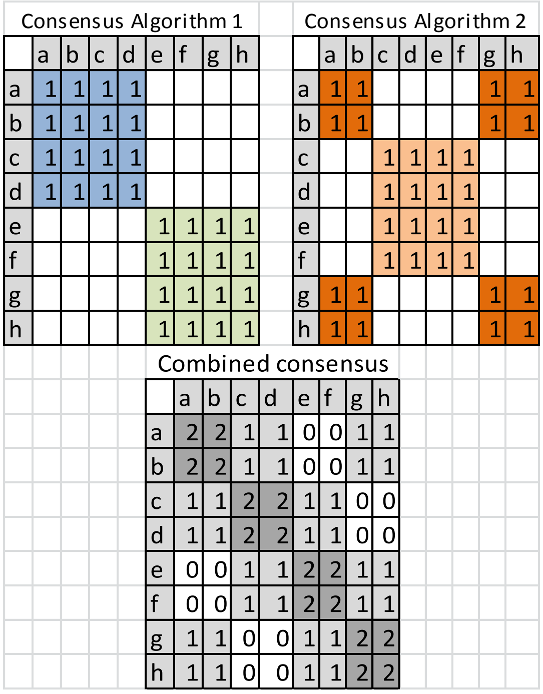
1) *Discussion of existing methods*

In [25], the authors propose the combination of partitions generated by different algorithms in the consensus clustering [1]. They propose to do so without any consideration about the algorithms used or the difference among partitions. The method suggests that the combination of a number of partitions with the same number of clusters will produce also a partition in the same number of clusters. This, as the experiments have revealed, is not always possible, despite applying HC methods if the consensus matrixes have tied weights.

Figure 14 shows an example in which two different algorithms generate perfect consensus matrixes for k=2 but the merge of the two matrixes cannot be clustered in a partition with k=2 without arbitrarily breaking ties. This situation cannot be avoided since the partitions are initially grouped based on the number of clusters they contain. Paradoxically, the more robust the independent algorithm consensus matrixes are, the more likely is that the combination of them cannot be separated unambiguously.

While is true that the case described is hiding that k is not the best option for the dataset, and in the example, k=4 is a much better option, the situation can also occur for the best k for the data when the two algorithms do not extract the same groups. Consider also that imposing a non-appropriate k to an algorithm does not mean the algorithm will generate more heterogeneous partitions; this will depend on the algorithm itself and the initialization parameters. While a properly designed algorithm should be able to behave inconsistently (generating different partitions) when k is not the natural for the dataset, and reducing AUC for k, this idea should be kept on mind when analyzing the results.

Coming back to the main thread, when the original k is not suitable for partitioning, it would be required to find the proper k what brings the research to the original problem, deciding the "best k" for clustering the dataset.



Consensus Algorithm 1

	a	b	c	d	e	f	g	h
a	1	1	1	1				
b	1	1	1	1				
c	1	1	1	1				
d	1	1	1	1				
e					1	1	1	1
f					1	1	1	1
g					1	1	1	1
h					1	1	1	1

Consensus Algorithm 2

	a	b	c	d	e	f	g	h
a	1	1					1	1
b	1	1					1	1
c			1	1	1	1		
d			1	1	1	1		
e			1	1	1	1		
f			1	1	1	1		
g	1	1					1	1
h	1	1					1	1

Combined consensus

	a	b	c	d	e	f	g	h
a	2	2	1	1	0	0	1	1
b	2	2	1	1	0	0	1	1
c	1	1	2	2	1	1	0	0
d	1	1	2	2	1	1	0	0
e	0	0	1	1	2	2	1	1
f	0	0	1	1	2	2	1	1
g	1	1	0	0	1	1	2	2
h	1	1	0	0	1	1	2	2

Figure 14 Separation not possible in the same K than original

2) *Proposed method for inter-method consensus: Robust clustering*

A more robust, and also conservative case, is robust clustering that permits to merge clustering results without any restrictions. Robust clustering obtains the intersection of the partitions provided

The method could be used for any number of initial partitions (one for each algorithm involved) just by applying it iteratively over the result of the previous iteration. The more partitions used the more micro-clusters will be obtained as the method breaks the entry partitions according to intersection.

The calculation of the intersection is the same that would be done to calculate the confusion matrix among the two partitions generating a cluster for each of the non-empty crossings. The number of clusters this way obtained can be up to the product of the number of clusters in the original partitions.

I. Feature extraction

From our final partition, we need now to extract those genes that explain the difference between clusters.

This can be done in a supervised or unsupervised way depending on whether the ground truth is considered or not.

Considering the ground truth is equivalent to applying the Robust Clustering method to the obtained partition as if the ground truth was the obtained partitioned by another method (what actually it is: the physicians and pathologists method).

It can also be the case that our dataset is only partially labelled and in this case the labelled samples can help to label the clusters based on their distribution.

In order to select the differentiating features, the CM-1 indicator is used. CM-1 can be understood as an indicator, for each feature, of the relative difference among the averages for each class in the dataset. Those features that show a more extreme value, either lower (under-expressed genes) or higher (over-expressed genes), will be the ones that have a higher contribution to explain the partition. CM1 can be calculated in a 1-VS-1 way, that is comparing individual clusters or, in a 1-VS-all way, comparing each individual cluster against all other clusters in the partition.

CM-1 is defined by the expression, as shown in [58]:

$$CM\_1(w,X,Y)=\frac{\frac{1}{|X|}\sum_{x\in X} x_w-\frac{1}{|Y|}\sum_{y\in Y} y_w}{1+max_{y\in Y}\{y_w\}-min_{y\in Y}\{y_w\}}.$$

Equation 17 CM-1

J. Solution analysis

In order to analyze the robustness of the solution, to measure the variability as the experiment is repeated a number of times, it is important to do it from different perspectives. In total three different measures have been done that are classified in two categories: external, where the ground truth is used, and internal, where the measure is done based on information in the own solutions.

In addition, the solution has been compared to a supervised method, decision trees, in order to see if there is any degree of agreement in the features that are obtained by CM1 and the features used to generate the classification tree.

Last but not least, the resulting list of probes will be analyzed using Domain Knowledge in two ways. First, in a quantitative manner by comparing the results obtained with those in other studies using the same dataset and also using public databases on genetic relationship to diseases. Second, by an expert, to be able to benchmark the results with those found relevant in other published studies related with brain tumors.

1) *External robustness*

Based on the concepts of Purity, Homogeneity and Completeness a new indicator has been defined. This indicator is defined from the partition being evaluated and compared with the available ground truth.

Since our ground truth is not only unreliable but also could become a limitation when trying to discover new diseases subtypes, supervised metrics for validating the goodness of the results are not fully appropriate and some modifications need to be introduced.

The definition performs the intersection of the partition obtained with the ground truth obtaining, the confusion matrix of the two partitions.

If the solution obtained for the first set of settings is taken, as shown in Figure 15 and Table 9, for each of the classes in the partition (1 to 6), whose size is above the resolution limit as defined in V Dataset (2, 3, 5 and 6), the majority class is identified.

In [59], some desirable objectives for cluster assignment are defined: (1) homogeneity, each cluster only contains members of a single class; (2) completeness, members of a given class are clustered in the same class. The formulas are explained in XI.E.10)

	1	**2**	**3**	**4**	**5**	**6**	**J**
C		23					0
GB		10	3	2	33	29	15
A	1	7	8	2	3	5	-
O		9	32	1	4	4	18
-		1			1	2	-
Total	-	50	43	-	41	40	

Table 9 Confusion Matrix for solution partition, including J factor used in EP

Based on this idea, two measures are calculated for each of the sub-clusters identified. The definition of the first one, reminiscent of Purity, and that we will call intra-Purity, is given by:

$$IP(C_i) = \frac{1}{|C_i|} max_j \, |C_{ij}|$$

$$IP(C) = average_{i\,:\,|C_i|>l} \, IP(C_i)$$

Equation 18 Intra-Purity

that represents the average of elements that are in the same class that the majority class of each cluster in the solution partition intersected with the ground truth.

The second measurement, called extra-Purity, is given, when classes in the ground truth are preserved, by:

$$EP(C_i) = max_j \, \frac{1}{\sum_k C_{kj}} \, | \, C_{ij} |$$

$$EP(C) = average_{i\,:\,|C_i|>l} \, EP(C_i)$$

Equation 19 Extra-Purity

In the general case, of having a class split in two or more clusters being majoritarian the definition becomes a bit more cumbersome:

$$J_k = \left\{ j : j \notin \ argmax \ _j \left| \, C_{jk} \, \right| \right\}$$

that is, the subclusters of the ground truth class k that are not majoritarian in their respective clusters in C. Using as example the confusion matrix in Table 9. For class G, J_G would be: 2, 3, 4 and from it the addition of corresponding cardinalities 10+3+2=15.

Then EP would be defined as:

$$EP(C_i) = max_j \, \frac{| \, C_{ij} |}{\left| C_{ij} \right| + \sum_{k:J_j} |C_{kj}|}$$

Equation 20 extra-Purity per cluster

where from the denominator we are excluding the samples belonging to the same cluster in the ground truth that belong to another majoritarian sub-cluster.

$$EP(C) = average_{i\,:\,|C_i|>l} \, EP(C_i)$$

Equation 21 Extra-Purity general case

An example, based on Table 9, will help us to clarify the definition. For IP, we obtain:

$$IP(C) = \frac{1}{4} * \left(\frac{23}{50} + \frac{32}{43} + \frac{33}{41} + \frac{29}{40}\right) = 0.6835$$

$$EP(C) = \frac{1}{4} * \left(\frac{23}{23+0} + \frac{32}{32+18} + \frac{33}{33+15} + \frac{29}{29+15}\right) = 0.746$$

$$P(C) = IP(C) + EP(C) = 0.6835 + 0.746 = 1.4295$$

Equation 22 Numerical example of the calculation of P, IP and EP

With this definition we are introducing two biases in the calculation of EP when for the same class in the ground truth there are two clusters where the class is majoritarian, GB case, corresponding to the scenario of discovery of a new subclass not represented in the ground truth. The first bias is that we are not considering classification errors between the two classes, we are then favoring the measure as our ratio will be higher or equal than the real. The second bias is that any number of samples in clusters where its class is not majoritarian will punish all clusters where it is majoritarian. In this case, we are calculating a strictly lower measure than the real. On the other hand, the small clusters with size below resolution limit are not considered for IP measure and only the classes that become majoritarian in one cluster of the solution partition are considered. Also the Unknown samples, (-) in the table, will count as incorrect when calculating IP. As the measurements start from the majoritarian clusters, having an unbalanced dataset favors that the classes with a higher number of samples become majoritarian more easily.

2) *Internal robustness*

To measure the robustness of the solutions two different approaches can be taken. The first approach consists on measuring the similarity of the different solutions obtained among them pairwise. Note that typically this is an external method because the ground truth is used, to measure accuracy. In this case, robustness is being measured using the different solutions generated, this is why it is being considered an internal method. As in the case of choosing the two more different partitions obtained from the individual clustering methods, the same battery of indicators has been chosen. As in each experiment 100 executions are run, the pairwise comparison will generate 100x99/2=4950 comparisons. The different indicators will be plotted separately in order to view the convergence of the results. Because the different indicators have different scales they have been separated in two groups when the results are shown.

The second method measures the convergence of each of the algorithms. This is done based on the average of the membership coefficients for each cluster, Equation 7, of the consensus matrixes generated. Then, for each set of settings we will obtain three distributions, one for each of the clustering algorithms. This will help to know where the variability originates in the whole process and can be used to modify the selection of clustering methods.

3) *Clustering vs Decision Tree*

In order to compare the two methods and given their different orientation the comparison has been made based on the features selected on the decision tree compared to the features extracted from the CM1 experiment.

For the Decision Tree, two different implementations have been chosen for two different tests.

The first implementation is bigML®, a commercial web service, that generates decision tree from datasets provided by the user. The bigML® implementation is based on CART decision trees but with modifications on the implementation in order to be able to deal with large volume data streaming based on algorithms by Tyree[60] and Ben-Haim[61], as explained by bigML® representatives, with whom I have been in contact in order to solve issues in their implementation that prevented the execution of the dataset.

The test done has consisted on generating the decision trees for the six experiments in the study and see if the features used in the decision tree have any matching with the features extracted from the CM1 feature extraction. In this case, the ground truth provided consists on the classification of the patients on 4 types (C, A, O, GB).

The second implementation is Python scikit-learn, that is based on an optimized version of the CART algorithm, no details are provided in the documentation. In this case, the samples provided where the clusters obtained as result from the study. The ground truth was the assignment to each cluster.

VII. EXPERIMENTS

In order to study the dataset object of this work, a series of experiments have been executed. The design of the experiments corresponds to the different options described in the data pre-processing and normalization, namely: filter or not outliers, use sqrtJSD or Pearson distance and if sqrtJSD is used normalize by column or not.

Parameters Set	**1**	**2**	**3**	**4**	**5**	**6**
Distance	Jensen-Shannon	Jensen-Shannon	Pearson	Jensen-Shannon	Jensen-Shannon	Pearson
Normalization	Column	Row and Column	n/a	Column	Row and Column	n/a
Outliers removed	No	No	No	MAD>5	MAD>5	MAD>5
Iterations	100 to 500	100 to 500	100 to 500	100 to 500	100 to 500	100 to 500
Repetitions	100	100	100	100	100	100
Label in figures	Raw-JS	Raw-Row JS	Raw-Pearson	WO-Outliers-JS	WO-Outliers-ROW JS	WO Outliers-Pearson

Table 10 Parameter Set Scenarios

The workflow formerly described in the former section has been executed 100 times and the results obtained have been aggregated in order to evaluate the robustness of the method proposed.

These experiments, as explained in E.2), have been limited due to limitations in computational resources. The time required for executing the complex networks method is very high, above 1 hour for each execution (for our parameters, as explained above). A total of more than 10 full days of execution are required to execute 100 subsamples with the 6 different sets of parameters used. Then the rest of the processing (execution of the other clustering methods and consensus) is required.

This limitation has been translated in two main aspects; first, only one dataset has been used in the study what does not permit to grant to the method any range of utility outside the current dataset.

Second, because of the limitations regarding the execution of the clustering based on complex networks modularity, 100 executions have been made and those same have been used for the different iterations by choosing 80 to be combined. Despite this way of calculating, given the nature of the method, the repetitiveness of the method is not compromised for two reasons. First, the subsamples of the other methods are calculated independently and the method to choose the partitions to be combined is based on distances so our process is, in the worst case, equivalent to fix one of the parameters of the algorithm (one of the partitions). Second, this method is based on the measure of modularity, and despite being calculated using stochastic methods, the results obtained are highly robust and therefore generating the subsamples for each iteration, instead of choosing among a set of generated ones, will not change the result. The membership coefficients obtained for the method confirm this reasoning, see Table 34.

The number of iterations performed in the experiments has been in the range 100 to 500 with increments of 100 iterations. Each experiment has been executed 100 times in order to measure the variability of the solutions obtained.

In addition, the CM1 features selected have been compared with the ones used in two decision trees.

Finally, the results have been analyzed by an expert on bio-markers in order to assess the relevance of the clusters found in addition to a quantitative analysis of the probesets to genes association.

VIII. RESULTS

Figure 14, shows the final solution obtained for the RAW-JS settings, the only configuration that shows convergence in the result generated. With 200 iterations in the consensus generation, this configuration can produce a solution that is very consistent across multiple repetitions of the execution, Table 32 and Figure 25. Note that the clusters may be shown in different order.

In Table 22,Table 23,Table 25,Table 26,Table 28,Table 29 and Table 32, the results for each of the configurations and different number of iterations is shown. In all cases, the result shown corresponds to the one with higher Purity as calculated for the study. In the case or RAW-JS, the method is converging at the higher Purity so this method is not biasing the result as we are analyzing the RAW-JS in parts of the analysis where this could be relevant. A different situation would have been if the convergence is not at the maximum Purity level, then the Median should have been taken, note that the Mean value may not correspond to any of the solutions generated.

The different result partitions generate though different qualitative aspects and not all of them divide the dataset in the same way. Taking the case where N_Iterations=500, Table 28 and Table 29, the result for each setting produces the following partitions:

	RAW	***WO-Outliers***
JS	1 Control group (high Completeness) 2 GB groups (high Homogeneity) 1 Oligoblastoma group 2 minor groups	1 Control group 2 GB groups (high Homogeneity) 1 Oligoblastoma group 1 minor group
ROW-JS	1 Control group (high Completeness) 2 GB groups (unbalanced) 1 Oligoblastoma group 2 minor groups	1 control group (high Completeness) 2 GB groups (high Homogeneity) 1 Oligoblastoma 1 minor group
Pearson	1 Control group 1 GB group (small with high homogeneity) 1 GB group (large but with low Homogeneity) 1 Oligoblastoma group (small) 2 small groups	1 control group 2 GB groups (1 high Homogeneity) 1 Oligo (small, low Completeness, high Homogeneity) 3 minor groups

Table 11 Qualitative analysis of the best result for each configuration

Many of the results agree on the existence of two GB groups although there is disagreement about the relative size of them. The best result, higher Purity, as can be expected from the qualitative description, occurs for the WO Outliers-ROW JS, but this value is an outlier in the opposite direction to how the convergence seems to be occurring. The convergent RAW-JS generates the fourth best result across configurations, very tied with RAW-ROW JS and WO Outliers-JS.

The number of iterations used in the algorithm is a sensitive parameter in order to obtain a consensus partition from the algorithm. In the explored ranged, from 100 to 500, only the first settings have achieved absolute convergence and it can be considered that the method generates a unique result (and a bunch of outliers), see Figure 23 and Figure 24.

An increase on the number of iterations only produces a noticeable improvement in the convergence of the results for some of the settings used Raw-JS and WO Outliers-ROW JS. In a lesser degree, also for Raw-Pearson there is an increase on the convergence of the solution partition but not enough to consider the method generates a unique solution.

The convergence is not only at the level of the final result but also at the individual clustering algorithms considered in the study. The Complex Networks modularity based method is by far the most stable method of the three used, with convergence at membership coefficient and partition generated. MSTKNN, generates the wider membership coefficient range, see Table 34. But the partitions generated show the same degree of convergence than Complex Networks, Table 39, and create the conditions for the method to converge: the two methods that generate the more different solutions converge separately for a number of iterations around 500.

The changes introduced in the existing algorithms, weighting the partitions, have a positive impact on the stability of the partitions generated without biasing the results to the number of partitions that appears more often (MSTKNN), has a larger r range(CN modularity) or larger cophenetic distance difference (HC), see Table 6 and Figure 13 as an example.

About the effect of normalization, column normalization or row and column normalization, only the first has generated a convergent solution and this only for the complete dataset (including outliers). This may indicate that this extra normalization by row is getting rid of part of the discriminating information required to consensuate a partition. Jensen-Shannon is then measuring not only the expression level of genes among samples but the difference across samples and genes, first term in JSD, Equation 10.

The fact that when the outliers are removed, the same convergence level is not achieved also indicates that the information on the outliers is required for the consensus and that despite the data distribution being wide this information is necessary. When the outliers are removed, the pairwise distances become more homogeneous and while the partitions obtained are still meaningful, the consensus is not achieved. The redundancy in the Microarray Data is not enough to compensate for the aggressive removal of features.

The Pearson correlation is not able to converge for none of the two datasets. This circumstance indicates that in order to measure the distance among samples is not only important the correlation among samples, as Pearson measures, but also the variation among genes in the samples and while Pearson considers σ_x of the distributions, the wide range of values in the gene expression matrix may not be able to properly represent it. JSD based on entropy, and using logarithms, is a better option.

The decision of using the two more different individual algorithm solutions, is not biasing the result as for different settings, the pair that receives the more votes is different, example Table 8. MSTKNN is always one of the algorithms involved.

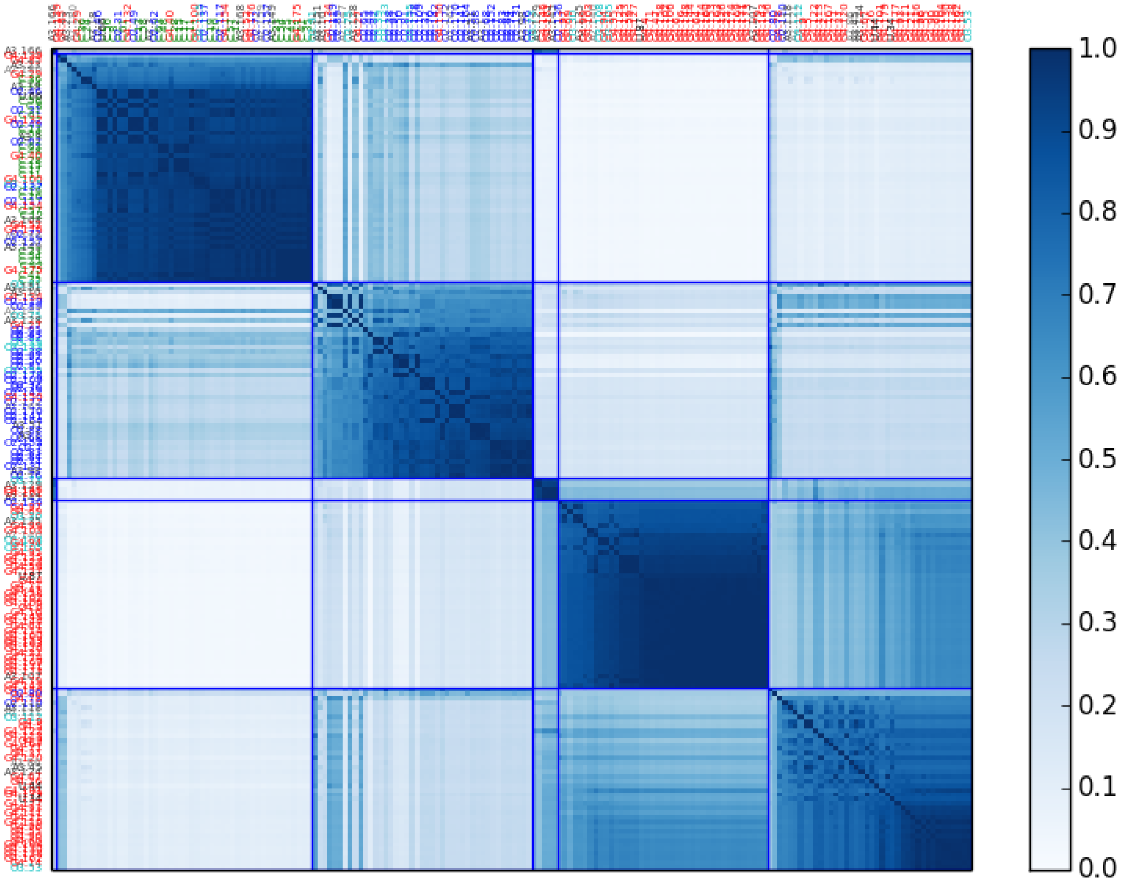


Figure 15 Consensus partition result for RAW-JS settings

The main question to be answered is how the consensus can be reached when one of the two components (MSTKNN) has such important variability, in the membership coefficient.

To verify that, analogously to how the final solution robustness has been assessed, the solutions generated for each of the methods have been compared pairwise, Table 39, there it can be observed that for Complex Networks algorithm and MSTKNN, for 500 iterations in the consensus, there is absolute convergence of the solution generated. This does not happen for Hierarchical Clustering that maintains high variability, as could be expected from the comments in IV.C.1)c) Disadvantages of Hierarchical Clustering. For the WO Outliers-Row JS configuration, the second closer to convergence, only the CN algorithm converges, Table 40.

The reason then for the high variability of the membership coefficient comes, because of the weighting system, not only from the variability of the consensus matrix that will become the individual solution, best K, also the other consensus matrixes have an influence as the weights change. To confirm this, we need to check how the best K is changing, what can be seen in Table 38; it can be observed that the best K, for Raw JS, is always the same, k=3. This happens even when the curves for different executions is very variable. It can also be observed that this is not the case for other configurations than RAW-JS for MSTKNN.

About the comparison of the selected features with the CM1 indicators, for the experiment using bigML, there has not been any match for any of the 165 unique probes, 1-vs-all, with the features used by bigML. It has been observed that many of the features in the bigML decision trees are used several times in the same tree. The tree has 40 internal nodes and the same feature is used up to 8 times in the tree, only 3 features are used only once. In this case, using the same feature many times can explain that there is no coincidence as the partitions are done in a more precise way. Also in DT, the selection of features is done at every level.

In the case of scikit-learn only one out of seven features, 212187_x_at (PTGDS), has appeared in the CM1 features extracted from the study. In this case, none of the features is repeated in the Decision Tree. What is interesting is that the gene appears in the G1X vs G2X group, down-regulated in second position, and G1X, down-regulated in second position, and in the Decision Tree it separates precisely the G1X and G2X groups (3rd and 4th group in Figure 17).

218618_s_at	FNDC3B	220984_s_at	SLCO5A1	243303_at	ECHDC1
236234_at	PDE1A	225075_at	PDRG1	220947_s_at	TBC1D10B
212187_x_at	PTGDS				

Figure 16 Probesets and genes used in Scikit Decision Tree for the cluster separation

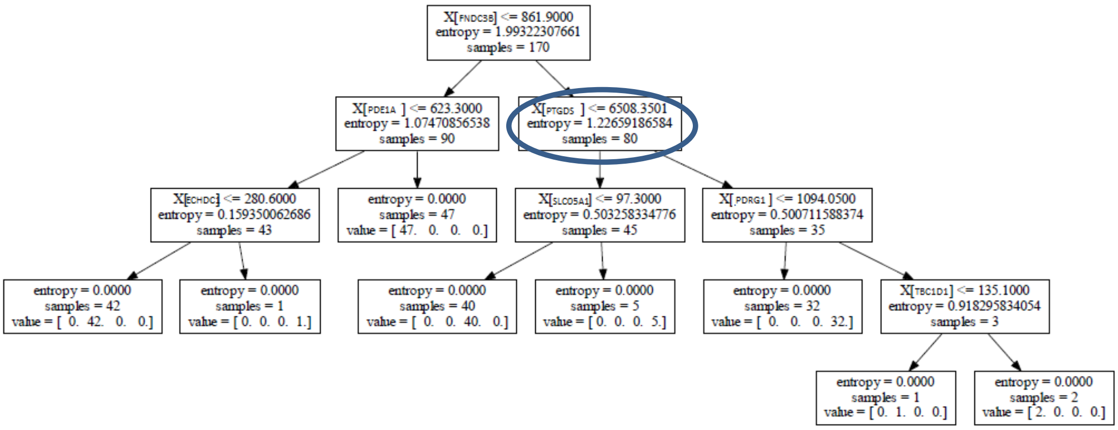


Figure 17 Scikit Decision Tree applied to the cluster solution

IX. DISCUSSION AND CONCLUSIONS

Based on the primary solution, the one with greatest convergence, Raw-JS, two groups with 33 and 29 samples have been partitioned, with only 15 samples being assigned to other clusters and 8 and 11 samples respectively being included in the groups with a different diagnose. The control group (C), the only one that can be granted absolute reliability, has been clearly partitioned although the group includes 27 other samples that are not controls, what is a downside of the solution. The Oligoblastoma group(O) has also been identified, with 32 out of 50 samples in a majoritarian group where only 11 samples do not share this diagnosis. The Astrocytoma group has not been clustered what is according to the low consensus ratio for the diagnose of this tumor subtype and its smaller representation in the dataset.

Based on the results obtained it becomes tempting to try to classify the Unclassified samples in the dataset based on the cluster they belong in the solution partition, and the majoritarian class on it. If that could be accepted, three out of the four unclassified would correspond to GB subtype, what considering that GB is the most reliably classified subtype (see Appendix XI.G) seems unlikely, but possible. The fourth subtype would be a Control sample what is unlikely to appear as unclassified. As stated in [19], this classification process would be biased and what is required is to apply supervised learning techniques that based on the clusters obtained select, without the Unclassified samples, the discriminant features and then the classification could be performed.

The statistical relevance of the results has been evaluated using two different methods. The first, internal, is based on the similarity of the results obtained when the experiments have been repeated several times and complemented by the measurement of the variation of the membership coefficients of the consensus matrixes of each of the clustering methods. The fact that robustness of the solution has been verified using internal methods is a solid indicator of the robustness of the result, as stated in [19]. The tables in section XI.B showing the Agreements and Indexes and Metrics witness this circumstance that is notably visible in Table 30, Table 31 and Figure 24 as well as Figure 25, Table 33 and Table 34.

The second, external, measures the variability of the intra- and extra- Purity metrics designed to consider the discovery of new classes. Both of them show highly consistent behavior in all 6 cases and (almost) absolute convergence for the Raw-JS settings, see Figure 23 and Table 35.

The value is around 1.43 (bounded 0-2), what is in the range of accuracy that physicians have in the best case, 70%, see Appendix XI.G, while the comparison is not strict as the measures used are different although strongly related. This level of accuracy is also in the same interval that other studies that have used the same dataset obtaining accuracies in the range of 40% to 72% depending on how the dataset is used (training or test), see XI.H Known results obtained from dataset GSE4290.

For the RAW-JS solution, the relevance of the genes discriminating found via CM1, have been validated by an expert on biomarkers. The detailed results can be read on

Appendix XI.A. The link to carcinogenic process is manifest as it is the number of publications that relate these genes to brain tumors. The results confirm the existence of Glioblastoma subtypes that the literature claims, see XI.A.2) for an expert validation of the results.

Verifying the statistical significance of the discovered genes, what is a common practice in the literature, as stated in [19], cannot be made based on conventional statistical tests, as the tests assume independence of the class definition and expression-profile data, what is not the case for cluster defined classes.

The lack of coincidence when trying to match the results to the features used in decision trees, gives support to [19], when claims that the classification of the unknown samples cannot be done based on the clustering results and a supervised method is required.

Also based on the Qualitative Analysis we find that from the 194 unique genes considered relevant for the study 119 genes are directly associated with Cancer and nervous diseases, what represents 61.3%. The number of unique genes is also an important factor to consider, as detailed in Appendix XI.C.1).

From a total theoretical of 453 after removing the probesets not in the SOURCE database, and counting only once the G1X-vs-G2X and its complementary, only 194 genes have remained what make for a very compact set that also adds, because of the many-to-1 association, another indicator of robustness of the solution.

The robustness happens also for each of the CM1 classes identified as can be seen in Table 49, where the number of unique genes consolidating by class (no coincidence of probeset since CM1 index is unique per each feature) is 417 out of 553.

In addition, the 19 out of 160 genes also found in other GSE4290 studies, are another proof of the interest of the method, as the low percentage cannot bring us to underestimate the result given the diversity of the studies considered, their techniques and their low mutual agreement (only 5 genes appear more than once), see Appendix XI.H.

Overall, the current study has contributed a method in order to extract a consensus clustering even when different algorithms have very divergent views of the dataset. The existing algorithms in the literature ignore some of the problems unveiled as part of the research. The proposed methodology can be adapted by using different algorithms and metrics to consider convergence.

x. Future Work

The results obtained are promising in order to continue working on the evolution of the methodology proposed. Obvious evolutions would be to include other clustering methods, and specifically methods that are based on the whole feature space, as the more different ones from the ones already considered.

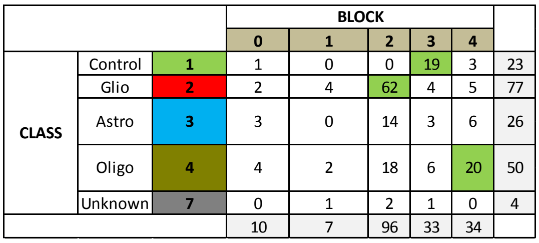
A comparison of the effect of the weights assigned to the different partitions would help to understand the main source of variability in the method and come out with new weighting parameters that can guarantee the convergence for different clustering algorithms or indicate that the lack of convergence shows the algorithm is not able to partition the dataset robustly.

The main question open by the study is how the clustering would be affected if the CM1 genes selected, both 1-vs-all and 1-vs-1, would be used as feature selection and the process re-run for the subset of probesets. The hypothesis to confirm is that the clusters will become stronger, requiring less iterations to reach strong consensus, but also that the results will improve as measuring only the relevant genes will get rid of the noise introduced by the indiscriminant genes (that are 2 orders of magnitude more numerous) producing a more reliable result. This process could be repeated iteratively until a certain stability criterion is met. This process corresponds to the red arrow in Figure 6.

The dataset used in the study is a quite large one, if compared to other Microarray Data studies, the effect of the size of the dataset should be considered and datasets with less samples be tested.

The new defined intra-/extra-Purity measure should be further evaluated in order to compensate for the biases already detected and explained in its definition. Additionally, it could be modified in order to be able to manage better cases where the size of the different classes is unbalanced. A first approach in this direction would be to consider the relative number of elements of each class instead of its absolute number.

This measure could also be improved if the membership coefficient in the solution partition for each of the samples is considered, instead of a 0/1, as it is the case now, that would connect our method with existing clustering methods based on fuzzy logic as well as Bayesian methods like Block Model [62] that when evaluated have produced quite interesting results in almost real time and where a better data modelling, to satisfy data distribution requirements for the method, could improve the overall results.



			BLOCK					
			0	1	2	3	4	
CLASS	Control	1	1	0	0	19	3	23
	Glio	2	2	4	62	4	5	77
	Astro	3	3	0	14	3	6	26
	Oligo	4	4	2	18	6	20	50
	Unknown	7	0	1	2	1	0	4
			10	7	96	33	34	

Table 12 Block model clustering confusion matrix

XI. Appendix

A. *Bio-interpretation of the results*

1) *Qualitative evidence of relevance*

For the lists of genes obtained, Table 47 and Table 48, its relationship to Cancer processes has been assessed. In order to do that, the genes have been queried against the database CTD (Comparative Toxicogenomics Database) and the Curated Disease Associations extracted: http://ctdbase.org/tools/batchQuery.go. All the associations obtained are made based on PubMed publications.

Based on that, the following conclusions have been extracted. The 8 genes on Table 14 correspond to genes that in CTD appear linked to varieties of Epilepsy (our control group is made up of Epilepsy patients). But only 2 of the genes (STXBP1 and SLC1A2) are found in the CX-vs-all CM1 selection. This happens because genes are usually involved in many biological processes. Also the information provided doesn't consider the degree of participation on the diseases so the results in this part must be considered as weak evidence.

Table 15 contains the 18 genes related directly to a variety of different tumor types (glioma, glioblastoma, gliosarcoma, astrocytoma, oligodendroglioma, lymphoma, neuroblastoma). Only 2 of the genes in the Epilepsy list (NTRK2, VEGFA) appear in this list.

Table 16 contains the 89 genes related to any kind of Cancer disease. Only 4 of the genes in the Epilepsy list appear in this list (NTRK2, VEGFA, SLC1A2, GAP43). Since the tissue being studied is brain tissue and the technique used is based on mRNA and therefore active genes, the gene can be considered related to the brain tumor.

Table 17 contains the 119 genes that the database associates with any kind of cancer plus nervous diseases, as advanced states of the disease cause other nervous problems to appear.

In addition to this, the KEGG database can be used to analyze the relationships among genes from different perspectives. This analysis requires a deeper understanding of the medical implications in order to be properly evaluated and can only be properly made by an expert in the topic.

Finally, comparing the genes identified in the study with others that have been considered relevant in studies that have also used the GSE4290 dataset, see Appendix H, we find that we have identified 19 out of 160 unique genes. This percentage although can be considered low, is not so low, as in the studies only 5 genes appear in more than one study. Also, some of the studies combine information from different datasets and the GSE4290 is used as test set and in other cases the studies are based also on biological experiments.

MBP	OLIG2	ID4
PLP1	VCAN	EGR1
BASP1	SOX8	APOD
EGFR	IGFBP2	GSN
PTN	CHI3L1	ANXA1
PTPRZ1	TIMP1	CD99
ENPP2		

Table 13 Genes also found in other GSE4290 studies

SPARCL1	Downregulated in G1X-vs-all	**NES**	Appears in G1X-vs-G2X
STXBP1	Upregulated in CX-vs-All	**NTRK2**	Downregulated in G1X Upregulated in OX Appears in G1X-vs-G2X
SLC1A2	Upregulated in CX-vs-All	**GRIA2**	Downregulated in G1X
VEGFA	Upregulated in G1X and G2X-vs-All. Downregulated in OX-vs-All	**GAP43**	Downregulated in OX

Table 14 Genes related to Epilepsy

FAM107A	FTH1	SOD2
NTRK2	HLA-C	HLA-DRB1
CHI3L1	HLA-B	HEY1
SPP1	HLA-A	GNAS
FN1	EGFR	JAG1
VEGFA	B2M	APOD

Table 15 Genes related to Glioma, Glioblastoma, Gliosarcoma, Astrocytoma, Oligodendroglioma, Limphoma, Neuroblastoma

FAM107A	CST3	CD44	CHI3L1	B2M	PEG3
TF	S100A6	CTSB	SERPINA3	GNB2L1	GNAS
BASP1	IGFBP5	HLA-DRB1	SPP1	EEF1A1	VOPP1
NTRK2	LGALS3	CD74	VIM	IGFBP7	HSPA8
NDRG2	ANXA1	RPL3	FABP7	CD99	PPIA
CRYAB	MAOB	PABPC1	IGFBP2	ACTG1	DBI
SCD	UCHL1	ZBTB20	LGALS1	MT2A	EGR1
OLFM1	HLA-C	MARCKS	FN1	POSTN	COL6A1
SLC1A2	MT3	VCAN	VEGFA	IGFBP3	MALAT1
ATP1B1	HLA-B	ID4	FTL	PTN	JAG1
EEF1A2	HLA-A	RPS3	FABP5	AQP1	TRIO
TSC22D1	RPS19	RPS6	CD63	LTF	APOD
FXYD6	RPL13	APOE	GAP43	A2M	GSN
TUBB2A	SPARC	HEY1	FTH1	CLU	SEPP1
PTPRO	EGFR	PRKACB	PGK1	SOD2	

Table 16 Genes related to Cancer

MBP	MT3	NPTN	MARCKS	OLFM1	POSTN	FABP5	HSPA8
PLP1	HLA-B	TUBB2A	VCAN	ADD3	IGFBP3	CD63	NES
FAM107A	HLA-A	PTPRO	ID4	QDPR	PTN	GAP43	PPIA
RTN1	RPS19	CHI3L1	RPS3	SPARCL1	AQP1	FTH1	DBI
TF	RPL13	LDHA	RPS6	SLC1A2	LTF	PGK1	EGR1
SNAP25	SPARC	SERPINA3	APOE	TSPAN7	A2M	COL1A1	COL6A1
BASP1	EGFR	SPP1	HEY1	SOX8	CLU	CST3	MALAT1
NTRK2	B2M	VIM	PRKACB	CALM1	SOD2	S100A6	JAG1
CHN1	GNB2L1	FABP7	PEG3	ATP1B1	CD44	IGFBP5	TRIO
NDRG2	EEF1A1	TIMP1	IDS	GABBR1	CTSB	LGALS3	APOD
CRYAB	IGFBP7	IGFBP2	SLC17A7	NEFL	HLA-DRB1	ANXA1	GSN
KIF5C	CD99	LGALS1	YWHAH	SYN2	CD74	COL4A1	CD24
SCD	RPL27A	FN1	SERPINI1	EEF1A2	RPL3	MAOB	PSAP
STXBP1	ACTG1	VEGFA	GNAS	TSC22D1	PABPC1	UCHL1	SEPP1
GRIA2	MT2A	FTL	VOPP1	FXYD6	ZBTB20	HLA-C	

Table 17 Genes related to Cancer and Nervous System Disease

2) *Expert evaluation: a bio perspective (by Prof. Pablo Moscato)*

Based on the CM1 features for the 4 1-vs-all separations and the G1X-vs-G2X and G2X-vs-G1X, Prof. Pablo Moscato, Co-Director of the Centre for Bioinformatics, Biomarker Discovery and Information-based Medicine at the Hunter Medical Research Institute (Newcastle, NSW, Australia), has analyzed the results. What follows in this subsection XI.A.2), is his analysis of the results applying its knowledge in the field of bioinformatics and cancer research. His collaboration has been independent of the rest of the study.

The bibliographical references in this section, numerous as they are, close to 200, are included in a separated References section: *References for bio-informatics study*.

a) *Preliminary analysis*

Figure 18 shows the Venn Diagram analysis of CM1 positive genes, upper-expressed, in the comparisons between G1-vs-all (G1X), G2-vs-all(G2X), O-vs-all(OX) and C-vs-all(CX). We have only used the probesets that have a positive value of CM1**. For each set, we listed gene names to which probes have been mapped and also the probe sets names**. This naturally explains why we have 91 objects in CX, which correspond to the 50 associated probe sets with the highest values of CM1 scores as well as the 41 gene names that have been associated to them (totalling 91 markers to compare).

The first observation is that, in general, there is no intersection between the sets, with the single exception of 39 markers that are in the intersection of G1X and G2X. In this intersection subset we found several markers which are common in glioblastoma. The list contains the following genes:

- *SPP1 (Secreted Phosphoprotein 1, Osteopontin),*
- *CHI3L1 (chitinase 3-like 1 (cartilage glycoprotein-39),*
- *VIM (Vimentin),*
- *FN1 (Fibronectin 1),*
- *VEGFA(vascular endothelial growth factor A),*
- *HLA-A (major histocompatibility complex, class I, A),*
- *HLA-B (major histocompatibility complex, class I, B),*
- *HLA-C (major histocompatibility complex, class I, C),*
- *TIMP1 (TIMP metallopeptidase inhibitor 1),*
- *MT2A (metallothionein 2A),*
- *LGALS1 (lectin, galactoside-binding, soluble, 1),*
- *TMSB10 (thymosin beta 10),*
- *IGFBP7 (insulin-like growth factor binding protein 7),*
- *B2M (beta-2-microglobulin)*
- *CD63 (CD63 molecule)*

A probe corresponding to 234989_at could not be mapped to a known gene.

There are also probesets that differentiate the group G1 from the rest. They total 39. They correspond to the following genes:

- *EGFR (epidermal growth factor receptor),*
- *FABP7 (fatty acid binding protein 7, brain),*
- *SEC61G (Sec61 gamma subunit),*
- *PTN (pleiotrophin),*
- *LDHA (lactate dehydrogenase A),*
- *IGFBP2 (insulin-like growth factor binding protein 2, 36kDa)*
- *IGFBP3 (insulin-like growth factor binding protein 3),*
- *COL4A1 (collagen, type IV, alpha 1),*
- *ACTG1 (actin, gamma 1),*
- *POSTN (periostin, osteoblast specific factor),*
- *LGALS3 (lectin, galactoside-binding, soluble, 3),*
- *RPS19 (ribosomal protein S19)*
- *RPS2 (ribosomal protein S2)*

The following probe sets, which are also in this group, are not mapped to a gene (232541_at, 216438_s_at).

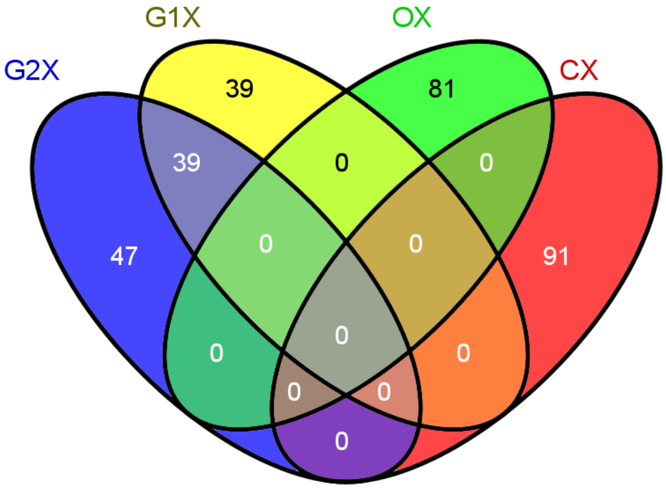


Figure 18 Probesets and genes from 1-vs-all CM1

The third group to discuss is the 47 probesets that correspond to G2X which are not in the other subsets. The genes associated to these probesets are:

- *SERPINA3 (serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3),*
- *C3 (complement component 3),*
- *FTL (ferritin, light polypeptide),*
- *FTH1 (ferritin, heavy polypeptide 1),*
- *LTF (lactotransferrin),*
- *CD74 (CD74 molecule, major histocompatibility complex, class II invariant chain), HLA-DRB1 (major histocompatibility complex, class II, DR beta 1),*
- *HLA-DPA1 (major histocompatibility complex, class II, DP alpha 1),*
- *CD44 (CD44 molecule (Indian blood group),*
- *SOD2 (superoxide dismutase 2, mitochondrial),*
- *CLU (Clusterin),*
- *A2M (alpha-2-macroglobulin),*
- *COL1A1 (collagen, type I, alpha 1),*
- *S100A6 (S100 calcium binding protein A6),*
- *C1QC (complement component 1, q subcomponent, C chain),*
- *CTSB (cathepsin B),*
- *MAOB (monoamine oxidase B),*

- *LAPTM5 (lysosomal protein transmembrane 5),*
- *AQP1 (Aquaporin 1),*
- *ANXA1 (annexin A1),*
- *IGFBP5 (insulin-like growth factor binding protein 5)*

Many of the 50 probesets in CX correspond to neuron markers like:

- *VSNL1 (visinin-like 1),*
- *RGS4 (regulator of G-protein signaling 4),*
- *SNAP25 (synaptosomal-associated protein, 25kDa),*
- *NEFL (neurofilament, light polypeptide), etc.,*

or oligodendrocyte specific markers like:

- *MBP (myelin basic protein),*
- *PLP1 (proteolipid protein 1), etc*

The increased value of CM1 in this group is clear as the comparison of the tumors against the controls indicate that there is a process of dedifferentiation from the brain architecture tissue and these neuron and oligodendrocyte specific markers are less abundant due to large majority of tumor cells in the samples. This said, we will concentrate our attention to the following three groups.

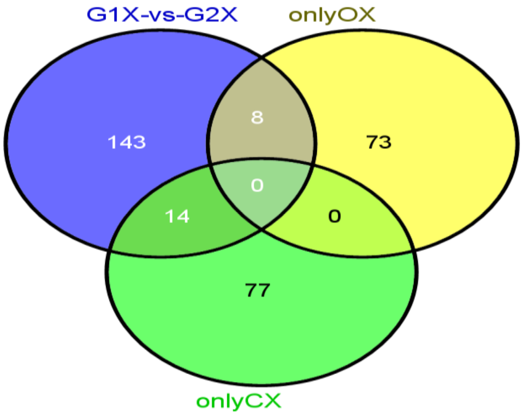


Figure 19 Venn-diagram for G1X-vs-G2X, OX, CX

We then proceed to calculate a signature of the top 100 probesets that best separate, according to the CM1 score, the group of samples labelled G1 from those labelled G2. As the CM1 score is not symmetric we calculated CM1 scores for all probes in two opportunities, i.e. having each of the two groups as the group of interest. However, in this case both results gave the same set of top positive and negative first 50 probesets. We call this group of probesets and the associated mapped genes as `G1X-vs-G2X' (see Figure 19). We compare this set with those that have appeared in the list for OX (that differentiate this group from the rest of samples) and also those that are in the group CX. We observe that there is a small intersection in the first case, five probesets (221796_at, 221795_at, 209283_at, 203381_s_at, 203382_s_at) mapping to the genes:

- *NTRK2 (neurotrophic tyrosine kinase, receptor, type 2),*
- *CRYAB (crystallin, alpha B)*
- *APOE (apolipoprotein E)*

In the second case we have matches, corresponding to the probesets: 209072_at, 211748_x_at, 210198_s_at, 212187_x_at, 207323_s_at, 227556_at, 209123_at, 201242_s_at, with matches to well-known oligodendrocyte markers:

- *MBP (Myelin Basic Protein) [1]*
- *PLP1 (Proteolipid Protein 1/ Myelin proteolipid protein) [1]*
- *PTGDS (prostaglandin D2 synthase 21kDa (brain))*
- *NME7 (non-metastatic cells 7, protein expressed in (nucleoside-diphosphate kinase)),*
- *QDPR (quinoid dihydropteridine reductase)*
- *ATP1B1 (ATPase, Na+/K+ transporting, beta 1 polypeptide)*

This relatively small intersection indicates that the differences between the groups G1 and G2 are, its top CM1 scoring, when computed independently of any other type of sample, very different to those of CX and OX as previously obtained.

