






Universitat Autònoma de Barcelona

**ADVERTIMENT.** L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons:  [http://cat.creativecommons.org/?page\\_id=184](http://cat.creativecommons.org/?page_id=184)

**ADVERTENCIA.** El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons:  <http://es.creativecommons.org/blog/licencias/>

**WARNING.** The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  <https://creativecommons.org/licenses/?lang=en>

# **SOLVING THE GLUCOCORTICOID PARADOX IN CANCER USING EXPRESSION DATA**

**MARIO HUERTA CASADO**

**2016**



# **SOLVING THE GLUCOCORTICOID PARADOX IN CANCER USING EXPRESSION DATA**

**Mario Huerta Casado**

**2016**

Submitted for the Degree of Doctor of Philosophy in Biotechnology by  
**Mario Huerta Casado**

The work was supervised by **Dr. Enric Querol Murillo** of the  
Institut de Biotecnologia i de Biomedicina- Universitat Autònoma de Barcelona  
and **Dr. Juan Cedano Rodriguez** of CENUR  
Litoral Norte-Salto, Universidad de la República, Uruguay

Mario Huerta Casado

Enric Querol Murillo

Juan Cedano Rodriguez

Universitat Autònoma de Barcelona, 2016



***Don't let the trees stop you from seeing the forest.***

Unknown author



# Abstract

Glucocorticoids are commonly used as an adjuvant treatment for side-effects and have anti-proliferative activity in several tumours. In spite of this fact, there has been reported a proliferative effect in several studies, that some of them involving the spread of cancer. These paradoxes are common in cancer research. Usually the same genes seem to be promoter and inhibitor of cancer. We shall attempt to reconcile these incongruities from the transcriptomic and tissue-physiology perspectives with our procedures and findings.

An accurate phenotype analysis of expression data can help to solve multiple paradoxes derived from tumour-progression models. We have developed several strategies, tools and databases to facilitate the study of interdependences among the phenotypes hidden inside expression data. These phenotypes can be identified by the sample clusters obtained by common clustering methods (HC, SOTA, SOM, PAM), but clustering methods do not provide information about their interdependence. We study these interdependences by the detection of non-linear expression-relationships where each fluctuation in the relationship implies a phenotype change and each relationship typology implies specific phenotype interdependence. In this way different strategies and tools have been developed to study phenotypic changes and the paradoxes subjected to them.

Using the tools developed we have studied the non-linear expression relationships related to glucocorticoid activity. The result is that we have found the possible reason for opposite effects of some stressor agents like dexamethasone on tumour progression as it has been confirmed by literature. This hidden reason has resulted in being linked with the type of tumour progression of the tissues. In the first type of tumour progression found, new cells can be stressed during proliferation and stressor agents increase tumour proliferation. In the second type, cell stress and tumour proliferation are antagonists so, therefore, stressor agents stop tumour proliferation in order to stress the cells.





# Resum

Els glucocorticoides s'utilitzen comunament com a tractament adjuvant per tractar efectes secundaris i tenen una activitat antiproliferativa en diversos tumors però, d'altra banda, el seu efecte proliferatiu s'ha documentat en diversos estudis i en alguns d'ells implica la propagació del càncer. Aquestes paradoxes són comuns en l'investigació del càncer. Molt sobint, els mateixos gens semblen ser a l'hora promotors i inhibidors del càncer. Amb el present treball pretenem conciliar aquestes incongruències des de la perspectiva de la genètica i de la fisiologia del teixit amb els nostres procediments i troballes.

Un anàlisi precís dels fenotips presents a les dades d'expressió gènica pot ajudar a resoldre múltiples paradoxes derivades dels models de progressió de tumors. En aquest sentit hem desenvolupat diverses estratègies, eines i bases de dades, per facilitar l'estudi de les interdependències entre els fenotips ocults dins de les dades d'expressió. Aquests fenotips poden ser identificats pels grups de mostres obtinguts mitjançant mètodes comuns d'agrupament (HC, SOTA, SOM, PAM), però els mètodes d'agrupament no proporcionen informació sobre la seva interdependència. Nosaltres estudiem aquestes interdependències per la detecció de relacions d'expressió no lineals, on cada fluctuació en la relació implica un canvi fenotípic i cada tipologia de la relació diferent implica una interdependència fenotípica específica. D'aquesta manera, s'han desenvolupat diferents estratègies i eines per estudiar els canvis fenotípics i les paradoxes sotmeses a ells.