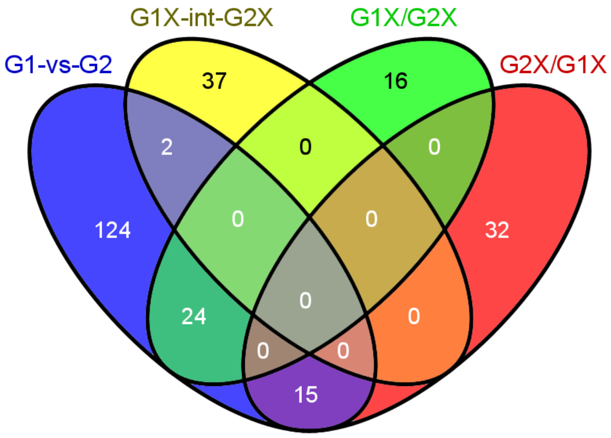


Figure 20 G1X and G2X Venn diagram analysis

With this information, we proceeded to compute another diagram, Figure 20. We include again this group of 100 probesets and their mapped genes and we label them as `G1-vs-G2'. We compare this lists with those of the intersection of the groups G1X and G2X (now labelled G1X-int-G2X) and the group obtained from G1X but eliminating the markers in G2X (labelled G1X/G2X) and those in G2X but eliminating the markers in G1X (analogously labelled [2, 3]).

In this case we are interested in some intersections. The intersection of G1-vs-G2 and G1X-int-G2X only brings as a marker probeset 201426_s_at, for VIM (Vimentin). This is an important finding that will be discussed later. The intersection of G1-vs-G2 and G1X/G2X brings the probesets 216438_s_at, 210095_s_at, 200869_at, 208949_s_at, 210809_s_at, 205029_s_at, 201984_s_at, 202718_at, 211737_x_at, [4, 5], 205030_at, 224999_at, 201983_s_at, corresponding to the genes: RPS2, IGFBP3, LGALS3, EGFR, POSTN, FABP7, IGFBP2, PTN, SEC61G.

The intersection between G1-vs-G2 and G2X/G1X is smaller with 200748_s_at, 202376_at, 217767_at, 213187_x_at, 212788_x_at, 200839_s_at, 201721_s_at, 225353_s_at corresponding to the genes: FTH1, SERPINA3, C3, FTL, CTSB, LAPTM5, C1QC.

b) *Annotation of the results*

In Figure 18 we pointed to the existence of several genes in both G1X and G2X that have the top CM1 scores. Since G1 and G2 are clusters that contain many samples labelled as glioblastoma, it is important to correlate the result with the literature. We start with the genes found in the intersection. The list includes:

Gene	Name and synonyms	References
SPP1	Secreted Phosphoprotein 1, Osteopontin	14: [6-19]
CHI3L1	Chitinase 3-like 1 (cartilage glycoprotein-39) (Synonims 39 kDa synovial protein, ASRT7, Cartilage glycoprotein 39, CGP-39, Chitinase-3-like protein 1, DKFZp686N19119, FLJ38139, GP39, GP-39, hCGP-39, HC-gp39, HCGP-3P, YKL40, YKL-40, YYL-40)	36: [6, 11, 20-53]
VIM	Vimentin	11: [23, 50-59]
FN1	Fibronectin 1	25: [11-13, 60-81]
VEGFA	Vascular endothelial growth factor A	6: [3, 5, 82-85],
TIMP1	TIMP metallopeptidase inhibitor 1	11: [14, 88-97]
MT2A	metallothionein 2A	1:[98]
HLA-A	major histocompatibility complex, class I, A	2: [86, 87]
CD63	CD63 molecule	1: [13]
HLA-B	major histocompatibility complex, class I, B	Related to HLA-A
HLA-C	major histocompatibility complex, class I, C	Related to HLA-A
LGALS1	lectin, galactoside-binding, soluble, 1	Related to LGALS3
TMSB10	thymosin beta 10	

IGFBP7	insulin-like growth factor binding protein 7	Related to IGFBP2
B2M	beta-2-microglobulin	

Table 18 Genes common to G1X and G2X, literature references

We now turn our attention to the probesets that differentiate the group G1 from the rest. They correspond to the genes in Table 19.

Gene	Name and synonyms	References
EGFR	epidermal growth factor receptor	19: [86, 99-116]
FABP7	fatty acid binding protein 7, brain	8: [117-124]
SEC61G	Sec61 gamma subunit	2:[86, 125]
PTN	pleiotrophin	12:[4, 5, 115, 126-134]
IGFBP3	insulin-like growth factor binding protein 3	13:[5, 135-146]
COL4A1	collagen, type IV, alpha 1	1:[116]
ACTG1	actin, gamma 1	1:[147]
POSTN	periostin, osteoblast specific factor	3:[16, 148, 149]
RPS2	ribosomal protein S2	2:[150, 151]
LGALS3	lectin, galactoside-binding, soluble, 3	See below
IGFBP2	insulin-like growth factor binding protein 2, 36kDa	See below
RPS19	ribosomal protein S19	
LDHA	lactate dehydrogenase A	

Table 19 Genes only in G1X

The following probe sets, which are also in this group, are not mapped to a gene (232541_at, 216438_s_at).

The intersection of G1-vs-G2 and G1X/G2X brings the genes in Table 20.

Gene	Name and synonyms	References
EGFR	epidermal growth factor receptor	13:[4, 5, 115, 126-134]
POSTN	periostin, osteoblast specific factor	3:[16, 148, 149]
FABP7	fatty acid binding protein 7, brain	8:[117-124]
IGFBP3	insulin-like growth factor binding protein 3	13:[5, 135-146],
PTN	pleiotrophin	13:[4, 5, 115, 126-134],
SEC61G	Sec61 gamma subunit	2: [86, 125]

LGALS3	lectin, galactoside-binding, soluble, 1	See below
IGFBP2	insulin-like growth factor binding protein 2, 36kDa	See below

Table 20 Genes G1-vs-G2 and G1X/G2X

The intersection between G1-vs-G2 and G2X/G1X contains the genes in Table 21.

Gene	Name and synonyms	References
SERPINA3	Serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3	2: [9, 152]
FTL	ferritin, light polypeptide	3: [153-155]
CTSB	Cathepsin B/APP secretase	20: [156-175]
LAPTM5	lysosomal protein transmembrane 5	1:[176]
FTH1	Ferritin, Heavy polypeptide 1/Cell proliferation-inducing gene 15 protein	
C3	Complement Component 3	
C1QC	complement component 1, q subcomponent, C chain	

Table 21 Genes G1-vs-G2 and G2X/G1X

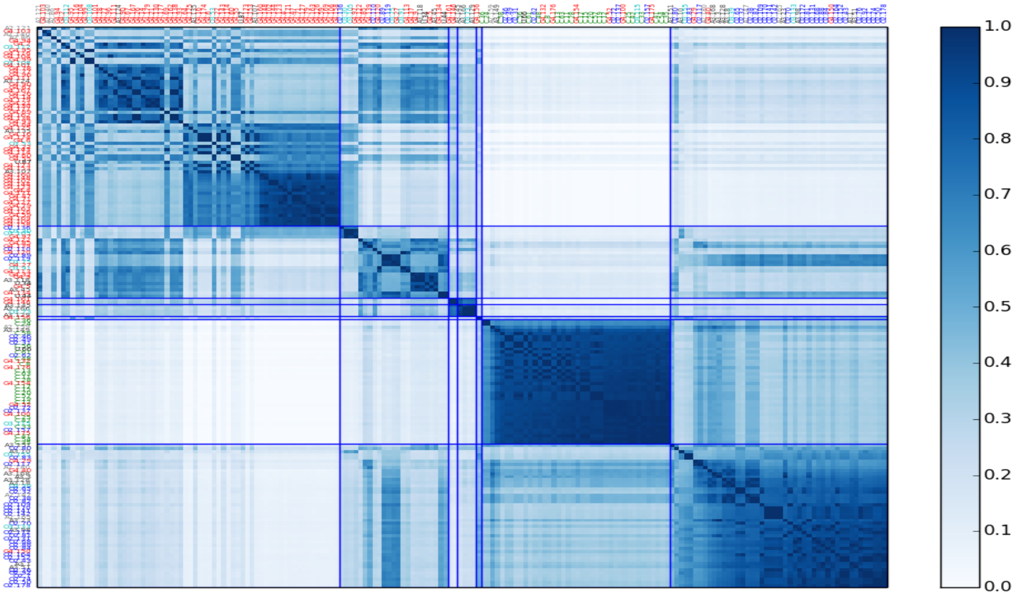
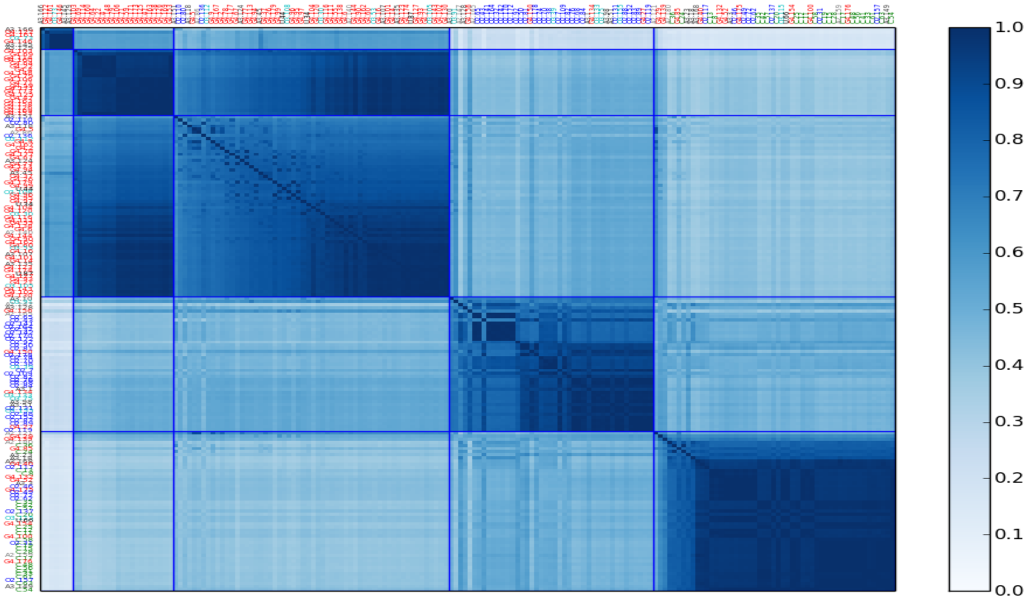
c) *Conclusions*

Overall, the number of genes that are presented in Tables 18, 19 and 20 that are related to glioblastoma is really impressive. It does seem to be that there are differences related to the expression of EGFR, IGFBP3 and PTN (to mention three which have been widely studied) while the method also brings to the attention IGFBP2 (from the same family) and LGALS3 (also known as Galectin-3) it has been proposed as a marker that can distinguish pilocytic astrocytomas from diffuse astrocytomas, and glioblastomas from anaplastic oligodendrogliomas in [177]. These results indicate that perhaps it should be consider as a marker to be used in relationship with the others in these panels. In [78], the authors analyzed 409 cases of surgically resected primary brain tumors and found its expression to be “definitely positive but heterogeneous” in glioblastoma and other types of tumors. **This is coherent with the observation here, which seems to indicate that LGALS3 may co-express with EGFR, IGFBP3 and PTN in one subtype of glioblastoma.** We refer to other papers on this gene for further references on reports on this gene in relationship with glioblastoma [177,178,179,180,181,182,183,184,185,186,187,188,189]. Plasma IGFB2 levels seem to correlate with prognosis of glioma patients [190] and anaplastic astrocytomas [191]. Immunohistochemistry for this gene was positive in 88.8% of the cases in a study involving 28 glioblastomas [192]. Several other studies have also linked it to cell proliferation and migration [193] and it deserves also to be included in the studies of function as has been reported elsewhere as being of interest [190,191,192,193,194,195,196,197,198,199,200,201,202].

B. *Experiments results*

This section contains all the graphs corresponding to the different experiments run for the study.

1) *All methods, N Iterations=100*

	Raw, N Iterations= 100
Column Normalization + Jensen-Shannon	
Row+Column Normalization + Jensen-Shannon	

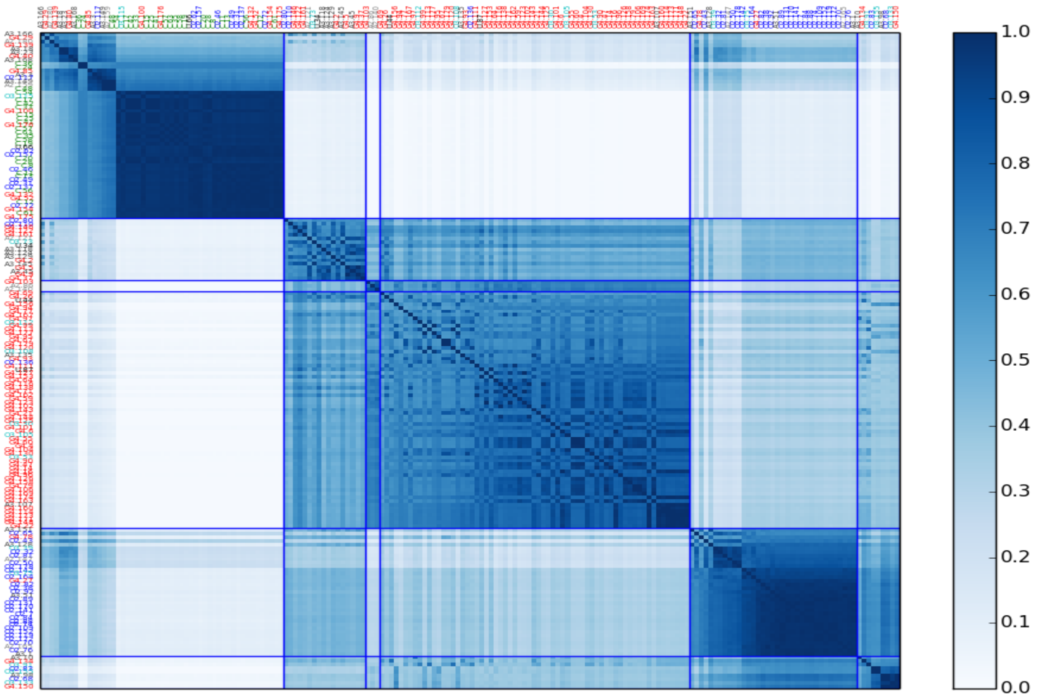
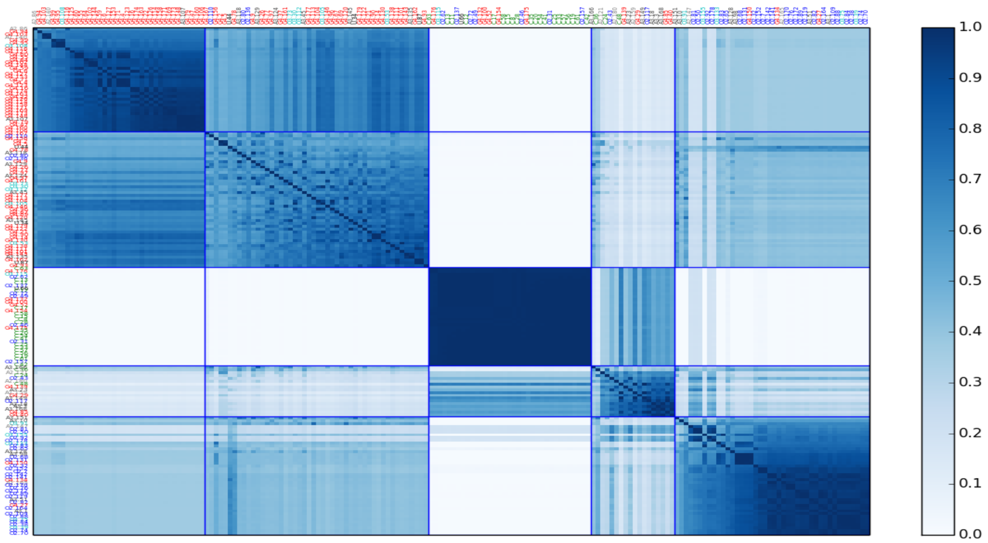
Pearson	

Table 22 Partition results for N Iterations=100, Raw Dataset

	WO Outliers, N Iterations=100
Column Normalization + Jensen-Shannon	

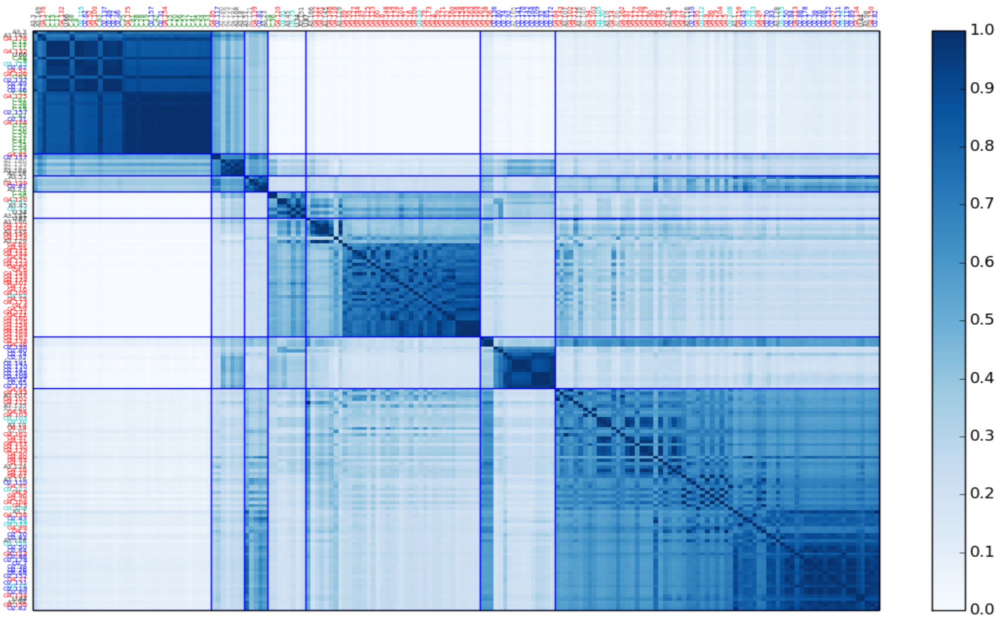
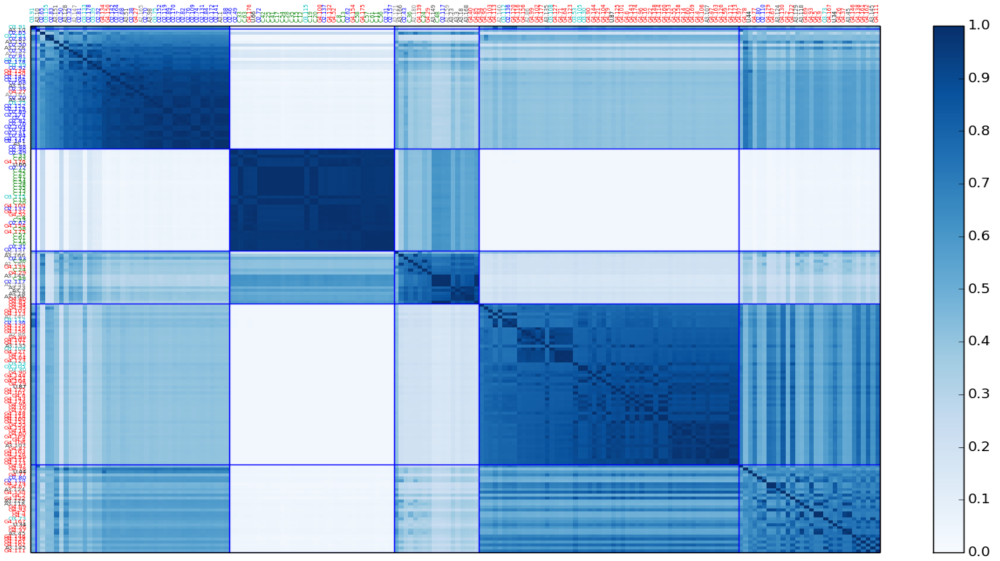
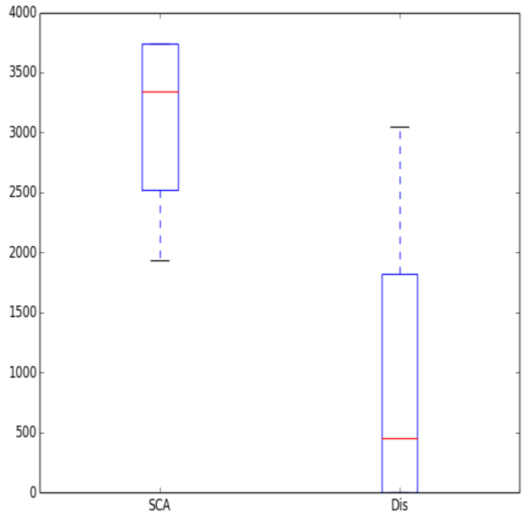
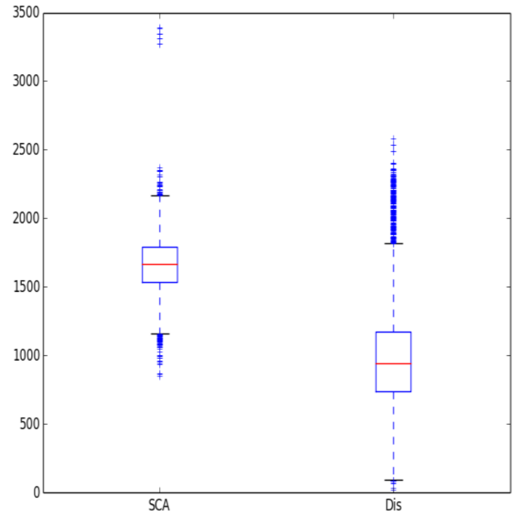
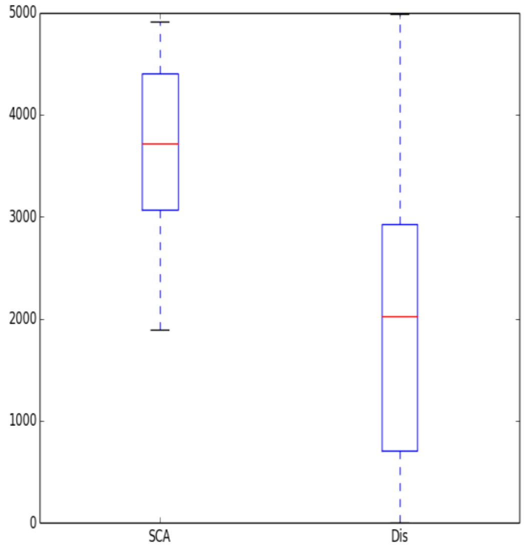
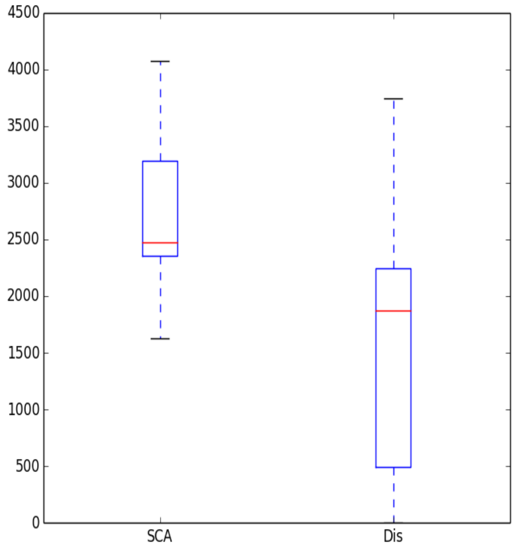
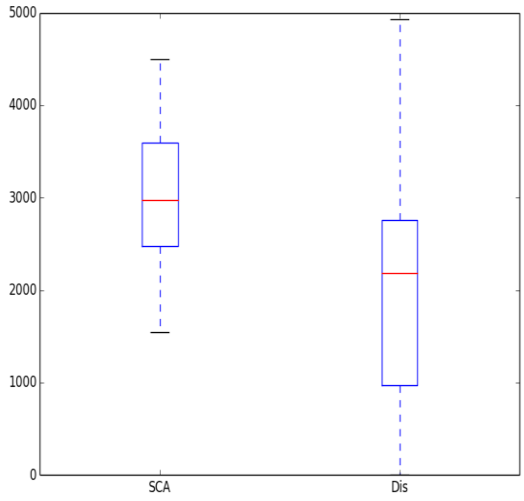
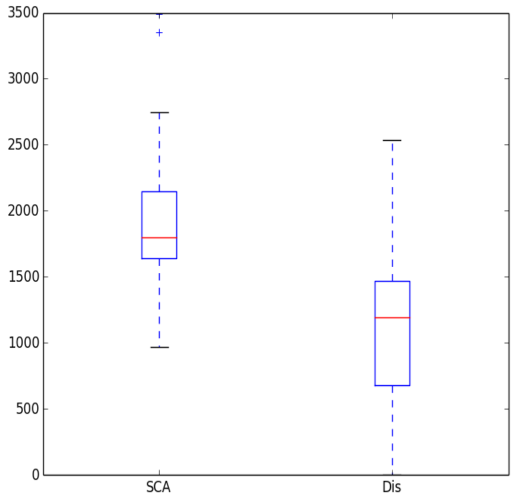
Row+Column Normalization + Jensen-Shannon	
Pearson	

Table 23 Partition results for N Iterations=100, WO Outliers

	Raw	WO Outliers
Column Normalization + Jensen-Shannon		
Row+Column Normalization + Jensen-Shannon		
Pearson		

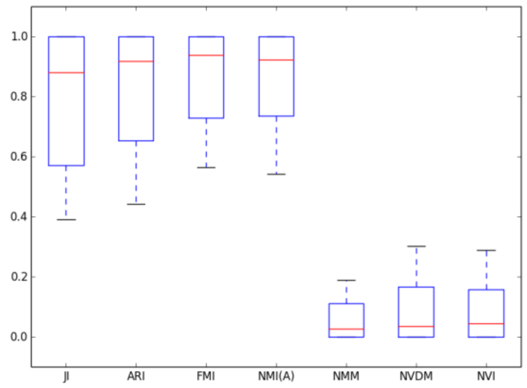
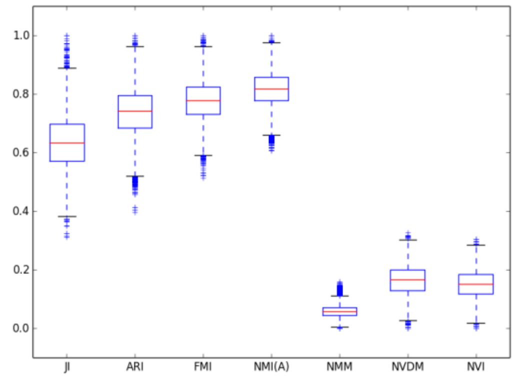
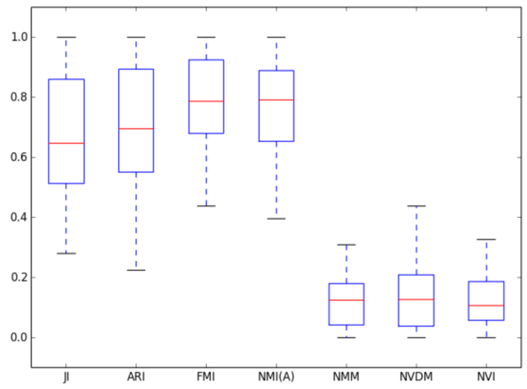
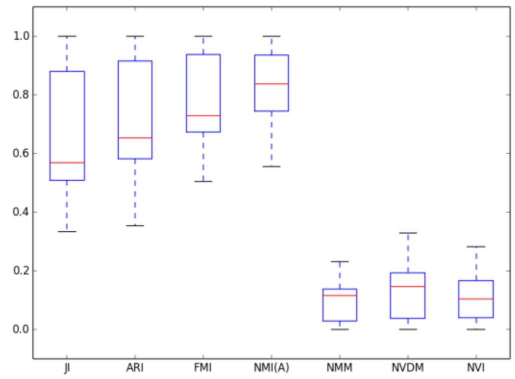
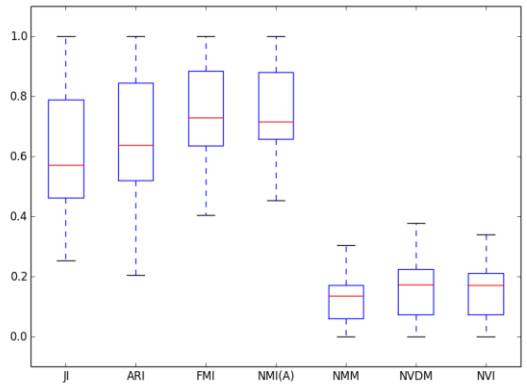
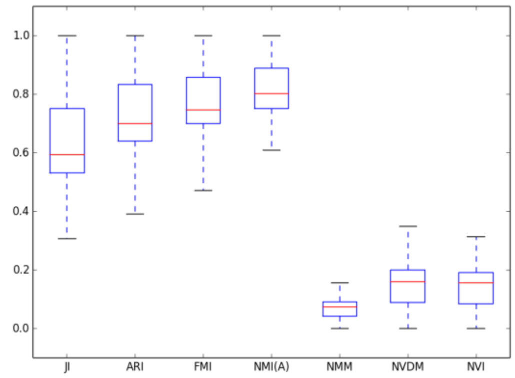
	Raw	WO Outliers
Column Normalization + Jensen-Shannon		
Row+Column Normalization + Jensen-Shannon		
Pearson		

Table 24 Indexes and Agreements for N Iterations=100

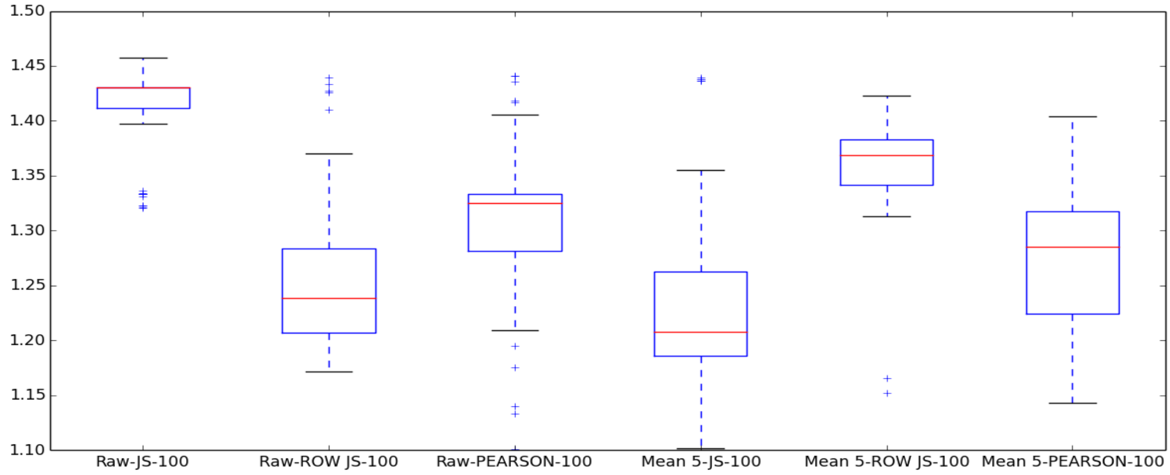
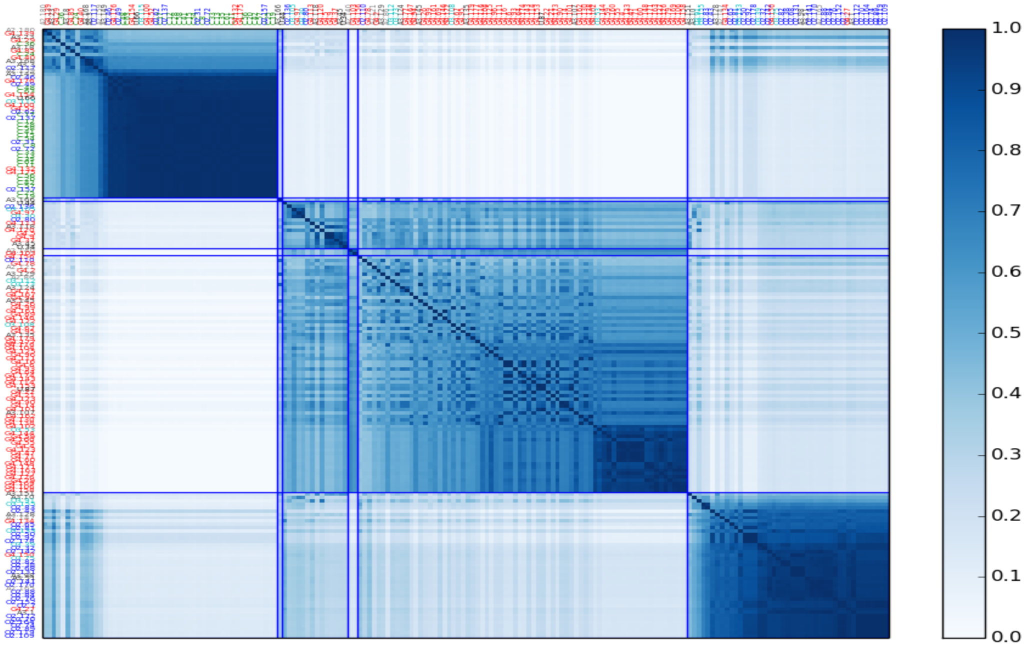


Figure 21 Accuracy for N Iterations=100

2) *All settings, N Iterations=200*

	Raw, N Iterations=200
Column Normalization + Jensen-Shannon	

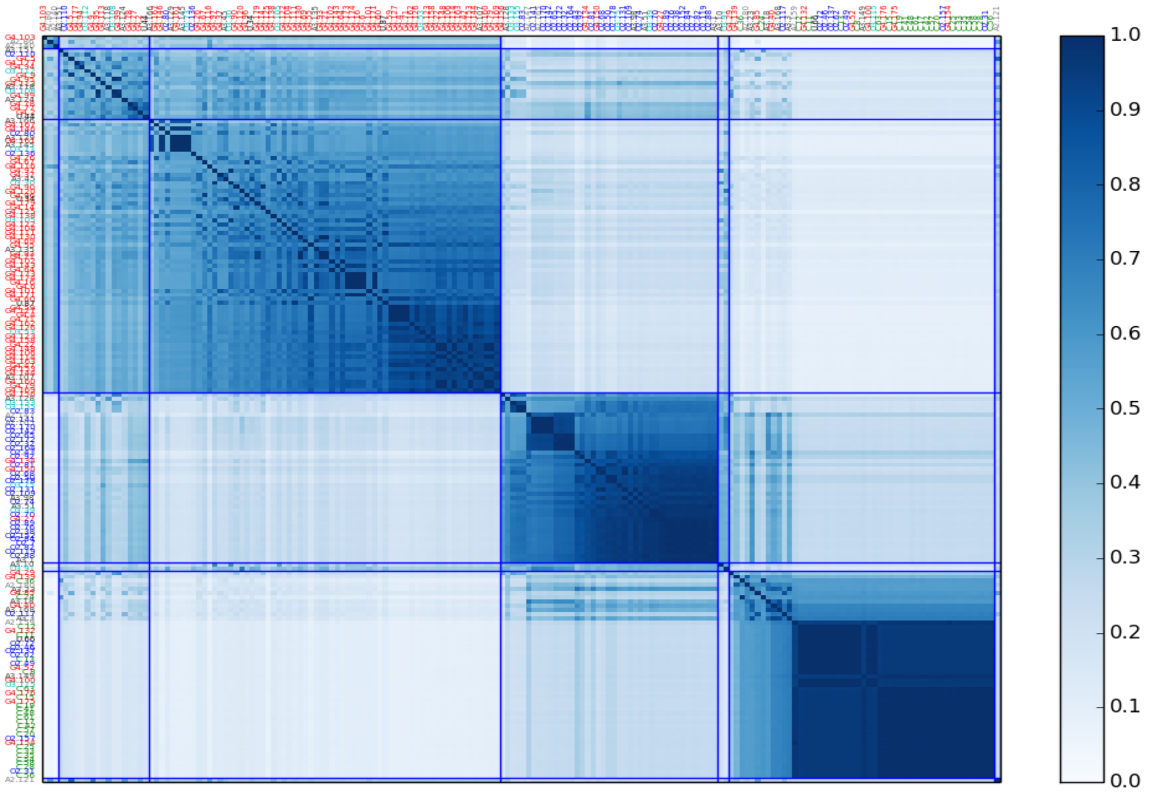
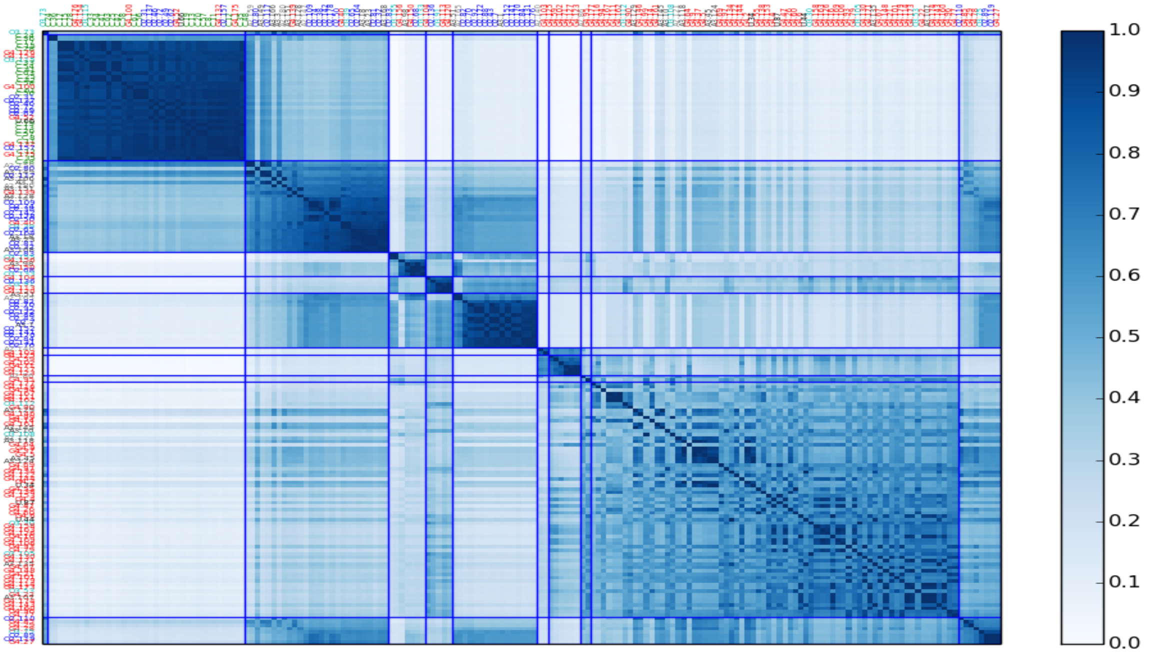
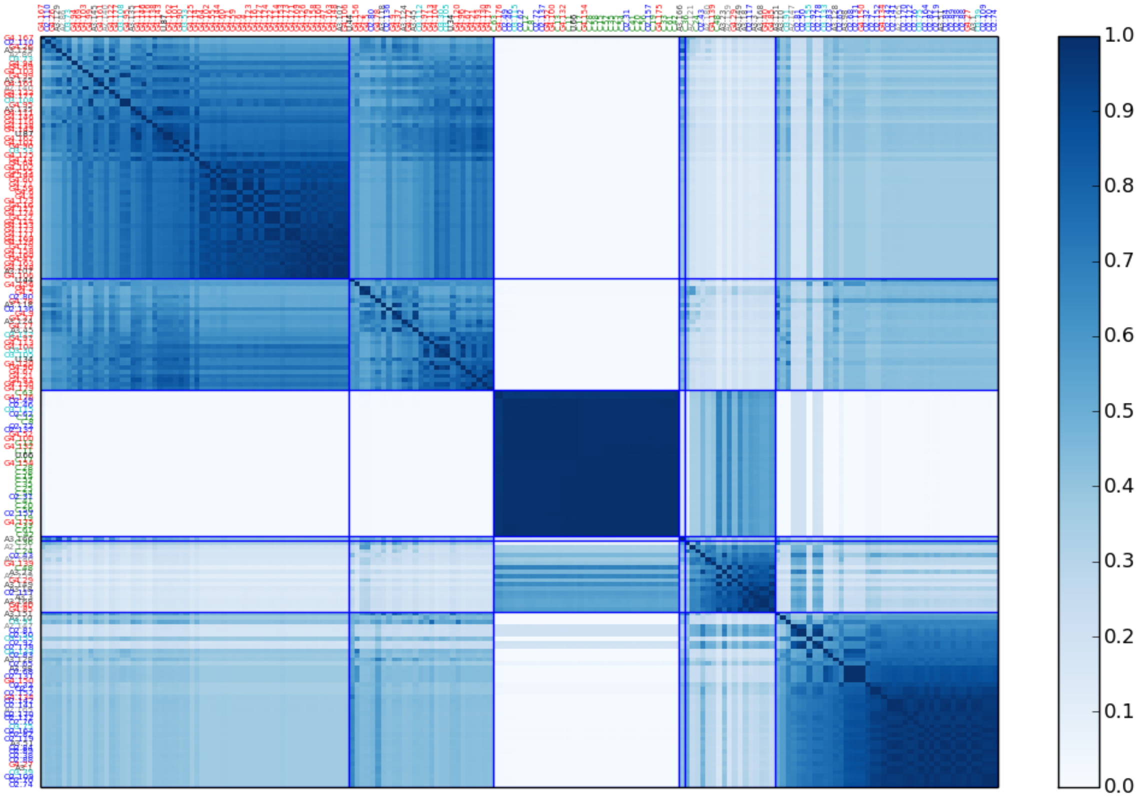
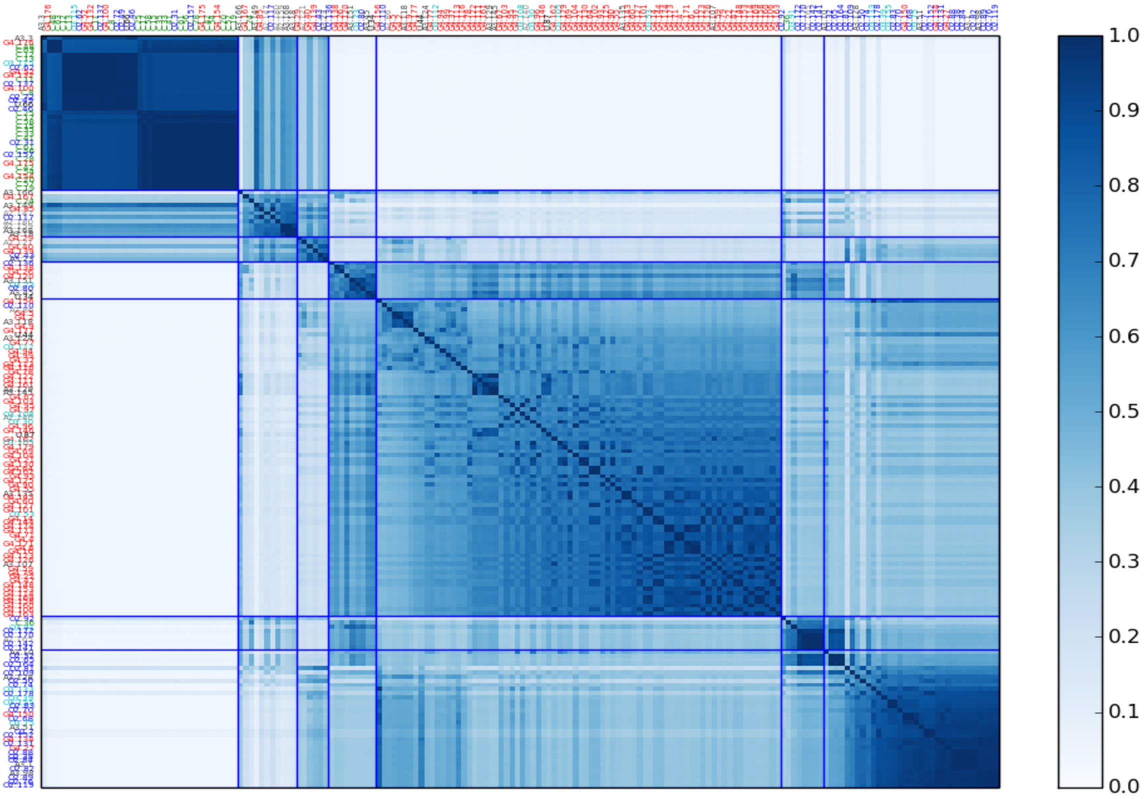
Row+Column Normalization + Jensen-Shannon	
Pearson	

Table 25 Partition results for N Iterations=200, Raw Dataset

	WO Outliers, N Iterations=200
Column Normalization + Jensen-Shannon	
Row+Column Normalization + Jensen-Shannon	

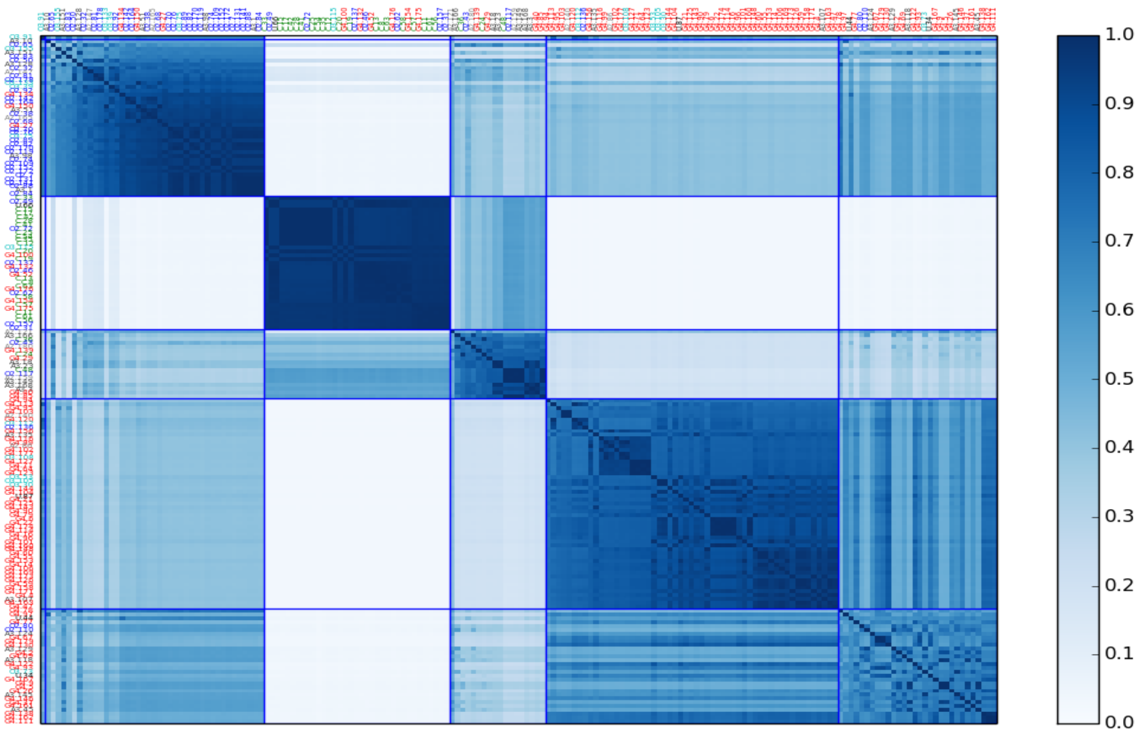
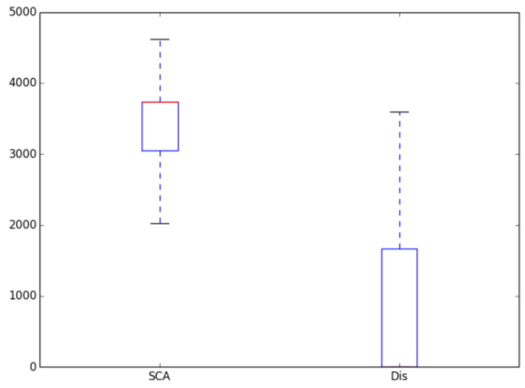
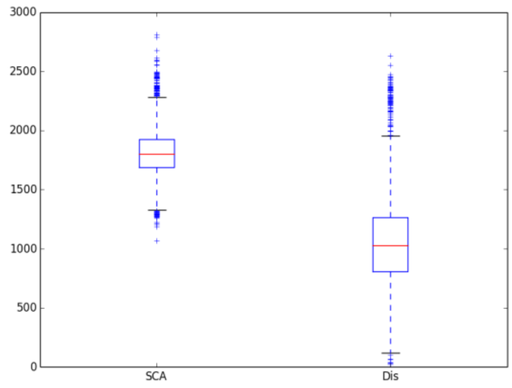
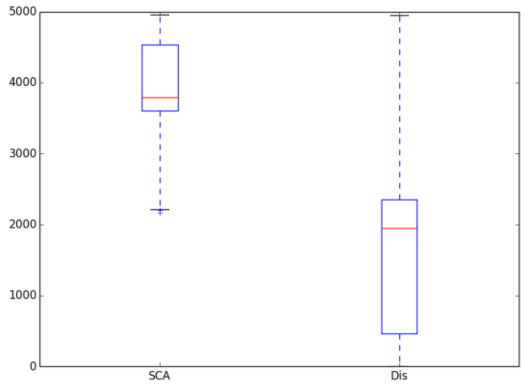
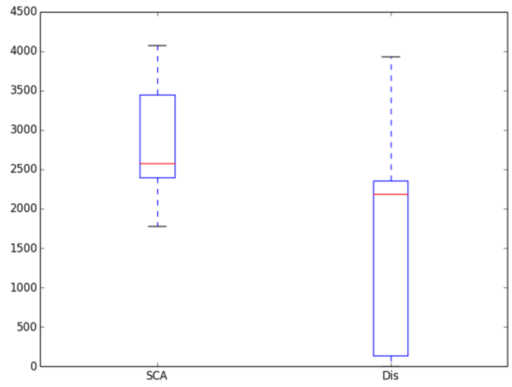
Pearson	

Table 26 Partition results for N Iterations=200, WO Outliers

	Raw	WO Outliers
Column Normalization + Jensen-Shannon		
Row+Column Normalization + Jensen-Shannon		

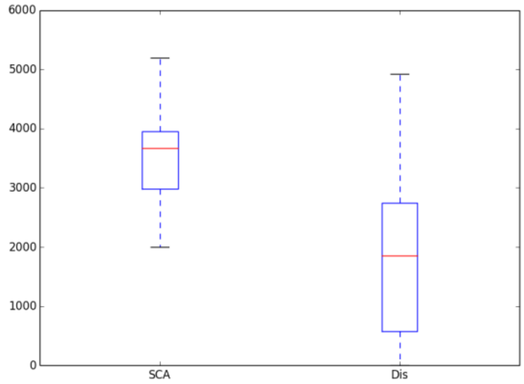
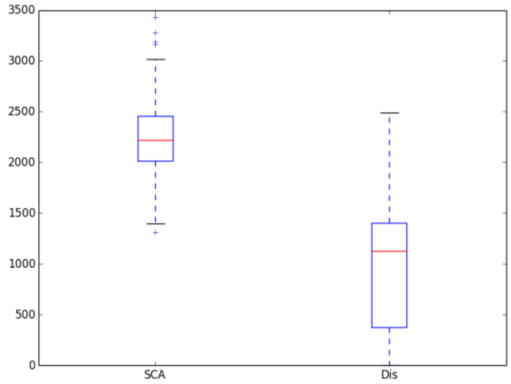
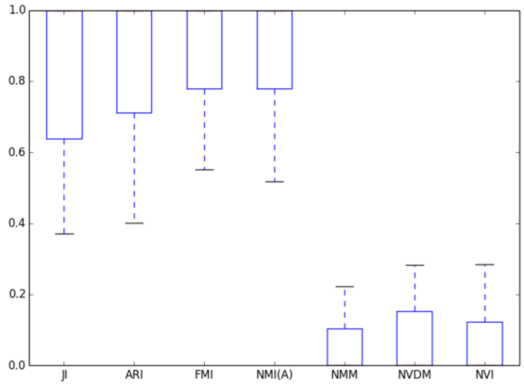
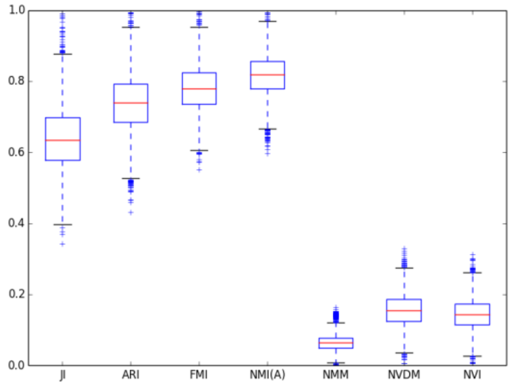
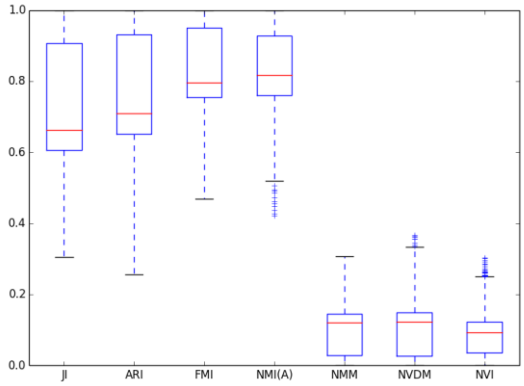
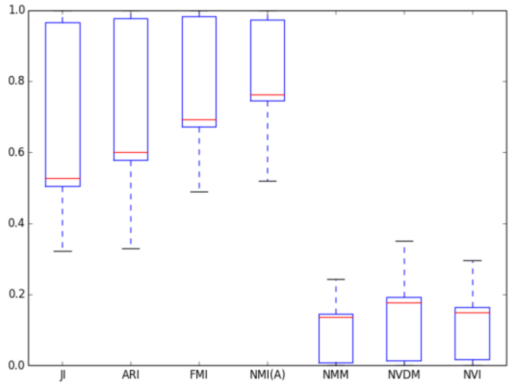
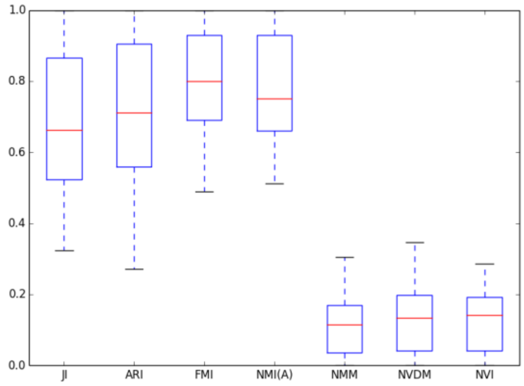
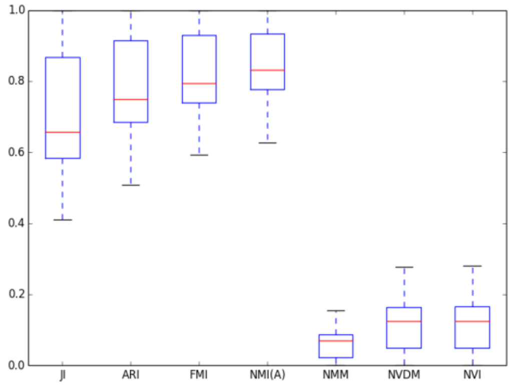
Pearson		
	Raw	WO Outliers
Column Normalization + Jensen-Shannon		
Row+Column Normalization + Jensen-Shannon		
Pearson		

Table 27 Indexes and Agreements for N Iterations=200

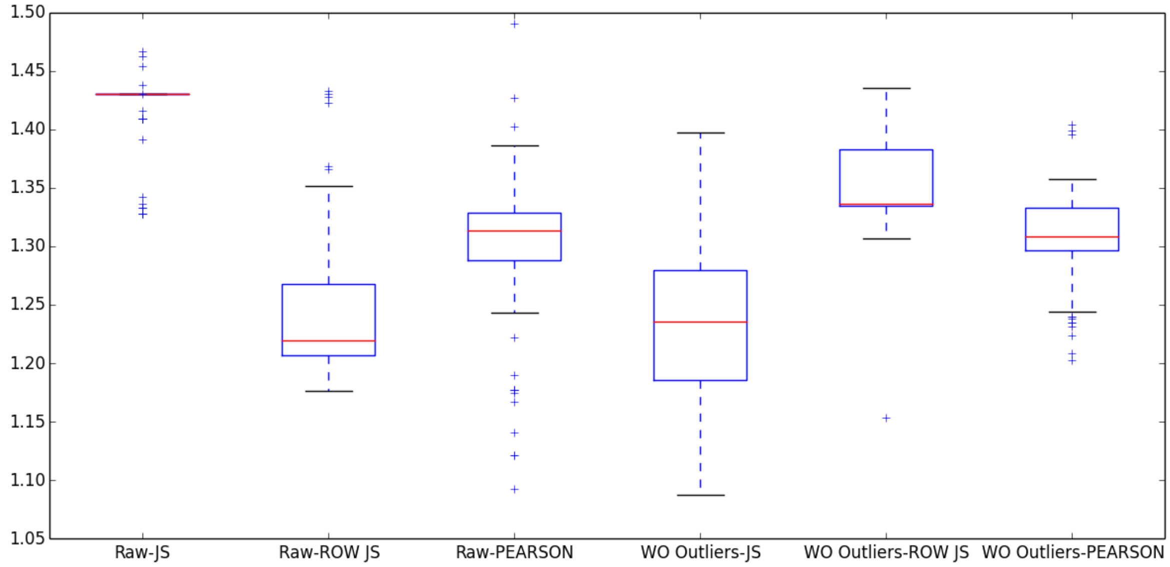
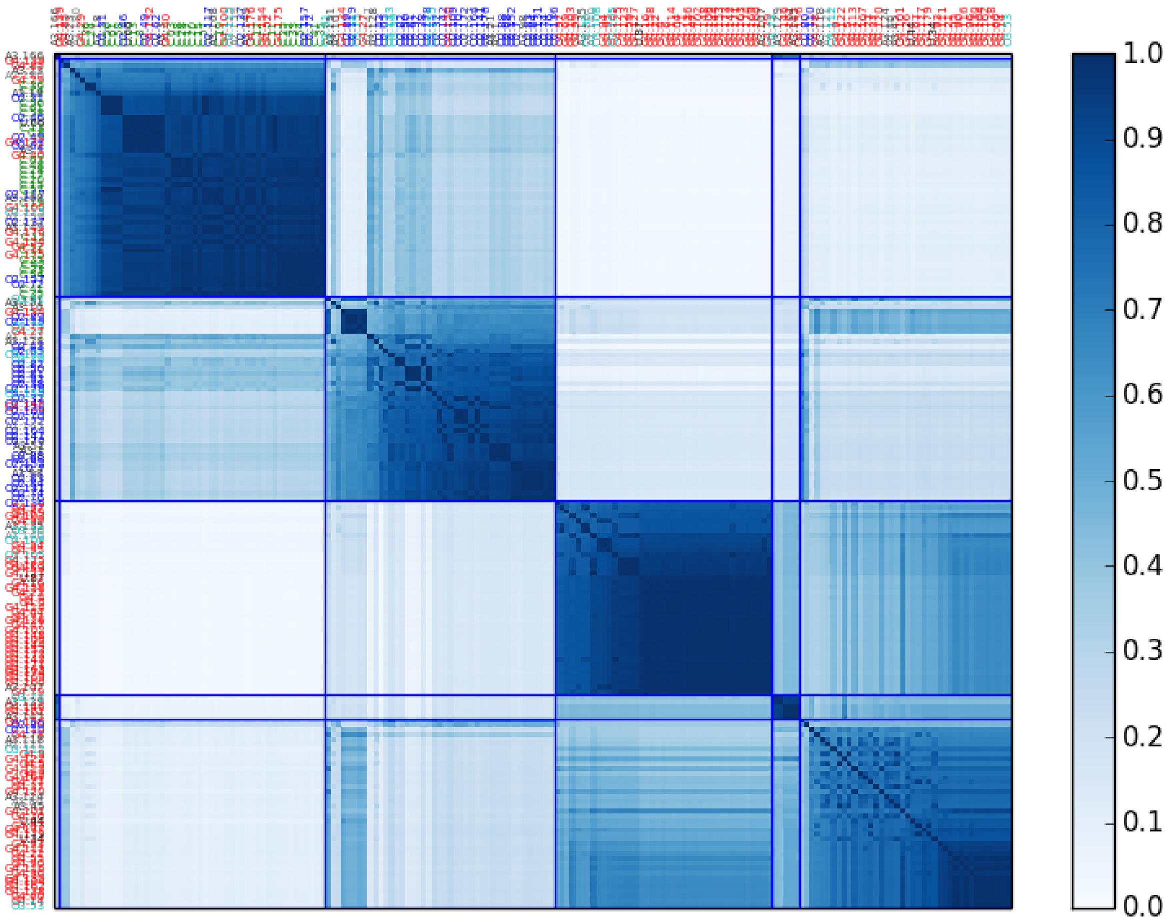


Figure 22 Accuracy for N Iterations=200

3) *All settings, N Iterations=500*

	Raw, N Iterations=500
Column Normalization + Jensen-Shannon	

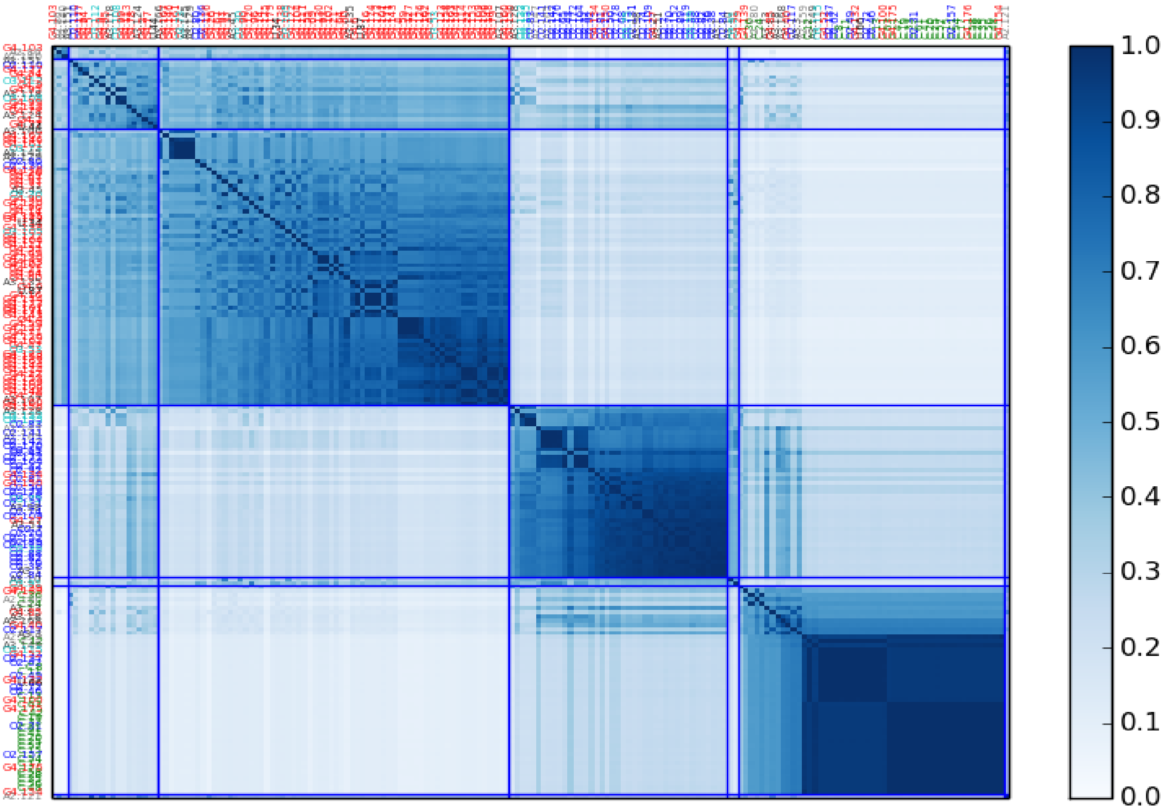
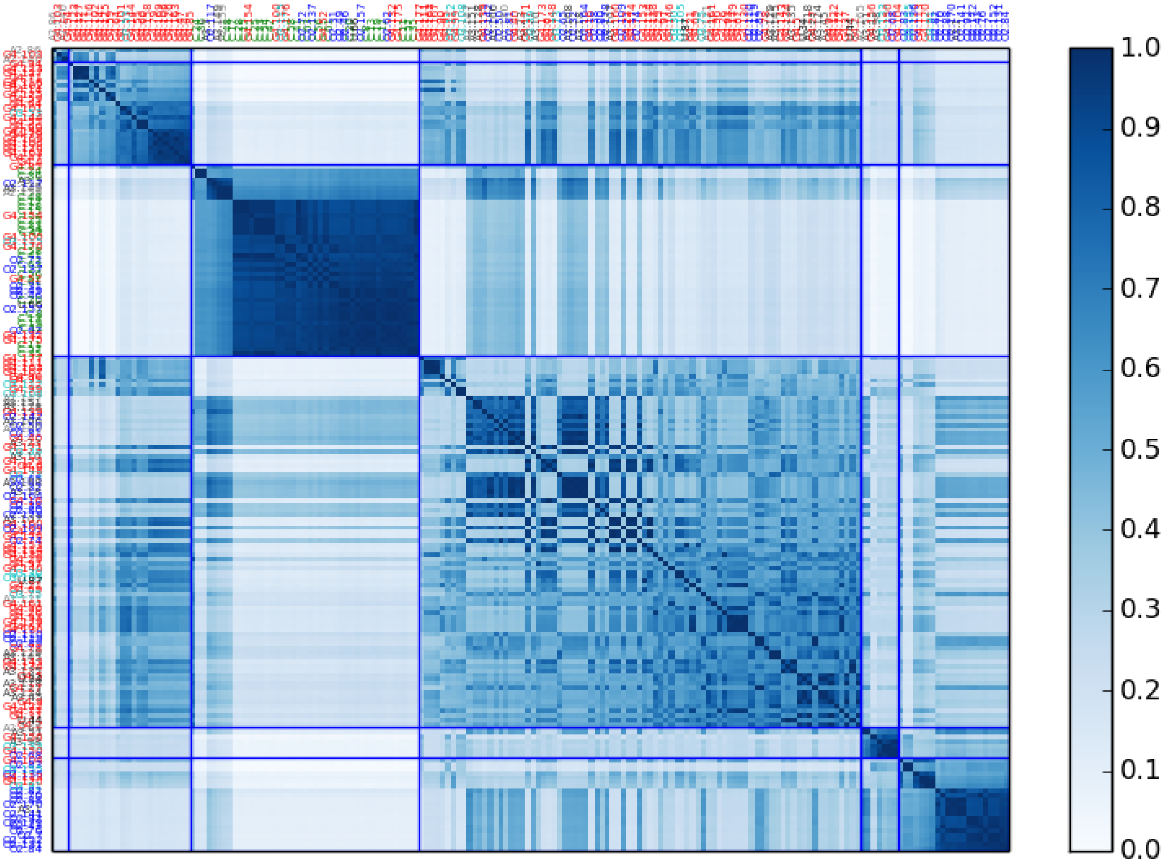
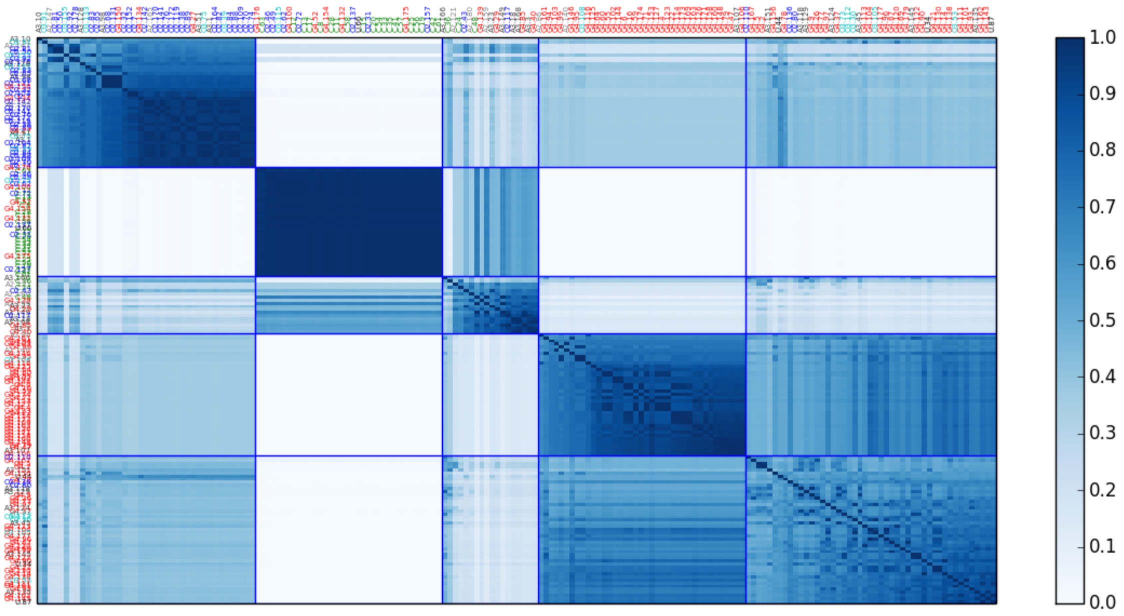
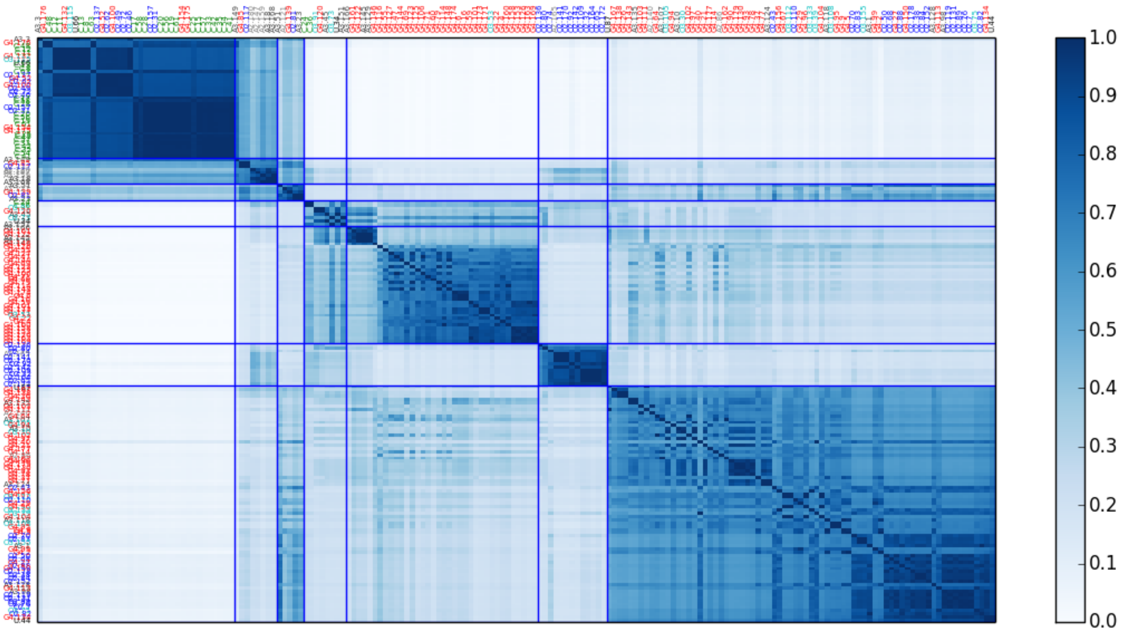
Row+Column Normalization + Jensen-Shannon	
Pearson	

Table 28 Partition results for N Iterations=500, Raw Dataset

	WO Outliers, N Iterations=500
Column Normalization + Jensen-Shannon	
Row+Column Normalization + Jensen-Shannon	

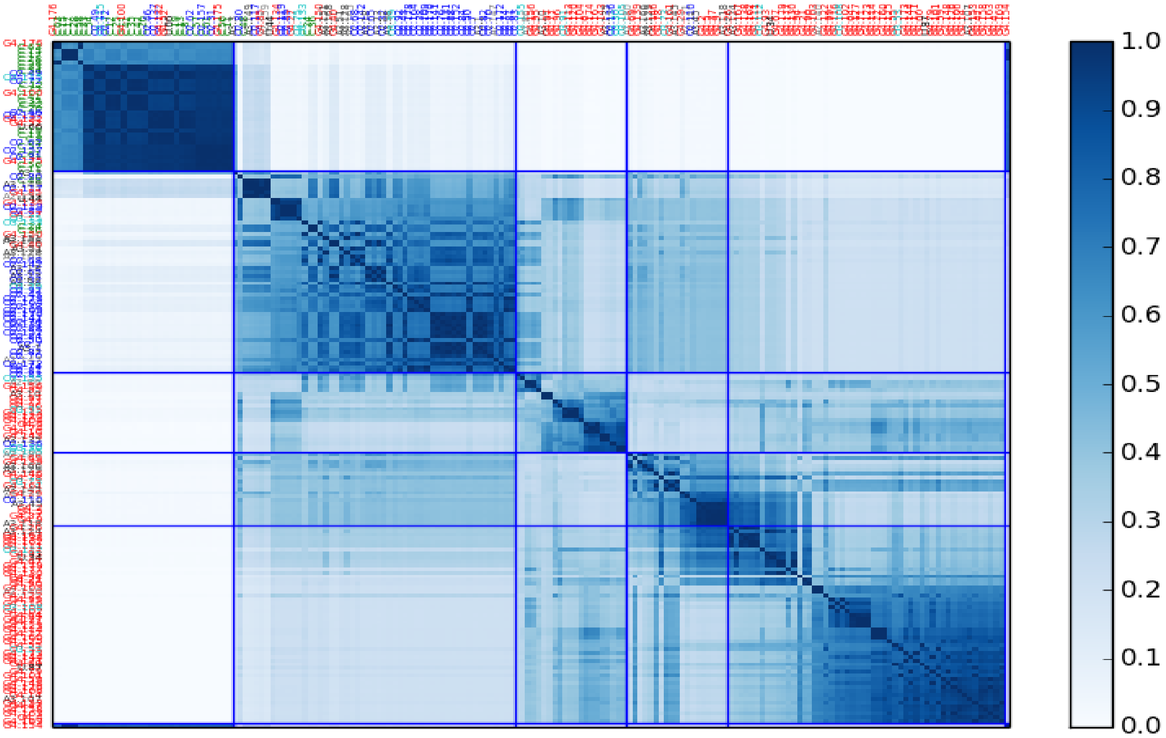
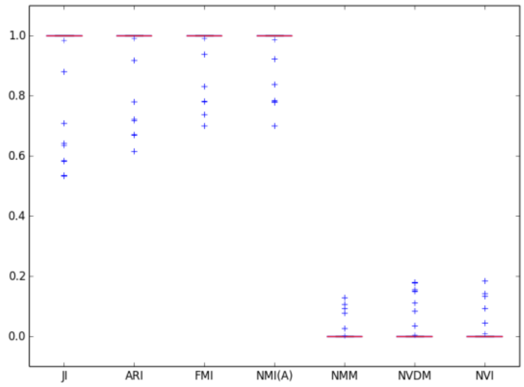
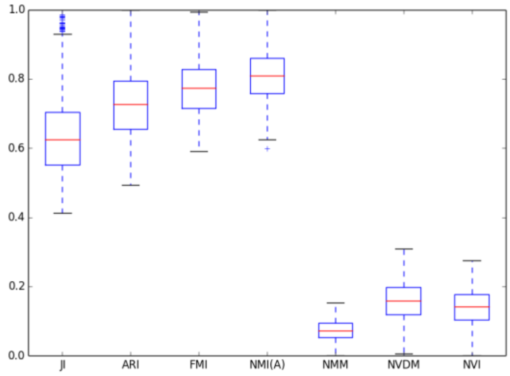
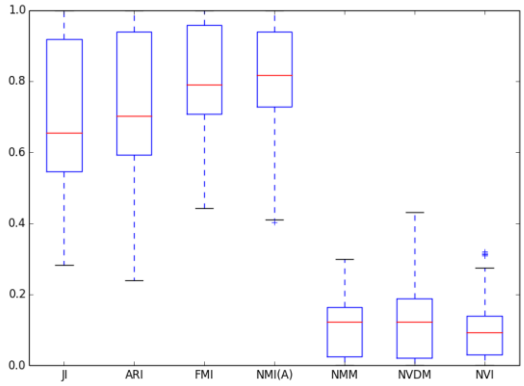
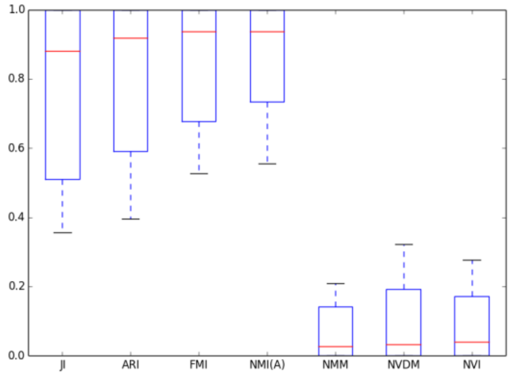
Pearson	

Table 29 Partition results for N Iterations=500, WO Outliers

	Raw	WO Outliers
Column Normalization + Jensen-Shannon		
Row+Column Normalization + Jensen-Shannon		

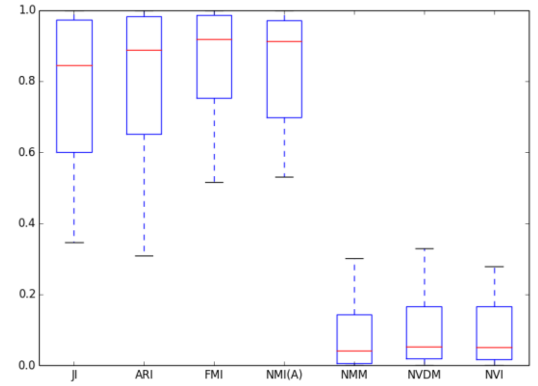
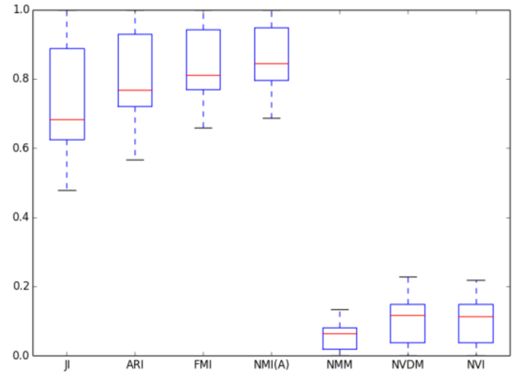
Pearson		

Table 30 Indexes for N Iterations=500

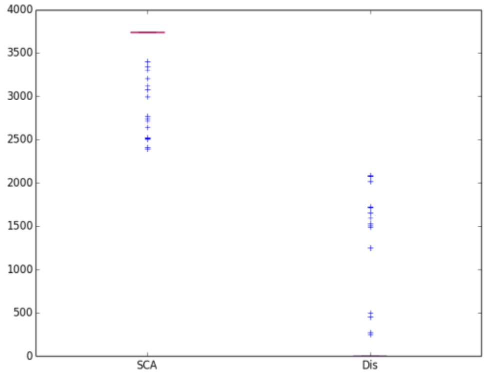
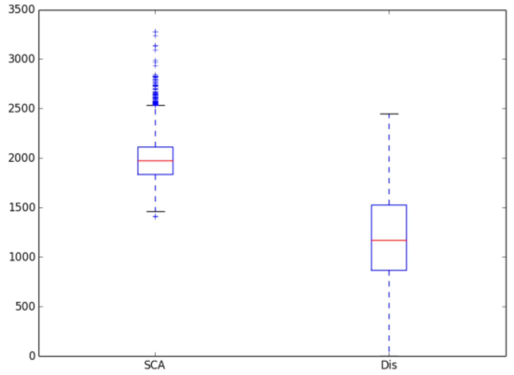
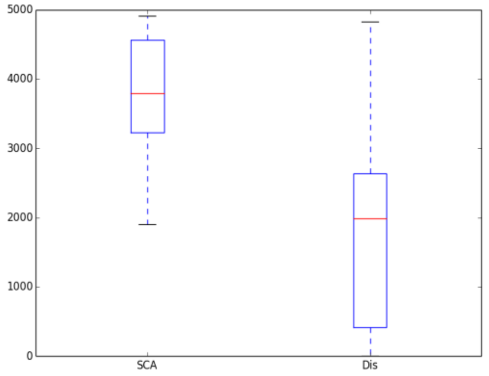
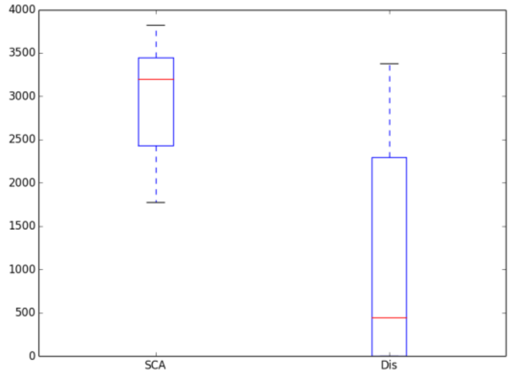
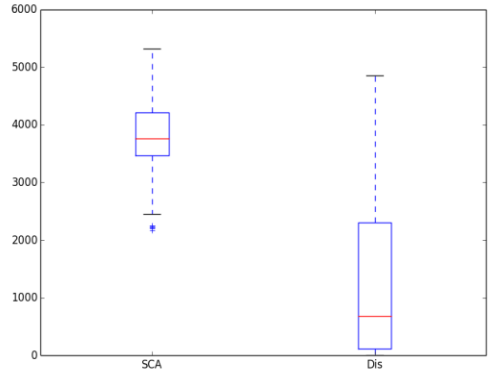
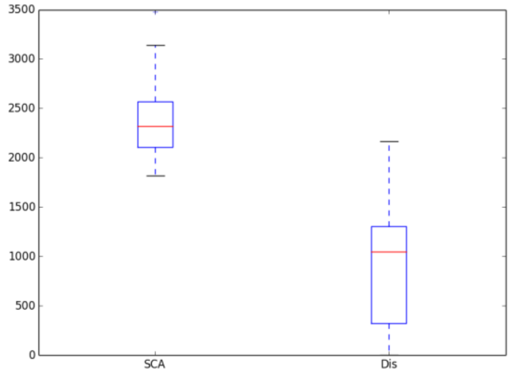
	Raw	WO Outliers(Mean 5)
Column Normalization + Jensen-Shannon		
Row+Column Normalization + Jensen-Shannon		
Pearson		

Table 31 Agreements for N Iterations=500

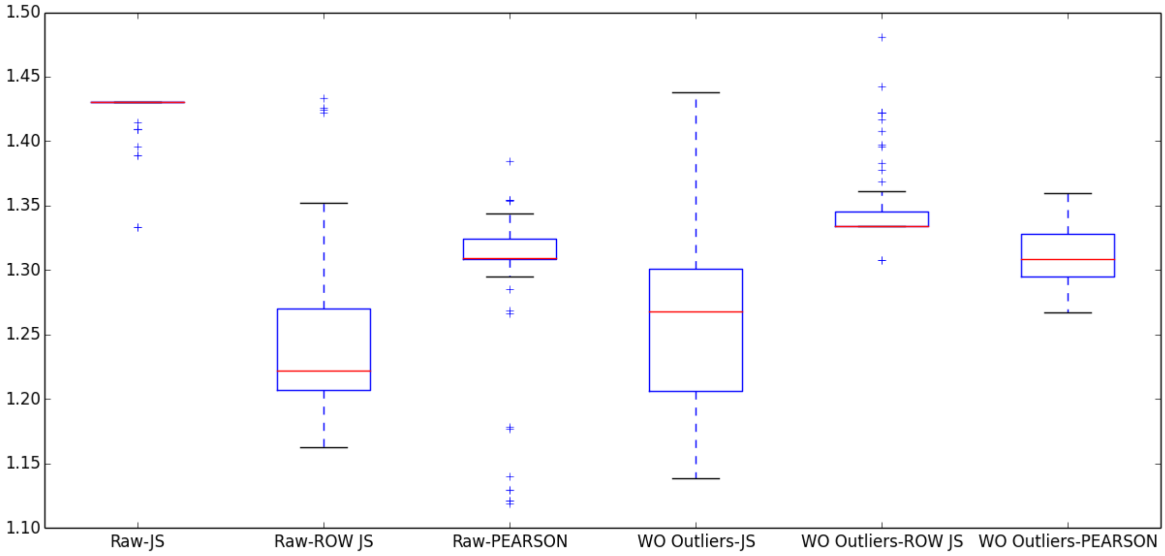


Figure 23 Accuracy for best partition N Iterations=500

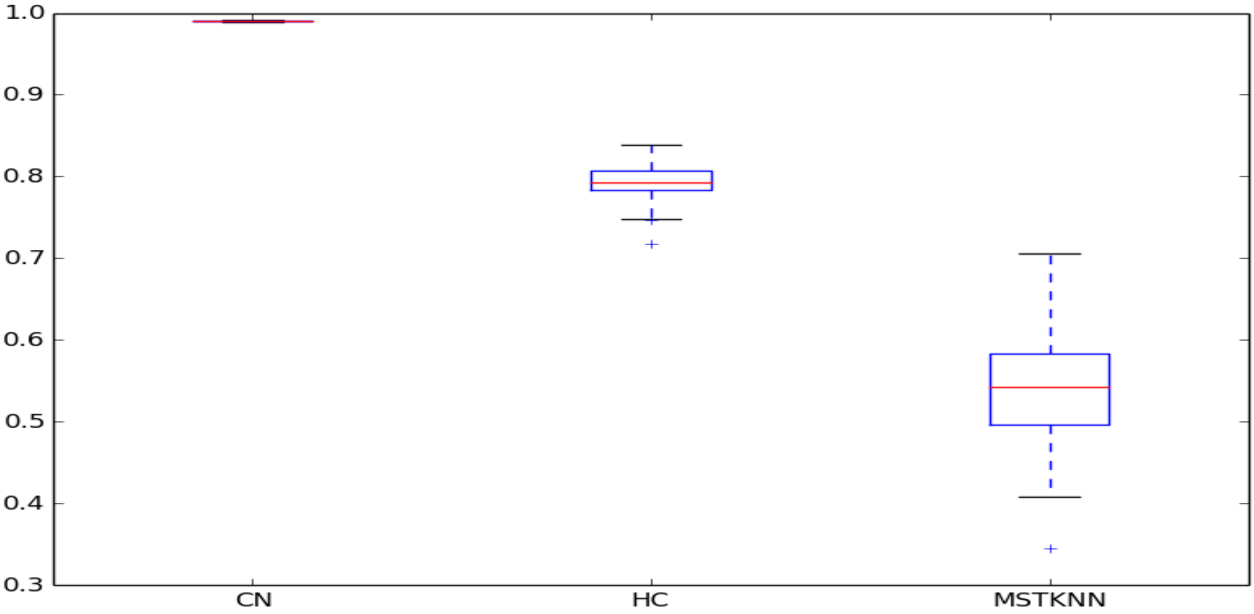
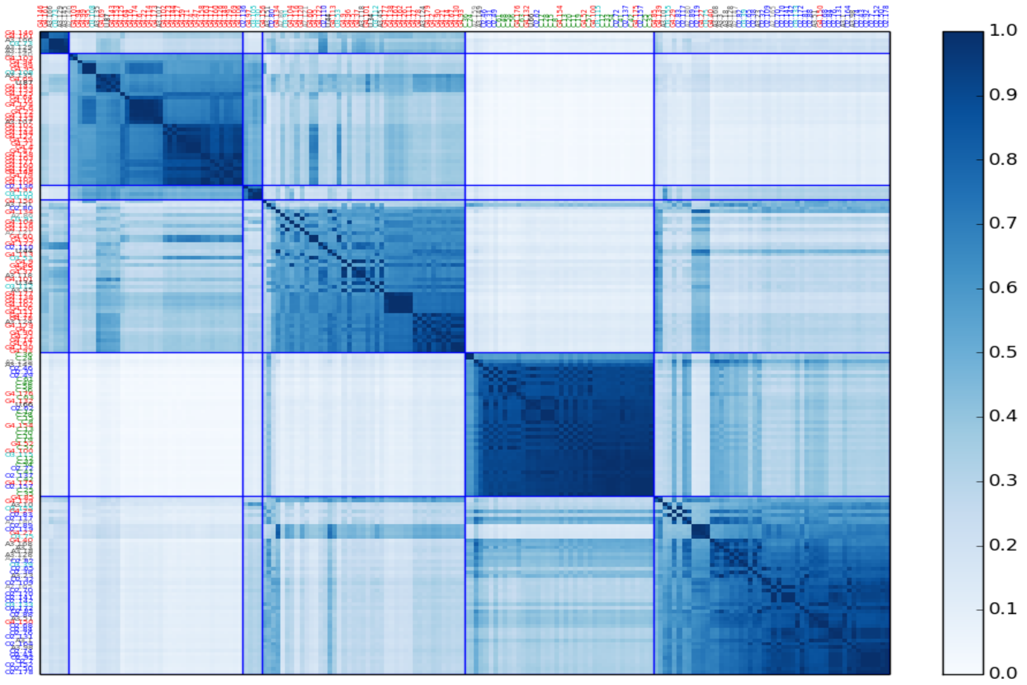
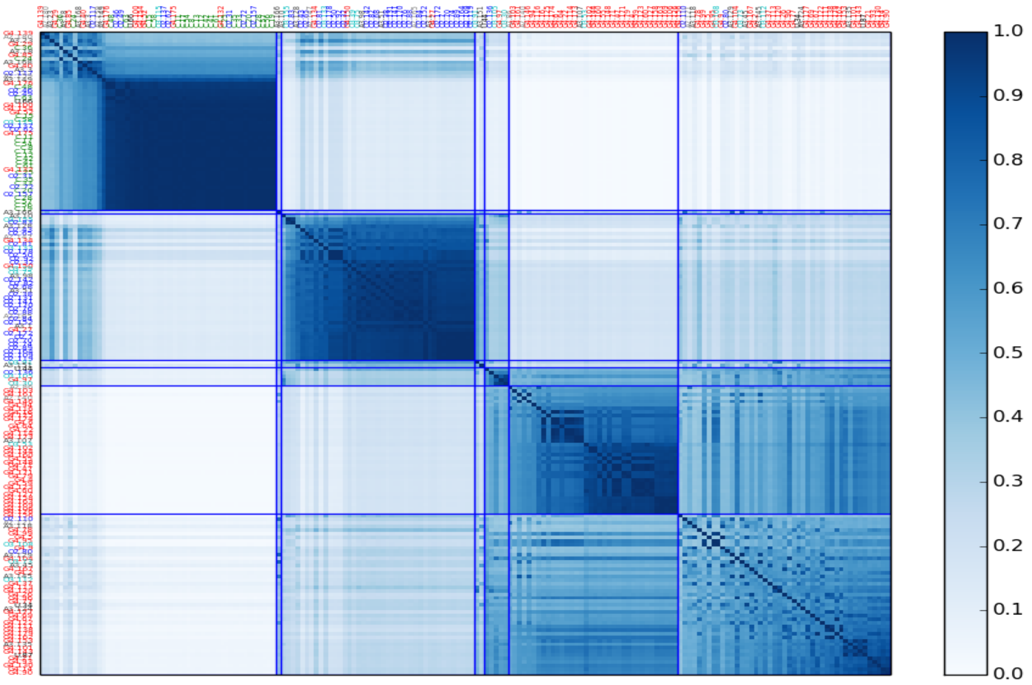


Figure 24 Membership coefficients for the clustering methods (RAW-JS)

4) *Raw, Column-Normalization, Jensen-Shannon, N Iterations=100 to 400*

Iterations	Raw, Column-Normalization, Jensen-Shannon
100	
200	

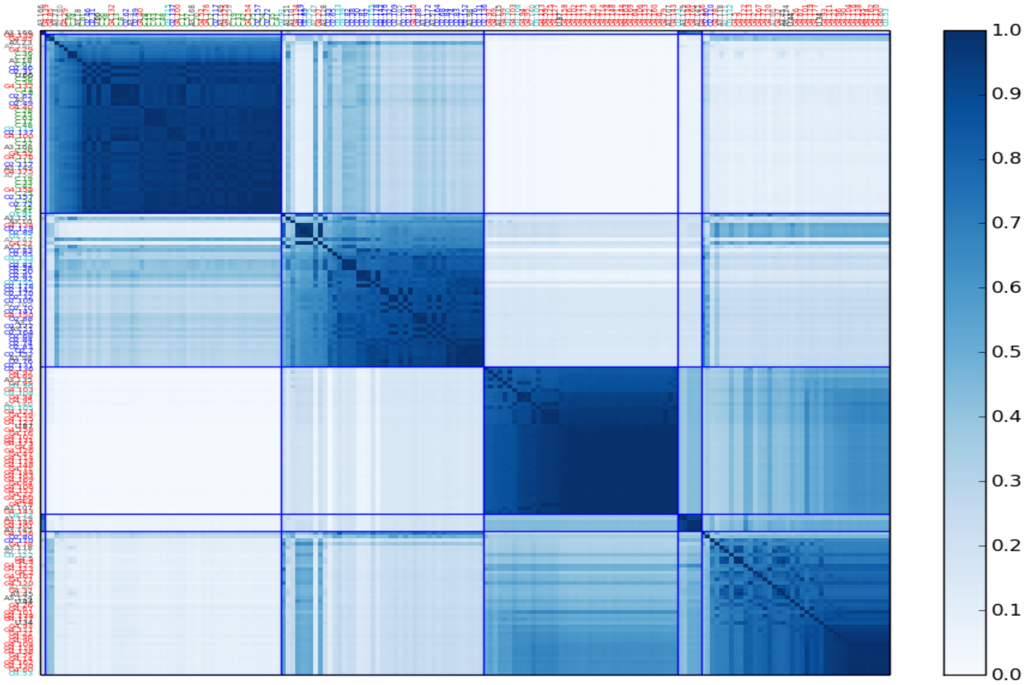
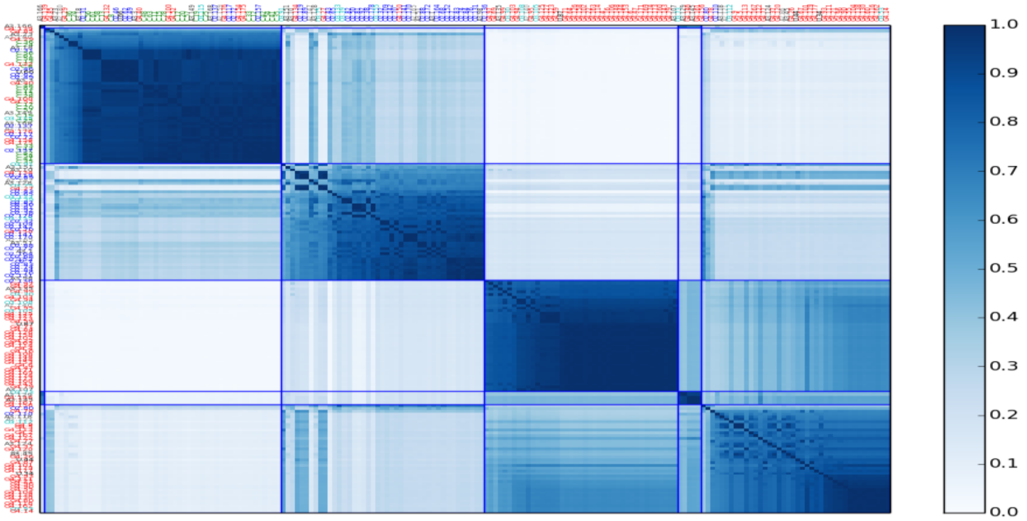
300	
400	

Table 32 Partition results for N Iterations=100 to 400, Raw Dataset, column normalization, Jensen-Shannon