Utilitzant les eines desenvolupades s'han estudiat les relacions d'expressió no lineals relacionades amb l'activitat dels glucocorticoides. Estudiant-les, hem trobat la possible raó dels efectes oposats d'alguns agents estressants com la dexametasona sobre la progressió tumoral, com ha estat confirmat per la literatura. Aquesta raó oculta ha resultat estar vinculada amb el tipus de progressió del tumor en cada teixit. En el primer tipus de progressió tumoral trobat, les noves cèl·lules poden ser estressades durant la proliferació i els agents estressants augmenten la proliferació tumoral. En el segon tipus, l'estrès cel·lular i la proliferació del tumor són antagonistes així, per tant, els agents estressants aturen la proliferació tumoral amb la finalitat d'estressar les cèl·lules.



# Resumen

Los glucocorticoides se utilizan comúnmente como tratamiento adyuvante para tratar efectos secundarios y tienen una actividad antiproliferativa en varios tumores, pero, por otra parte, su efecto proliferativo se ha documentado en varios estudios que en algunos de ellos implica la propagación del cáncer. Estas paradojas son comunes en la investigación del cáncer. A menudo los mismos genes parecen ser promotores e inhibidores del cáncer. En el presente trabajo vamos a tratar de conciliar estas incongruencias desde la perspectiva genética y de la fisiológica del tejido con nuestros procedimientos y hallazgos.

Un análisis preciso de los fenotipos presentes en los datos de expresión génica puede ayudar a resolver múltiples paradojas derivadas de los modelos de progresión de tumores. Con ese fin hemos desarrollado diversas estrategias, herramientas y bases de datos para facilitar el estudio de las interdependencias entre los fenotipos ocultos dentro de los datos de expresión. Estos fenotipos pueden ser identificados por los grupos de muestras obtenidos mediante métodos comunes de agrupación (HC, SO, SOM, PAM), pero los métodos de agrupación no proporcionan información sobre su interdependencia. Estudiamos estas interdependencias mediante la detección de relaciones de expresión no lineales, donde cada fluctuación en la relación implica un cambio fenotípico y cada tipología de la relación implica una interdependencia fenotípica específica. De este modo, se han desarrollado diferentes estrategias y herramientas para estudiar los cambios fenotípicos y las paradojas sometidas a ellos.

Utilizando las herramientas desarrolladas se han estudiado las relaciones de expresión no lineales relacionadas con la actividad de los glucocorticoides. Estudiándolas, hemos encontrado la posible razón de los efectos opuestos de algunos agentes estresantes como la dexametasona sobre la progresión tumoral, como ha sido confirmado por la literatura. Esta razón oculta resultó estar vinculada con el tipo de progresión del tumor de los diferentes tejidos. En el primer tipo de progresión tumoral encontrado, las nuevas células pueden ser estresadas durante la proliferación y los agentes estresantes aumentan la proliferación tumoral. En el segundo tipo, el estrés celular y la proliferación del tumor son antagonistas así, por tanto, los agentes estresantes detienen la proliferación tumoral con el fin de estresar las células.



# Table of Contents

<b>1. Background</b>	<b>21</b>
1.1. The glucocorticoid paradox	24
1.2. What is the hidden reason for this dual behaviour of glucocorticoids?	24
1.3. To search for the hidden reason for glucocorticoids' dual behaviour we need tools to study dual behaviours in cancer findings.	25
1.4. Studying phenotypes from expression data using sample clustering	26
1.5. Studying phenotypic changes from expression data.	29
1.6. Using non-linear expression relationships to study phenotypes and phenotypic changes	30
<b>2. Objectives</b>	<b>35</b>
<b>3. The Methods and tools developed and the data analysed</b>	<b>39</b>
3.1. THE EXPRESSION DATA ANALYSED	41
3.2. NON-LINEAL EXPRESSION RELATIONSHIPS VS LINEAL EXPRESSION RELATIONSHIPS	45
3.2.1. The steps followed by our procedure	47
3.2.2. The Principal Curves of Oriented-Points (PCOP) calculation	47
3.2.3. The minimum-spanning tree	50
3.2.4. The gene-selection algorithm	52
3.2.4.1. Hierarchical clustering	52
3.2.4.2. Gene selection following the minimum-spanning tree	53
3.2.5. Obtaining the intra-set behaviour pattern of the generated gene subset	55
3.3. CLUSTERING THE SAMPLES ALONG THE PCOP	61
3.3.1. Defining sample classes	61
3.3.2. Clustering the sample from a non-linear expression relationship	61
3.3.3. Colouring the sample clusters on a gene-expression relationship	62
3.3.4. Gene search based on the arrangement of the sample classes along the gene-expression ranges	64
3.3.5. Basic analysis procedure	65
3.3.6. Contextualisation consulting GEO database	65
3.3.7. Sample-classes definition use cases	65
3.3.7.1. Case 1: Defining the sample classes from a non-linear expression relationship	66
3.3.7.2. Case 2: Defining the sample classes by means of selecting gene-expression ranges	67
3.3.7.3. Case 3: Defining the sample classes by means of classifying the experiments using previous knowledge	68
3.3.7.4. Marker-gene search based on the arrangement of the sample classes along their expression range	69
3.4. PCOPGENE-NET : THE INTERACTIVE GENE NETWORK	71
3.4.1. The detailed view of the gene-expression relationship	72
3.4.2. The gene network, gene clusters, minimum-spanning tree, and graph layout	73
3.4.3. Tool in use	74
3.4.4. Biomedical-database remote access	75
3.4.5. Assigning attributes to the sample classes	75
3.4.6. Obtaining marker genes from continuous analysis	76
3.4.7. Obtaining marker genes from non-continuous analysis	76
3.4.8. Defining sample subclasses and obtaining more marker genes	76
3.4.9. Obtaining attributes by accessing remote databases	76