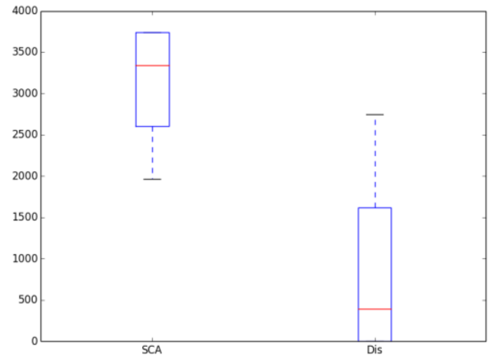
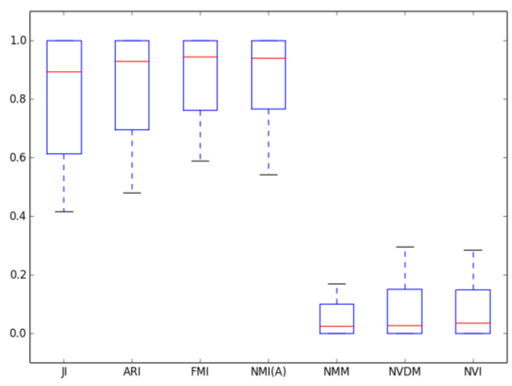
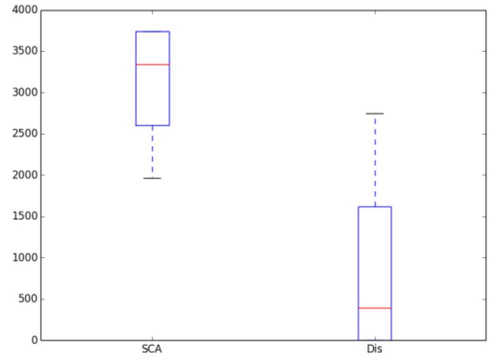
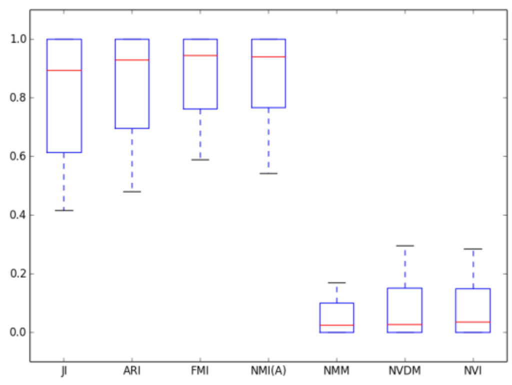
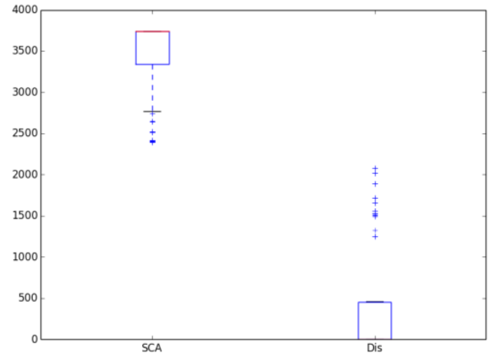
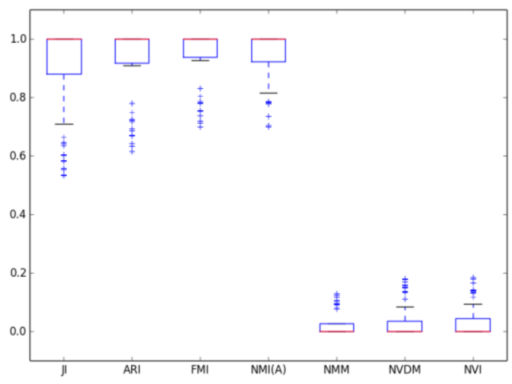
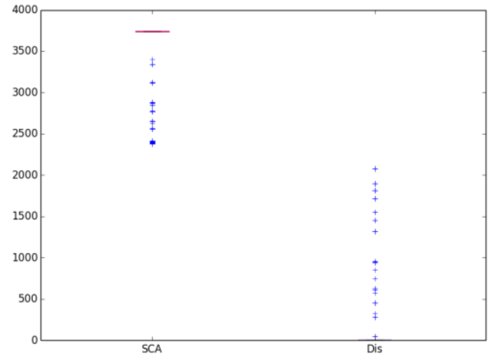
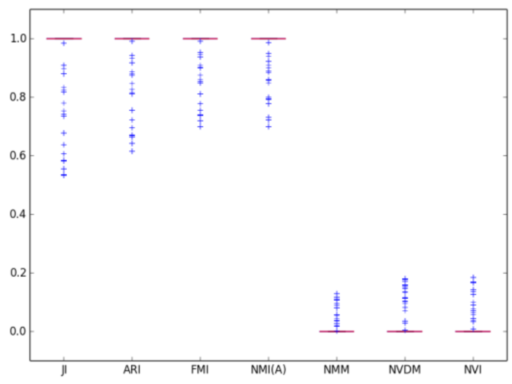
N Iterations	Agreements	Indexes
100		
200		
300		
400		

Table 33 Agreements and Indexes for Raw Jensen-Shannon data

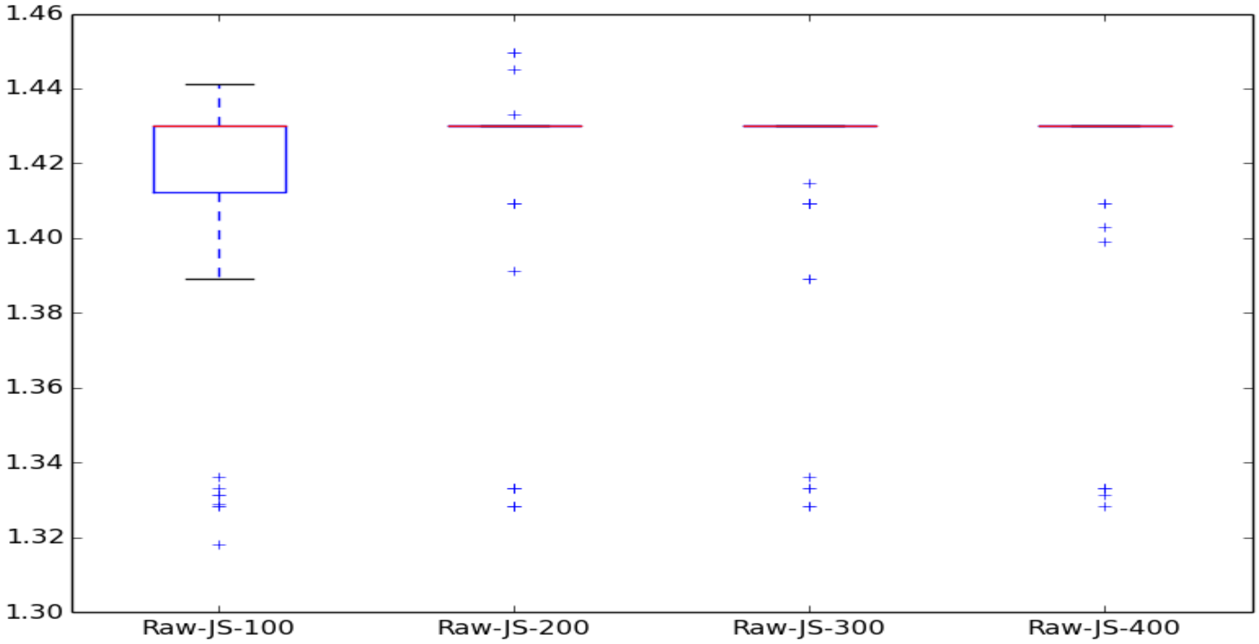
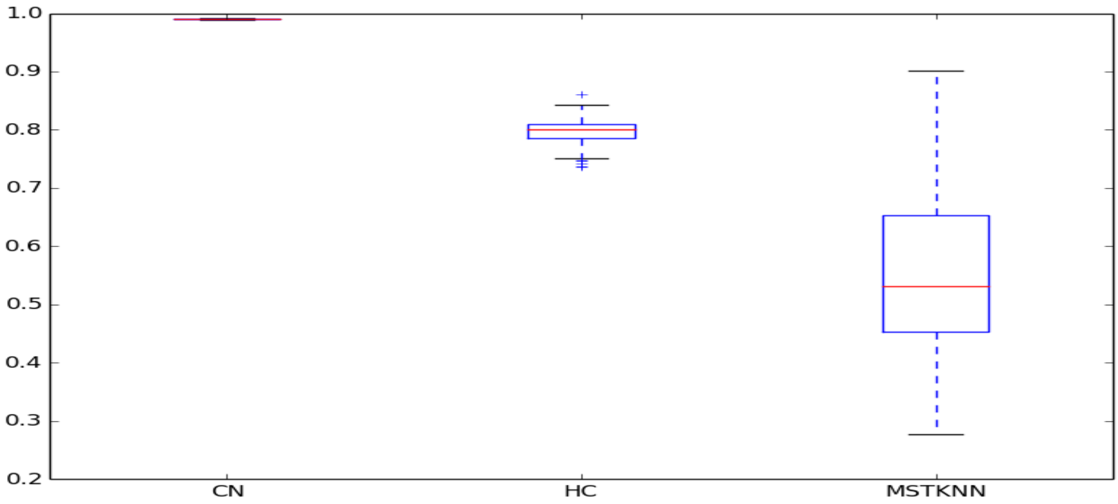


Figure 25 Accuracy for different Nr of Iterations for Raw Jensen-Shannon settings

Iterations	Raw, Column-Normalization, Jensen-Shannon
100	

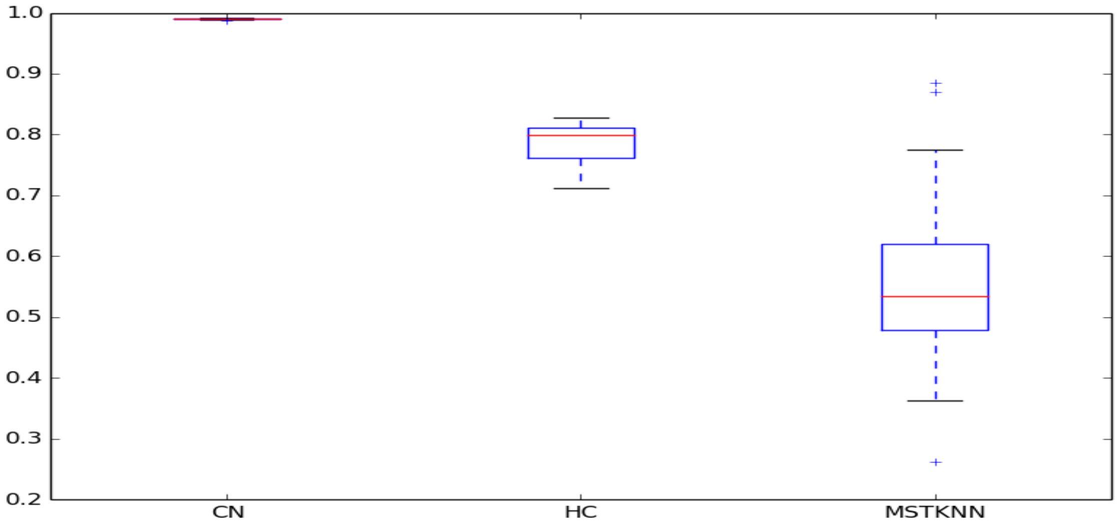
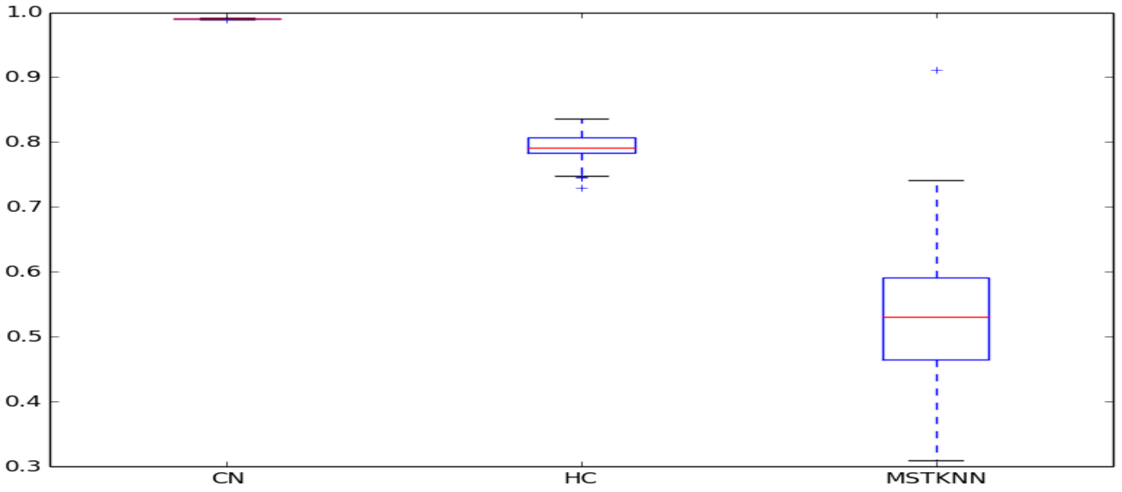
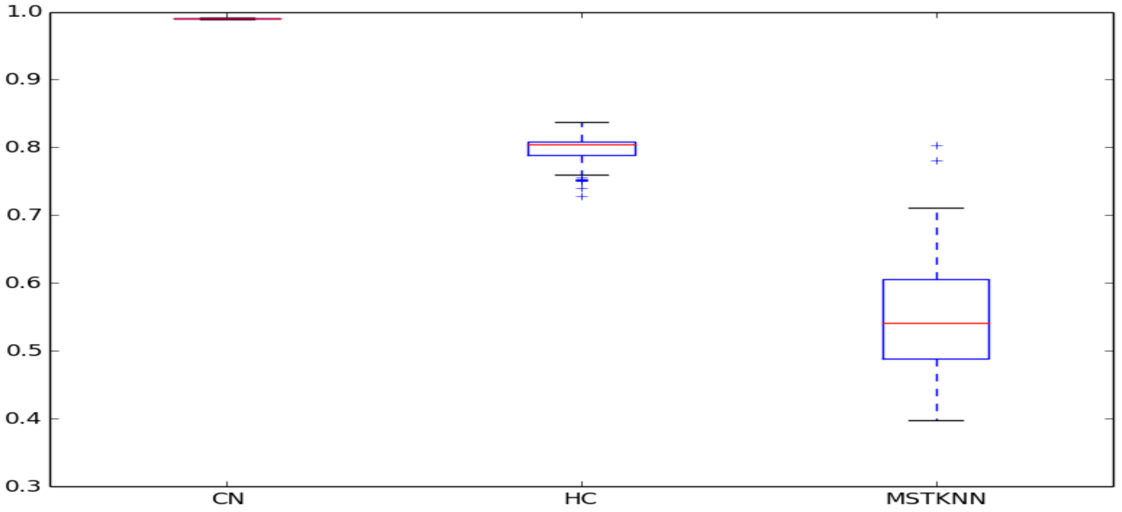
200	
300	
400	

Table 34 Membership coefficients for the different clustering algorithms

5) *Comparison of Accuracy for all methods*

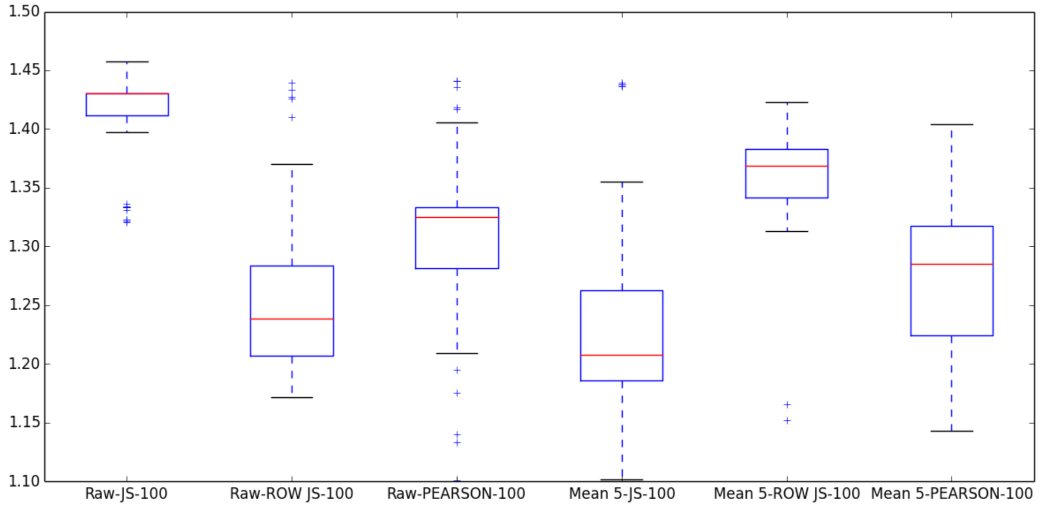
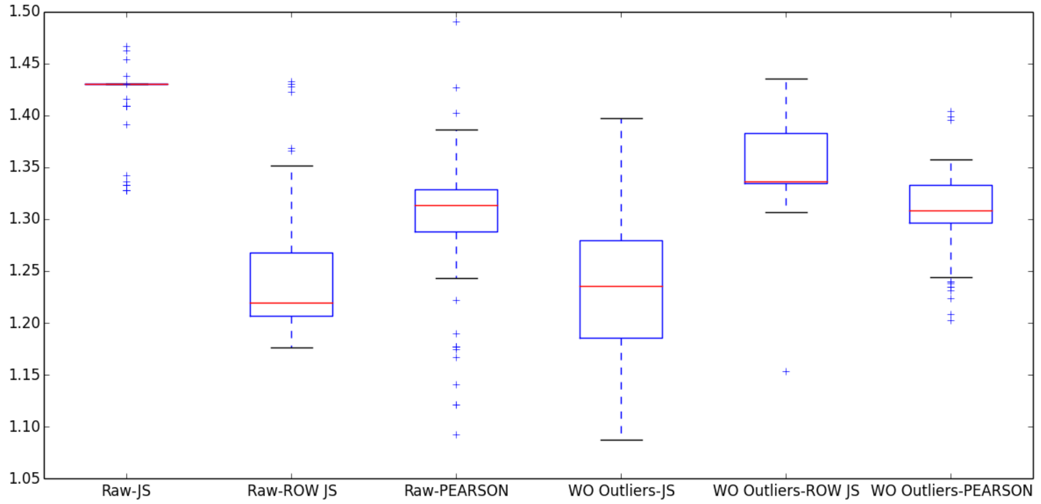
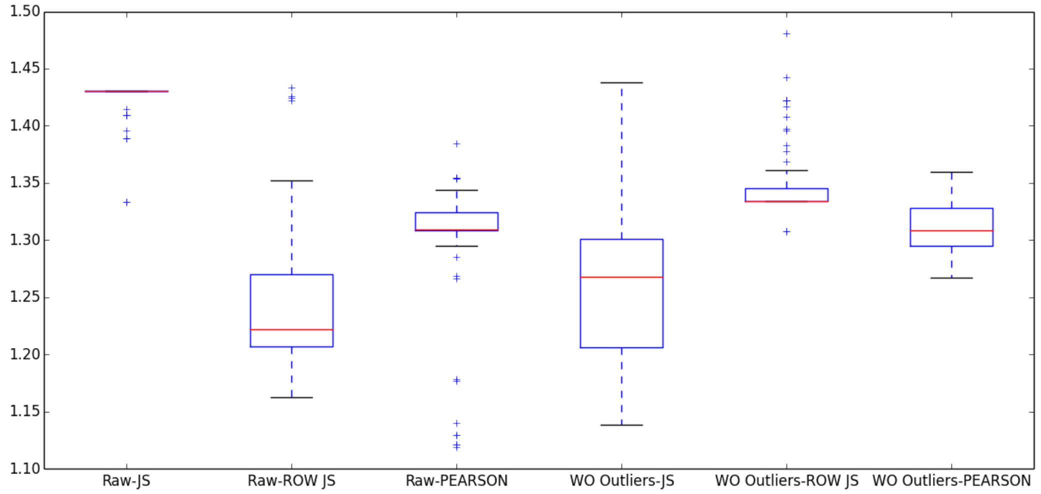
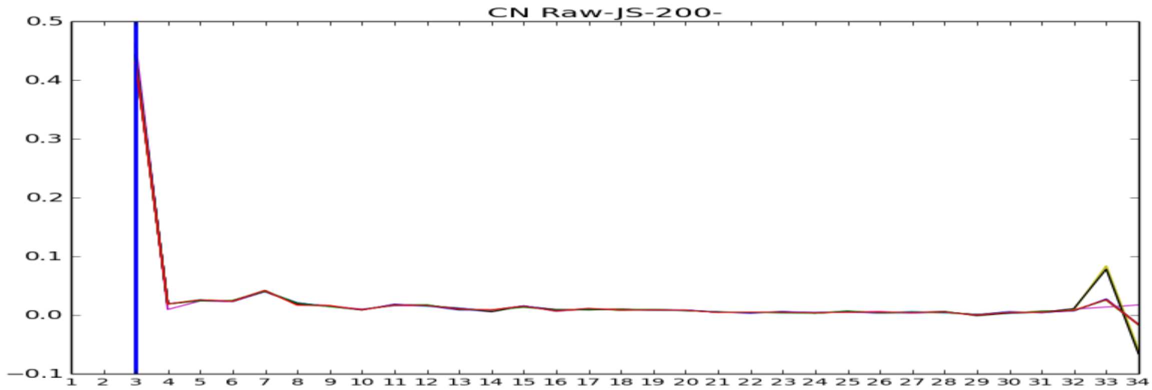
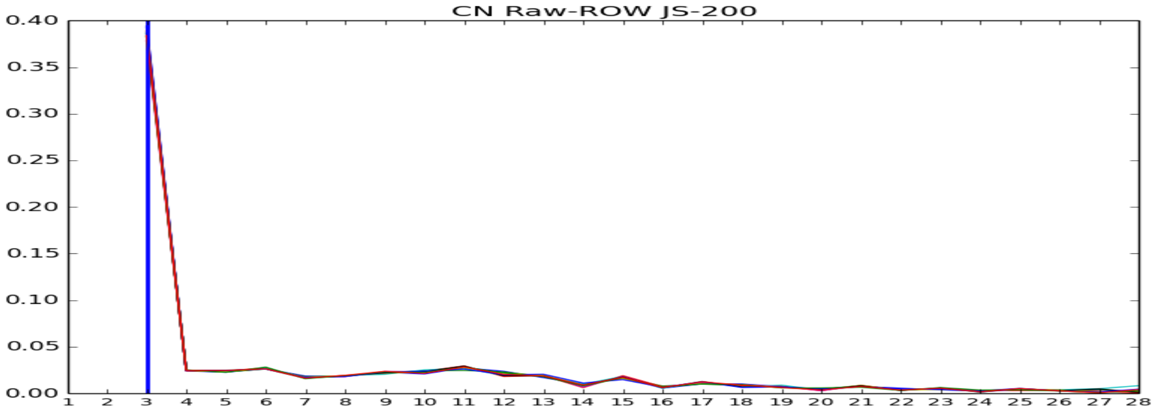
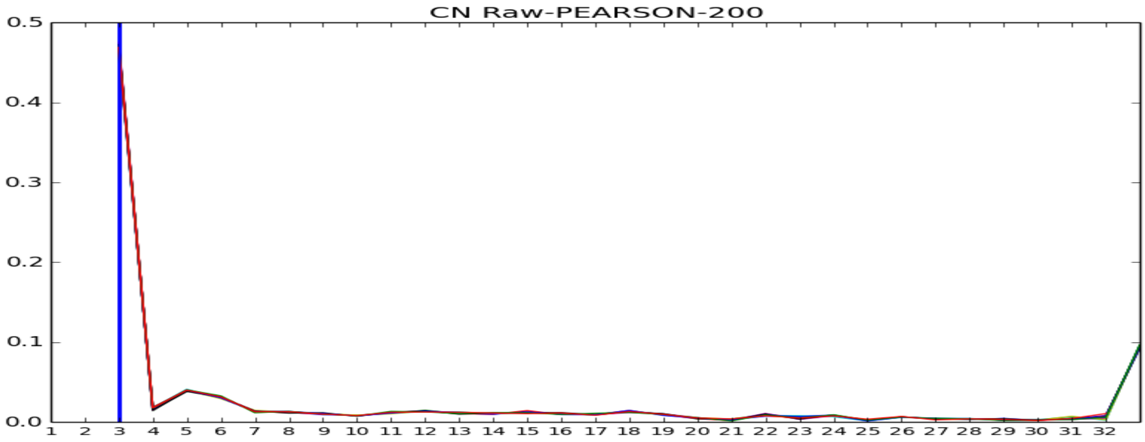
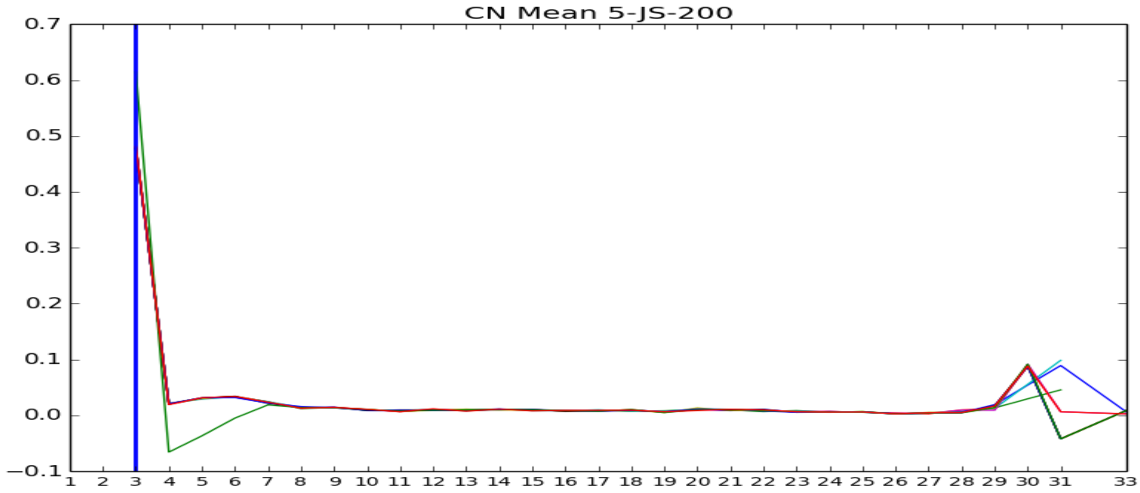
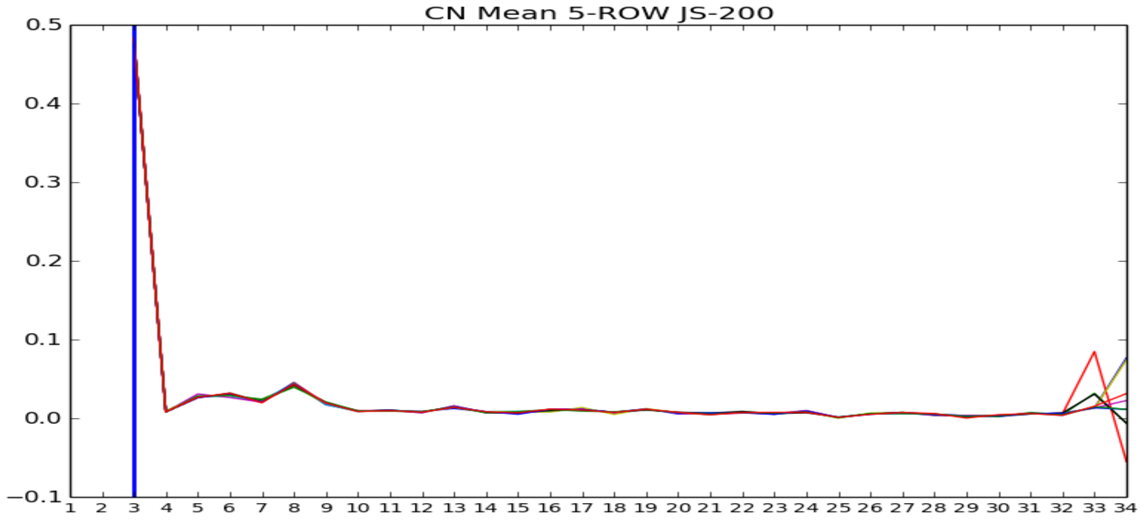
Iterations	All methods
100	
200	
500	

Table 35 Evolution of accuracy convergence for different N Iterations

6) *Selection of Best K*

In order to show the variability of the Best K for each algorithm 10 executions have been performed and the AUC and best K plotted for each of them. The thickness of the vertical lines is proportional to the number of times the K has been the one with best AUC improvement. In some occasions the area for K=2 is not represented as it corresponds to ∞ an increase of as it compares to 0. The evolution of the AUC for each execution is shown in a different color helping to have an idea of the stability of the process.

	CN
Raw JS	
Raw Row JS	

Raw Pearson	
Outliers JS	
Outliers Row JS	

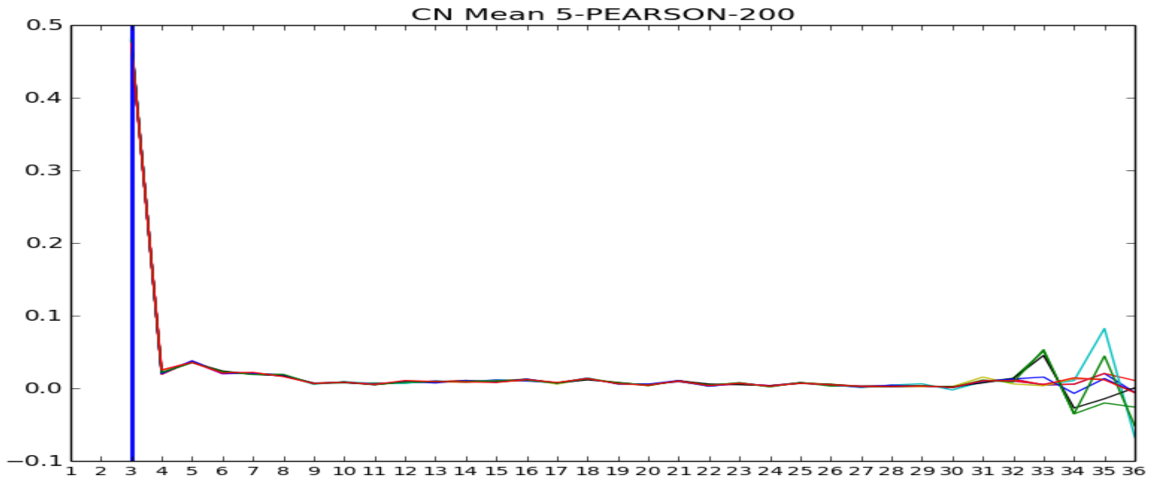
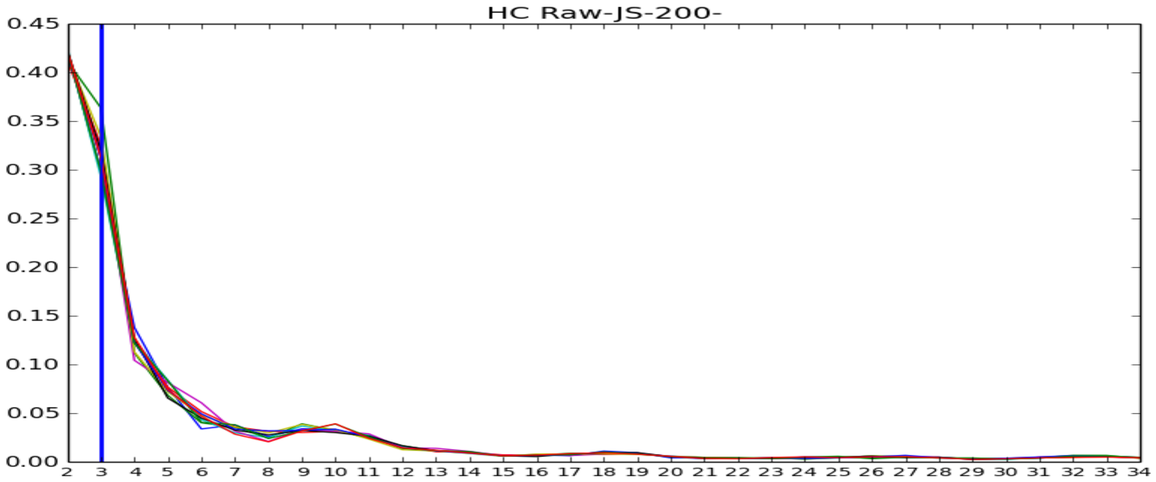
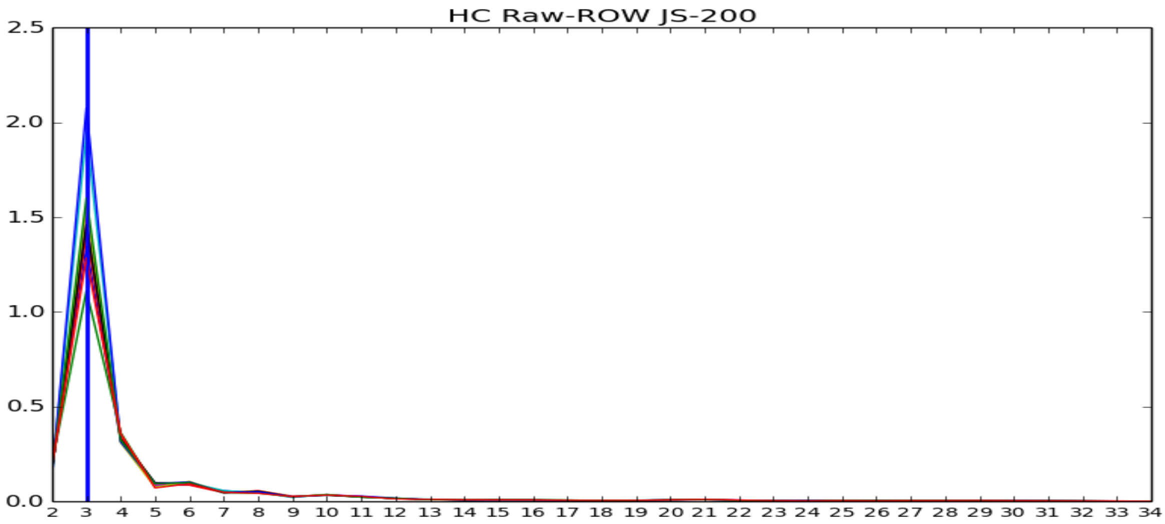
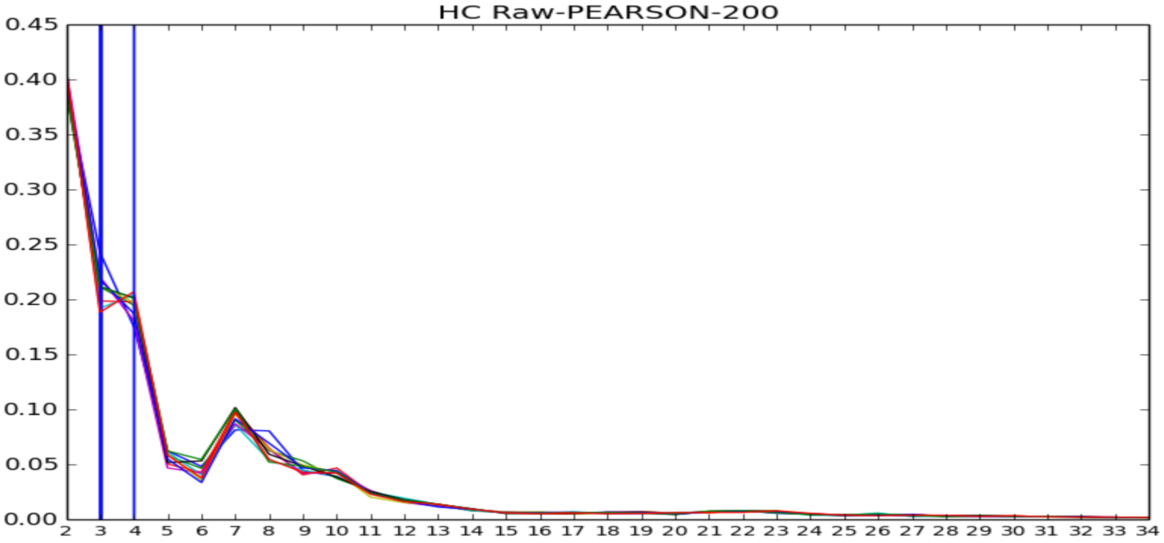
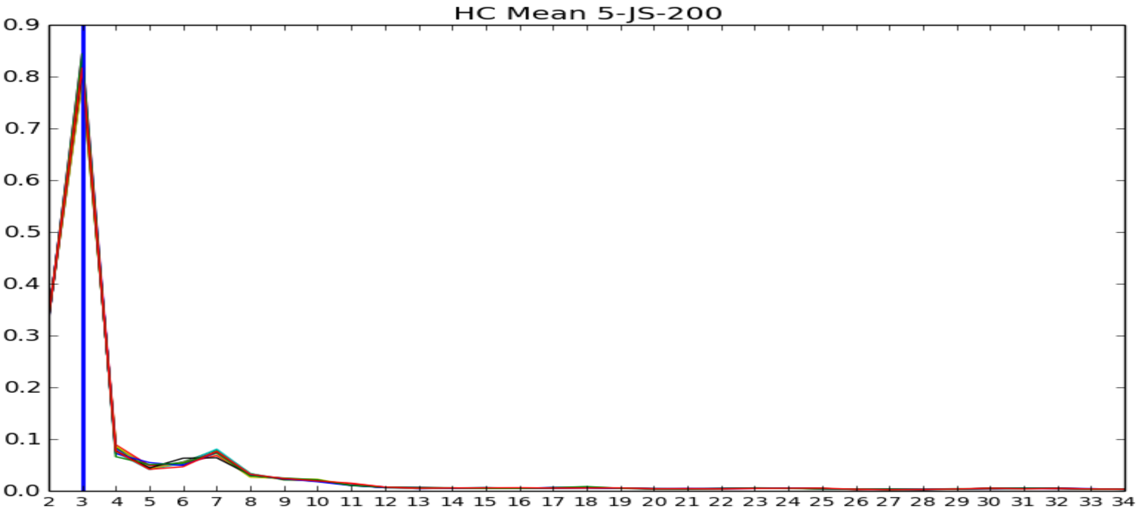
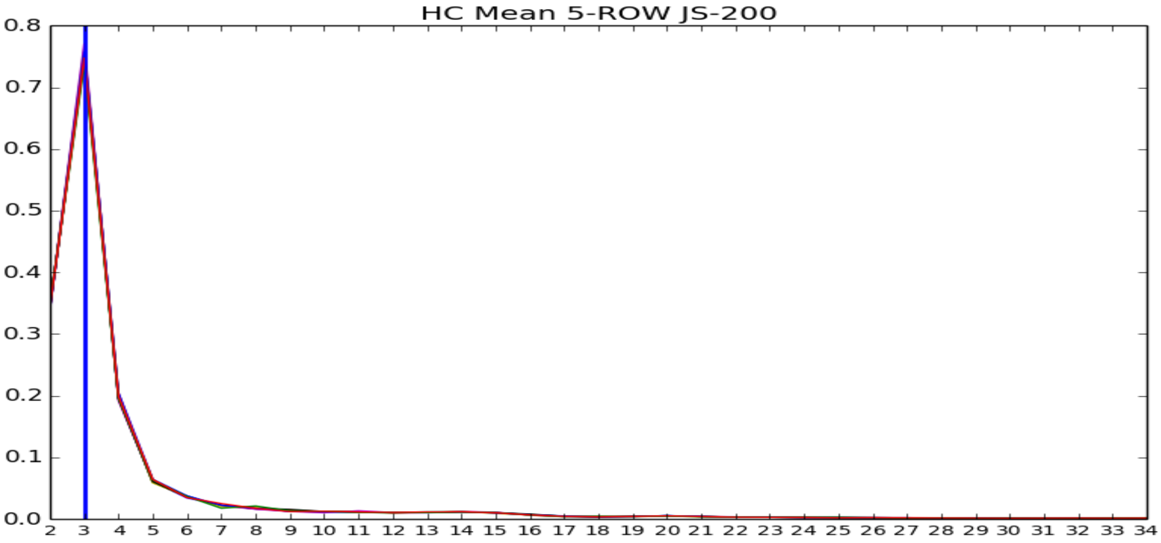
Outliers Pearson	

Table 36 Selection of best K for CN

	HC
Raw JS	
Raw Row JS	

Raw Pearson	
Outliers JS	
Outliers Row JS	

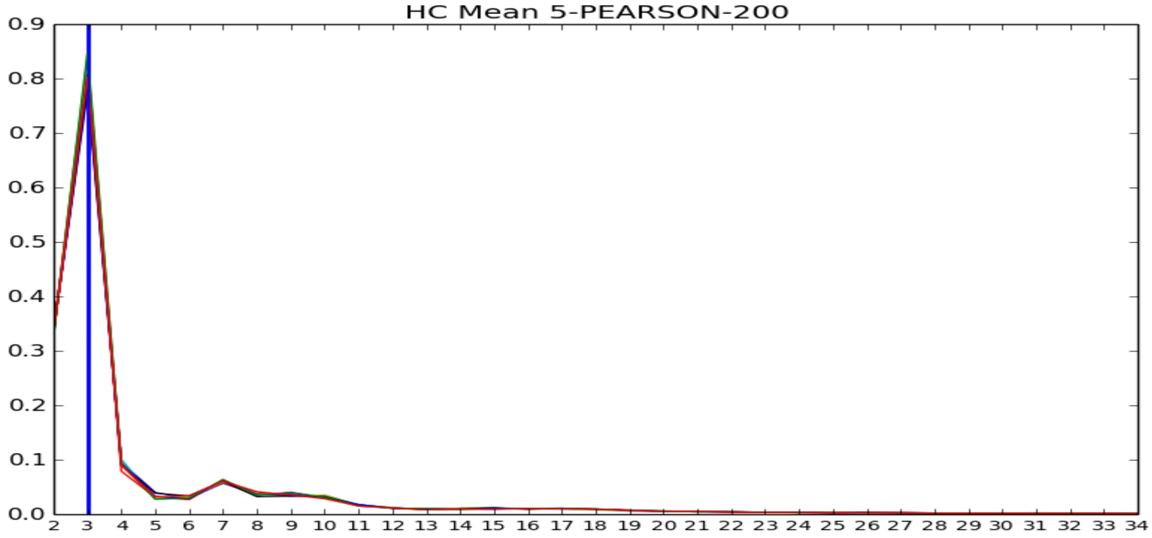
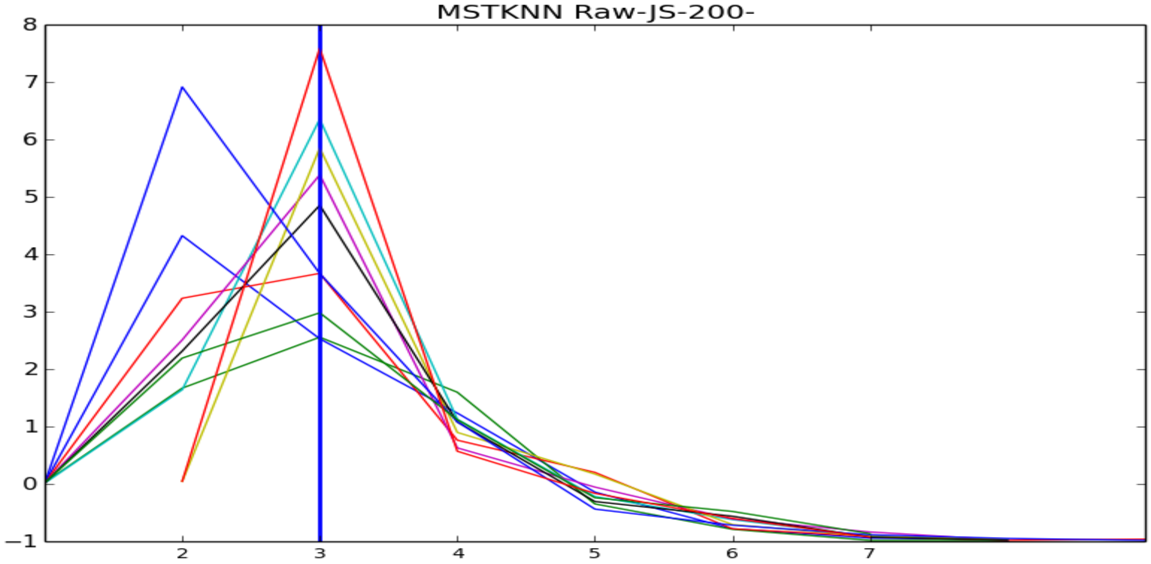
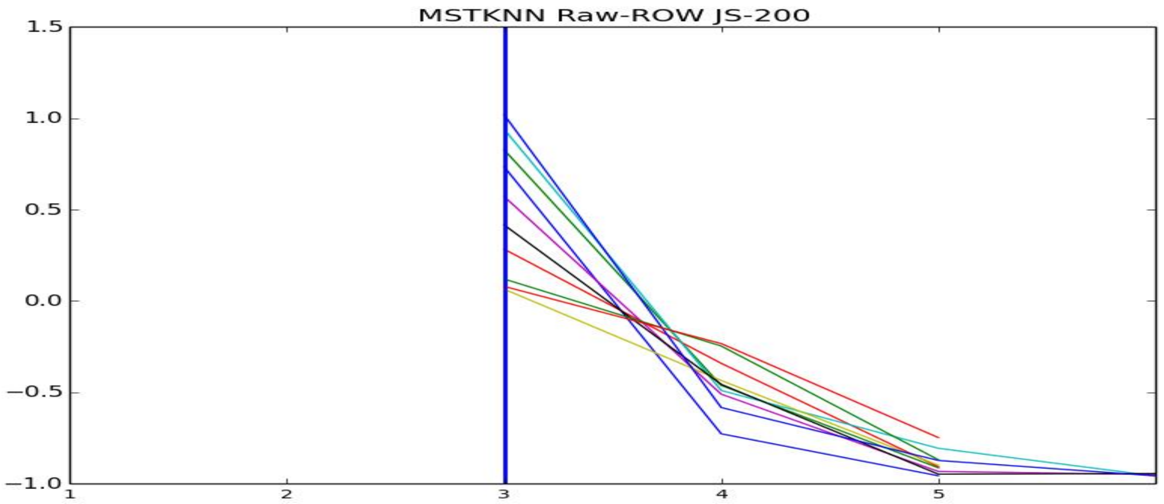
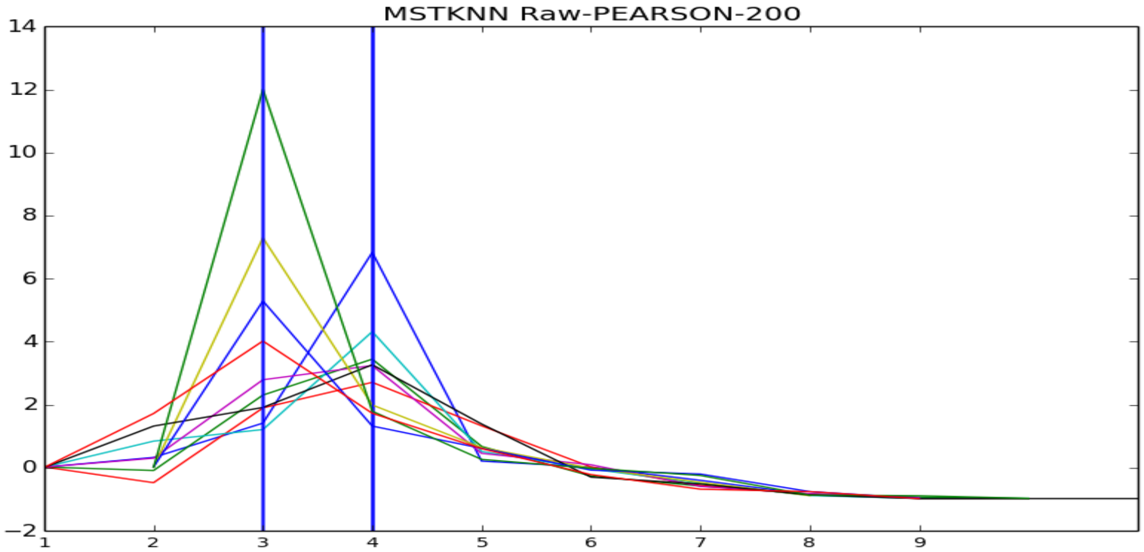
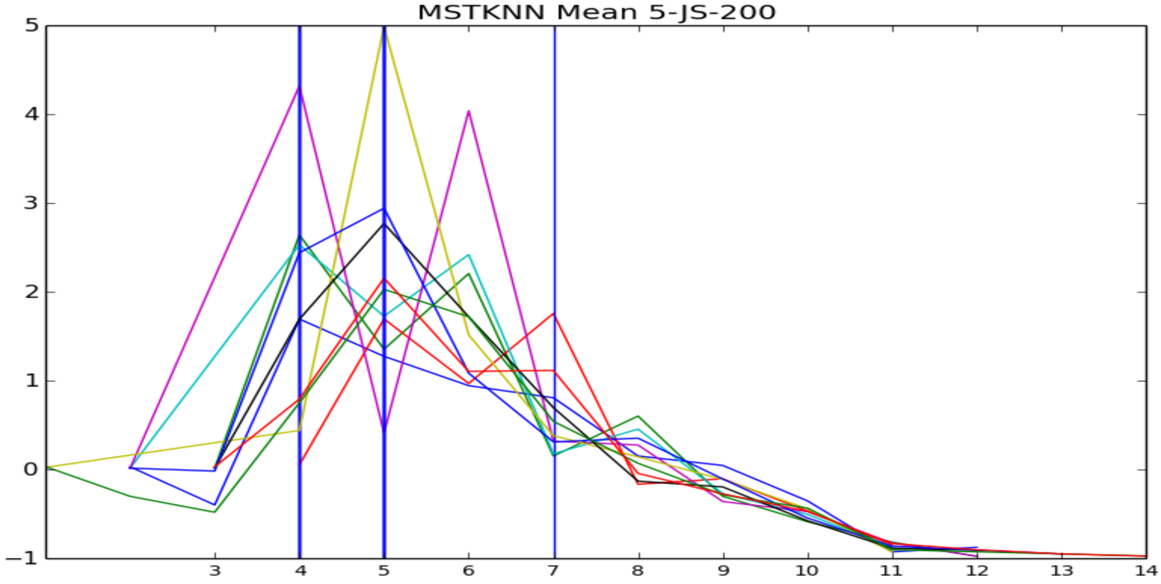
Outliers Pearson	

Table 37 Selection of best K for HC

	MSTKNN
Raw JS	

Raw Row JS	 MSTKNN Raw-ROW JS-200
Raw Pearson	 MSTKNN Raw-PEARSON-200
Outliers JS	 MSTKNN Mean 5-JS-200

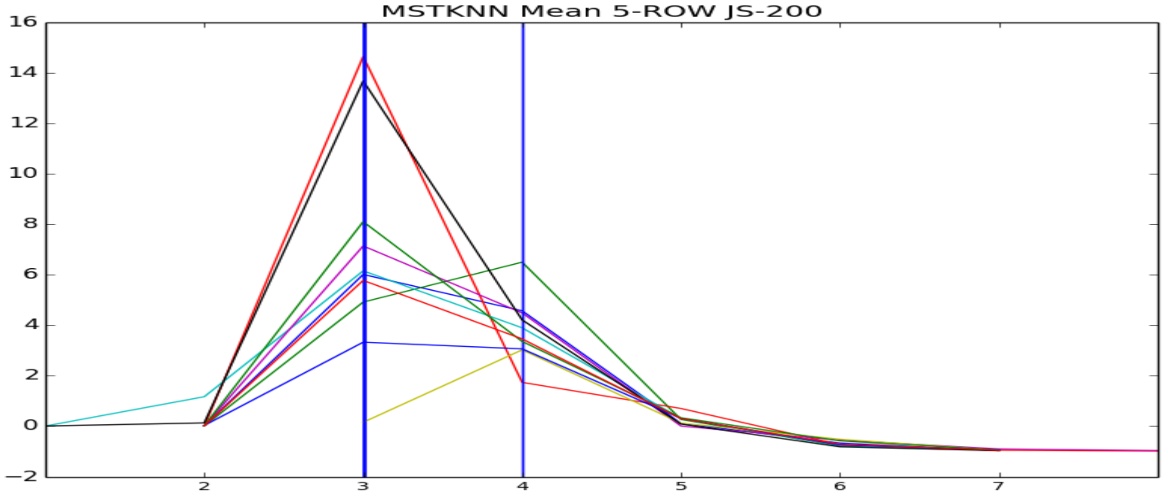
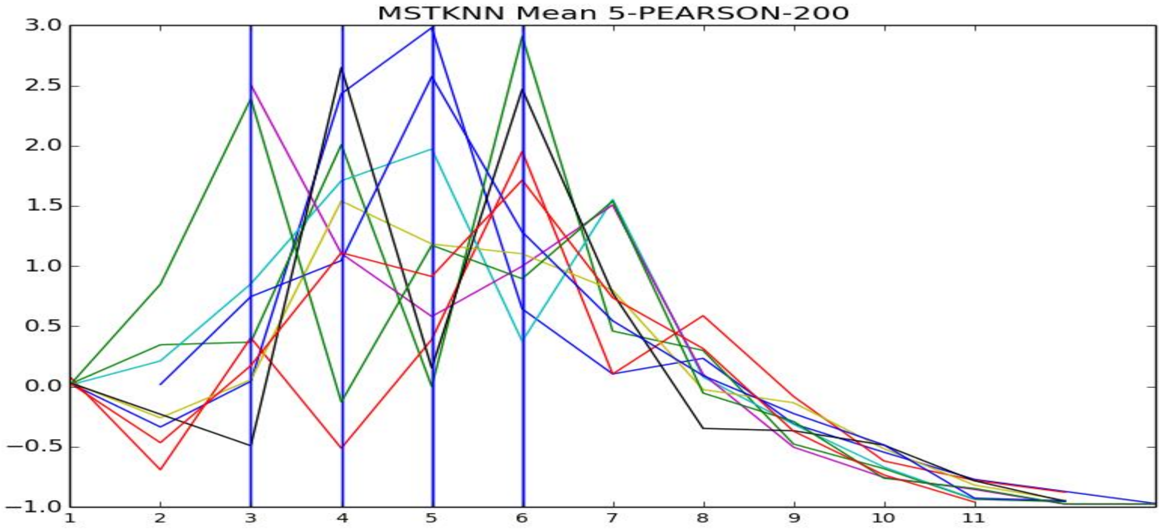
<table>
<tr><td>Outliers Row JS</td><td>
</td></tr>
<tr><td>Outliers Pearson</td><td>
</td></tr>
</table>

Table 38 Selection of best K for MSTKNN

7) *Solution variability for each algorithm for RAW-JS configuration*

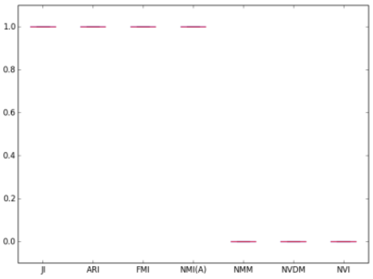
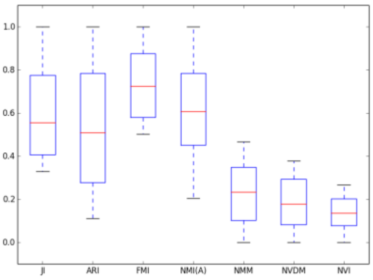
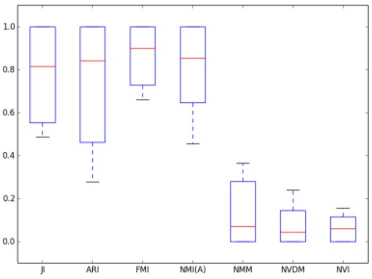
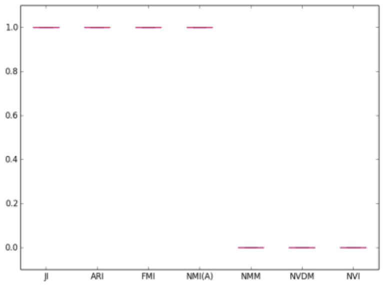
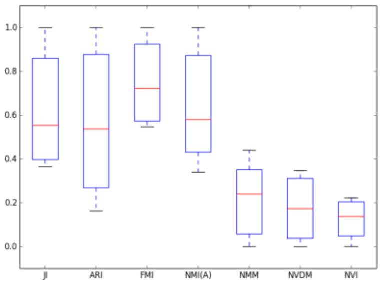
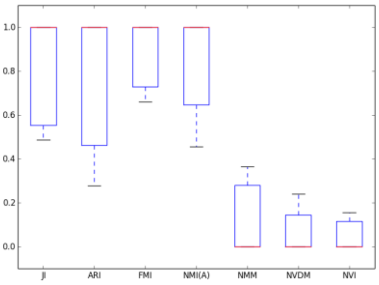
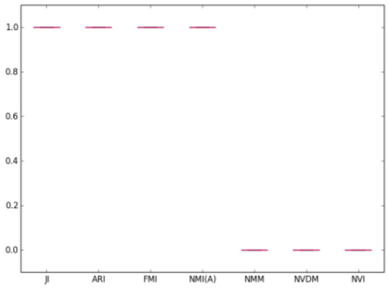
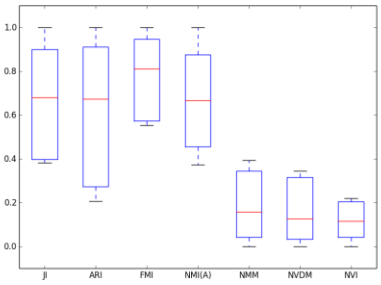
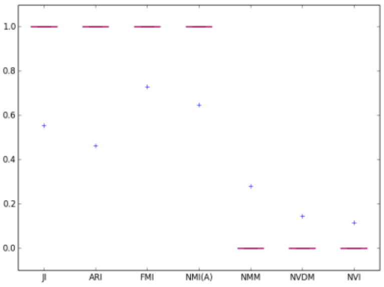
	CN	HC	MSTKNN
100			
200			
500			

Table 39 Variability of the solutions for individual algorithms in RAW-JS configuration

8) *Solution variability for each algorithm for WO Outliers-JS*

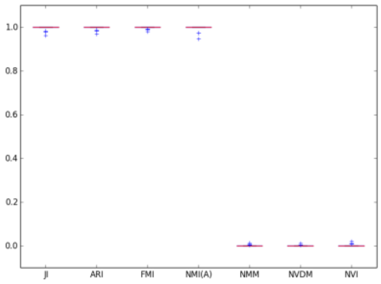
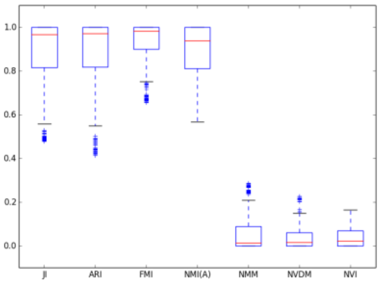
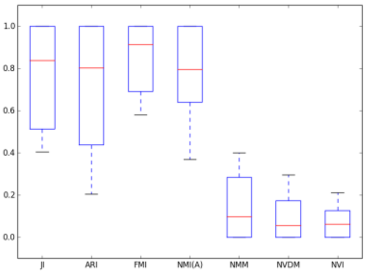
	CN	HC	MSTKNN
500			

Table 40 Variability of the solutions for individual algorithms in WO Outliers-Row-JS configuration

C. Feature selection using CM1

1) Probeset to Gene matching from CM1 extracted features

Table 44, Table 45 and Table 46 show the list of features obtained from the CM1 indicator for the main scenarios in the study, 1-VS-all for all the classes and 1-vs-1 for the two Glioblastoma types detected, over-expressed (green), under-expressed(red).

The initial list of 100 probesets for case (50 over-expressed, 50 under-expressed) has been mapped to their corresponding genes using the SOURCE public database in a many-to-1 process (more than one probeset correspond to the same gene):

http://smd.princeton.edu/cgi-bin/source/sourceBatchSearch

From the whole list of probesets shown in the aforementioned tables, only 26 unique probesets, appearing 47 times (repetitions), are not found on the SOURCE database, Table 41.

213841_at	232541_at	218094_s_at	215963_x_at	213828_x_at
200869_at	224999_at	1554007_at	227984_at	200834_s_at
221798_x_at	201522_x_at	209312_x_at	1569872_a_at	204018_x_at
234989_at	229606_at	215193_x_at	211927_x_at	209458_x_at
216438_s_at	211458_s_at	213158_at	211940_x_at	211745_x_at

Table 41 Probes IDs not found on SOURCE database

If the found genes are consolidated to eliminate duplicate assignments, we obtain the lists of genes in Table 47 and Table 48, where the name in brackets shows the number of genes in the column, respectively for up-regulated and down-regulated.

It is also relevant to mention that despite CM1 not being a symmetric indicator, the Genes obtained for G1X-vs-G2X up-regulated match completely the G2X-vs-G1X down-regulated and vice versa, columns 5 and 6, compared crossway, in Table 47 and Table 48. Table 49 summarizes the number of unique genes per class, no duplication of probesets since CM1 score is unique for each probeset and also the "Not found" probesets for each class. The low total indicates high coherence of the results.

If all the genes for 1-vs-all are consolidated, a total of 165 genes are obtained, Table 42, the G1X-vs-G2X contribute 29 additional unique genes, Table 43, up to 194 genes. The maximum theoretical is 453, so it represents a ratio of appearance of 2.33 times per gene, what indicates a very compact set of genes.

MBP	NDRG2	MLLT11	SYN2	PTPRO	VEGFA	YWHAG	GNB2L1	IGFBP3	RPL34	OLIG2
PTGDS	CRYAB	SPARCL1	GPM6A	NGFRAP1	FTL	MAOB	RPL13A	PTN	RPS3A	SCD5
PLP1	KIF5C	FAM123A	EEF1A2	AK5	FABP5	UCHL1	RPLP2	AQP1	RPL7	SMOC1
FAM107A	NRGN	SLC1A2	TSC22D1	CHI3L1	CD63	HLA-C	RPL39	LAPTM5	RPL3	LRIG1
RTN1	SCD	TSPAN7	GABBR2	TMSB10	GAP43	MT3	RPS16	LTF	PABPC1	BCAN
FBXL16	STXBP1	SOX8	RGS4	LDHA	FTH1	HLA-B	EEF1A1	A2M	ZBTB20	C1orf61
TF	GRIA2	CAMK2N1	MAP2	SERPINA3	PGK1	HLA-A	RPS15A	CLU	MARCKS	OLIG1
SNAP25	OLFM1	RTN3	FXYD6	SPP1	COL1A1	RPLP0	IGFBP7	HLA-DPA1	VCAN	NME7
BASP1	MAP1A	CALM1	NAPB	VIM	CST3	RPS2	RPS11	SOD2	ID4	PRKACB
SYT1	ADD3	ATP1B1	DNM1	FABP7	S100A6	RPS19	RPLP1	CD44	RPS3	PEG3
NTRK2	C7orf41	PEA15	NPTN	TIMP1	IGFBP5	RPL13	CD99	CTSB	RPS6	IDS
ALDOC	QDPR	GABBR1	TUBB2A	IGFBP2	LGALS3	SPARC	RPL27A	C1QC	TCF12	SLC17A7
VSNL1	EDIL3	BEX1	NDRG4	LGALS1	ANXA1	PTPRZ1	ACTG1	HLA-DRB1	APOE	MDH1
PLEKHB1	GPRC5B	NEFL	AGXT2L1	SEC61G	ACTN1	EGFR	MT2A	CD74	HEY1	YWHAH
CHN1	STMN2	ENC1	CPE	FN1	COL4A1	B2M	POSTN	C3	RPL30	SERPINI1

Table 42 Unique genes 1-vs-All

TRIB2	HSPA8	LANCL2	COL6A1	APOD	PFN2	ENPP2	SEPP1
GNAS	NES	EGR1	MALAT1	GSN	DNER	RNASE1	
VOPP1	PPIA	S100A16	JAG1	SLC44A1	CD24	CLDN11	
HOPX	DBI	TMEM158	TRIO	CLDND1	CNP	PSAP	

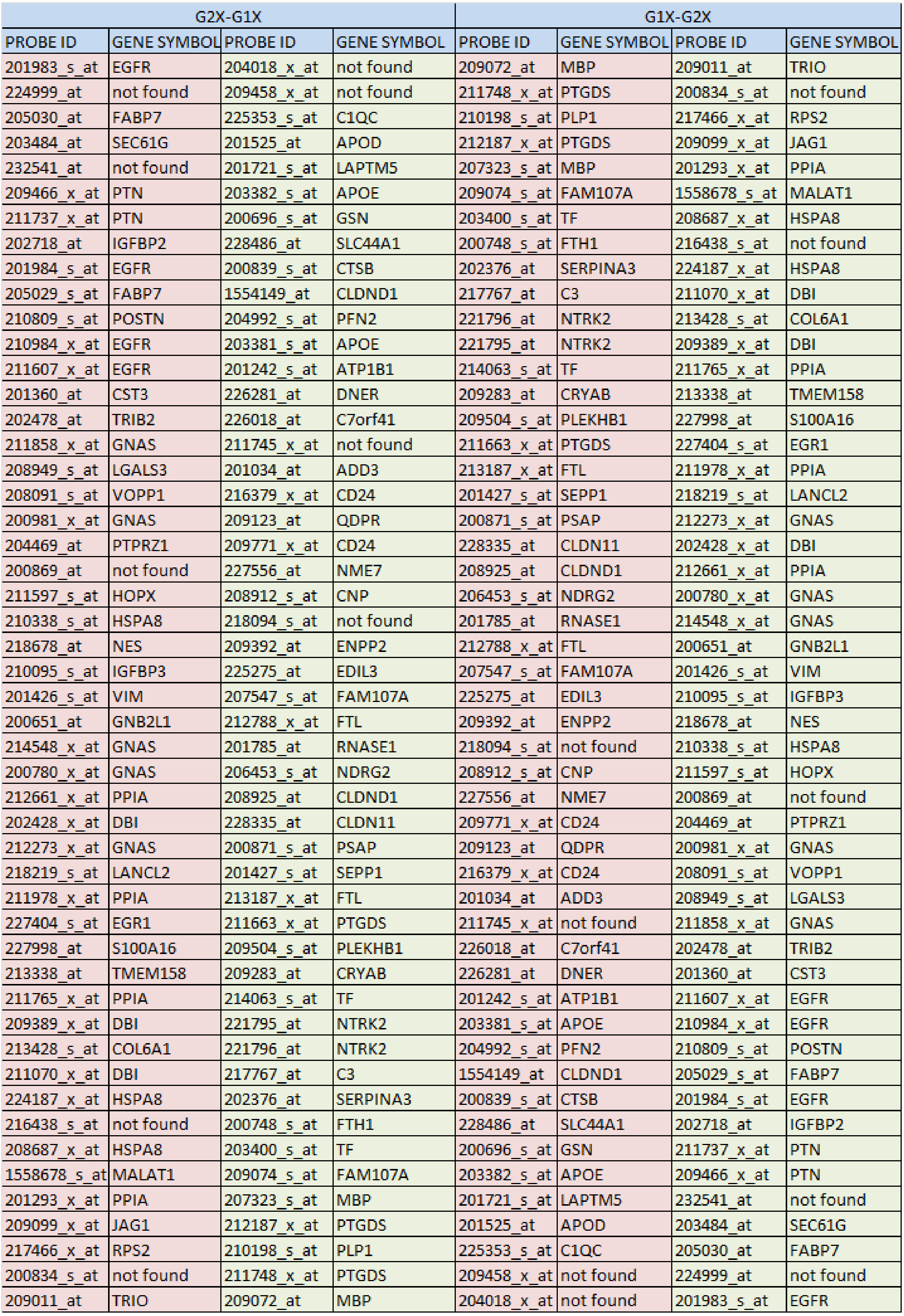
Table 43 Unique Gene IDs in G1X-vs-G2X, G2X-vs-G1X not in 1-VS-all

	G1X		G2X		OX		CX	
	PROBE ID	GENE SYMBOL	PROBE ID	GENE SYMBOL	PROBE ID	GENE SYMBOL	PROBE ID	GENE SYMBOL
C	209072_at	MBP	203485_at	RTN1	209395_at	CHI3L1	201426_s_at	VIM
M	211748_x_at	PTGDS	203999_at	SYT1	217733_s_at	TMSB10	211720_x_at	RPLP0
1	210198_s_at	PLP1	203797_at	VSNL1	209396_s_at	CHI3L1	208856_x_at	RPLP0
	212187_x_at	PTGDS	202507_s_at	SNAP25	200650_s_at	LDHA	201033_x_at	RPLP0
U	207323_s_at	MBP	227641_at	FBXL16	202376_at	SERPINA3	212433_x_at	RPS2
N	209074_s_at	FAM107A	210222_s_at	RTN1	209875_s_at	SPP1	213414_s_at	RPS19
D	203485_at	RTN1	204081_at	NRGN	201426_s_at	VIM	203107_x_at	RPS2
E	227641_at	FBXL16	211985_s_at	CALM1	205030_at	FABP7	221798_x_at	not found
R	203400_s_at	TF	225491_at	SLC1A2	201666_at	TIMP1	214351_x_at	RPL13
E	202508_s_at	SNAP25	212624_s_at	CHN1	234989_at	not found	200665_s_at	SPARC
X	202391_at	BASP1	202508_s_at	SNAP25	202718_at	IGFBP2	211972_x_at	RPLP0
P	203999_at	SYT1	202022_at	ALDOC	216438_s_at	not found	202649_x_at	RPS19
R	221796_at	NTRK2	202391_at	BASP1	201105_at	LGALS1	215313_x_at	HLA-A
E	202022_at	ALDOC	203146_s_at	GABBR1	203484_at	SEC61G	209395_at	CHI3L1
S	203797_at	VSNL1	202260_s_at	STXBP1	211719_x_at	FN1	204469_at	PTPRZ1
S	209504_s_at	PLEKHB1	218332_at	BEX1	216442_x_at	FN1	209396_s_at	CHI3L1
E	221795_at	NTRK2	205591_at	OLFM1	210512_s_at	VEGFA	201983_s_at	EGFR
D	212624_s_at	CHN1	221805_at	NEFL	204141_at	TUBB2A	216231_s_at	B2M
	202507_s_at	SNAP25	211984_at	CALM1	210495_x_at	FN1	202376_at	SERPINA3
G	206453_s_at	NDRG2	202242_at	TSPAN7	218309_at	CAMK2N1	200869_at	not found
E	209283_at	CRYAB	203000_at	STMN2	213187_x_at	FTL	200651_at	GNB2L1
N	203130_s_at	KIF5C	201341_at	ENC1	202345_s_at	FABP5	210646_x_at	RPL13A
E	214063_s_at	TF	229039_at	SYN2	200663_at	CD63	200909_s_at	RPLP2
S	204081_at	NRGN	219549_s_at	RTN3	212464_s_at	FN1	213932_x_at	HLA-A
	200832_s_at	SCD	209469_at	GPM6A	204471_at	GAP43	208695_s_at	RPL39
	202260_s_at	STXBP1	201522_x_at	not found	201341_at	ENC1	212790_x_at	RPL13A
	211663_x_at	PTGDS	204540_at	EEF1A2	203797_at	VSNL1	226131_s_at	RPS16
	205358_at	GRIA2	215111_s_at	TSC22D1	203999_at	SYT1	209140_x_at	HLA-B
	205591_at	OLFM1	209990_s_at	GABBR2	200748_s_at	FTH1	216526_x_at	HLA-C
	203151_at	MAP1A	229606_at	not found	200738_s_at	PGK1	200716_x_at	RPL13A
	201034_at	ADD3	221916_at	NEFL	1556499_s_at	COL1A1	212734_x_at	RPL13
	226018_at	C7orf41	204337_at	RGS4	212788_x_at	FTL	204892_x_at	EEF1A1
	209123_at	QDPR	225540_at	MAP2	201360_at	CST3	200781_s_at	RPS15A
	225275_at	EDIL3	209470_s_at	GPM6A	217728_at	S100A6	211719_x_at	FN1
	203632_s_at	GPRC5B	217897_at	FXYD6	211959_at	IGFBP5	214459_x_at	HLA-C
	203000_at	STMN2	225111_s_at	NAPB	208949_s_at	LGALS3	211927_x_at	not found
	213841_at	not found	215116_s_at	DNM1	212624_s_at	CHN1	201162_at	IGFBP7
	211071_s_at	MLLT11	218309_at	CAMK2N1	201012_at	ANXA1	211911_x_at	HLA-B
	210222_s_at	RTN1	202228_s_at	NPTN	208636_at	ACTN1	208812_x_at	HLA-C
	200795_at	SPARCL1	204141_at	TUBB2A	211980_at	COL4A1	200031_s_at	RPS11
	230496_at	FAM123A	200832_s_at	SCD	222985_at	YWHAG	200763_s_at	RPLP1
	225491_at	SLC1A2	211458_s_at	not found	204041_at	MAOB	216442_x_at	FN1
	207547_s_at	FAM107A	213841_at	not found	201387_s_at	UCHL1	211940_x_at	not found
	202242_at	TSPAN7	209159_s_at	NDRG4	204081_at	NRGN	201029_s_at	CD99
	226913_s_at	SOX8	221008_s_at	AGXT2L1	216526_x_at	HLA-C	213614_x_at	EEF1A1
	218309_at	CAMK2N1	201116_s_at	CPE	205970_at	MT3	203034_s_at	RPL27A
	219549_s_at	RTN3	211600_at	PTPRO	209140_x_at	HLA-B	201891_s_at	B2M
	211984_at	CALM1	217963_s_at	NGFRAP1	205029_s_at	FABP7	213828_x_at	not found
	201242_s_at	ATP1B1	219308_s_at	AK5	202507_s_at	SNAP25	210495_x_at	FN1
	200788_s_at	PEA15	1554007_at	not found	215313_x_at	HLA-A	212363_x_at	ACTG1

Table 44 CM1 Underexpressed genes, 1-VS-ALL

	G1X		G2X		OX		CX	
	PROBE ID	GENE SYMBOL	PROBE ID	GENE SYMBOL	PROBE ID	GENE SYMBOL	PROBE ID	GENE SYMBOL
C	209875_s_at	SPP1	211959_at	IGFBP5	200665_s_at	SPARC	227556_at	NME7
M	224585_x_at	ACTG1	201163_s_at	IGFBP7	201033_x_at	RPLP0	203798_s_at	VSNL1
1	213214_x_at	ACTG1	201012_at	ANXA1	200026_at	RPL34	202741_at	PRKACB
	200869_at	not found	216231_s_at	B2M	200099_s_at	RPS3A	209242_at	PEG3
O	212185_x_at	MT2A	201162_at	IGFBP7	214680_at	NTRK2	212221_x_at	IDS
V	210809_s_at	POSTN	217733_s_at	TMSB10	208856_x_at	RPLP0	204229_at	SLC17A7
E	208729_x_at	HLA-B	200663_at	CD63	211720_x_at	RPLP0	210222_s_at	RTN1
R	212363_x_at	ACTG1	211911_x_at	HLA-B	200717_x_at	RPL7	209123_at	QDPR
E	221798_x_at	not found	209047_at	AQP1	211666_x_at	RPL3	200978_at	MDH1
X	210512_s_at	VEGFA	201721_s_at	LAPTM5	215157_x_at	PABPC1	204141_at	TUBB2A
P	234989_at	not found	204041_at	MAOB	235308_at	ZBTB20	229606_at	not found
R	203107_x_at	RPS2	1556499_s_at	COL1A1	201670_s_at	MARCKS	204337_at	RGS4
E	210095_s_at	IGFBP3	202018_s_at	LTF	221731_x_at	VCAN	202228_s_at	NPTN
S	211980_at	COL4A1	201105_at	LGALS1	209292_at	ID4	219308_s_at	AK5
S	201984_s_at	EGFR	217757_at	A2M	212391_x_at	RPS3A	201020_at	YWHAH
E	210495_x_at	FN1	208792_s_at	CLU	214351_x_at	RPL13	211458_s_at	not found
D	212464_s_at	FN1	211990_at	HLA-DPA1	217897_at	FXYD6	222985_at	YWHAG
	202649_x_at	RPS19	212185_x_at	MT2A	208692_at	RPS3	205352_at	SERPINI1
G	212433_x_at	RPS2	209312_x_at	not found	201254_x_at	RPS6	225491_at	SLC1A2
E	201105_at	LGALS1	214459_x_at	HLA-C	213158_at	not found	203151_at	MAP1A
N	200650_s_at	LDHA	210512_s_at	VEGFA	208986_at	TCF12	215116_s_at	DNM1
E	208949_s_at	LGALS3	208812_x_at	HLA-C	225897_at	MARCKS	201242_s_at	ATP1B1
S	216442_x_at	FN1	221477_s_at	SOD2	203382_s_at	APOE	225111_s_at	NAPB
	216231_s_at	B2M	215193_x_at	not found	44783_s_at	HEY1	211984_at	CALM1
	217733_s_at	TMSB10	212063_at	CD44	200062_s_at	RPL30	204540_at	EEF1A2
	208812_x_at	HLA-C	200838_at	CTSB	212039_x_at	RPL3	210198_s_at	PLP1
	211719_x_at	FN1	216526_x_at	HLA-C	213825_at	OLIG2	207323_s_at	MBP
	205029_s_at	FABP7	225353_s_at	C1QC	215963_x_at	not found	221916_at	NEFL
	201162_at	IGFBP7	201666_at	TIMP1	211073_x_at	RPL3	201387_s_at	UCHL1
	213932_x_at	HLA-A	213932_x_at	HLA-A	227984_at	not found	203000_at	STMN2
	213414_s_at	RPS19	209140_x_at	HLA-B	205383_s_at	ZBTB20	229039_at	SYN2
	214459_x_at	HLA-C	200839_s_at	CTSB	224901_at	SCD5	203130_s_at	KIF5C
	209140_x_at	HLA-B	217728_at	S100A6	222784_at	SMOC1	202391_at	BASP1
	200663_at	CD63	212464_s_at	FN1	200787_s_at	PEA15	211985_s_at	CALM1
	211911_x_at	HLA-B	208306_x_at	HLA-DRB1	202022_at	ALDOC	201341_at	ENC1
	216438_s_at	not found	209619_at	CD74	221795_at	NTRK2	221805_at	NEFL
	216526_x_at	HLA-C	215313_x_at	HLA-A	1569872_a_at	not found	203485_at	RTN1
	215313_x_at	HLA-A	210495_x_at	FN1	201217_x_at	RPL3	202260_s_at	STXBP1
	232541_at	not found	213187_x_at	FTL	211596_s_at	LRIG1	212187_x_at	PTGDS
	201666_at	TIMP1	212788_x_at	FTL	203381_s_at	APOE	205591_at	OLFM1
	211737_x_at	PTN	234989_at	not found	221796_at	NTRK2	218309_at	CAMK2N1
	209466_x_at	PTN	216442_x_at	FN1	200788_s_at	PEA15	202508_s_at	SNAP25
	224999_at	not found	200748_s_at	FTH1	209291_at	ID4	211748_x_at	PTGDS
	202718_at	IGFBP2	211719_x_at	FN1	212667_at	SPARC	209072_at	MBP
	209396_s_at	CHI3L1	217767_at	C3	219107_at	BCAN	204081_at	NRGN
	203484_at	SEC61G	201426_s_at	VIM	213841_at	not found	212624_s_at	CHN1
	205030_at	FABP7	209396_s_at	CHI3L1	205103_at	C1orf61	202507_s_at	SNAP25
	209395_at	CHI3L1	209395_at	CHI3L1	228170_at	OLIG1	227641_at	FBXL16
	201983_s_at	EGFR	209875_s_at	SPP1	209283_at	CRYAB	203797_at	VSNL1
	201426_s_at	VIM	202376_at	SERPINA3	226913_s_at	SOX8	203999_at	SYT1

Table 45 CM1 Overexpressed genes, 1-VS-ALL



G2X-G1X				G1X-G2X			
PROBE ID	GENE SYMBOL	PROBE ID	GENE SYMBOL	PROBE ID	GENE SYMBOL	PROBE ID	GENE SYMBOL
201983_s_at	EGFR	204018_x_at	not found	209072_at	MBP	209011_at	TRIO
224999_at	not found	209458_x_at	not found	211748_x_at	PTGDS	200834_s_at	not found
205030_at	FABP7	225353_s_at	C1QC	210198_s_at	PLP1	217466_x_at	RPS2
203484_at	SEC61G	201525_at	APOD	212187_x_at	PTGDS	209099_x_at	JAG1
232541_at	not found	201721_s_at	LAPTM5	207323_s_at	MBP	201293_x_at	PPIA
209466_x_at	PTN	203382_s_at	APOE	209074_s_at	FAM107A	1558678_s_at	MALAT1
211737_x_at	PTN	200696_s_at	GSN	203400_s_at	TF	208687_x_at	HSPA8
202718_at	IGFBP2	228486_at	SLC44A1	200748_s_at	FTH1	216438_s_at	not found
201984_s_at	EGFR	200839_s_at	CTSB	202376_at	SERPINA3	224187_x_at	HSPA8
205029_s_at	FABP7	1554149_at	CLDND1	217767_at	C3	211070_x_at	DBI
210809_s_at	POSTN	204992_s_at	PFN2	221796_at	NTRK2	213428_s_at	COL6A1
210984_x_at	EGFR	203381_s_at	APOE	221795_at	NTRK2	209389_x_at	DBI
211607_x_at	EGFR	201242_s_at	ATP1B1	214063_s_at	TF	211765_x_at	PPIA
201360_at	CST3	226281_at	DNER	209283_at	CRYAB	213338_at	TMEM158
202478_at	TRIB2	226018_at	C7orf41	209504_s_at	PLEKHB1	227998_at	S100A16
211858_x_at	GNAS	211745_x_at	not found	211663_x_at	PTGDS	227404_s_at	EGR1
208949_s_at	LGALS3	201034_at	ADD3	213187_x_at	FTL	211978_x_at	PPIA
208091_s_at	VOPP1	216379_x_at	CD24	201427_s_at	SEPP1	218219_s_at	LANCL2
200981_x_at	GNAS	209123_at	QDPR	200871_s_at	PSAP	212273_x_at	GNAS
204469_at	PTPRZ1	209771_x_at	CD24	228335_at	CLDN11	202428_x_at	DBI
200869_at	not found	227556_at	NME7	208925_at	CLDND1	212661_x_at	PPIA
211597_s_at	HOPX	208912_s_at	CNP	206453_s_at	NDRG2	200780_x_at	GNAS
210338_s_at	HSPA8	218094_s_at	not found	201785_at	RNASE1	214548_x_at	GNAS
218678_at	NES	209392_at	ENPP2	212788_x_at	FTL	200651_at	GNB2L1
210095_s_at	IGFBP3	225275_at	EDIL3	207547_s_at	FAM107A	201426_s_at	VIM
201426_s_at	VIM	207547_s_at	FAM107A	225275_at	EDIL3	210095_s_at	IGFBP3
200651_at	GNB2L1	212788_x_at	FTL	209392_at	ENPP2	218678_at	NES
214548_x_at	GNAS	201785_at	RNASE1	218094_s_at	not found	210338_s_at	HSPA8
200780_x_at	GNAS	206453_s_at	NDRG2	208912_s_at	CNP	211597_s_at	HOPX
212661_x_at	PPIA	208925_at	CLDND1	227556_at	NME7	200869_at	not found
202428_x_at	DBI	228335_at	CLDN11	209771_x_at	CD24	204469_at	PTPRZ1
212273_x_at	GNAS	200871_s_at	PSAP	209123_at	QDPR	200981_x_at	GNAS
218219_s_at	LANCL2	201427_s_at	SEPP1	216379_x_at	CD24	208091_s_at	VOPP1
211978_x_at	PPIA	213187_x_at	FTL	201034_at	ADD3	208949_s_at	LGALS3
227404_s_at	EGR1	211663_x_at	PTGDS	211745_x_at	not found	211858_x_at	GNAS
227998_at	S100A16	209504_s_at	PLEKHB1	226018_at	C7orf41	202478_at	TRIB2
213338_at	TMEM158	209283_at	CRYAB	226281_at	DNER	201360_at	CST3
211765_x_at	PPIA	214063_s_at	TF	201242_s_at	ATP1B1	211607_x_at	EGFR
209389_x_at	DBI	221795_at	NTRK2	203381_s_at	APOE	210984_x_at	EGFR
213428_s_at	COL6A1	221796_at	NTRK2	204992_s_at	PFN2	210809_s_at	POSTN
211070_x_at	DBI	217767_at	C3	1554149_at	CLDND1	205029_s_at	FABP7
224187_x_at	HSPA8	202376_at	SERPINA3	200839_s_at	CTSB	201984_s_at	EGFR
216438_s_at	not found	200748_s_at	FTH1	228486_at	SLC44A1	202718_at	IGFBP2
208687_x_at	HSPA8	203400_s_at	TF	200696_s_at	GSN	211737_x_at	PTN
1558678_s_at	MALAT1	209074_s_at	FAM107A	203382_s_at	APOE	209466_x_at	PTN
201293_x_at	PPIA	207323_s_at	MBP	201721_s_at	LAPTM5	232541_at	not found
209099_x_at	JAG1	212187_x_at	PTGDS	201525_at	APOD	203484_at	SEC61G
217466_x_at	RPS2	210198_s_at	PLP1	225353_s_at	C1QC	205030_at	FABP7
200834_s_at	not found	211748_x_at	PTGDS	209458_x_at	not found	224999_at	not found
209011_at	TRIO	209072_at	MBP	204018_x_at	not found	201983_s_at	EGFR

Table 46 CM1 indicators for 1-vs-1, G1X-vs-G2X

	G1X	G2X	OX	CX	G2X-G1X	G2X-G1X
	GENE SYMBOL (41)	GENE SYMBOL(40)	GENE SYMBOL(42)	GENE SYMBOL(28)	GENESYMBOL (29)	GENESYMBOL (36)
CM1 UNDEREXPRESSED GENES	MBP	RTN1	CHI3L1	VIM	EGFR	MBP
	PTGDS	SYT1	TMSB10	RPLP0	FABP7	PTGDS
	PLP1	VSNL1	LDHA	RPS2	SEC61G	PLP1
	FAM107A	SNAP25	SERPINA3	RPS19	PTN	FAM107A
	RTN1	FBXL16	SPP1	RPL13	IGFBP2	TF
	FBXL16	NRGN	VIM	SPARC	POSTN	FTH1
	TF	CALM1	FABP7	HLA-A	CST3	SERPINA3
	SNAP25	SLC1A2	TIMP1	CHI3L1	TRIB2	C3
	BASP1	CHN1	IGFBP2	PTPRZ1	GNAS	NTRK2
	SYT1	ALDOC	LGALS1	EGFR	LGALS3	CRYAB
	NTRK2	BASP1	SEC61G	B2M	VOPP1	PLEKHB1
	ALDOC	GABBR1	FN1	SERPINA3	PTPRZ1	FTL
	VSNL1	STXBP1	VEGFA	GNB2L1	HOPX	SEPP1
	PLEKHB1	BEX1	TUBB2A	RPL13A	HSPA8	PSAP
	CHN1	OLFM1	CAMK2N1	RPLP2	NES	CLDN11
	NDRG2	NEFL	FTL	RPL39	IGFBP3	CLDND1
	CRYAB	TSPAN7	FABP5	RPS16	VIM	NDRG2
	KIF5C	STMN2	CD63	HLA-B	GNB2L1	RNASE1
	NRGN	ENC1	GAP43	HLA-C	PPIA	EDIL3
	SCD	SYN2	ENC1	EEF1A1	DBI	ENPP2
	STXBP1	RTN3	VSNL1	RPS15A	LANCL2	CNP
	GRIA2	GPM6A	SYT1	FN1	EGR1	NME7
	OLFM1	EEF1A2	FTH1	IGFBP7	S100A16	CD24
	MAP1A	TSC22D1	PGK1	RPS11	TMEM158	QDPR
	ADD3	GABBR2	COL1A1	RPLP1	COL6A1	ADD3
	C7orf41	RGS4	CST3	CD99	MALAT1	C7orf41
	QDPR	MAP2	S100A6	RPL27A	JAG1	DNER
	EDIL3	FXYD6	IGFBP5	ACTG1	RPS2	ATP1B1
	GPRC5B	NAPB	LGALS3		TRIO	APOE
	STMN2	DNM1	CHN1			PFN2
	MLLT11	CAMK2N1	ANXA1			CTSB
	SPARCL1	NPTN	ACTN1			SLC44A1
	FAM123A	TUBB2A	COL4A1			GSN
	SLC1A2	SCD	YWHAG			LAPTM5
	TSPAN7	NDRG4	MAOB			APOD
	SOX8	AGXT2L1	UCHL1			C1QC
	CAMK2N1	CPE	NRGN			
	RTN3	PTPRO	HLA-C			
	CALM1	NGFRAP1	MT3			
	ATP1B1	AK5	HLA-B			
	PEA15		SNAP25			
			HLA-A			

Table 47 Unique Gene IDs, down-regulated

	G1X	G2X	OX	CX	G2X-G1X	G2X-G1X
	GENE SYMBOL(28)	GENE SYMBOL(36)	GENE SYMBOL(31)	GENE SYMBOL(41)	UP (36)	UP (29)
C	SPP1	IGFBP5	SPARC	NME7	C1QC	TRIO
M	ACTG1	IGFBP7	RPLP0	VSNL1	APOD	RPS2
1	MT2A	ANXA1	RPL34	PRKACB	LAPTM5	JAG1
	POSTN	B2M	RPS3A	PEG3	APOE	PPIA
O	HLA-B	TMSB10	NTRK2	IDS	GSN	MALAT1
V	VEGFA	CD63	RPL7	SLC17A7	SLC44A1	HSPA8
E	RPS2	HLA-B	RPL3	RTN1	CTSB	DBI
R	IGFBP3	AQP1	PABPC1	QDPR	CLDND1	COL6A1
E	COL4A1	LAPTM5	ZBTB20	MDH1	PFN2	TMEM158
X	EGFR	MAOB	MARCKS	TUBB2A	ATP1B1	S100A16
P	FN1	COL1A1	VCAN	RGS4	DNER	EGR1
R	RPS19	LTF	ID4	NPTN	C7orf41	LANCL2
E	LGALS1	LGALS1	RPL13	AK5	ADD3	GNAS
S	LDHA	A2M	FXYD6	YWHAH	CD24	GNB2L1
S	LGALS3	CLU	RPS3	YWHAG	QDPR	VIM
E	B2M	HLA-DPA1	RPS6	SERPINI1	NME7	IGFBP3
D	TMSB10	MT2A	TCF12	SLC1A2	CNP	NES
	HLA-C	HLA-C	APOE	MAP1A	ENPP2	HOPX
G	FABP7	VEGFA	HEY1	DNM1	EDIL3	PTPRZ1
E	IGFBP7	SOD2	RPL30	ATP1B1	FAM107A	VOPP1
N	HLA-A	CD44	OLIG2	NAPB	FTL	LGALS3
E	CD63	CTSB	SCD5	CALM1	RNASE1	TRIB2
S	TIMP1	C1QC	SMOC1	EEF1A2	NDRG2	CST3
	PTN	TIMP1	PEA15	PLP1	CLDN11	EGFR
	IGFBP2	HLA-A	ALDOC	MBP	PSAP	POSTN
	CHI3L1	S100A6	LRIG1	NEFL	SEPP1	FABP7
	SEC61G	FN1	BCAN	UCHL1	PTGDS	IGFBP2
	VIM	HLA-DRB1	C1orf61	STMN2	PLEKHB1	PTN
		CD74	OLIG1	SYN2	CRYAB	SEC61G
		FTL	CRYAB	KIF5C	TF	
		FTH1	SOX8	BASP1	NTRK2	
		C3		ENC1	C3	
		VIM		STXBP1	SERPINA3	
		CHI3L1		PTGDS	FTH1	
		SPP1		OLFM1	MBP	
		SERPINA3		CAMK2N1	PLP1	
				SNAP25		
				NRGN		
				CHN1		
				FBXL16		
				SYT1		

Table 48 Unique Gene IDs, upregulated

Genes		**G1X**	**G2X**	**OX**	**CX**	**G1X-G2X**	**G2X-G1X**	**Total**
Downregulated	**Unique genes**	41	40	42	28	29	36	216
	Not found	1	5	2	5	4	5	22
Upregulated	**Unique genes**	28	36	31	41	36	29	201
	Not found	6	3	5	2	5	4	25

Table 49 Number of unique genes per class

2) *CM1 feature extraction diagrams*

The CM1 scoring for the 1-vs-all classes are in the following table.

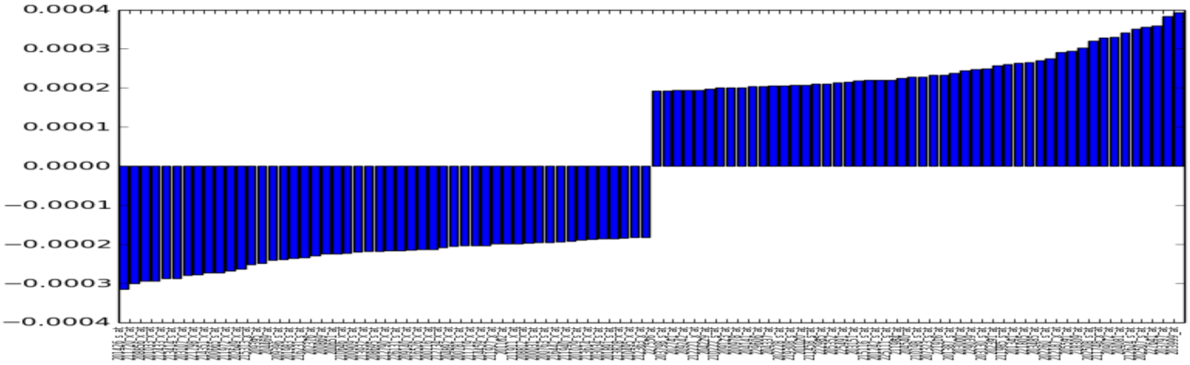
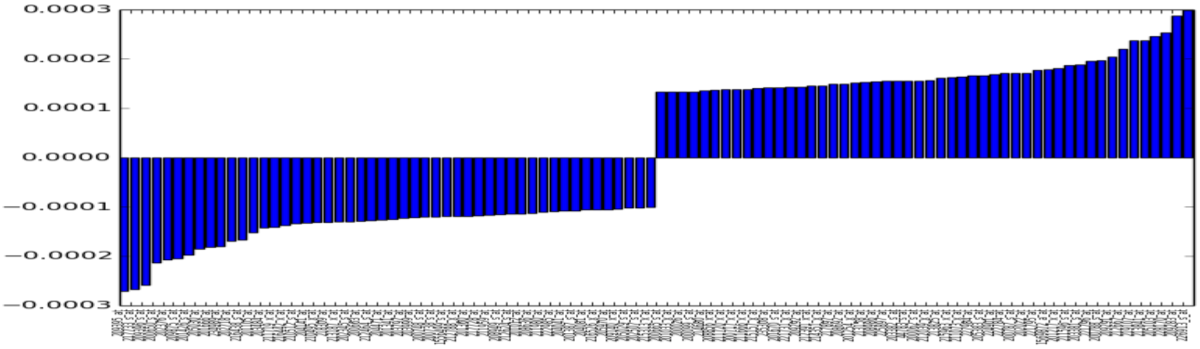
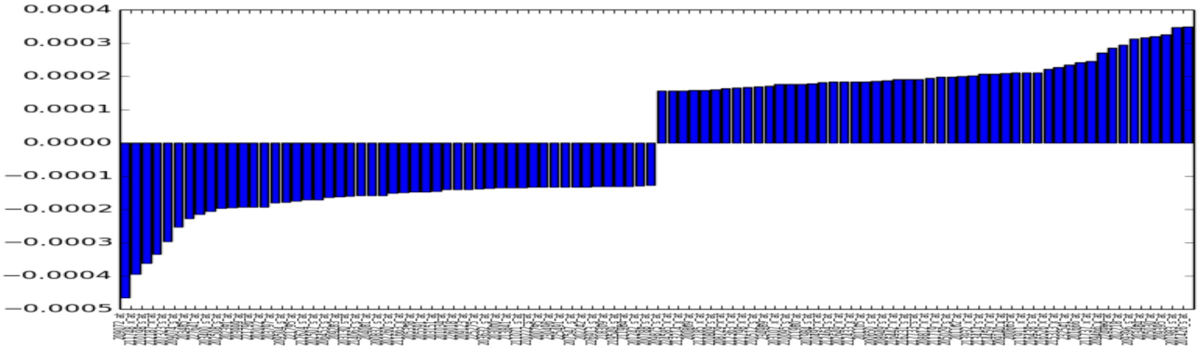
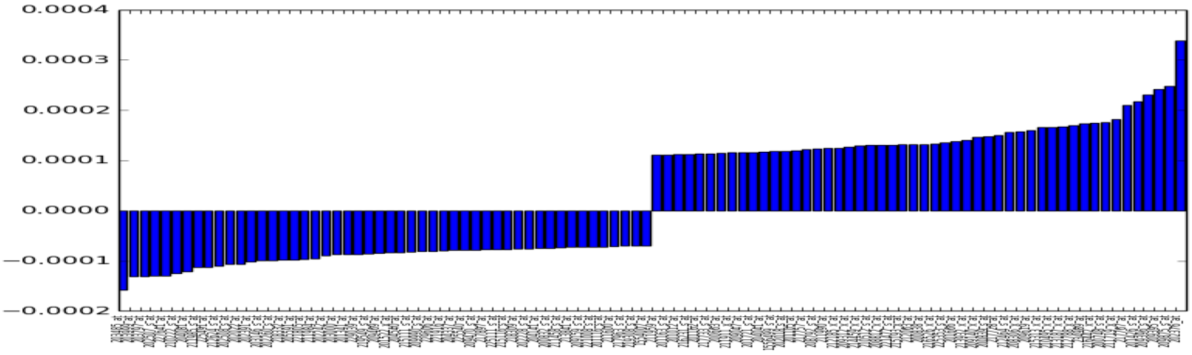
CX vs all	
OX vs all	
G1X vs all	
G2X vs all	

Table 50 CM1 weights for top/bottom 50 1-vs-All

Although all the CM1 1-vs-1 sets have been extracted, only the G1X-vs-G2X has been considered relevant for the study by the bioinformatics expert doing the analysis. Hence, only that subset has been considered in the study also for the quantitative gene analysis.

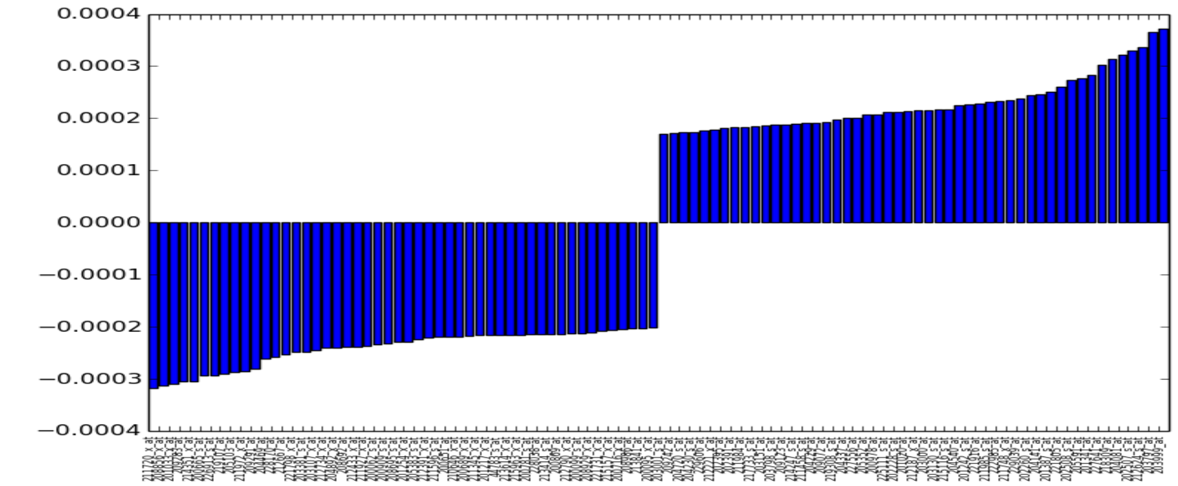
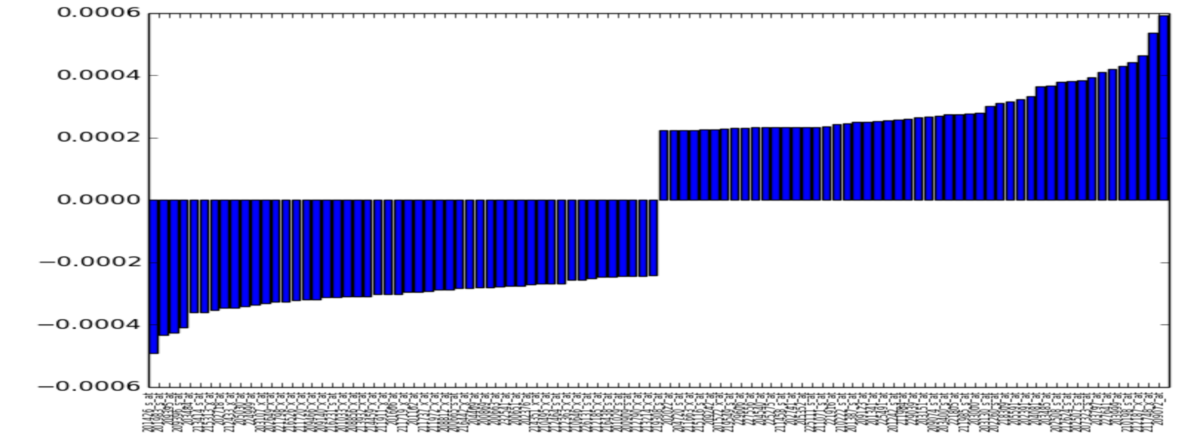
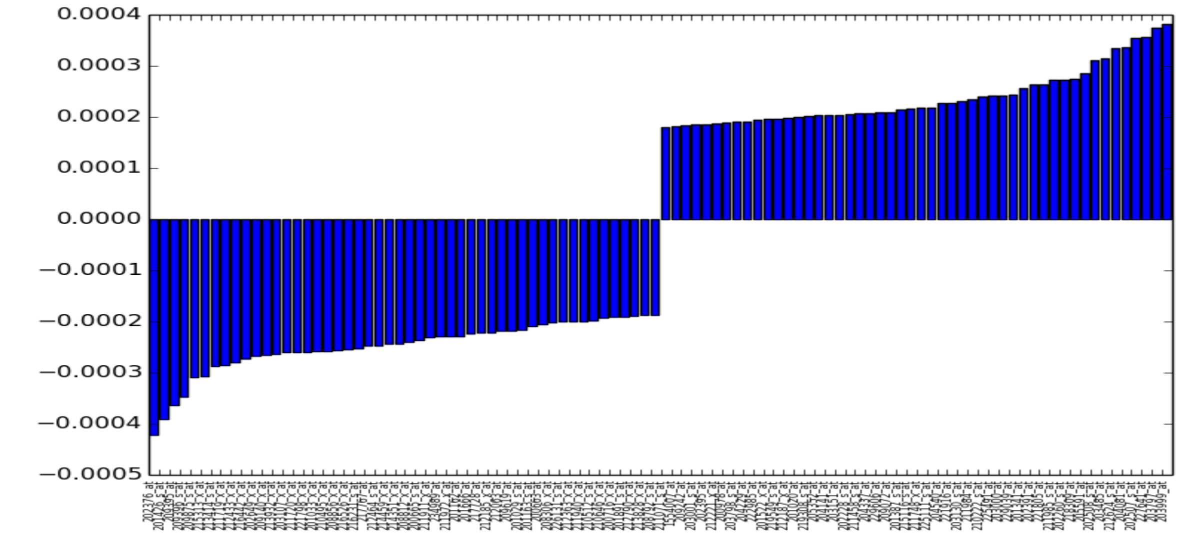
CX vs OX	
CX vs G1X	
CX vs G2X	

Table 51 CM1 weights for CX-vs-1

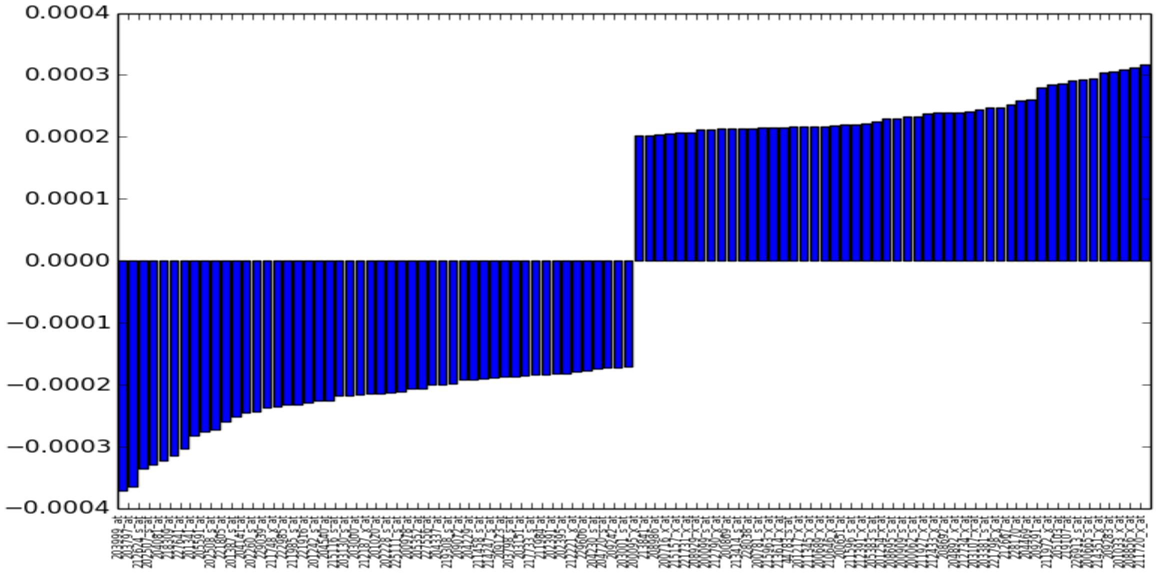
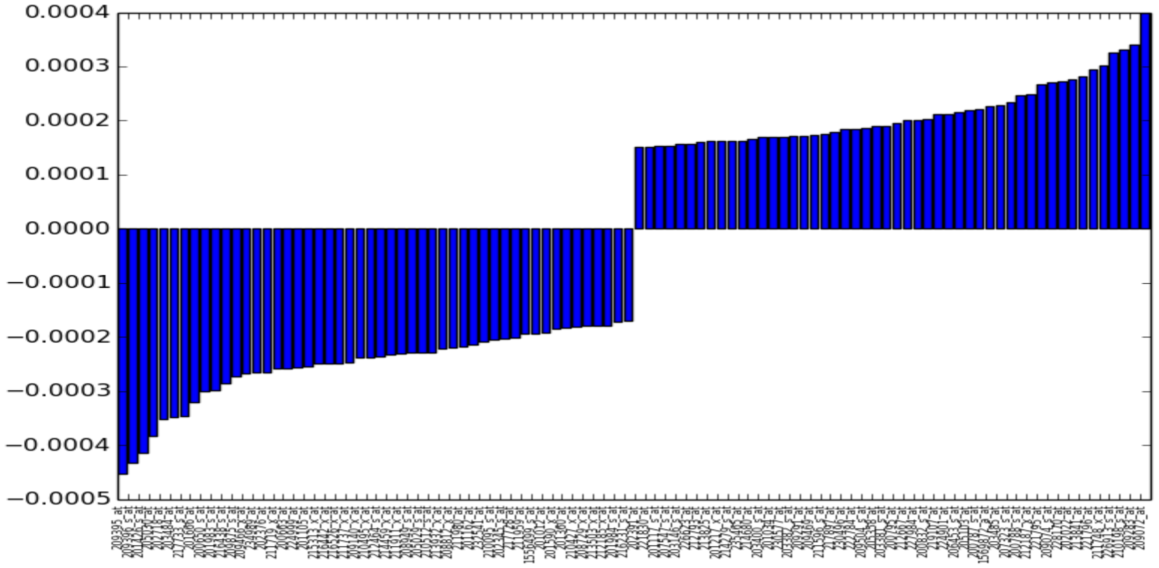
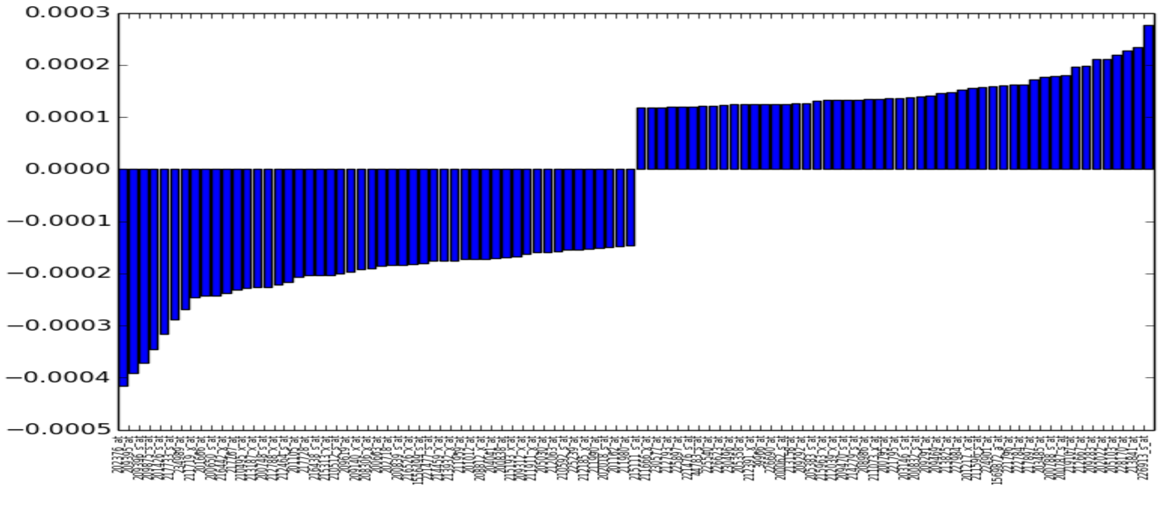
OX vs CX	
OX vs G1X	
OX vs G2X	

Table 52 CM1 weights for OX-vs-1

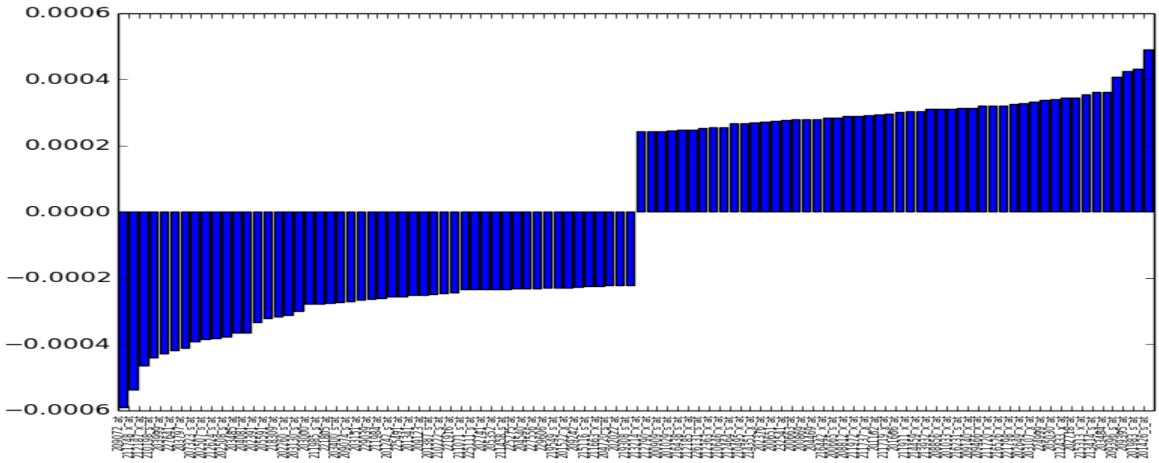
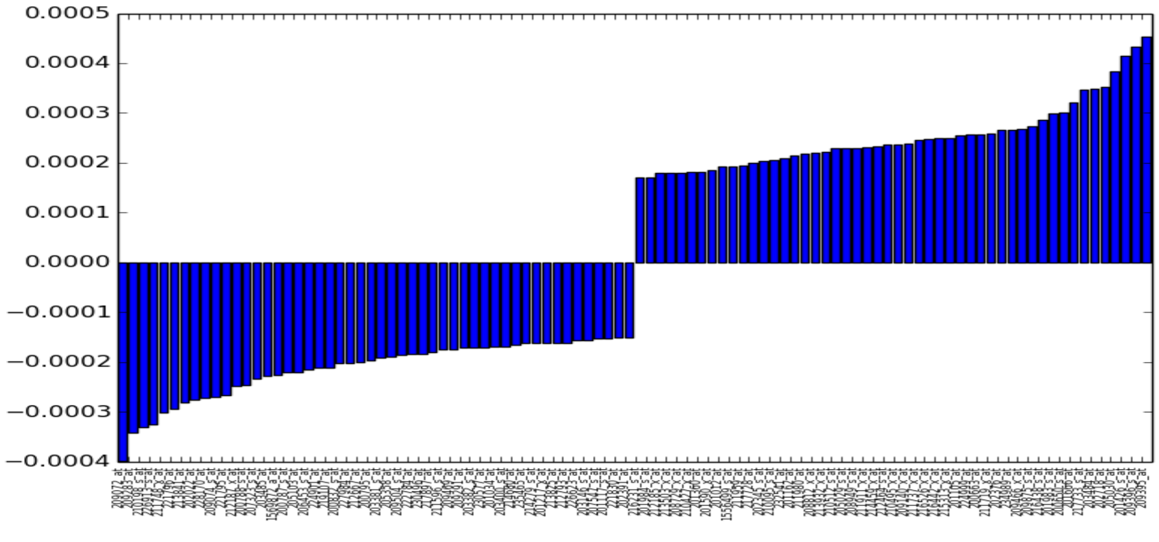
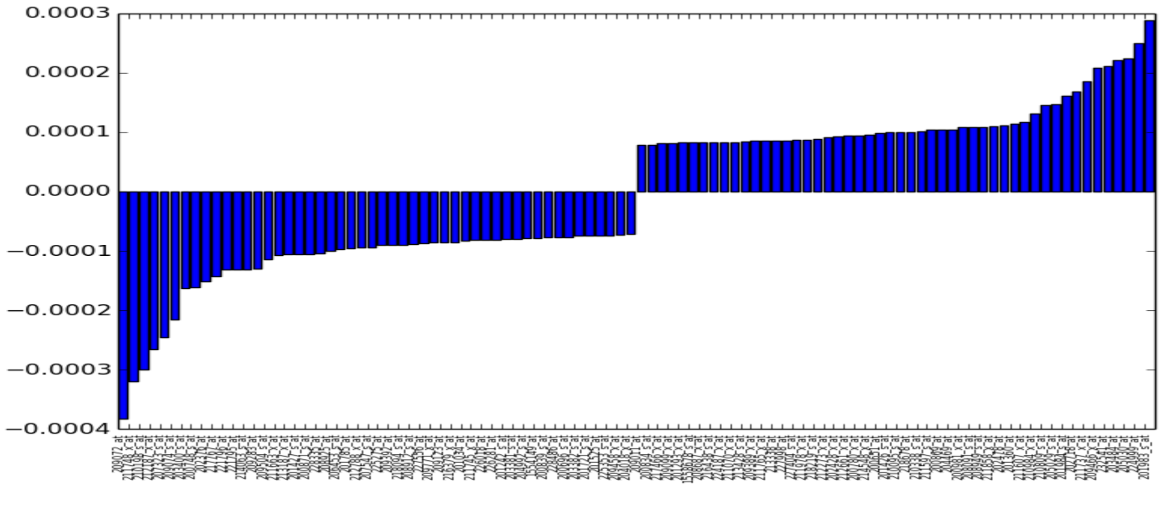
G1X vs CX	
G1X vs OX	
G1X vs G2X	

Table 53 CM1 weights for G1X-vs-1

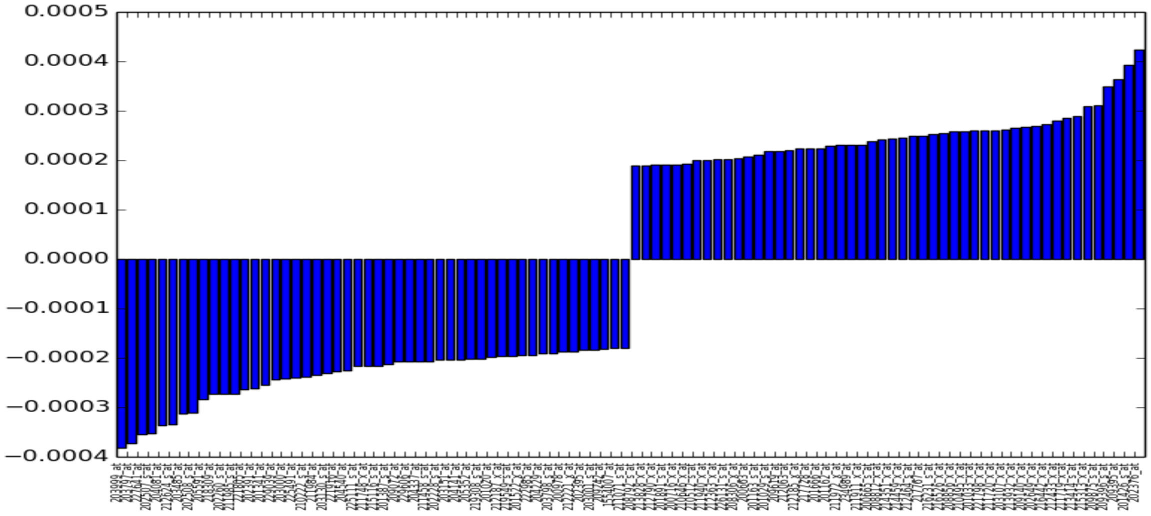
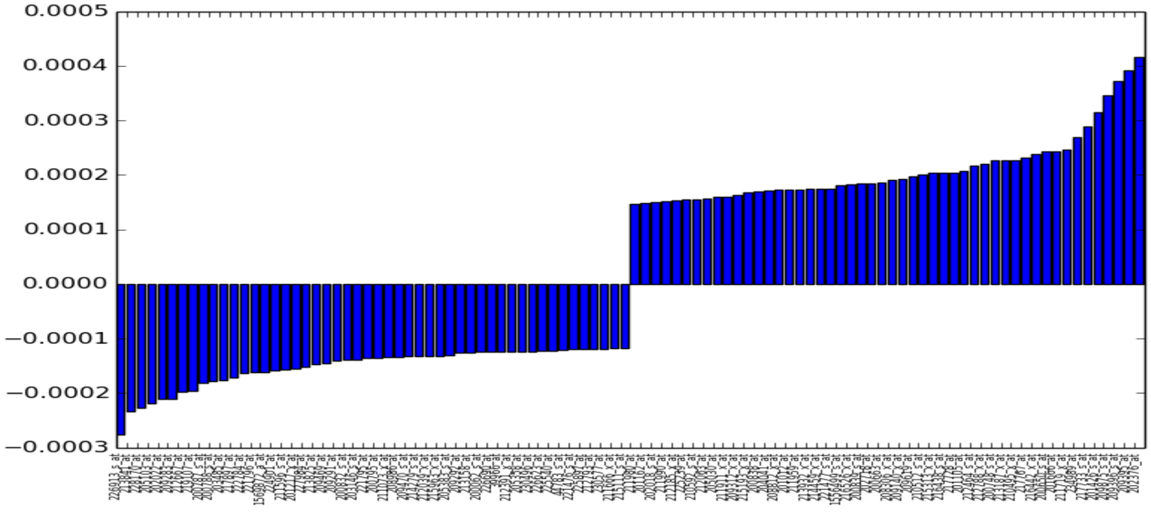
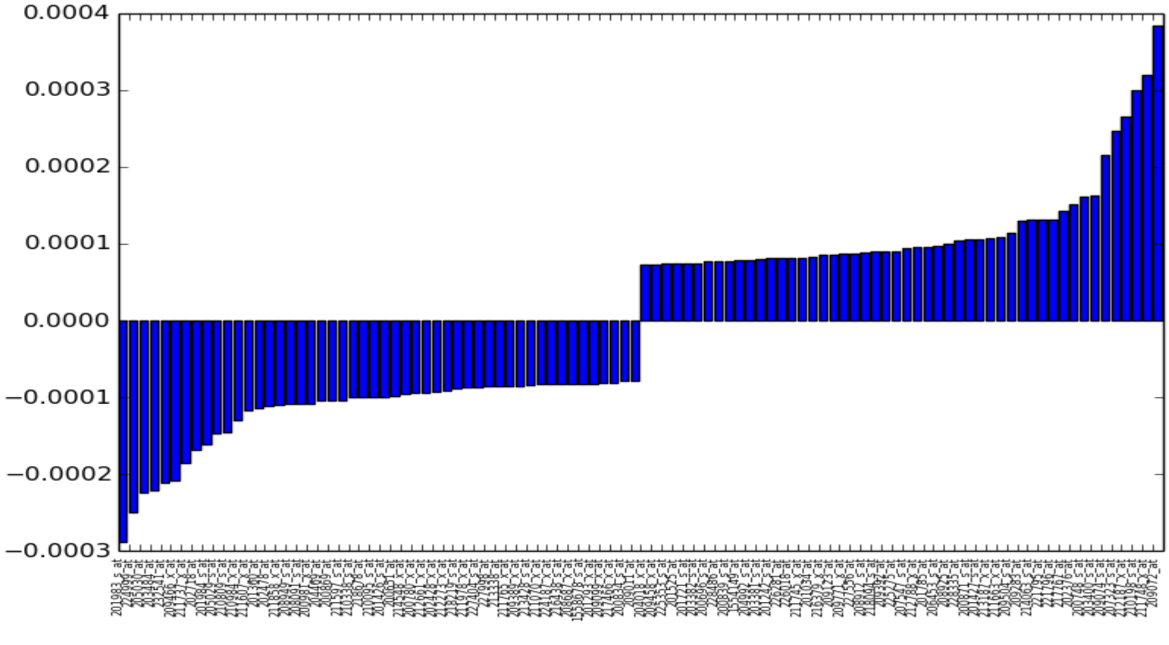
G2X vs CX	
G2X vs OX	
G2X vs G1X	

Table 54 CM1 weights for G2X-vs-1

D. *Hierarchical clustering linkage functions*

The linkage functions most commonly used in Hierarchical Clustering follows:

- *Single linkage:* The new distance is the minimum distance among the two elements creating the cluster.

$$d(u,v) = \min(dist(u[i], v[j]))$$

Equation 23 Single linkage

- *Complete linkage:* The new distance is the maximum distance among the two elements creating the cluster.

$$d(u,v) = \max(dist(u[i], v[j]))$$

Equation 24 Complete linkage

- *Average linkage:* The new distance is the average pairwise distance among all the samples included in the two originating clusters.

$$d(u,v) = \sum_{ij} \frac{d(u[i], v[j])}{(|u| * |v|)}$$

Equation 25 Average linkage

This method inspired the below-mentioned Centroid linkage method and has been used successfully in [63] in analysis of gene expression data.

- *Weighted linkage:* The new distance is the average distance from the two clusters (s, t) creating the new cluster (u) to the cluster (v) the distance is being calculated to.

$$d(u,v) = (dist(s,v) + dist(t,v))/2$$

Equation 26 Weighted linkage

A second group of linkage methods would include (among others) as defined in [64]:

- *Centroids linkage:* The new distance is calculated as the distance among the centroids where the centroid is calculated from all the samples in the cluster.

$$d(r,s) = \|\bar{x}_r - \bar{x}_s\|_2 \qquad \bar{x}_r = \frac{1}{n_r}\sum_{i=1}^{n_r} x_{ri}$$

Equation 27 Centroid linkage

- *Median linkage:* The new distance is calculated based on the distance of centroids but the new centroid is calculated based on the average of the two clusters (p, q) that were joined to create the cluster.

$$d(r,s) = \left\| \tilde{x}_r - \tilde{x}_s \right\|_2 \qquad \tilde{x}_r = \frac{1}{2}(\tilde{x}_p + \tilde{x}_q)$$

Equation 28 Median linkage

- *Ward linkage:* The new distance, according to [26] "*minimizes the information loss associated with clustering. Ward used an error sum-of-squares criterion to define information loss. At each step, union of every possible pair of clusters is considered and the two clusters whose fusion results in the smallest increase in 'information loss' are combined. ".The formula as defined in [64]:*

$$d(r,s) = \sqrt{\frac{2n_r n_s}{(n_r + n_s)}} \left\| \bar{x}_r - \bar{x}_s \right\|_2 ,$$

Equation 29 Ward linkage

E. *Similarity and dissimilarity measures*

A common way of comparing the results of a classifier compared to the known true classification is the confusion matrix. It gives the number of samples that have been classified in each of the classes and this is intersected with its known class. Based on this, the values that are correctly classified will appear on the diagonal of the matrix while incorrect classified will be off-diagonal. Each Y, Z intersection will indicate how many class Y elements have been classified as class Z. While this concept is very intuitive and clear for classification problems it requires a further sophistication in order to be useful for clustering. The reason is that since clustering is some kind of unlabeled classifier, when there are disagreements it may not be possible to identify what is the part of the cluster that is correctly clustered. As an example, consider how to measure the correctness when for example one cluster has been split in two different clusters otherwise perfectly identified.

To solve this problem, many of the similarity measures used for clustering are based on counting pair matches that is the number of pairs of samples that are in the same or different partition in the two partitions being compared. Four cases are considered, being P1 and P2 the compared partitions:

	Same class in P2	Diff class in P2
Same class in P1	a	b
Diff class in P2	c	d

where a is the number of pairs of samples in the same class in the two compared partitions, d is the number of pairs of samples in different classes and b, c correspond respectively to the pairs in the same class in one partition that are not in the same class in the other.

Indexes defined will have value 1 for matching partitions, meaning indexes measure the similarity of the two partitions.

Metrics are the opposite concept, and measure the dissimilarity; therefore, they will have value 0 for matching partitions.

1) *Agreements and Disagreements*

Based on the “a, b, c, d” concepts defined previously the following indicators have been used:

Same Class Agreements: corresponds to a, the number of pairs that are in the same class in the two partitions.

Agreements: Corresponds to a + d, that is, pairs that are either in the same class in both partitions or elements that are not in the same partition in neither of the compared partitions.

Disagreements: Corresponds to b + c, that is, the sum of the number of pairs that are in the same class in one partition but not in the other.

The three indicators considered together are redundant since:

$$a + b + c + d = \binom{N}{2}$$

N is the number of samples in the dataset.

2) *Jaccard Index*

The Jaccard Index was defined in [65]. Based on our introductory definition, the Jaccard Index can be defined as:

$$JI = \frac{a}{a + b + c}$$

Equation 30 Jaccard Index

3) *Fowlkes Mallows Index*

First defined in [66], based on our baseline definition, the formula is given by:

$$FMI = \sqrt{\frac{a}{a+b} * \frac{a}{a+c}}$$

Equation 31 Fowlkes-Mallows Index

4) *Normalized Mutual Information Index*

It was defined in [67] and has two different normalizations possible, arithmetic and geometric. The arithmetic version is given by:

$$NMI(X,Y) = \frac{I(X,Y)}{H(X) + H(Y)}$$

Equation 32 Normalized Mutual Information arithmetic

The geometric version is defined in information theory terms as:

$$NMI(X,Y) = \frac{I(X,Y)}{\sqrt{H(X) * H(Y)}}$$

Equation 33 Normalized Mutual Information geometric

where I(X,Y) is the mutual information:

$$I(X,Y) = H(X) - H(X|Y)$$

Equation 34 Mutual Information

and H(X) corresponds to entropy as defined in Equation 11.

Alternativately, in clustering terms:

$$I(X,Y) = -2 \sum_{i=1}^{N} \sum_{j=1}^{M} x_{i,j} * \log \frac{x_{i,j}/N}{x_i x_j}$$

Equation 35 Mutual Information for Clustering

And x_i is the number of samples in class i divided by the total number of samples, and $x_{i,j}$ is the number of samples in class i in the first partition and class j in the second partition.

5) *Normalized Mirkin Metric*

The Mirkin Metric was defined in [68], being his formula, as shown in [69]:

$$\mathcal{M}(\mathcal{C}, \mathcal{C}') = \sum_{k} n_k^2 + \sum_{k'} n_{k'}^2 - 2 \sum_{k} \sum_{k'} n_{kk'}^2$$

Equation 36 Mirkin Metric

where n_x corresponds to the cluster x of the partition and n_{xy} corresponds to the intersections of the two partitions.

Its invariant (normalized) version is:

$$\mathcal{M}_{\text{inv}}(\mathcal{C}, \mathcal{C}') = \frac{\mathcal{M}(\mathcal{C}, \mathcal{C}')}{n^2}$$

Equation 37 Normalized Mirkin Metric

being n the number of samples in the dataset.

6) *Normalized Van Dongen Metric*

In [69], it is reported to be defined in . The formula is:

$$\mathcal{D}(\mathcal{C},\mathcal{C}') = 2n - \sum_{k} \max_{k'} n_{kk'} - \sum_{k'} \max_{k} n_{kk'}$$

Equation 38 Van Dongen Metric

where the definition of n_{xy} is analogous to the previous section.

Again, the normalized version is given by:

$$\mathcal{D}_{\mathrm{inv}}(\mathcal{C},\mathcal{C}') = \frac{\mathcal{D}(\mathcal{C},\mathcal{C}')}{2n}$$

Equation 39 Normalized Van Dongen Metric

7) *Normalized Variation of Information Metric*

Variation of Information is defined in [70]:

$$VI = H(X|Y) + H(Y|X)$$

Equation 40 Variation of Information

It can be normalized, to make it bounded and not dependent on dataset size, in different ways, being 2 log N (VI upper bound) the more advantageous as it is proved to be a metric (symmetric and satisfying the triangular inequality).

$$NVI(X,Y) = \frac{H(X|Y) + H(Y|X)}{2\log N}$$

Equation 41 Normalized Variation of Information

8) *Adjusted Rand Index*

The Rand Index(RI) [71], whose name is due to William Rand not to random numbers or anything alike, is defined as:

$$RI = \frac{a+d}{a+b+c+d}$$

Equation 42 Rand Index

It is bounded between 0 and 1, being 1 the value for two matching partitions.

The Adjusted Rand Index (ARI) was defined in [72] as referred in [73]. Its formula is given by:

$$\frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right] - \left[\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}\right] / \binom{n}{2}}$$

Equation 43 Adjusted Rand Index

$n_{i,j}$ corresponds to the number of samples in class i on the first partition and in class j in the second partition, $n_{i.}$ and $n_{.j}$ correspond to all the samples in class i of the first partition or class j of the second partition regardless the class on the other partition. If the confusion matrix of the two partitions was used, they would correspond to a whole row or column of the matrix.

The concept is motivated in order to solve one limitation the Rand Index has. The limitation is that it would be expected Rand Index for two random partitions is 0 what is not the case.

9) *Purity*

While the concept is an old one and has appeared in many publications, it is best explained in [74]. It is defined formally as:

$$Purity(X,Y) = \sum_i \frac{n_{i+}}{n} \, max_j \, \frac{n_{ij}}{n_{i+}} = \sum_i \frac{1}{n} \max{}_j \, n_{ij}$$

Equation 44 Purity

In plain words, purity is the addition of the ratios of the dominant classes in each partition compared to a reference partition, typically the ground truth partition.

Purity is [0, 1] bounded with 1 meaning perfect matching, and 0 complete dissimilarity. Purity is not a symmetric measure.

Based on Purity and in order to obtain a symmetric measure, the F-measure is defined:

$$F\_Measure(X,Y) = \frac{2 * Purity(X,Y) * Purity(Y,X)}{Purity(X,Y) + Purity\ (Y,X)}$$

Equation 45 F-Measure

10) *Homogeneity and Completeness*

Other information theory based measures are Homogeneity(h) and Completeness(c) defined in [59] as:

$$h(C,K) = \begin{cases} 1 & H(C) = 0 \\ 1 - \frac{H(C|K)}{H(C)} & H(C) \neq 0 \end{cases}$$

Equation 46 Homogeneity

$$c(C,K) = \begin{cases} 1 & H(K) = 0 \\ 1 - \frac{H(K|C)}{H(K)} & H(K) \neq 0 \end{cases}$$

Equation 47 Completeness

where H(x) corresponds to the entropy as previously defined.

Homogeneity provides a measure of the content of each cluster, and is roughly proportional to the number of elements of the majority class in each cluster, conceptually similar to Purity.

Completeness is the complementary concept, indicating the number of elements of one class that are not in other clusters.

Still one more measure can be defined from h and c, the V-measure, defined as the harmonic mean of h and c:

$$V = \frac{2\,h\,c}{h + c}$$

Equation 48 V-measure

As explained in [70], the V measure has important drawbacks and favors partitions with a large number of clusters in particular the singleton partition (each element is a cluster).

F. *Data Normalization tests*

The normalization to sum to unity modifies it is not a linear one when considered among samples as it executed independently for each of the samples. Each data point is divided by the sum of the expressions of the probes in the sample, guaranteeing that the sum of the sample is the unity.

To have an estimation of how this transformation affects the data a simple dataset has been generated. The dataset has 4 samples and 4 genes:

	p1	p2	p3	p4
g1	10	20	30	40
g2	4	4	4	4
g3	1000	2000	3000	4000
g4	500	500	500	500

Table 55 Synthetic dataset

As can be observed the range for each of the genes are very different from each other. Note that g2 and g4 have the same value for all the samples. It would be desirable that any transformation applied maintain this.

If the sum to unity normalization to each sample (p1-p4) is applied we obtain the following transformed dataset:

	p1	p2	p3	p4
g1	0.006605	0.007924	0.008489	0.008803
g2	0.002642	0.001585	0.001132	0.00088
g3	0.660502	0.792393	0.848896	0.880282
g4	0.330251	0.198098	0.141483	0.110035

Table 56 Column normalized dataset

Note that for g2, the values for p1 and p4, maximum and minimum respectively, have now a ratio of 3. Despite having a much higher initial value the ratio is the same for g4. The ratio for g1 compared with its original value ratio is also 3.

If prior to applying the column normalization row normalization is applied, we obtain after the first transformation:

	p1	p2	p3	p4
g1	0.1	0.2	0.3	0.4
g2	0.25	0.25	0.25	0.25
g3	0.1	0.2	0.3	0.4
g4	0.25	0.25	0.25	0.25

Table 57 Dataset after row normalization

And finally:

	p1	p2	p3	p4
g1	0.142857	0.222222	0.272727	0.307692
g2	0.357143	0.277778	0.227273	0.192308
g3	0.142857	0.222222	0.272727	0.307692
g4	0.357143	0.277778	0.227273	0.192308

Table 58 Dataset after subsequent row and column normalization

In this case, the ratio between p1 and p4 for g2 and g4 is 1.85, much lower than when the data was only column normalized. Also for g1, compared to the original ratio 1:4 as the initial values were different, is 1.85.

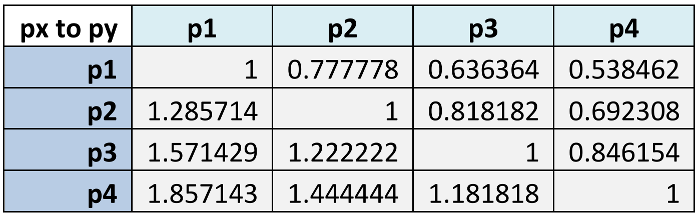
Despite this evidence, without any formal validation to support further conclusions, it cannot be denied that the normalization is not an inconsequent transformation and may influence the final results, particularly, depending on the distance function used.

When the distance function only considers the features in a 1-by-1 basis, such as Euclidean distance, the distance calculated will be affected. In the example, for the first case the distance is 0.311 and for the second case the distance is 0.329. The more important drawback is that distances that should be zero will not.

In each case, all the pairwise comparisons among samples generate the same ratio for all the genes, indicating that the transformation applied while not affecting all the data points in the same way makes it consistently when pairwise considered.

px to py	p1	p2	p3	p4
p1	1	0.599842	0.42841	0.333187
p2	1.667107	1	0.714205	0.555458
p3	2.334214	1.400158	1	0.777729
p4	3.001321	1.800317	1.285795	1

Table 59 Sample to sample factor for any gene with column normalization



px to py	p1	p2	p3	p4
p1	1	0.777778	0.636364	0.538462
p2	1.285714	1	0.818182	0.692308
p3	1.571429	1.222222	1	0.846154
p4	1.857143	1.444444	1.181818	1

Table 60 Sample to sample norm factor for any gene with row-column normalization

In cases like Pearson distance, what is being measured is not the distance among individual features but more the distance among features when these features are considered as part of the whole sample. A trivial example, the Pearson distance of any distribution respect to the same distribution linearly transformed is equal to 0.

In the case of Jensen-Shannon divergence, what are being compared are the probability distributions, what keeps more similarities with the Pearson scenario that with the Euclidean distance scenario. In addition, one of the properties said that the JSD was zero only if the two distributions were identical, therefore, we may expect different results for the two normalizations described.

Finally, computing the distance matrix for the two cases it can be observed that the matrixes have different values and also slightly different distributions, Table 3. Hence, generating different results for the two normalizations may depend not only on the data but also on the clustering algorithm in place.

G. *Brain tumour diagnosis reliability*

The reliability of tumor diagnosis has been subject of study. Two recent studies are [75] and [76]. On the studies, the consensus of diagnosis of glioma patients is evaluated based on observed significant inter-observer variation of glioma. The diagnosis is variant in both typing and grading of the gliomas.

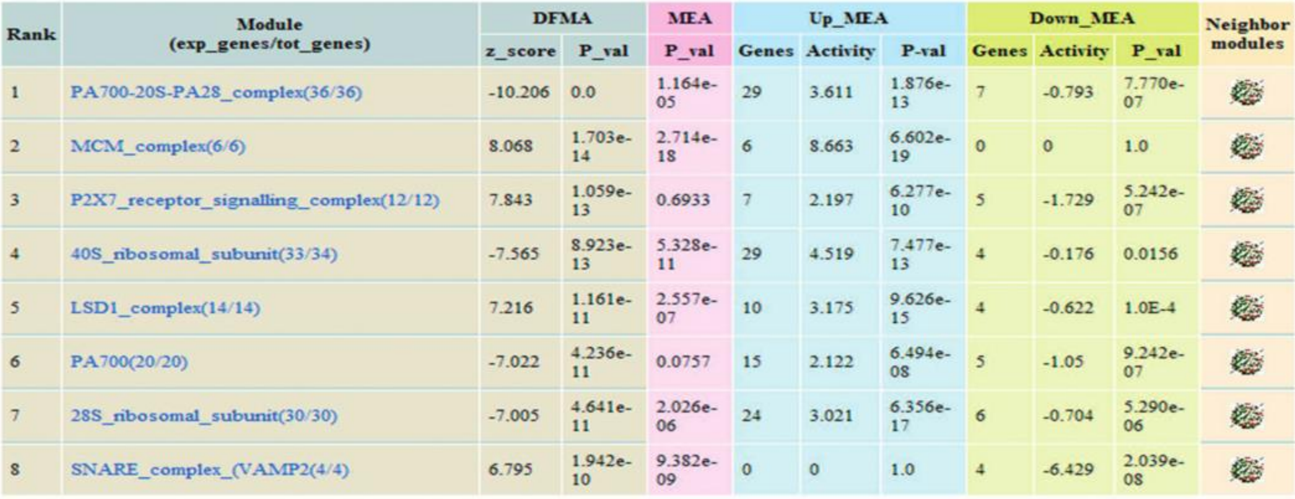
The first study mentions that from 500 brain tumors reviewed, 42.8% show some diagnosis disagreement that can be considered serious in 8.8%. The study refers to other studies mentioning that this misdiagnose is higher when the patient proceeds from local community hospitals as opposed to academic hospitals. 16% of the discordant diagnoses were clinically significant as were affecting treatment and/or prognosis. A study of 244 cases with intervention of four pathologists showed 52% initial agreement that grew to 62% after the fourth round of reviews. Oligodendrogliomas became a 25% of the diagnosis while initially they were only 5%. The list of studies reviewed goes mentioning that among the different know glioma types there are mixed types such as oligoastrocytomas making the diagnosis criteria blurrier with only 13% agreement for those cases. One preliminary conclusion of the study is that among 20 to 30% of gliomas are reclassified based on independent reviews.

The second study, focused on different diagnosis criteria and its changing definition along the years, explains the difficulty of diagnosis in the small series of oligodendroglioma tumors available. The same study mentions that some criteria are universally accepted as indicator of a glioma. About diagnosis variability, the study mentions that only 36% cases of Astrocytoma (AA) are confirmed. For Glioblastoma this ratio raised to 73% indicating more robust clinic diagnostic criteria. Only 32% majority and 8% consensus was achieved for Oligoastrocytoma (OA).

H. *Known results obtained from dataset GSE4290*

In [42], the original study that made available the dataset, the main conclusions are related to only one gene, SCF(Stem Cell Factor). The study combines the analysis of the genetic information in Micro Array data with other tests so the information is of limited use for comparing with the current results.

In [77], the same dataset has been used obtaining a list of modules (MEA) and the corresponding up-regulated or down-regulated groups.



Rank	Module (exp_genes/tot_genes)	DFMA		MEA	Up_MEA			Down_MEA			Neighbor modules
		z_score	P_val	P_val	Genes	Activity	P-val	Genes	Activity	P_val	
1	PA700-20S-PA28_complex(36/36)	-10.206	0.0	1.164e-05	29	3.611	1.876e-13	7	-0.793	7.770e-07	🕸
2	MCM_complex(6/6)	8.068	1.703e-14	2.714e-18	6	8.663	6.602e-19	0	0	1.0	🕸
3	P2X7_receptor_signalling_complex(12/12)	7.843	1.059e-13	0.6933	7	2.197	6.277e-10	5	-1.729	5.242e-07	🕸
4	40S_ribosomal_subunit(33/34)	-7.565	8.923e-13	5.328e-11	29	4.519	7.477e-13	4	-0.176	0.0156	🕸
5	LSD1_complex(14/14)	7.216	1.161e-11	2.557e-07	10	3.175	9.626e-15	4	-0.622	1.0E-4	🕸
6	PA700(20/20)	-7.022	4.236e-11	0.0757	15	2.122	6.494e-08	5	-1.05	9.242e-07	🕸
7	28S_ribosomal_subunit(30/30)	-7.005	4.641e-11	2.026e-06	24	3.021	6.356e-17	6	-0.704	5.290e-06	🕸
8	SNARE_complex_(VAMP2(4/4)	6.795	1.942e-10	9.382e-09	0	0	1.0	4	-6.429	2.039e-08	🕸

Table 61 Results in [77] showing MEA

In [78], CLIC, the algorithm developed as part of the study is not capable of dealing with all genetic information(>40K genes), while for others algorithms compared the threshold was even lower.

In [41], the method permits to obtain a classification of the samples with 56% accuracy on average using a set of 44 genes selected with a different dataset (training set) and capable of dealing with higher accuracy if the training is performed with two or more datasets from independent studies. Only GBM (77 samples) and OLG (50 samples) are considered. The study lists 11 GBM serum markers present in the marker-panel, Table 62.

APOD	CHI3L1	IGFBP2	PSG9
CALU	CSF1	NID1	PTN
CD163	EGFR	PDGFC	

Table 62 11 GB serum markers

When the dataset was used to train a model and the obtained model tested with other datasets, the accuracy was an average 40.66%. When other datasets were used for training and the dataset used to test the accuracy was 72.08%.

To differentiate GBM from OLG the cut of the genes 1p and 19q becomes a marker as it is the Olig2 over-expression. GBM and A3 are differentiated based on the expression of FLNA, ANXA1, and bHLH. IDH is another biomarker for GBM. EPN gliomas are differentiated from other types based on TLE4, OLIG2.

In [79], the GSE4290 dataset is used in a classification of the patients in 1-vs-1 groups (GBM-vs-OLG and GBM-vs-AC), Figure 26. The experiment is done as part of a study to improve the prognosis profiles for gliomas, Figure 27. The study, that is based on several datasets, concludes that 42 probes are relevant based on a multivariate Cox statistical. In the case of our dataset the 42 probes are specific for lower grade gliomas and group 3 GBM.

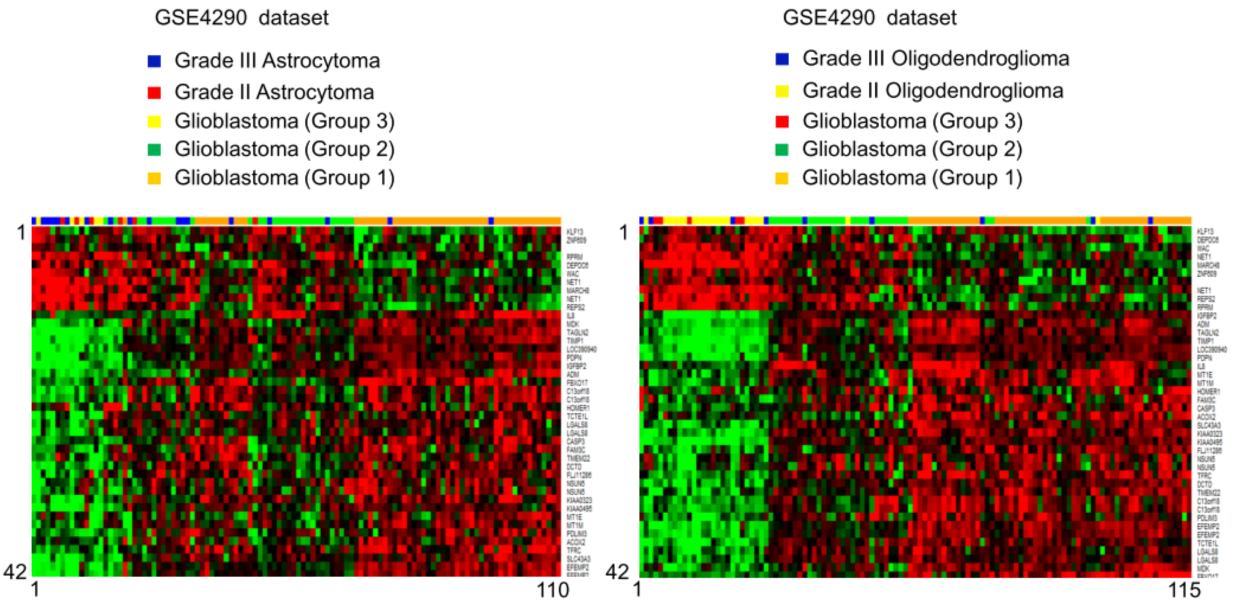


Figure 26 Clustering of GBM-vs-AC and GBM-vs-OL, from [77]

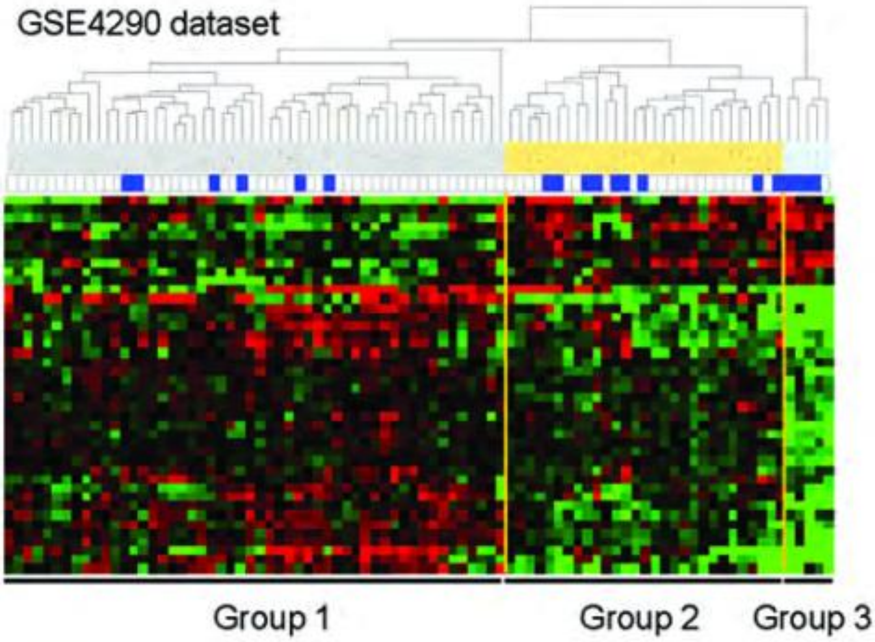
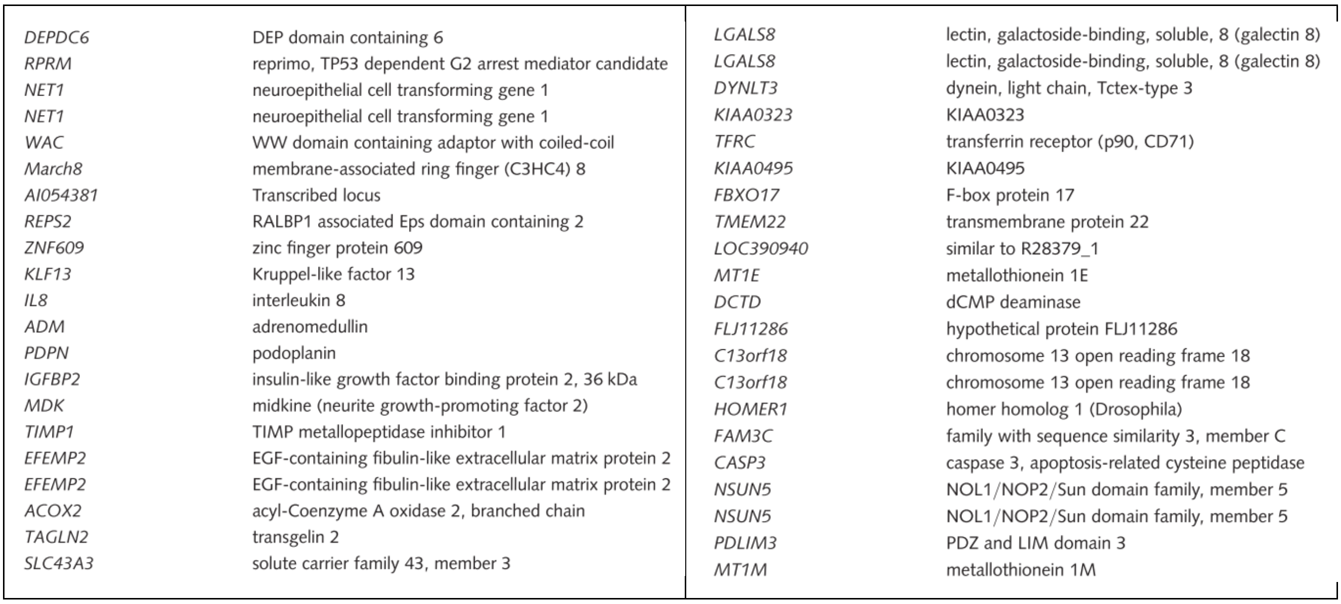


Figure 27 GSE4290 Prognosis classification



DEPDC6	DEP domain containing 6	*LGALS8*	lectin, galactoside-binding, soluble, 8 (galectin 8)
RPRM	reprimo, TP53 dependent G2 arrest mediator candidate	*LGALS8*	lectin, galactoside-binding, soluble, 8 (galectin 8)
NET1	neuroepithelial cell transforming gene 1	*DYNLT3*	dynein, light chain, Tctex-type 3
NET1	neuroepithelial cell transforming gene 1	*KIAA0323*	KIAA0323
WAC	WW domain containing adaptor with coiled-coil	*TFRC*	transferrin receptor (p90, CD71)
March8	membrane-associated ring finger (C3HC4) 8	*KIAA0495*	KIAA0495
AI054381	Transcribed locus	*FBXO17*	F-box protein 17
REPS2	RALBP1 associated Eps domain containing 2	*TMEM22*	transmembrane protein 22
ZNF609	zinc finger protein 609	*LOC390940*	similar to R28379_1
KLF13	Kruppel-like factor 13	*MT1E*	metallothionein 1E
IL8	interleukin 8	*DCTD*	dCMP deaminase
ADM	adrenomedullin	*FLJ11286*	hypothetical protein FLJ11286
PDPN	podoplanin	*C13orf18*	chromosome 13 open reading frame 18
IGFBP2	insulin-like growth factor binding protein 2, 36 kDa	*C13orf18*	chromosome 13 open reading frame 18
MDK	midkine (neurite growth-promoting factor 2)	*HOMER1*	homer homolog 1 (Drosophila)
TIMP1	TIMP metallopeptidase inhibitor 1	*FAM3C*	family with sequence similarity 3, member C
EFEMP2	EGF-containing fibulin-like extracellular matrix protein 2	*CASP3*	caspase 3, apoptosis-related cysteine peptidase
EFEMP2	EGF-containing fibulin-like extracellular matrix protein 2	*NSUN5*	NOL1/NOP2/Sun domain family, member 5
ACOX2	acyl-Coenzyme A oxidase 2, branched chain	*NSUN5*	NOL1/NOP2/Sun domain family, member 5
TAGLN2	transgelin 2	*PDLIM3*	PDZ and LIM domain 3
SLC43A3	solute carrier family 43, member 3	*MT1M*	metallothionein 1M

Table 63 42 Probesets relevant for glioma prognosis

In [80], 78 genes, shown in Table 64, where selected to differentiate glioma-relevant gliogenesis genes. The genes where selected based on differential expression after excluding genes with standard deviation below 1.5% of the maximal standard deviation.

AGT,	DAG1,	EXOC4,	HMGA2,	NAB2,	SOX11,
ANXA1,	DLL3,	FGF2,	ITGAM,	NF1,	SOX2,
ASCL1,	EGR1,	FOXD1,	KLF15,	NFIB,	SOX4,
ATF5,	CDKN2C,	GFAP,	LAMB2,	NOG,	SOX5,
BCL2,	CSPG4,	GLI3,	LIF,	NOTCH1,	SOX6,
BMP2,	CTNNB1,	GPC1,	LYN,	NR2E1,	SOX8,
C1S,	CXCR4,	GSN,	MET,	OLIG2,	SOX9,
CCL2,	EGR2,	HDAC2,	MMP14,	PAX2,	STAT3,
CD86,	EIF2B1	HES1,	MPP5,	PRKCH,	TCF7L2,
CD9,	EIF2B4,	HES5,	MYT1,	PTEN,	TGFB2,
CDK1,	ERBB2,	HEXB,	PAX6,	PTPRC,	TSPO,
CDK6,	ID2,	HMBS,	PDGFRA,	RELA,	VCAN,
CDKN1A,	ID4,	HOXA2,	POU3F2,	SMARCA4,	ZNF226.

Table 64 List of gliogenesis related genes as in [80]

In [81], while a different dataset has been used, a 90 genes profile has been defined in order to differentiate among different types of glioblastomas.

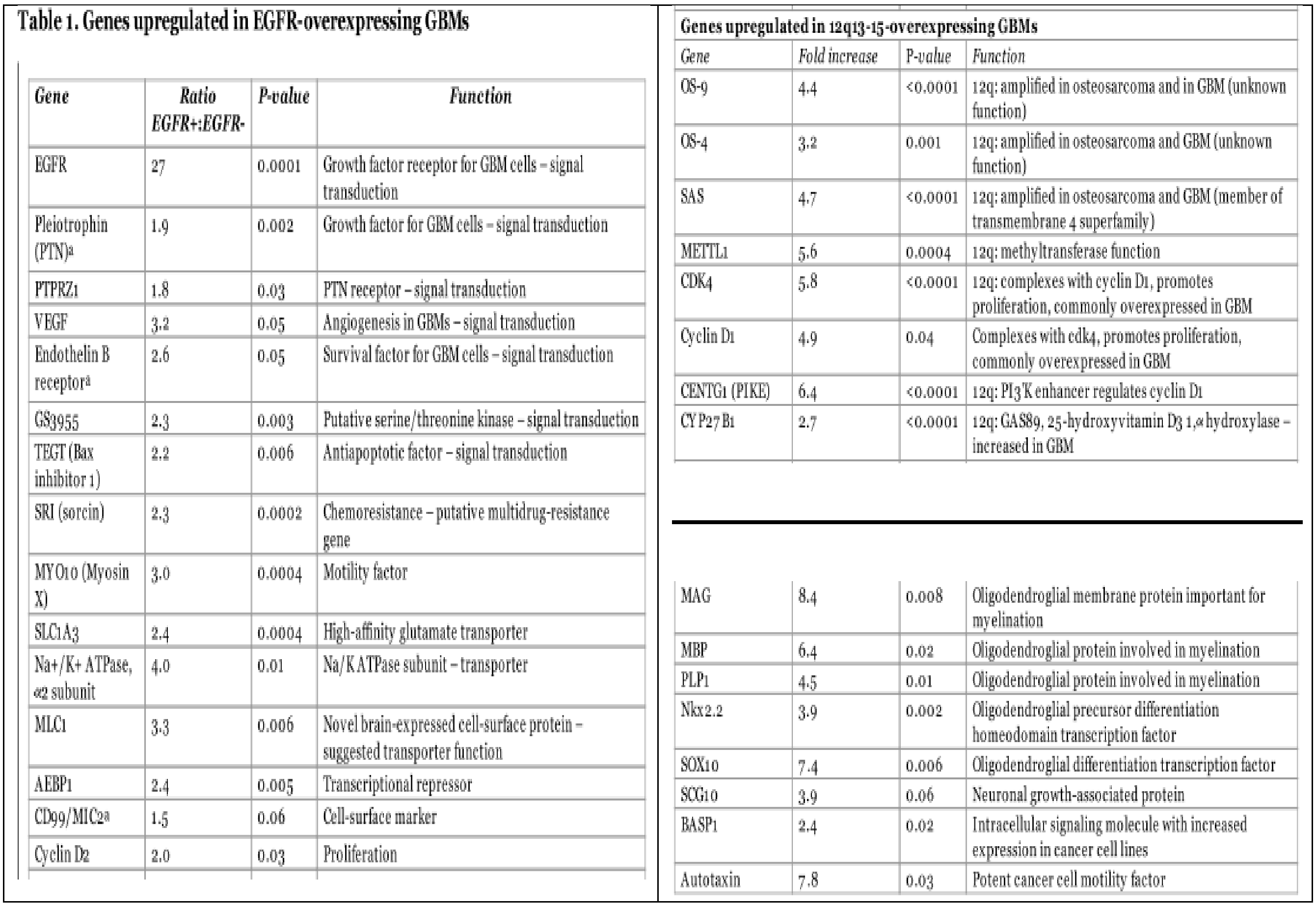


Table 1. Genes upregulated in EGFR-overexpressing GBMs

Gene	*Ratio EGFR+:EGFR-*	*P-value*	*Function*
EGFR	27	0.0001	Growth factor receptor for GBM cells – signal transduction
Pleiotrophin (PTN)[a]	1.9	0.002	Growth factor for GBM cells – signal transduction
PTPRZ1	1.8	0.03	PTN receptor – signal transduction
VEGF	3.2	0.05	Angiogenesis in GBMs – signal transduction
Endothelin B receptor[a]	2.6	0.05	Survival factor for GBM cells – signal transduction
GS3955	2.3	0.003	Putative serine/threonine kinase – signal transduction
TEGT (Bax inhibitor 1)	2.2	0.006	Antiapoptotic factor – signal transduction
SRI (sorcin)	2.3	0.0002	Chemoresistance – putative multidrug-resistance gene
MYO10 (Myosin X)	3.0	0.0004	Motility factor
SLC1A3	2.4	0.0004	High-affinity glutamate transporter
Na+/K+ ATPase, α2 subunit	4.0	0.01	Na/K ATPase subunit – transporter
MLC1	3.3	0.006	Novel brain-expressed cell-surface protein – suggested transporter function
AEBP1	2.4	0.005	Transcriptional repressor
CD99/MIC2[a]	1.5	0.06	Cell-surface marker
Cyclin D2	2.0	0.03	Proliferation

Genes upregulated in 12q13-15-overexpressing GBMs

Gene	*Fold increase*	*P-value*	*Function*
OS-9	4.4	<0.0001	12q: amplified in osteosarcoma and in GBM (unknown function)
OS-4	3.2	0.001	12q: amplified in osteosarcoma and GBM (unknown function)
SAS	4.7	<0.0001	12q: amplified in osteosarcoma and GBM (member of transmembrane 4 superfamily)
METTL1	5.6	0.0004	12q: methyltransferase function
CDK4	5.8	<0.0001	12q: complexes with cyclin D1, promotes proliferation, commonly overexpressed in GBM
Cyclin D1	4.9	0.04	Complexes with cdk4, promotes proliferation, commonly overexpressed in GBM
CENTG1 (PIKE)	6.4	<0.0001	12q: PI3'K enhancer regulates cyclin D1
CYP27B1	2.7	<0.0001	12q: GAS89, 25-hydroxyvitamin D3 1,α hydroxylase – increased in GBM

MAG	8.4	0.008	Oligodendroglial membrane protein important for myelination
MBP	6.4	0.02	Oligodendroglial protein involved in myelination
PLP1	4.5	0.01	Oligodendroglial protein involved in myelination
Nkx2.2	3.9	0.002	Oligodendroglial precursor differentiation homeodomain transcription factor
SOX10	7.4	0.006	Oligodendroglial differentiation transcription factor
SCG10	3.9	0.06	Neuronal growth-associated protein
BASP1	2.4	0.02	Intracellular signaling molecule with increased expression in cancer cell lines
Autotaxin	7.8	0.03	Potent cancer cell motility factor

Table 65 Genes overexpressed in GBM for EGFR and 12q13-15

XII. REFERENCES

1 Monti, S., Tamayo, P., Mesirov, J., and Golub, T.: 'Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data', Machine learning, 2003, 52, (1-2), pp. 91-118

2 Wikipedia: 'Bioinformatics'

3 Mangubat, A.: 'New Challenges in the Fast Changing Landscape of Bioinformatics'

4 Fernald, G.H., Capriotti, E., Daneshjou, R., Karczewski, K.J., and Altman, R.B.: 'Bioinformatics challenges for personalized medicine', Bioinformatics, 2011, 27, (13), pp. 1741-1748

5 Altman, R.B., and Raychaudhuri, S.: 'Whole-genome expression analysis: challenges beyond clustering', Current opinion in structural biology, 2001, 11, (3), pp. 340-347

6 Speed, T.: 'Statistical analysis of gene expression microarray data' (CRC Press, 2004. 2004)

7 Kleinberg, J.: 'An impossibility theorem for clustering', Advances in neural information processing systems, 2003, pp. 463-470

8 Zadeh, R.B., and Ben-David, S.: 'A uniqueness theorem for clustering'. Proc. Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, Quebec, Canada2009 pp. Pages

9 University, U.

10 Berrar, D.P., Dubitzky, W., and Granzow, M.: 'A practical approach to microarray data analysis' (Springer, 2003. 2003)

11 Lin, S.M., and Johnson, K.F.: 'Methods of microarray data analysis III' (Springer, 2003. 2003)

12 Shoemaker, J.S., and Lin, S.M.: 'Methods of microarray data analysis IV' (Springer, 2005. 2005)

13 Simon, R.M.: 'Design and analysis of DNA microarray investigations' (Springer, 2003. 2003)

14 Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., and Yu, X.: 'Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling', Nature, 2000, 403, (6769), pp. 503-511

15 Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., and Caligiuri, M.A.: 'Molecular classification of cancer: class discovery and class prediction by gene expression monitoring', science, 1999, 286, (5439), pp. 531-537

16 Fortunato, S.: 'Community detection in graphs', Physics Reports, 2010, 486, (3), pp. 75-174

17 Newman, M.: 'Communities, modules and large-scale structure in networks', Nature Physics, 2012, 8, (1), pp. 25-31

18 Aas, K.: 'Microarray data mining: A survey', NR Note, SAMBA, Norwegian Computing Center, 2001

19 Dupuy, A., and Simon, R.M.: 'Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting', Journal of the National Cancer Institute, 2007, 99, (2), pp. 147-157

20 Ben-Dor, A., Shamir, R., and Yakhini, Z.: 'Clustering gene expression patterns', Journal of computational biology, 1999, 6, (3-4), pp. 281-297

21 Inostroza, M.: 'An Integrated and Scalable Approach Based on Combinatorial Optimization Techniques for the Analysis of Microarray Data', 2008

22 Sharan, R., and Shamir, R.: 'CLICK: a clustering algorithm with applications to gene expression analysis', in Editor (Ed.)^(Eds.): 'Book CLICK: a clustering algorithm with applications to gene expression analysis' (2000, edn.), pp. 16

23 Guha, S., Rastogi, R., and Shim, K.: 'CURE: an efficient clustering algorithm for large databases', in Editor (Ed.)^(Eds.): 'Book CURE: an efficient clustering algorithm for large databases' (ACM, 1998, edn.), pp. 73-84

24 Swift, S., Tucker, A., Vinciotti, V., Martin, N., Orengo, C., Liu, X., and Kellam, P.: 'Consensus clustering and functional interpretation of gene-expression data', Genome biology, 2004, 5, (11), pp. R94

25 Simpson, T.I., Armstrong, J.D., and Jarman, A.P.: 'Merged consensus clustering to assess and improve class discovery with microarray data', BMC bioinformatics, 2010, 11, (1), pp. 590

26 Lee, M.-L.T.: 'Analysis of microarray gene expression data', 2004

27 Wesolowska, M.: 'A study on feature selection based on AICc and its application to microarray data', 2009

28 Arenas, A., Gómez, S., and Fernández, A.: 'Multiple resolution of the modular structure of complex networks', in Editor (Ed.)^(Eds.): 'Book Multiple resolution of the modular structure of complex networks' (2007, edn.), pp.

29 Vega-Pons, S., Correa-Morris, J., and Ruiz-Shulcloper, J.: 'Weighted partition consensus via kernels', Pattern Recognition, 2010, 43, (8), pp. 2712-2724

30 Castells Domingo, X.: 'Integration of high throughput data to detect groups of glioblastoma multiforme', 2013

31 Speer, N., Frohlich, H., Spieth, C., and Zell, A.: 'Functional grouping of genes using spectral clustering and Gene Ontology', in Editor (Ed.)^(Eds.): 'Book Functional grouping of genes using spectral clustering and Gene Ontology' (2005, edn.), pp. 298-303 vol. 291

32 Maulik, U., and Bandyopadhyay, S.: 'Performance evaluation of some clustering algorithms and validity indices', Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2002, 24, (12), pp. 1650-1654

33 Vega-Pons, S., and Ruiz-Shulcloper, J.: 'Partition selection approach for hierarchical clustering based on clustering ensemble': 'Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications' (Springer, 2010), pp. 525-532

34 Halkidi, M., Batistakis, Y., and Vazirgiannis, M.: 'Clustering validity checking methods: part II', ACM Sigmod Record, 2002, 31, (3), pp. 19-27

35 Granell, C., Gómez, S., and Arenas, A.: 'Mesoscopic analysis of networks: Applications to exploratory analysis and data clustering', Chaos: An Interdisciplinary Journal of Nonlinear Science, 2011, 21, (1), pp. 016102

36 Granell Martorell, C.: 'Exploratory data analysis using network based techniques', 2012

37 Grotkjær, T., Winther, O., Regenberg, B., Nielsen, J., and Hansen, L.K.: 'Robust multi-scale clustering of large DNA microarray datasets with the consensus algorithm', Bioinformatics, 2006, 22, (1), pp. 58-67

38 Goder, A., and Filkov, V.: 'Consensus Clustering Algorithms: Comparison and Refinement', in Editor (Ed.)^(Eds.): 'Book Consensus Clustering Algorithms: Comparison and Refinement' (SIAM, 2008, edn.), pp. 109-117

39 Lancichinetti, A., and Fortunato, S.: 'Consensus clustering in complex networks', Scientific reports, 2012, 2

40 Fortunato, S., and Barthelemy, M.: 'Resolution limit in community detection', Proceedings of the National Academy of Sciences, 2007, 104, (1), pp. 36-41

41 Sung, J., Kim, P.-J., Ma, S., Funk, C.C., Magis, A.T., Wang, Y., Hood, L., Geman, D., and Price, N.D.: 'Multi-study integration of brain cancer transcriptomes reveals organ-level molecular signatures', PLoS computational biology, 2013, 9, (7), pp. e1003148

42 Sun, L., Hui, A.-M., Su, Q., Vortmeyer, A., Kotliarov, Y., Pastorino, S., Passaniti, A., Menon, J., Walling, J., Bailey, R., Rosenblum, M., Mikkelsen, T., and Fine, H.A.: 'Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain', Cancer Cell, 2006, 9, (4), pp. 287-300

43 Berretta, R., and Moscato, P.: 'Cancer Biomarker Discovery: The Entropic Hallmark', PLoS ONE, 2010, 5, (8), pp. e12262

44 Ravetti, M.G., Rosso, O.A., Berretta, R., and Moscato, P.: 'Uncovering molecular biomarkers that correlate cognitive decline with the changes of hippocampus' gene expression profiles in Alzheimer's disease', PloS one, 2010, 5, (4), pp. e10153

45 Rosso, O.A., Craig, H., and Moscato, P.: 'Shakespeare and other English Renaissance authors as characterized by Information Theory complexity quantifiers', Physica A: Statistical Mechanics and its Applications, 2009, 388, (6), pp. 916-926

46 de Souto, M.C., Costa, I.G., de Araujo, D.S., Ludermir, T.B., and Schliep, A.: 'Clustering cancer gene expression data: a comparative study', BMC bioinformatics, 2008, 9, (1), pp. 497

47 Rao, C.R.: 'Diversity and dissimilarity coefficients: a unified approach', Theoretical Population Biology, 1982, 21, (1), pp. 24-43

48 Grosse, I., Bernaola-Galván, P., Carpena, P., Román-Roldán, R., Oliver, J., and Stanley, H.E.: 'Analysis of symbolic sequences using the Jensen-Shannon divergence', Physical Review E, 2002, 65, (4), pp. 041905

49 Olshen, R.A., and Rajaratnam, B.: 'Successive normalization of rectangular arrays', Annals of statistics, 2010, 38, (3), pp. 1638

50 Pearson, K.: 'Note on regression and inheritance in the case of two parents', Proceedings of the Royal Society of London, 1895, 58, (347-352), pp. 240-242

51 Taylor, J.R.: 'An Introduction To Error Analysis: The Study Of Uncertainties In Physical Measurements Author: John R. Taylor, Publisher', 1996

52 Gómez, S.A., Alex; Fernández, Albert: 'Radatools', 2007

53 Arenas, A., Fernandez, A., and Gomez, S.: 'Analysis of the structure of complex networks at different resolution levels', New Journal of Physics, 2008, 10, (5), pp. 053039

54 Kernighan, B.W., and Lin, S.: 'An efficient heuristic procedure for partitioning graphs', Bell system technical journal, 1970, 49, (2), pp. 291-307

55 Vahid, F., and Le, T.D.: 'Extending the Kernighan/Lin heuristic for hardware and software functional partitioning', Design Automation for Embedded Systems, 1997, 2, (2), pp. 237-261

56 Newman, M.E.: 'Fast algorithm for detecting community structure in networks', Physical review E, 2004, 69, (6), pp. 066133

57 Fernández, A., and Gómez, S.: 'Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms', Journal of Classification, 2008, 25, (1), pp. 43-65

58 Marsden, J., Budden, D., Craig, H., and Moscato, P.: 'Language Individuation and Marker Words: Shakespeare and His Maxwell's Demon', PloS one, 2013, 8, (6), pp. e66813

59 Rosenberg, A., and Hirschberg, J.: 'V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure', in Editor (Ed.)^(Eds.): 'Book V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure' (2007, edn.), pp. 410-420

60 Tyree, S., Weinberger, K.Q., Agrawal, K., and Paykin, J.: 'Parallel boosted regression trees for web search ranking'. Proc. Proceedings of the 20th international conference on World wide web, Hyderabad, India2011 pp. Pages

61 Ben-Haim, Y., and Tom-Tov, E.: 'A streaming parallel decision tree algorithm', The Journal of Machine Learning Research, 2010, 11, pp. 849-872

62 Peixoto, T.P.: 'Hierarchical block structures and high-resolution model selection in large networks', arXiv preprint arXiv:1310.4377, 2013

63 Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D.: 'Cluster analysis and display of genome-wide expression patterns', Proceedings of the National Academy of Sciences, 1998, 95, (25), pp. 14863-14868

64 MathWorks: 'Linkage', 2014

65 Jaccard, P.: 'The distribution of the flora in the alpine zone. 1', New phytologist, 1912, 11, (2), pp. 37-50

66 Fowlkes, E.B., and Mallows, C.L.: 'A method for comparing two hierarchical clusterings', Journal of the American statistical association, 1983, 78, (383), pp. 553-569

67 Strehl, A., and Ghosh, J.: 'Cluster ensembles---a knowledge reuse framework for combining multiple partitions', The Journal of Machine Learning Research, 2003, 3, pp. 583-617

68 Mirkin, B.: 'Mathematical classification and clustering: From how to what and why' (Springer, 1998. 1998)

69 Meilă, M.: 'Comparing clusterings—an information based distance', Journal of Multivariate Analysis, 2007, 98, (5), pp. 873-895

70 Reichart, R., and Rappoport, A.: 'The NVI clustering evaluation measure', in Editor (Ed.)^(Eds.): 'Book The NVI clustering evaluation measure' (Association for Computational Linguistics, 2009, edn.), pp. 165-173

71 Rand, W.M.: 'Objective criteria for the evaluation of clustering methods', Journal of the American Statistical association, 1971, 66, (336), pp. 846-850

72 Hubert, L., and Arabie, P.: 'Comparing partitions', Journal of classification, 1985, 2, (1), pp. 193-218

73 Yeung, K.Y., and Ruzzo, W.L.: 'Details of the adjusted Rand index and clustering algorithms, supplement to the paper "An empirical study on principal component analysis for clustering gene expression data"', Bioinformatics, 2001, 17, (9), pp. 763-774

74 Labatut, V.: 'Generalized Measures for the Evaluation of Community Detection Methods', arXiv preprint arXiv:1303.5441, 2013

75 van den Bent, M.J.: ‘Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician’s perspective’, Acta neuropathologica, 2010, 120, (3), pp. 297-304

76 Kros, J.M.: ‘Grading of gliomas: the road from eminence to evidence’, Journal of Neuropathology & Experimental Neurology, 2011, 70, (2), pp. 101-109

77 Sun, C.-H., Hwang, T., Oh, K., and Yi, G.-S.: ‘DynaMod: dynamic functional modularity analysis’, Nucleic acids research, 2010, 38, (suppl 2), pp. W103-W108

78 Yun, T., Hwang, T., Cha, K., and Yi, G.-S.: ‘CLIC: clustering analysis of large microarray datasets with individual dimension-based clustering’, Nucleic acids research, 2010, 38, (suppl 2), pp. W246-W253

79 Kim, Y.-W., Koul, D., Kim, S.H., Lucio-Eterovic, A.K., Freire, P.R., Yao, J., Wang, J., Almeida, J.S., Aldape, K., and Yung, W.A.: ‘Identification of prognostic gene signatures of glioblastoma: a study based on TCGA data analysis’, Neuro-oncology, 2013, 15, (7), pp. 829-839

80 Chen, D., Persson, A., Sun, Y., Salford, L.G., Nord, D.G., Englund, E., Jiang, T., and Fan, X.: ‘Better Prognosis of Patients with Glioma Expressing FGF2-Dependent PDGFRA Irrespective of Morphological Diagnosis’, PLoS ONE, 2013, 8, (4), pp. e61556

81 Mischel, P.S., Shai, R., Shi, T., Horvath, S., Lu, K.V., Choe, G., Seligson, D., Kremen, T.J., Palotie, A., and Liau, L.M.: ‘Identification of molecular subtypes of glioblastoma by gene expression profiling’, Oncogene, 2003, 22, (15), pp. 2361-2373

XIII. REFERENCES FOR BIO-INFORMATICS STUDY

1 Kong, J., Cooper, L.A., Wang, F., Gao, J., Teodoro, G., Scarpace, L., Mikkelsen, T., Schniederjan, M.J., Moreno, C.S., Saltz, J.H., and Brat, D.J.: 'Machine-based morphologic analysis of glioblastoma using whole-slide pathology images uncovers clinically relevant molecular correlates', PLoS One, 2013, 8, (11), pp. e81049

2 Saidi, A., Javerzat, S., Bellahcene, A., De Vos, J., Bello, L., Castronovo, V., Deprez, M., Loiseau, H., Bikfalvi, A., and Hagedorn, M.: 'Experimental anti-angiogenesis causes upregulation of genes associated with poor survival in glioblastoma', International journal of cancer. Journal international du cancer, 2008, 122, (10), pp. 2187-2198

3 Yue, X., Wang, P., Xu, J., Zhu, Y., Sun, G., Pang, Q., and Tao, R.: 'MicroRNA-205 functions as a tumor suppressor in human glioblastoma cells by targeting VEGF-A', Oncology reports, 2012, 27, (4), pp. 1200-1206

4 Grzelinski, M., Steinberg, F., Martens, T., Czubayko, F., Lamszus, K., and Aigner, A.: 'Enhanced antitumorigenic effects in glioblastoma on double targeting of pleiotrophin and its receptor ALK', Neoplasia, 2009, 11, (2), pp. 145-156

5 Heo, J.C., Jung, T.H., Jung, D.Y., Park, W.K., and Cho, H.: 'Indatraline inhibits Rho- and calcium-mediated glioblastoma cell motility and angiogenesis', Biochemical and biophysical research communications, 2014, 443, (2), pp. 749-755

6 Formolo, C.A., Williams, R., Gordish-Dressman, H., MacDonald, T.J., Lee, N.H., and Hathout, Y.: 'Secretome signature of invasive glioblastoma multiforme', J Proteome Res, 2011, 10, (7), pp. 3149-3159

7 Atai, N.A., Bansal, M., Lo, C., Bosman, J., Tigchelaar, W., Bosch, K.S., Jonker, A., De Witt Hamer, P.C., Troost, D., McCulloch, C.A., Everts, V., Van Noorden, C.J., and Sodek, J.: 'Osteopontin is up-regulated and associated with neutrophil and macrophage infiltration in glioblastoma', Immunology, 2011, 132, (1), pp. 39-48

8 Sreekanthreddy, P., Srinivasan, H., Kumar, D.M., Nijaguna, M.B., Sridevi, S., Vrinda, M., Arivazhagan, A., Balasubramaniam, A., Hegde, A.S., Chandramouli, B.A., Santosh, V., Rao, M.R., Kondaiah, P., and Somasundaram, K.: 'Identification of potential serum biomarkers of glioblastoma: serum osteopontin levels correlate with poor prognosis', Cancer Epidemiol Biomarkers Prev, 2010, 19, (6), pp. 1409-1422

9 Schuhmann, M.U., Zucht, H.D., Nassimi, R., Heine, G., Schneekloth, C.G., Stuerenburg, H.J., and Selle, H.: 'Peptide screening of cerebrospinal fluid in patients with glioblastoma multiforme', Eur J Surg Oncol, 2010, 36, (2), pp. 201-207

10 Lamour, V., Le Mercier, M., Lefranc, F., Hagedorn, M., Javerzat, S., Bikfalvi, A., Kiss, R., Castronovo, V., and Bellahcene, A.: 'Selective osteopontin knockdown exerts anti-tumoral activity in a human glioblastoma model', International journal of cancer. Journal international du cancer, 2010, 126, (8), pp. 1797-1805

11 Colin, C., Baeza, N., Bartoli, C., Fina, F., Eudes, N., Nanni, I., Martin, P.M., Ouafik, L., and Figarella-Branger, D.: 'Identification of genes differentially expressed in glioblastoma versus pilocytic astrocytoma using Suppression Subtractive Hybridization', Oncogene, 2006, 25, (19), pp. 2818-2826

12 Wei, K.C., Huang, C.Y., Chen, P.Y., Feng, L.Y., Wu, T.W., Chen, S.M., Tsai, H.C., Lu, Y.J., Tsang, N.M., Tseng, C.K., Pai, P.C., and Shin, J.W.: 'Evaluation of the prognostic value of CD44 in glioblastoma multiforme', Anticancer Res, 2010, 30, (1), pp. 253-259

13 Wei, K.C., Huang, C.Y., Chen, P.Y., Feng, L.Y., Wu, T.W., Chen, S.M., Tsai, H.C., Lu, Y.J., Tsang, N.M., Tseng, C.K., Pai, P.C., and Shin, J.W.: 'Evaluation of the prognostic value of CD44 in glioblastoma multiforme', Anticancer research, 2010, 30, (1), pp. 253-259

14 Sreekanthreddy, P., Srinivasan, H., Kumar, D.M., Nijaguna, M.B., Sridevi, S., Vrinda, M., Arivazhagan, A., Balasubramaniam, A., Hegde, A.S., Chandramouli, B.A., Santosh, V., Rao, M.R., Kondaiah, P., and Somasundaram, K.: 'Identification of potential serum biomarkers of glioblastoma: serum osteopontin levels correlate with poor prognosis', Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 2010, 19, (6), pp. 1409-1422

15 Velpula, K.K., Dasari, V.R., and Rao, J.S.: 'The homing of human cord blood stem cells to sites of inflammation: unfolding mysteries of a novel therapeutic paradigm for glioblastoma multiforme', Cell cycle, 2012, 11, (12), pp. 2303-2313

16 Formolo, C.A., Williams, R., Gordish-Dressman, H., MacDonald, T.J., Lee, N.H., and Hathout, Y.: 'Secretome signature of invasive glioblastoma multiforme', Journal of proteome research, 2011, 10, (7), pp. 3149-3159

17 Schuhmann, M.U., Zucht, H.D., Nassimi, R., Heine, G., Schneekloth, C.G., Stuerenburg, H.J., and Selle, H.: 'Peptide screening of cerebrospinal fluid in patients with glioblastoma multiforme', European journal of surgical

oncology : the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology, 2010, 36, (2), pp. 201-207

18 Lamour, V., Le Mercier, M., Lefranc, F., Hagedorn, M., Javerzat, S., Bikfalvi, A., Kiss, R., Castronovo, V., and Bellahcene, A.: 'Selective osteopontin knockdown exerts anti-tumoral activity in a human glioblastoma model', International journal of cancer. Journal international du cancer, 2010, 126, (8), pp. 1797-1805

19 Natarajan, M., Stewart, J.E., Golemis, E.A., Pugacheva, E.N., Alexandropoulos, K., Cox, B.D., Wang, W., Grammer, J.R., and Gladson, C.L.: 'HEF1 is a necessary and specific downstream effector of FAK that promotes the migration of glioblastoma cells', Oncogene, 2006, 25, (12), pp. 1721-1732

20 Venugopal, C., Wang, X.S., Manoranjan, B., McFarlane, N., Nolte, S., Li, M., Murty, N., Siu, K.W., and Singh, S.K.: 'GBM secretome induces transient transformation of human neural precursor cells', J Neurooncol, 2012, 109, (3), pp. 457-466

21 Eurich, K., Segawa, M., Toei-Shimizu, S., and Mizoguchi, E.: 'Potential role of chitinase 3-like-1 in inflammation-associated carcinogenic changes of epithelial cells', World J Gastroenterol, 2009, 15, (42), pp. 5249-5259

22 Shao, R., Francescone, R., Ngernyuang, N., Bentley, B., Taylor, S.L., Moral, L., and Yan, W.: 'Anti-YKL-40 antibody and ionizing irradiation synergistically inhibit tumor vascularization and malignancy in glioblastoma', Carcinogenesis, 2014, 35, (2), pp. 373-382

23 Ma, X., Yoshimoto, K., Guan, Y., Hata, N., Mizoguchi, M., Sagata, N., Murata, H., Kuga, D., Amano, T., Nakamizo, A., and Sasaki, T.: 'Associations between microRNA expression and mesenchymal marker gene expression in glioblastoma', Neuro Oncol, 2012, 14, (9), pp. 1153-1162

24 Salvati, M., Pichierri, A., Piccirilli, M., Floriana Brunetto, G.M., D'Elia, A., Artizzu, S., Santoro, F., Arcella, A., Giangaspero, F., Frati, A., Simione, L., and Santoro, A.: 'Extent of tumor removal and molecular markers in cerebral glioblastoma: a combined prognostic factors study in a surgical series of 105 patients', J Neurosurg, 2012, 117, (2), pp. 204-211

25 Desantis, S.M., Houseman, E.A., Coull, B.A., Nutt, C.L., and Betensky, R.A.: 'Supervised Bayesian latent class models for high-dimensional data', Stat Med, 2012, 31, (13), pp. 1342-1360

26 Bernardi, D., Padoan, A., Ballin, A., Sartori, M., Manara, R., Scienza, R., Plebani, M., and Della Puppa, A.: 'Serum YKL-40 following resection for cerebral glioblastoma', J Neurooncol, 2012, 107, (2), pp. 299-305

27 Singh, S.K., Bhardwaj, R., Wilczynska, K.M., Dumur, C.I., and Kordula, T.: 'A complex of nuclear factor I-X3 and STAT3 regulates astrocyte and glioma migration through the secreted glycoprotein YKL-40', J Biol Chem, 2011, 286, (46), pp. 39893-39903

28 Kavsan, V.M., Baklaushev, V.P., Balynska, O.V., Iershov, A.V., Areshkov, P.O., Yusubalieva, G.M., Grinenko, N.P., Victorov, I.V., Rymar, V.I., Sanson, M., and Chekhonin, V.P.: 'Gene Encoding Chitinase 3-Like 1 Protein (CHI3L1) is a Putative Oncogene', Int J Biomed Sci, 2011, 7, (3), pp. 230-237

29 Iwamoto, F.M., Hottinger, A.F., Karimi, S., Riedel, E., Dantis, J., Jahdi, M., Panageas, K.S., Lassman, A.B., Abrey, L.E., Fleisher, M., DeAngelis, L.M., Holland, E.C., and Hormigo, A.: 'Serum YKL-40 is a marker of prognosis and disease status in high-grade gliomas', Neuro Oncol, 2011, 13, (11), pp. 1244-1251

30 Serao, N.V., Delfino, K.R., Southey, B.R., Beever, J.E., and Rodriguez-Zas, S.L.: 'Cell cycle and aging, morphogenesis, and response to stimuli genes are individualized biomarkers of glioblastoma progression and survival', BMC Med Genomics, 2011, 4, pp. 49

31 Francescone, R.A., Scully, S., Faibish, M., Taylor, S.L., Oh, D., Moral, L., Yan, W., Bentley, B., and Shao, R.: 'Role of YKL-40 in the angiogenesis, radioresistance, and progression of glioblastoma', J Biol Chem, 2011, 286, (17), pp. 15332-15343

32 Zhang, W., Murao, K., Zhang, X., Matsumoto, K., Diah, S., Okada, M., Miyake, K., Kawai, N., Fei, Z., and Tamiya, T.: 'Resveratrol represses YKL-40 expression in human glioma U87 cells', BMC cancer, 2010, 10, pp. 593

33 Lin, B., Madan, A., Yoon, J.G., Fang, X., Yan, X., Kim, T.K., Hwang, D., Hood, L., and Foltz, G.: 'Massively parallel signature sequencing and bioinformatics analysis identifies up-regulation of TGFBI and SOX4 in human glioblastoma', PLoS One, 2010, 5, (4), pp. e10210

34 Horbinski, C., Wang, G., and Wiley, C.A.: 'YKL-40 is directly produced by tumor cells and is inversely linked to EGFR in glioblastomas', Int J Clin Exp Pathol, 2010, 3, (3), pp. 226-237

35 Antonelli, M., Buttarelli, F.R., Arcella, A., Nobusawa, S., Donofrio, V., Oghaki, H., and Giangaspero, F.: 'Prognostic significance of histological grading, p53 status, YKL-40 expression, and IDH1 mutations in pediatric high-grade gliomas', J Neurooncol, 2010, 99, (2), pp. 209-215

36 Ducray, F., Idbaih, A., de Reynies, A., Bieche, I., Thillet, J., Mokhtari, K., Lair, S., Marie, Y., Paris, S., Vidaud, M., Hoang-Xuan, K., Delattre, O., Delattre, J.Y., and Sanson, M.: 'Anaplastic oligodendrogliomas with 1p19q codeletion have a proneural gene expression profile', Mol Cancer, 2008, 7, pp. 41

37 Saidi, A., Javerzat, S., Bellahcene, A., De Vos, J., Bello, L., Castronovo, V., Deprez, M., Loiseau, H., Bikfalvi, A., and Hagedorn, M.: 'Experimental anti-angiogenesis causes upregulation of genes associated with poor survival in glioblastoma', Int J Cancer, 2008, 122, (10), pp. 2187-2198

38 Kroes, R.A., Dawson, G., and Moskal, J.R.: 'Focused microarray analysis of glyco-gene expression in human glioblastomas', J Neurochem, 2007, 103 Suppl 1, pp. 14-24

39 Pelloski, C.E., Ballman, K.V., Furth, A.F., Zhang, L., Lin, E., Sulman, E.P., Bhat, K., McDonald, J.M., Yung, W.K., Colman, H., Woo, S.Y., Heimberger, A.B., Suki, D., Prados, M.D., Chang, S.M., Barker, F.G., 2nd, Buckner, J.C., James, C.D., and Aldape, K.: 'Epidermal growth factor receptor variant III status defines clinically distinct subtypes of glioblastoma', J Clin Oncol, 2007, 25, (16), pp. 2288-2294

40 Hormigo, A., Gu, B., Karimi, S., Riedel, E., Panageas, K.S., Edgar, M.A., Tanwar, M.K., Rao, J.S., Fleisher, M., DeAngelis, L.M., and Holland, E.C.: 'YKL-40 and matrix metalloproteinase-9 as potential serum biomarkers for patients with high-grade gliomas', Clin Cancer Res, 2006, 12, (19), pp. 5698-5704

41 Pelloski, C.E., Lin, E., Zhang, L., Yung, W.K., Colman, H., Liu, J.L., Woo, S.Y., Heimberger, A.B., Suki, D., Prados, M., Chang, S., Barker, F.G., 3rd, Fuller, G.N., and Aldape, K.D.: 'Prognostic associations of activated mitogen-activated protein kinase and Akt pathways in glioblastoma', Clin Cancer Res, 2006, 12, (13), pp. 3935-3941

42 Johansen, J.S., Jensen, B.V., Roslind, A., Nielsen, D., and Price, P.A.: 'Serum YKL-40, a new prognostic biomarker in cancer patients?', Cancer Epidemiol Biomarkers Prev, 2006, 15, (2), pp. 194-202

43 Tso, C.L., Freije, W.A., Day, A., Chen, Z., Merriman, B., Perlina, A., Lee, Y., Dia, E.Q., Yoshimoto, K., Mischel, P.S., Liau, L.M., Cloughesy, T.F., and Nelson, S.F.: 'Distinct transcription profiles of primary and secondary glioblastoma subgroups', Cancer Res, 2006, 66, (1), pp. 159-167

44 Pelloski, C.E., Mahajan, A., Maor, M., Chang, E.L., Woo, S., Gilbert, M., Colman, H., Yang, H., Ledoux, A., Blair, H., Passe, S., Jenkins, R.B., and Aldape, K.D.: 'YKL-40 expression is associated with poorer response to radiation and shorter overall survival in glioblastoma', Clin Cancer Res, 2005, 11, (9), pp. 3326-3334

45 Nutt, C.L., Betensky, R.A., Brower, M.A., Batchelor, T.T., Louis, D.N., and Stemmer-Rachamimov, A.O.: 'YKL-40 is a differential diagnostic marker for histologic subtypes of high-grade gliomas', Clin Cancer Res, 2005, 11, (6), pp. 2258-2264

46 Junker, N., Johansen, J.S., Hansen, L.T., Lund, E.L., and Kristjansen, P.E.: 'Regulation of YKL-40 expression during genotoxic or microenvironmental stress in human glioblastoma cells', Cancer Sci, 2005, 96, (3), pp. 183-190

47 Nigro, J.M., Misra, A., Zhang, L., Smirnov, I., Colman, H., Griffin, C., Ozburn, N., Chen, M., Pan, E., Koul, D., Yung, W.K., Feuerstein, B.G., and Aldape, K.D.: 'Integrated array-comparative genomic hybridization and expression array profiles identify clinically relevant molecular subtypes of glioblastoma', Cancer Res, 2005, 65, (5), pp. 1678-1686

48 Shostak, K., Labunskyy, V., Dmitrenko, V., Malisheva, T., Shamayev, M., Rozumenko, V., Zozulya, Y., Zehetner, G., and Kavsan, V.: 'HC gp-39 gene is upregulated in glioblastomas', Cancer Lett, 2003, 198, (2), pp. 203-210

49 Tanwar, M.K., Gilbert, M.R., and Holland, E.C.: 'Gene expression microarray analysis reveals YKL-40 to be a potential serum marker for malignant character in human glioma', Cancer Res, 2002, 62, (15), pp. 4364-4368

50 Balzeau, J., Peterson, A., and Eyer, J.: 'The vimentin-tubulin binding site peptide (Vim-TBS.58-81) crosses the plasma membrane and enters the nuclei of human glioma cells', Int J Pharm, 2012, 423, (1), pp. 77-83

51 Fortin, S., Le Mercier, M., Camby, I., Spiegl-Kreinecker, S., Berger, W., Lefranc, F., and Kiss, R.: 'Galectin-1 is implicated in the protein kinase C epsilon/vimentin-controlled trafficking of integrin-beta1 in glioblastoma cells', Brain Pathol, 2010, 20, (1), pp. 39-49

52 Sembritzki, O., Hagel, C., Lamszus, K., Deppert, W., and Bohn, W.: 'Cytoplasmic localization of wild-type p53 in glioblastomas correlates with expression of vimentin and glial fibrillary acidic protein', Neuro Oncol, 2002, 4, (3), pp. 171-178

53 Perzelcova, A., Macikova, Tardy, M., Mraz, P., Steno, J., and Bizik, I.: 'Co-expression of GFAP, vimentin and cytokeratins in GL-15 glioblastoma cell line', Neoplasma, 2000, 47, (6), pp. 362-366

54 Skalli, O., Wilhelmsson, U., Orndahl, C., Fekete, B., Malmgren, K., Rydenhag, B., and Pekny, M.: 'Astrocytoma grade IV (glioblastoma multiforme) displays 3 subtypes with unique expression profiles of intermediate filament proteins', Hum Pathol, 2013, 44, (10), pp. 2081-2088

55 Sun, S., Wong, T.S., Zhang, X.Q., Pu, J.K., Lee, N.P., Day, P.J., Ng, G.K., Lui, W.M., and Leung, G.K.: 'Protein alterations associated with temozolomide resistance in subclones of human glioblastoma cell lines', J Neurooncol, 2012, 107, (1), pp. 89-100

56 Svendsen, A., Verhoeff, J.J., Immervoll, H., Brogger, J.C., Kmiecik, J., Poli, A., Netland, I.A., Prestegarden, L., Planaguma, J., Torsvik, A., Kjersem, A.B., Sakariassen, P.O., Heggdal, J.I., Van Furth, W.R., Bjerkvig, R., Lund-Johansen, M., Enger, P.O., Felsberg, J., Brons, N.H., Tronstad, K.J., Waha, A., and Chekenya, M.: 'Expression of the progenitor marker NG2/CSPG4 predicts poor survival and resistance to ionising radiation in glioblastoma', Acta Neuropathol, 2011, 122, (4), pp. 495-510

57 Wolanczyk, M., Hulas-Bigoszewska, K., Witusik-Perkowska, M., Papierz, W., Jaskolski, D., Liberski, P.P., and Rieske, P.: 'Imperfect oligodendrocytic and neuronal differentiation of glioblastoma cells', Folia Neuropathol, 2010, 48, (1), pp. 27-34

58 Zheng, L.T., Lee, S., Yin, G.N., Mori, K., and Suk, K.: 'Down-regulation of lipocalin 2 contributes to chemoresistance in glioblastoma cells', J Neurochem, 2009, 111, (5), pp. 1238-1251

59 Hoelzinger, D.B., Mariani, L., Weis, J., Woyke, T., Berens, T.J., McDonough, W.S., Sloan, A., Coons, S.W., and Berens, M.E.: 'Gene expression profile of glioblastoma multiforme invasive phenotype points to new therapeutic targets', Neoplasia, 2005, 7, (1), pp. 7-16

60 Westhoff, M.A., Zhou, S., Nonnenmacher, L., Karpel-Massler, G., Jennewein, C., Schneider, M., Halatsch, M.E., Carragher, N.O., Baumann, B., Krause, A., Simmet, T., Bachem, M.G., Wirtz, C.R., and Debatin, K.M.: 'Inhibition of NF-kappaB signaling ablates the invasive phenotype of glioblastoma', Molecular cancer research : MCR, 2013, 11, (12), pp. 1611-1623

61 Yang, W., and Yee, A.J.: 'Versican V2 isoform enhances angiogenesis by regulating endothelial cell activities and fibronectin expression', FEBS letters, 2013, 587, (2), pp. 185-192

62 Pedron, S., and Harley, B.A.: 'Impact of the biophysical features of a 3D gelatin microenvironment on glioblastoma malignancy', Journal of biomedical materials research. Part A, 2013, 101, (12), pp. 3404-3415

63 DeLay, M., Jahangiri, A., Carbonell, W.S., Hu, Y.L., Tsao, S., Tom, M.W., Paquette, J., Tokuyasu, T.A., and Aghi, M.K.: 'Microarray analysis verifies two distinct phenotypes of glioblastomas resistant to antiangiogenic therapy', Clin Cancer Res, 2012, 18, (10), pp. 2930-2942

64 Sengupta, S., Nandi, S., Hindi, E.S., Wainwright, D.A., Han, Y., and Lesniak, M.S.: 'Short hairpin RNA-mediated fibronectin knockdown delays tumor growth in a mouse glioma model', Neoplasia, 2010, 12, (10), pp. 837-847

65 Rieske, P., Golanska, E., Zakrzewska, M., Piaskowski, S., Hulas-Bigoszewska, K., Wolanczyk, M., Szybka, M., Witusik-Perkowska, M., Jaskolski, D.J., Zakrzewski, K., Biernat, W., Krynska, B., and Liberski, P.P.: 'Arrested neural and advanced mesenchymal differentiation of glioblastoma cells-comparative study with neural progenitors', BMC cancer, 2009, 9, pp. 54

66 Ulrich, T.A., de Juan Pardo, E.M., and Kumar, S.: 'The mechanical rigidity of the extracellular matrix regulates the structure, motility, and proliferation of glioma cells', Cancer Res, 2009, 69, (10), pp. 4167-4174

67 Mikheeva, S.A., Mikheev, A.M., Petit, A., Beyer, R., Oxford, R.G., Khorasani, L., Maxwell, J.P., Glackin, C.A., Wakimoto, H., Gonzalez-Herrero, I., Sanchez-Garcia, I., Silber, J.R., Horner, P.J., and Rostomily, R.C.: 'TWIST1 promotes invasion through mesenchymal change in human glioblastoma', Mol Cancer, 2010, 9, pp. 194

68 Kim, H., Lee, Y., Lee, I.H., Kim, S., Kim, D., Saw, P.E., Lee, J., Choi, M., Kim, Y.C., and Jon, S.: 'Synthesis and therapeutic evaluation of an aptide-docetaxel conjugate targeting tumor-associated fibronectin', Journal of controlled release : official journal of the Controlled Release Society, 2014, 178, pp. 118-124

69 Serres, E., Debarbieux, F., Stanchi, F., Maggiorella, L., Grall, D., Turchi, L., Burel-Vandenbos, F., Figarella-Branger, D., Virolle, T., Rougon, G., and Van Obberghen-Schilling, E.: 'Fibronectin expression in glioblastomas promotes cell cohesion, collective invasion of basement membrane in vitro and orthotopic tumor growth in mice', Oncogene, 2013

70 Sabari, J., Lax, D., Connors, D., Brotman, I., Mindrebo, E., Butler, C., Entersz, I., Jia, D., and Foty, R.A.: 'Fibronectin matrix assembly suppresses dispersal of glioblastoma cells', PLoS One, 2011, 6, (9), pp. e24810

71 Lee, H.K., Seo, I.A., Shin, Y.K., Lee, S.H., Seo, S.Y., Suh, D.J., and Park, H.T.: 'Netrin-1 specifically enhances cell spreading on fibronectin in human glioblastoma cells', The Korean journal of physiology & pharmacology : official journal of the Korean Physiological Society and the Korean Society of Pharmacology, 2008, 12, (5), pp. 225-230

72 Yuan, L., Siegel, M., Choi, K., Khosla, C., Miller, C.R., Jackson, E.N., Piwnica-Worms, D., and Rich, K.M.: 'Transglutaminase 2 inhibitor, KCC009, disrupts fibronectin assembly in the extracellular matrix and sensitizes orthotopic glioblastomas to chemotherapy', Oncogene, 2007, 26, (18), pp. 2563-2573

73 Huang, J.M., Tian, X.X., Zhong, Y.F., Ma, D.L., Ma, Y., You, J.F., and Zhang, Y.: '[Effects of beta1-integrin, fibronectin and laminin on invasive behavior of human gliomas]', Zhonghua bing li xue za zhi Chinese journal of pathology, 2006, 35, (8), pp. 478-482

74 Lo, K.M., Lan, Y., Lauder, S., Zhang, J., Brunkhorst, B., Qin, G., Verma, R., Courtenay-Luck, N., and Gillies, S.D.: 'huBC1-IL12, an immunocytokine which targets EDB-containing oncofetal fibronectin in tumors and tumor vasculature, shows potent anti-tumor activity in human tumor models', Cancer immunology, immunotherapy : CII, 2007, 56, (4), pp. 447-457

75 Spaeth, N., Wyss, M.T., Pahnke, J., Biollaz, G., Trachsel, E., Drandarov, K., Treyer, V., Weber, B., Neri, D., and Buck, A.: 'Radioimmunotherapy targeting the extra domain B of fibronectin in C6 rat gliomas: a preliminary study about the therapeutic efficacy of iodine-131-labeled SIP(L19)', Nuclear medicine and biology, 2006, 33, (5), pp. 661-666

76 Caffo, M., Germano, A., Caruso, G., Meli, F., Galatioto, S., Sciacca, M.P., and Tomasello, F.: 'An immunohistochemical study of extracellular matrix proteins laminin, fibronectin and type IV collagen in paediatric glioblastoma multiforme', Acta neurochirurgica, 2004, 146, (10), pp. 1113-1118; discussion 1118

77 Huang, W., Chiquet-Ehrismann, R., Moyano, J.V., Garcia-Pardo, A., and Orend, G.: 'Interference of tenascin-C with syndecan-4 binding to fibronectin blocks cell adhesion and stimulates tumor cell proliferation', Cancer Res, 2001, 61, (23), pp. 8586-8594

78 Liu, C., Yao, J., Mercola, D., and Adamson, E.: 'The transcription factor EGR-1 directly transactivates the fibronectin gene and enhances attachment of human glioblastoma cell line U251', J Biol Chem, 2000, 275, (27), pp. 20315-20323

79 Ohnishi, T., Hiraga, S., Izumoto, S., Matsumura, H., Kanemura, Y., Arita, N., and Hayakawa, T.: 'Role of fibronectin-stimulated tumor cell migration in glioma invasion in vivo: clinical significance of fibronectin and fibronectin receptor expressed in human glioma tissues', Clinical & experimental metastasis, 1998, 16, (8), pp. 729-741

80 Castellani, P., Viale, G., Dorcaratto, A., Nicolo, G., Kaczmarek, J., Querze, G., and Zardi, L.: 'The fibronectin isoform containing the ED-B oncofetal domain: a marker of angiogenesis', International journal of cancer. Journal international du cancer, 1994, 59, (5), pp. 612-618

81 Mapstone, T.B., and Galloway, P.G.: 'Expression of glial fibrillary acidic protein, vimentin, fibronectin, and N-myc oncoprotein in primary human brain tumor cell explants', Pediatric neurosurgery, 1991, 17, (4), pp. 169-174

82 Wang, E., Zhang, C., Polavaram, N., Liu, F., Wu, G., Schroeder, M.A., Lau, J.S., Mukhopadhyay, D., Jiang, S.W., O'Neill, B.P., Datta, K., and Li, J.: 'The role of factor inhibiting HIF (FIH-1) in inhibiting HIF-1 transcriptional activity in glioblastoma multiforme', PLoS One, 2014, 9, (1), pp. e86102

83 Dieterich, L.C., Mellberg, S., Langenkamp, E., Zhang, L., Zieba, A., Salomaki, H., Teichert, M., Huang, H., Edqvist, P.H., Kraus, T., Augustin, H.G., Olofsson, T., Larsson, E., Soderberg, O., Molema, G., Ponten, F., Georgii-Hemming, P., Alafuzoff, I., and Dimberg, A.: 'Transcriptional profiling of human glioblastoma vessels indicates a key role of VEGF-A and TGFbeta2 in vascular abnormalization', The Journal of pathology, 2012, 228, (3), pp. 378-390

84 Sie, M., de Bont, E.S., Scherpen, F.J., Hoving, E.W., and den Dunnen, W.F.: 'Tumour vasculature and angiogenic profile of paediatric pilocytic astrocytoma; is it much different from glioblastoma?', Neuropathology and applied neurobiology, 2010, 36, (7), pp. 636-647

85 Dreyfuss, J.M., Johnson, M.D., and Park, P.J.: 'Meta-analysis of glioblastoma multiforme versus anaplastic astrocytoma identifies robust gene markers', Mol Cancer, 2009, 8, pp. 71

86 Neidert, M.C., Schoor, O., Trautwein, C., Trautwein, N., Christ, L., Melms, A., Honegger, J., Rammensee, H.G., Herold-Mende, C., Dietrich, P.Y., and Stevanovic, S.: 'Natural HLA class I ligands from glioblastoma: extending the options for immunotherapy', J Neurooncol, 2013, 111, (3), pp. 285-294

87 Myers, C.E., Hanavan, P., Antwi, K., Mahadevan, D., Nadeem, A.J., Cooke, L., Scheck, A.C., Laughrey, Z., and Lake, D.F.: 'CTL recognition of a novel HLA-A*0201-binding peptide derived from glioblastoma multiforme tumor cells', Cancer immunology, immunotherapy : CII, 2011, 60, (9), pp. 1319-1332

88 Raghu, H., Nalla, A.K., Gondi, C.S., Gujrati, M., Dinh, D.H., and Rao, J.S.: 'uPA and uPAR shRNA inhibit angiogenesis via enhanced secretion of SVEGFR1 independent of GM-CSF but dependent on TIMP-1 in endothelial and glioblastoma cells', Molecular oncology, 2012, 6, (1), pp. 33-47

89 Li, S.C., Vu, L.T., Ho, H.W., Yin, H.Z., Keschrumrus, V., Lu, Q., Wang, J., Zhang, H., Ma, Z., Stover, A., Weiss, J.H., Schwartz, P.H., and Loudon, W.G.: 'Cancer stem cells from a rare form of glioblastoma multiforme involving the neurogenic ventricular wall', Cancer cell international, 2012, 12, (1), pp. 41

90 Polisetty, R.V., Gupta, M.K., Nair, S.C., Ramamoorthy, K., Tiwary, S., Shiras, A., Chandak, G.R., and Sirdeshmukh, R.: ‘Glioblastoma cell secretome: analysis of three glioblastoma cell lines reveal 148 non-redundant proteins’, Journal of proteomics, 2011, 74, (10), pp. 1918-1925

91 Kolenda, J., Jensen, S.S., Aaberg-Jessen, C., Christensen, K., Andersen, C., Brunner, N., and Kristensen, B.W.: ‘Effects of hypoxia on expression of a panel of stem cell and chemoresistance markers in glioblastoma-derived spheroids’, J Neurooncol, 2011, 103, (1), pp. 43-58

92 Crocker, M., Ashley, S., Giddings, I., Petrik, V., Hardcastle, A., Aherne, W., Pearson, A., Bell, B.A., Zacharoulis, S., and Papadopoulos, M.C.: ‘Serum angiogenic profile of patients with glioblastoma identifies distinct tumor subtypes and shows that TIMP-1 is a prognostic factor’, Neuro Oncol, 2011, 13, (1), pp. 99-108

93 Aaberg-Jessen, C., Christensen, K., Offenberg, H., Bartels, A., Dreehsen, T., Hansen, S., Schroder, H.D., Brunner, N., and Kristensen, B.W.: ‘Low expression of tissue inhibitor of metalloproteinases-1 (TIMP-1) in glioblastoma predicts longer patient survival’, J Neurooncol, 2009, 95, (1), pp. 117-128

94 Zhao, Y., Lyons, C.E., Jr., Xiao, A., Templeton, D.J., Sang, Q.A., Brew, K., and Hussaini, I.M.: ‘Urokinase directly activates matrix metalloproteinases-9: a potential role in glioblastoma invasion’, Biochemical and biophysical research communications, 2008, 369, (4), pp. 1215-1220

95 Tsatas, D., Kanagasundaram, V., Kaye, A., and Novak, U.: ‘EGF receptor modifies cellular responses to hyaluronan in glioblastoma cell lines’, Journal of clinical neuroscience : official journal of the Neurosurgical Society of Australasia, 2002, 9, (3), pp. 282-288

96 Park, M.J., Park, I.C., Hur, J.H., Rhee, C.H., Choe, T.B., Yi, D.H., Hong, S.I., and Lee, S.H.: ‘Protein kinase C activation by phorbol ester increases in vitro invasion through regulation of matrix metalloproteinases/tissue inhibitors of metalloproteinases system in D54 human glioblastoma cells’, Neuroscience letters, 2000, 290, (3), pp. 201-204

97 Mohanam, S., Wang, S.W., Rayford, A., Yamamoto, M., Sawaya, R., Nakajima, M., Liotta, L.A., Nicolson, G.L., Stetler-Stevenson, W.G., and Rao, J.S.: ‘Expression of tissue inhibitors of metalloproteinases: negative regulators of human glioblastoma invasion in vivo’, Clinical & experimental metastasis, 1995, 13, (1), pp. 57-62

98 Falnoga, I., Zelenik Pevec, A., Slejkovec, Z., Znidaric, M.T., Zajc, I., Mlakar, S.J., and Marc, J.: ‘Arsenic trioxide (ATO) influences the gene expression of metallothioneins in human glioblastoma cells’, Biological trace element research, 2012, 149, (3), pp. 331-339

99 Zahonero, C., and Sanchez-Gomez, P.: ‘EGFR-dependent mechanisms in glioblastoma: towards a better therapeutic strategy’, Cellular and molecular life sciences : CMLS, 2014

100 Zadeh, G., Bhat, K.P., and Aldape, K.: ‘EGFR and EGFRvIII in glioblastoma: partners in crime’, Cancer cell, 2013, 24, (4), pp. 403-404

101 Verreault, M., Weppler, S.A., Stegeman, A., Warburton, C., Strutt, D., Masin, D., and Bally, M.B.: ‘Combined RNAi-mediated suppression of Rictor and EGFR resulted in complete tumor regression in an orthotopic glioblastoma tumor model’, PLoS One, 2013, 8, (3), pp. e59597

102 Lee, J.C., Vivanco, I., Beroukhim, R., Huang, J.H., Feng, W.L., DeBiasi, R.M., Yoshimoto, K., King, J.C., Nghiemphu, P., Yuza, Y., Xu, Q., Greulich, H., Thomas, R.K., Paez, J.G., Peck, T.C., Linhart, D.J., Glatt, K.A., Getz, G., Onofrio, R., Ziaugra, L., Levine, R.L., Gabriel, S., Kawaguchi, T., O'Neill, K., Khan, H., Liau, L.M., Nelson, S.F., Rao, P.N., Mischel, P., Pieper, R.O., Cloughesy, T., Leahy, D.J., Sellers, W.R., Sawyers, C.L., Meyerson, M., and Mellinghoff, I.K.: ‘Epidermal growth factor receptor activation in glioblastoma through novel missense mutations in the extracellular domain’, PLoS medicine, 2006, 3, (12), pp. e485

103 Bienkowski, M., Piaskowski, S., Stoczynska-Fidelus, E., Szybka, M., Banaszczyk, M., Witusik-Perkowska, M., Jesien-Lewandowicz, E., Jaskolski, D.J., Radomiak-Zaluska, A., Jesionek-Kupnicka, D., Sikorska, B., Papierz, W., Rieske, P., and Liberski, P.P.: ‘Screening for EGFR amplifications with a novel method and their significance for the outcome of glioblastoma patients’, PLoS One, 2013, 8, (6), pp. e65444

104 Dasari, V.R., Velpula, K.K., Alapati, K., Gujrati, M., and Tsung, A.J.: ‘Cord blood stem cells inhibit epidermal growth factor receptor translocation to mitochondria in glioblastoma’, PLoS One, 2012, 7, (2), pp. e31884

105 Nitta, M., Kozono, D., Kennedy, R., Stommel, J., Ng, K., Zinn, P.O., Kushwaha, D., Kesari, S., Inda, M.M., Wykosky, J., Furnari, F., Hoadley, K.A., Chin, L., DePinho, R.A., Cavenee, W.K., D'Andrea, A., and Chen, C.C.: ‘Targeting EGFR induced oxidative stress by PARP1 inhibition in glioblastoma therapy’, PLoS One, 2010, 5, (5), pp. e10767

106 Rao, S.A., Arimappamagan, A., Pandey, P., Santosh, V., Hegde, A.S., Chandramouli, B.A., and Somasundaram, K.: ‘miR-219-5p inhibits receptor tyrosine kinase pathway by targeting EGFR in glioblastoma’, PLoS One, 2013, 8, (5), pp. e63164

107 Nagane, M., Levitzki, A., Gazit, A., Cavenee, W.K., and Huang, H.J.: 'Drug resistance of human glioblastoma cells conferred by a tumor-specific mutant epidermal growth factor receptor through modulation of Bcl-XL and caspase-3-like proteases', Proceedings of the National Academy of Sciences of the United States of America, 1998, 95, (10), pp. 5724-5729

108 Szerlip, N.J., Pedraza, A., Chakravarty, D., Azim, M., McGuire, J., Fang, Y., Ozawa, T., Holland, E.C., Huse, J.T., Jhanwar, S., Leversha, M.A., Mikkelsen, T., and Brennan, C.W.: 'Intratumoral heterogeneity of receptor tyrosine kinases EGFR and PDGFRA amplification in glioblastoma defines subpopulations with distinct growth factor response', Proceedings of the National Academy of Sciences of the United States of America, 2012, 109, (8), pp. 3041-3046

109 Layfield, L.J., Willmore, C., Tripp, S., Jones, C., and Jensen, R.L.: 'Epidermal growth factor receptor gene amplification and protein expression in glioblastoma multiforme: prognostic significance and relationship to other prognostic factors', Applied immunohistochemistry & molecular morphology : AIMM / official publication of the Society for Applied Immunohistochemistry, 2006, 14, (1), pp. 91-96

110 Dorward, N.L., Hawkins, R.A., and Whittle, I.R.: 'Epidermal growth factor receptor activity and clinical outcome in glioblastoma and meningioma', British journal of neurosurgery, 1993, 7, (2), pp. 197-199

111 Kapoor, G.S., Christie, A., and O'Rourke, D.M.: 'EGFR inhibition in glioblastoma cells induces G2/M arrest and is independent of p53', Cancer biology & therapy, 2007, 6, (4), pp. 571-579

112 Zawrocki, A., and Biernat, W.: 'Epidermal growth factor receptor in glioblastoma', Folia Neuropathol, 2005, 43, (3), pp. 123-132

113 Palumbo, S., Tini, P., Toscano, M., Allavena, G., Angeletti, F., Manai, F., Miracco, C., Comincini, S., and Pirtoli, L.: 'Combined EGFR and Autophagy Modulation Impairs Cell Migration and Enhances Radiosensitivity in Human Glioblastoma Cells', Journal of cellular physiology, 2014

114 Howard, B.M., Gursel, D.B., Bleau, A.M., Beyene, R.T., Holland, E.C., and Boockvar, J.A.: 'EGFR signaling is differentially activated in patient-derived glioblastoma stem cells', Journal of experimental therapeutics & oncology, 2010, 8, (3), pp. 247-260

115 Brockmann, M.A., Ulbricht, U., Gruner, K., Fillbrandt, R., Westphal, M., and Lamszus, K.: 'Glioblastoma and cerebral microvascular endothelial cell migration in response to tumor-associated growth factors', Neurosurgery, 2003, 52, (6), pp. 1391-1399; discussion 1399

116 Kunkle, B.W., Yoo, C., and Roy, D.: 'Reverse engineering of modified genes by Bayesian network analysis defines molecular determinants critical to the development of glioblastoma', PLoS One, 2013, 8, (5), pp. e64140

117 Liang, Y., Diehn, M., Watson, N., Bollen, A.W., Aldape, K.D., Nicholas, M.K., Lamborn, K.R., Berger, M.S., Botstein, D., Brown, P.O., and Israel, M.A.: 'Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme', Proceedings of the National Academy of Sciences of the United States of America, 2005, 102, (16), pp. 5814-5819

118 De Rosa, A., Pellegatta, S., Rossi, M., Tunici, P., Magnoni, L., Speranza, M.C., Malusa, F., Miragliotta, V., Mori, E., Finocchiaro, G., and Bakker, A.: 'A radial glia gene marker, fatty acid binding protein 7 (FABP7), is involved in proliferation and invasion of glioblastoma cells', PLoS One, 2012, 7, (12), pp. e52113

119 Cataltepe, O., Arikan, M.C., Ghelfi, E., Karaaslan, C., Ozsurekci, Y., Dresser, K., Li, Y., Smith, T.W., and Cataltepe, S.: 'Fatty acid binding protein 4 is expressed in distinct endothelial and non-endothelial cell populations in glioblastoma', Neuropathology and applied neurobiology, 2012, 38, (5), pp. 400-410

120 Qiang, L., Wu, T., Zhang, H.W., Lu, N., Hu, R., Wang, Y.J., Zhao, L., Chen, F.H., Wang, X.T., You, Q.D., and Guo, Q.L.: 'HIF-1alpha is critical for hypoxia-mediated maintenance of glioblastoma stem cells by activating Notch signaling pathway', Cell death and differentiation, 2012, 19, (2), pp. 284-294

121 Barbus, S., Tews, B., Karra, D., Hahn, M., Radlwimmer, B., Delhomme, N., Hartmann, C., Felsberg, J., Krex, D., Schackert, G., Martinez, R., Reifenberger, G., and Lichter, P.: 'Differential retinoic acid signaling in tumors of long- and short-term glioblastoma survivors', Journal of the National Cancer Institute, 2011, 103, (7), pp. 598-606

122 Etcheverry, A., Aubry, M., de Tayrac, M., Vauleon, E., Boniface, R., Guenot, F., Saikali, S., Hamlat, A., Riffaud, L., Menei, P., Quillien, V., and Mosser, J.: 'DNA methylation in glioblastoma: impact on gene expression and clinical outcome', BMC genomics, 2010, 11, pp. 701

123 Lin, Y.C., Hung, C.M., Tsai, J.C., Lee, J.C., Chen, Y.L., Wei, C.W., Kao, J.Y., and Way, T.D.: 'Hispidulin potently inhibits human glioblastoma multiforme cells through activation of AMP-activated protein kinase (AMPK)', Journal of agricultural and food chemistry, 2010, 58, (17), pp. 9511-9517

124 Liang, Y., Bollen, A.W., Aldape, K.D., and Gupta, N.: 'Nuclear FABP7 immunoreactivity is preferentially expressed in infiltrative glioma and is associated with poor prognosis in EGFR-overexpressing glioblastoma', BMC cancer, 2006, 6, pp. 97

125 Lu, Z., Zhou, L., Killela, P., Rasheed, A.B., Di, C., Poe, W.E., McLendon, R.E., Bigner, D.D., Nicchitta, C., and Yan, H.: 'Glioblastoma proto-oncogene SEC61gamma is required for tumor cell survival and response to endoplasmic reticulum stress', Cancer Res, 2009, 69, (23), pp. 9105-9111

126 Poimenidi, E., Hatziapostolou, M., and Papadimitriou, E.: 'Serum stimulates Pleiotrophin gene expression in an AP-1-dependent manner in human endothelial and glioblastoma cells', Anticancer Res, 2009, 29, (1), pp. 349-354

127 Grzelinski, M., Bader, N., Czubayko, F., and Aigner, A.: 'Ribozyme-targeting reveals the rate-limiting role of pleiotrophin in glioblastoma', International journal of cancer. Journal international du cancer, 2005, 117, (6), pp. 942-951

128 Wellstein, A.: 'ALK receptor activation, ligands and therapeutic targeting in glioblastoma and in other cancers', Frontiers in oncology, 2012, 2, pp. 192

129 Dos Santos, C., Karaky, R., Renoir, D., Hamma-Kourbali, Y., Albanese, P., Gobbo, E., Griscelli, F., Opolon, P., Dalle, S., Perricaudet, M., Courty, J., and Delbe, J.: 'Antitumorigenic effects of a mutant of the heparin affin regulatory peptide on the U87 MG glioblastoma cell line', International journal of cancer. Journal international du cancer, 2010, 127, (5), pp. 1038-1051

130 Lu, K.V., Jong, K.A., Kim, G.Y., Singh, J., Dia, E.Q., Yoshimoto, K., Wang, M.Y., Cloughesy, T.F., Nelson, S.F., and Mischel, P.S.: 'Differential induction of glioblastoma migration and growth by two forms of pleiotrophin', J Biol Chem, 2005, 280, (29), pp. 26953-26964

131 Chang, Y., Berenson, J.R., Wang, Z., and Deuel, T.F.: 'Dominant negative pleiotrophin induces tetraploidy and aneuploidy in U87MG human glioblastoma cells', Biochemical and biophysical research communications, 2006, 351, (2), pp. 336-339

132 Muller, S., Lamszus, K., Nikolich, K., and Westphal, M.: 'Receptor protein tyrosine phosphatase zeta as a therapeutic target for glioblastoma therapy', Expert opinion on therapeutic targets, 2004, 8, (3), pp. 211-220

133 Ulbricht, U., Eckerich, C., Fillbrandt, R., Westphal, M., and Lamszus, K.: 'RNA interference targeting protein tyrosine phosphatase zeta/receptor-type protein tyrosine phosphatase beta suppresses glioblastoma growth in vitro and in vivo', J Neurochem, 2006, 98, (5), pp. 1497-1506

134 Grzelinski, M., Urban-Klein, B., Martens, T., Lamszus, K., Bakowsky, U., Hobel, S., Czubayko, F., and Aigner, A.: 'RNA interference-mediated gene silencing of pleiotrophin through polyethylenimine-complexed small interfering RNAs in vivo exerts antitumoral effects in glioblastoma xenografts', Human gene therapy, 2006, 17, (7), pp. 751-766

135 Li, X., Liu, Y., Granberg, K.J., Wang, Q., Moore, L.M., Ji, P., Gumin, J., Sulman, E.P., Calin, G.A., Haapasalo, H., Nykter, M., Shmulevich, I., Fuller, G.N., Lang, F.F., and Zhang, W.: 'Two mature products of MIR-491 coordinate to suppress key cancer hallmarks in glioblastoma', Oncogene, 2014

136 Arscott, W.T., Tandle, A.T., Zhao, S., Shabason, J.E., Gordon, I.K., Schlaff, C.D., Zhang, G., Tofilon, P.J., and Camphausen, K.A.: 'Ionizing radiation and glioblastoma exosomes: implications in tumor biology and cell migration', Translational oncology, 2013, 6, (6), pp. 638-648

137 Ahani, N., Karimi Arzenani, M., Shirkoohi, R., Rokouei, M., Alipour Eskandani, M., and Nikravesh, A.: 'Expression of insulin-like growth factor binding protein-2 (IGFBP-2) gene in negative and positive human cytomegalovirus glioblastoma multiforme tissues', Medical oncology, 2014, 31, (2), pp. 812

138 Li, Y., Hu, J., Guan, F., Song, L., Fan, R., Zhu, H., Hu, X., Shen, E., and Yang, B.: 'Copper induces cellular senescence in human glioblastoma multiforme cells through downregulation of Bmi-1', Oncology reports, 2013, 29, (5), pp. 1805-1810

139 Zhao, Z., Liu, Y., He, H., Chen, X., Chen, J., and Lu, Y.C.: 'Candidate genes influencing sensitivity and resistance of human glioblastoma to Semustine', Brain research bulletin, 2011, 86, (3-4), pp. 189-194

140 Elstner, A., Stockhammer, F., Nguyen-Dobinsky, T.N., Nguyen, Q.L., Pilgermann, I., Gill, A., Guhr, A., Zhang, T., von Eckardstein, K., Picht, T., Veelken, J., Martuza, R.L., von Deimling, A., and Kurtz, A.: 'Identification of diagnostic serum protein profiles of glioblastoma patients', J Neurooncol, 2011, 102, (1), pp. 71-80

141 Santosh, V., Arivazhagan, A., Sreekanthreddy, P., Srinivasan, H., Thota, B., Srividya, M.R., Vrinda, M., Sridevi, S., Shailaja, B.C., Samuel, C., Prasanna, K.V., Thennarasu, K., Balasubramaniam, A., Chandramouli, B.A., Hegde, A.S., Somasundaram, K., Kondaiah, P., and Rao, M.R.: 'Grade-specific expression of insulin-like growth factor-binding proteins-2, -3, and -5 in astrocytomas: IGFBP-3 emerges as a strong predictor of survival in patients with newly diagnosed glioblastoma', Cancer Epidemiol Biomarkers Prev, 2010, 19, (6), pp. 1399-1408

142 Mehrian-Shai, R., Chen, C.D., Shi, T., Horvath, S., Nelson, S.F., Reichardt, J.K., and Sawyers, C.L.: 'Insulin growth factor-binding protein 2 is a candidate biomarker for PTEN status and PI3K/Akt pathway activation in glioblastoma and prostate cancer', Proceedings of the National Academy of Sciences of the United States of America, 2007, 104, (13), pp. 5563-5568

143 Wang, H., Wang, H., Shen, W., Huang, H., Hu, L., Ramdas, L., Zhou, Y.H., Liao, W.S., Fuller, G.N., and Zhang, W.: 'Insulin-like growth factor binding protein 2 enhances glioblastoma invasion by activating invasion-enhancing genes', Cancer Res, 2003, 63, (15), pp. 4315-4321

144 Fuller, G.N., Rhee, C.H., Hess, K.R., Caskey, L.S., Wang, R., Bruner, J.M., Yung, W.K., and Zhang, W.: 'Reactivation of insulin-like growth factor binding protein 2 expression in glioblastoma multiforme: a revelation by parallel gene expression profiling', Cancer Res, 1999, 59, (17), pp. 4228-4232

145 Shen, L., Dean, N.M., and Glazer, R.I.: 'Induction of p53-dependent, insulin-like growth factor-binding protein-3-mediated apoptosis in glioblastoma multiforme cells by a protein kinase Calpha antisense oligonucleotide', Molecular pharmacology, 1999, 55, (2), pp. 396-402

146 Shen, L., and Glazer, R.I.: 'Induction of apoptosis in glioblastoma cells by inhibition of protein kinase C and its association with the rapid accumulation of p53 and induction of the insulin-like growth factor-1-binding protein-3', Biochemical pharmacology, 1998, 55, (10), pp. 1711-1719

147 Luo, Y., Kong, F., Wang, Z., Chen, D., Liu, Q., Wang, T., Xu, R., Wang, X., and Yang, J.Y.: 'Loss of ASAP3 destabilizes cytoskeletal protein ACTG1 to suppress cancer cell migration', Molecular medicine reports, 2014, 9, (2), pp. 387-394

148 Mustafa, D.A., Dekker, L.J., Stingl, C., Kremer, A., Stoop, M., Sillevis Smitt, P.A., Kros, J.M., and Luider, T.M.: 'A proteome comparison between physiological angiogenesis and angiogenesis in glioblastoma', Molecular & cellular proteomics : MCP, 2012, 11, (6), pp. M111 008466

149 Zinn, P.O., Mahajan, B., Sathyan, P., Singh, S.K., Majumder, S., Jolesz, F.A., and Colen, R.R.: 'Radiogenomic mapping of edema/cellular invasion MRI-phenotypes in glioblastoma multiforme', PLoS One, 2011, 6, (10), pp. e25451

150 Hunecke, D., Spanel, R., Langer, F., Nam, S.W., and Borlak, J.: 'MYC-regulated genes involved in liver cell dysplasia identified in a transgenic model of liver cancer', The Journal of pathology, 2012

151 Safaeian, M., Hildesheim, A., Gonzalez, P., Yu, K., Porras, C., Li, Q., Rodriguez, A.C., Sherman, M.E., Schiffman, M., Wacholder, S., Burk, R., Herrero, R., Burdette, L., Chanock, S.J., and Wang, S.S.: 'Single nucleotide polymorphisms in the PRDX3 and RPS19 and risk of HPV persistence and cervical precancer/cancer', PLoS One, 2012, 7, (4), pp. e33619

152 Kroh, H., and Cervos-Navarro, J.: 'Alpha-1-antichymotrypsin in human glioblastoma multiforme cells and its relation to GFAP immunostaining', Clin Neuropathol, 1991, 10, (4), pp. 181-186

153 Hoelscher, M., Richter, N., Melle, C., von Eggeling, F., Schaenzer, A., and Nestler, U.: 'SELDI-TOF analysis of glioblastoma cyst fluid is an approach for assessing cellular protein expression', Neurol Res, 2013, 35, (10), pp. 993-1001

154 Hoelscher, M., Richter, N., Melle, C., Eggeling, F.V., Schaenzer, A., and Nestler, U.: 'SELDI-TOF analysis of glioblastoma cyst fluid is an approach for assessing cellular protein expression', Neurol Res, 2013

155 Gautam, P., Nair, S.C., Gupta, M.K., Sharma, R., Polisetty, R.V., Uppin, M.S., Sundaram, C., Puligopu, A.K., Ankathi, P., Purohit, A.K., Chandak, G.R., Harsha, H.C., and Sirdeshmukh, R.: 'Proteins with altered levels in plasma from glioblastoma patients as revealed by iTRAQ-based quantitative proteomic analysis', PLoS One, 2012, 7, (9), pp. e46153

156 Alapati, K., Gopinath, S., Malla, R.R., Dasari, V.R., and Rao, J.S.: 'uPAR and cathepsin B knockdown inhibits radiation-induced PKC integrated integrin signaling to the cytoskeleton of glioma-initiating cells', Int J Oncol, 2012, 41, (2), pp. 599-610

157 Malla, R., Gopinath, S., Alapati, K., Gondi, C.S., Gujrati, M., Dinh, D.H., Mohanam, S., and Rao, J.S.: 'Downregulation of uPAR and cathepsin B induces apoptosis via regulation of Bcl-2 and Bax and inhibition of the PI3K/Akt pathway in gliomas', PLoS One, 2010, 5, (10), pp. e13731

158 Kast, R.E.: 'Glioblastoma invasion, cathepsin B, and the potential for both to be inhibited by auranofin, an old anti-rheumatoid arthritis drug', Cent Eur Neurosurg, 2010, 71, (3), pp. 139-142

159 Colin, C., Voutsinos-Porche, B., Nanni, I., Fina, F., Metellus, P., Intagliata, D., Baeza, N., Bouvier, C., Delfino, C., Loundou, A., Chinot, O., Lah, T., Kos, J., Martin, P.M., Ouafik, L., and Figarella-Branger, D.: 'High expression of cathepsin B and plasminogen activator inhibitor type-1 are strong predictors of survival in glioblastomas', Acta Neuropathol, 2009, 118, (6), pp. 745-754

160 Gole, B., Duran Alonso, M.B., Dolenc, V., and Lah, T.: ‘Post-translational regulation of cathepsin B, but not of other cysteine cathepsins, contributes to increased glioblastoma cell invasiveness in vitro’, Pathol Oncol Res, 2009, 15, (4), pp. 711-723

161 Wang, M., Tang, J., Liu, S., Yoshida, D., and Teramoto, A.: ‘Expression of cathepsin B and microvascular density increases with higher grade of astrocytomas’, J Neurooncol, 2005, 71, (1), pp. 3-7

162 Yan, S., and Sloane, B.F.: ‘Isolation of a novel USF2 isoform: repressor of cathepsin B expression’, Gene, 2004, 337, pp. 199-206

163 Gondi, C.S., Lakka, S.S., Yanamandra, N., Olivero, W.C., Dinh, D.H., Gujrati, M., Tung, C.H., Weissleder, R., and Rao, J.S.: ‘Adenovirus-mediated expression of antisense urokinase plasminogen activator receptor and antisense cathepsin B inhibits tumor growth, invasion, and angiogenesis in gliomas’, Cancer Res, 2004, 64, (12), pp. 4069-4077

164 Lakka, S.S., Gondi, C.S., Yanamandra, N., Olivero, W.C., Dinh, D.H., Gujrati, M., and Rao, J.S.: ‘Inhibition of cathepsin B and MMP-9 gene expression in glioblastoma cell line via RNA interference reduces tumor cell invasion, tumor growth and angiogenesis’, Oncogene, 2004, 23, (27), pp. 4681-4689

165 Yanamandra, N., Gumidyala, K.V., Waldron, K.G., Gujrati, M., Olivero, W.C., Dinh, D.H., Rao, J.S., and Mohanam, S.: ‘Blockade of cathepsin B expression in human glioblastoma cells is associated with suppression of angiogenesis’, Oncogene, 2004, 23, (12), pp. 2224-2230

166 Yan, S., Jane, D.T., Dufresne, M.J., and Sloane, B.F.: ‘Transcription of cathepsin B in glioma cells: regulation by an E-box adjacent to the transcription initiation site’, Biol Chem, 2003, 384, (10-11), pp. 1421-1427

167 Mai, J., Sameni, M., Mikkelsen, T., and Sloane, B.F.: ‘Degradation of extracellular matrix protein tenascin-C by cathepsin B: an interaction involved in the progression of gliomas’, Biol Chem, 2002, 383, (9), pp. 1407-1413

168 Konduri, S., Lakka, S.S., Tasiou, A., Yanamandra, N., Gondi, C.S., Dinh, D.H., Olivero, W.C., Gujrati, M., and Rao, J.S.: ‘Elevated levels of cathepsin B in human glioblastoma cell lines’, Int J Oncol, 2001, 19, (3), pp. 519-524

169 Mohanam, S., Jasti, S.L., Kondraganti, S.R., Chandrasekar, N., Lakka, S.S., Kin, Y., Fuller, G.N., Yung, A.W., Kyritsis, A.P., Dinh, D.H., Olivero, W.C., Gujrati, M., Ali-Osman, F., and Rao, J.S.: ‘Down-regulation of cathepsin B expression impairs the invasive and tumorigenic potential of human glioblastoma cells’, Oncogene, 2001, 20, (28), pp. 3665-3673

170 Osmak, M., Svetic, B., Gabrijelcic-Geiger, D., and Skrk, J.: ‘Drug-resistant human laryngeal carcinoma cells have increased levels of cathepsin B’, Anticancer Res, 2001, 21, (1A), pp. 481-483

171 Yan, S., Berquin, I.M., Troen, B.R., and Sloane, B.F.: ‘Transcription of human cathepsin B is mediated by Sp1 and Ets family factors in glioma’, DNA Cell Biol, 2000, 19, (2), pp. 79-91

172 Strojnik, T., Zajc, I., Bervar, A., Zidanik, B., Golouh, R., Kos, J., Dolenc, V., and Lah, T.: ‘Cathepsin B and its inhibitor stefin A in brain tumors’, Pflugers Arch, 2000, 439, (3 Suppl), pp. R122-123

173 Strojnik, T., Kos, J., Zidanik, B., Golouh, R., and Lah, T.: ‘Cathepsin B immunohistochemical staining in tumor and endothelial cells is a new prognostic factor for survival in patients with brain tumors’, Clin Cancer Res, 1999, 5, (3), pp. 559-567

174 Sivaparvathi, M., Sawaya, R., Wang, S.W., Rayford, A., Yamamoto, M., Liotta, L.A., Nicolson, G.L., and Rao, J.S.: ‘Overexpression and localization of cathepsin B during the progression of human gliomas’, Clinical & experimental metastasis, 1995, 13, (1), pp. 49-56

175 Rempel, S.A., Rosenblum, M.L., Mikkelsen, T., Yan, P.S., Ellis, K.D., Golembieski, W.A., Sameni, M., Rozhin, J., Ziegler, G., and Sloane, B.F.: ‘Cathepsin B expression and localization in glioma progression and invasion’, Cancer Res, 1994, 54, (23), pp. 6027-6031

176 Andreopoulos, B., and Anastassiou, D.: ‘Integrated Analysis Reveals hsa-miR-142 as a Representative of a Lymphocyte-Specific Gene Expression and Methylation Signature’, Cancer Inform, 2012, 11, pp. 61-75

177 Neder L, Marie SK, Carlotti CG, Jr., Gabbai AA, Rosemberg S, et al. (2004) Galectin-3 as an immunohistochemical tool to distinguish pilocytic astrocytomas from diffuse astrocytomas, and glioblastomas from anaplastic oligodendrogliomas. Brain pathology 14: 399-405.

178 Park SH, Min HS, Kim B, Myung J, Paek SH (2008) Galectin-3: a useful biomarker for differential diagnosis of brain tumors. Neuropathology : official journal of the Japanese Society of Neuropathology 28: 497-506.

179 Byeon SJ, Myung JK, Kim SH, Kim SK, Phi JH, et al. (2012) Distinct genetic alterations in pediatric glioblastomas. Child's nervous system : ChNS : official journal of the International Society for Pediatric Neurosurgery 28: 1025-1032.

180 Wei J, Barr J, Kong LY, Wang Y, Wu A, et al. (2010) Glioma-associated cancer-initiating cells induce immunosuppression. Clinical cancer research : an official journal of the American Association for Cancer Research 16: 461-473.

181 Le Mercier M, Fortin S, Mathieu V, Kiss R, Lefranc F (2010) Galectins and gliomas. Brain pathology 20: 17-27.

182 Vladimirova V, Waha A, Luckerath K, Pesheva P, Probstmeier R (2008) Runx2 is expressed in human glioma cells and mediates the expression of galectin-3. Journal of neuroscience research 86: 2450-2461.

183 Debray C, Vereecken P, Belot N, Teillard P, Brion JP, et al. (2004) Multifaceted role of galectin-3 on human glioblastoma cell motility. Biochemical and biophysical research communications 325: 1393-1398.

184 Deininger MH, Trautmann K, Meyermann R, Schluesener HJ (2002) Galectin-3 labeling correlates positively in tumor cells and negatively in endothelial cells with malignancy and poor prognosis in oligodendroglioma patients. Anticancer research 22: 1585-1592.

185 Camby I, Belot N, Rorive S, Lefranc F, Maurage CA, et al. (2001) Galectins are differentially expressed in supratentorial pilocytic astrocytomas, astrocytomas, anaplastic astrocytomas and glioblastomas, and significantly modulate tumor astrocyte migration. Brain pathology 11: 12-26.

186 Dumic J, Barisic K, Flogel M, Lauc G (2000) Galectin-3 decreases in mice exposed to immobilization stress. Stress 3: 241-246.

187 Dumic J, Lauc G, Flogel M (2000) Expression of galectin-3 in cells exposed to stress-roles of jun and NF-kappaB. Cellular physiology and biochemistry : international journal of experimental cellular physiology, biochemistry, and pharmacology 10: 149-158.

188 Tews DS, Nissen A (1998) Expression of adhesion factors and degrading proteins in primary and secondary glioblastomas and their precursor tumors. Invasion & metastasis 18: 271-284.

189 Bresalier RS, Yan PS, Byrd JC, Lotan R, Raz A (1997) Expression of the endogenous galactose-binding protein galectin-3 correlates with the malignant potential of tumors in the central nervous system. Cancer 80: 776-787.

190 Han S, Meng L, Wang Y, Wu A (2014) Plasma IGFBP-2 levels after postoperative combined radiotherapy and chemotherapy predict prognosis in elderly glioblastoma patients. PloS one 9: e93791.

191 Kulkarni A, Thota B, Srividya MR, Thennarasu K, Arivazhagan A, et al. (2012) Expression pattern and prognostic significance of IGFBP isoforms in anaplastic astrocytoma. Pathology oncology research : POR 18: 961-967.

192 Marucci G, Morandi L, Magrini E, Farnedi A, Franceschi E, et al. (2008) Gene expression profiling in glioblastoma and immunohistochemical evaluation of IGFBP-2 and CDC20. Virchows Archiv : an international journal of pathology 453: 599-609.

193 Schlenska-Lange A, Knupfer H, Lange TJ, Kiess W, Knupfer M (2008) Cell proliferation and migration in glioblastoma multiforme cell lines are influenced by insulin-like growth factor I in vitro. Anticancer research 28: 1055-1060.

194 Ahani N, Karimi Arzenani M, Shirkoohi R, Rokouei M, Alipour Eskandani M, et al. (2014) Expression of insulin-like growth factor binding protein-2 (IGFBP-2) gene in negative and positive human cytomegalovirus glioblastoma multiforme tissues. Medical oncology 31: 812.

195 Santosh V, Arivazhagan A, Sreekanthreddy P, Srinivasan H, Thota B, et al. (2010) Grade-specific expression of insulin-like growth factor-binding proteins-2, -3, and -5 in astrocytomas: IGFBP-3 emerges as a strong predictor of survival in patients with newly diagnosed glioblastoma. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 19: 1399-1408.

196 Lin Y, Jiang T, Zhou K, Xu L, Chen B, et al. (2009) Plasma IGFBP-2 levels predict clinical outcomes of patients with high-grade gliomas. Neuro-oncology 11: 468-476.

197 Fukushima T, Kataoka H (2007) Roles of insulin-like growth factor binding protein-2 (IGFBP-2) in glioblastoma. Anticancer research 27: 3685-3692.

198 Fukushima T, Tezuka T, Shimomura T, Nakano S, Kataoka H (2007) Silencing of insulin-like growth factor-binding protein-2 in human glioblastoma cells reduces both invasiveness and expression of progression-associated gene CD24. The Journal of biological chemistry 282: 18634-18644.

199 Mehrian-Shai R, Chen CD, Shi T, Horvath S, Nelson SF, et al. (2007) Insulin growth factor-binding protein 2 is a candidate biomarker for PTEN status and PI3K/Akt pathway activation in glioblastoma and prostate cancer. Proceedings of the National Academy of Sciences of the United States of America 104: 5563-5568.

201 Jiang R, Mircean C, Shmulevich I, Cogdell D, Jia Y, et al. (2006) Pathway alterations during glioma progression revealed by reverse phase protein lysate arrays. Proteomics 6: 2964-2971.

202 Song SW, Fuller GN, Khan A, Kong S, Shen W, et al. (2003) IIp45, an insulin-like growth factor binding protein 2 (IGFBP-2) binding protein, antagonizes IGFBP-2 stimulation of glioma cell invasion. Proceedings of the National Academy of Sciences of the United States of America 100: 13970-13975.

203 Elmlinger MW, Deininger MH, Schuett BS, Meyermann R, Duffner F, et al. (2001) In vivo expression of insulin-like growth factor-binding protein-2 in human gliomas increases with the tumor grade. Endocrinology 142: 1652-1658.