<b>3.5. THE BIOLOGICAL SIGNIFICANCE OF THE DIFFERENT EXPRESSION-RELATIONSHIP CURVE TYPES .....</b>	<b>79</b>
<b>3.6. NON-LINEAR EXPRESSION RELATIONSHIPS BETWEEN SETS OF COEXPRESSED GENES.....</b>	<b>87</b>
3.6.1. Genes coexpressed with a pair of genes with a non linear expression relationship between them .....	89
3.6.2. Cliques of non-linear expression relationships between genes .....	89
3.6.3. Pairs of isomorphic and linear Cliques of non linear expression relationships.....	90
3.6.4. Cliques of isomorphic and linear cliques of non-linear expression relationships between genes .....	90
3.6.5. Tool's output interfaces .....	92
3.6.5.1. Studying the complex expression relationships between a target gene and sets of coexpressed genes .....	92
3.6.5.2. The curve type indicates the type of activation and deactivation relationship between sets of coexpressed genes .....	93
3.6.5.3. Studying the complex expression relationships between sets of coexpressed genes .....	96
<b>3.7. DECONSTRUCTING SAMPLE CLUSTERS IN MULTIPLE CONCURRENT PHENOTYPIC CHANGES.....</b>	<b>99</b>
3.7.1. Automatic detection of sudden changes in expression relationships .....	103
3.7.2. Automatic detection of phenotypic changes .....	104
3.7.3. The showing of the intersections between transverse phenotypes .....	105
3.7.4. Crossing the sample clusters obtained by common clustering methods with the transverse-phenotypes intersection.....	106
3.7.5. Tool's output interfaces .....	106
3.7.5.1. The phenotypic-changes hierarchy view .....	106
3.7.5.2. The gene-relationships list view.....	108
<b>3.8. CROSSING CLUSTERS: STUDYING THE RELATIONSHIPS AMONG THE SAMPLE CLUSTERS OBTAINED BY CLUSTERING METHODS FROM EXPRESSION DATA .....</b>	<b>121</b>
3.8.1. Detection of curvature points and non-linear expression relationships .....	122
3.8.2. Arrangement of sample clusters along the PCOP .....	122
3.8.3. Classification of the non-linear expression relationships that cross each sample-cluster arrangement.....	122
3.8.4. Calculation of the sample-cluster arrangement by different clustering methods .....	122
3.8.5. Tool's output interfaces .....	123
3.8.5.1. List of sample-cluster arrangements view .....	123
3.8.5.2. List of non-linear expression relationships crossing the sample clusters view.....	124
<b>3.9. MGDB: COMPARING DIFFERENTIALLY EXPRESSED GENES FROM DIFFERENT MICROARRAYS TO COMPARE PHENOTYPES AND SAMPLE CLUSTERS FROM DIFFERENT MICROARRAYS .....</b>	<b>127</b>
3.9.1. Database of Microarray marker genes .....	128
3.9.2. Marker-gene search in the user's expression matrix .....	128
3.9.3. Crossing the user's microarray marker genes with the database of microarray marker genes .....	129
3.9.4. Tool's output interfaces .....	129
3.9.4.1. List-of-microarrays view .....	129
3.9.4.2. List-of-marker-genes view .....	129
3.9.5. Marker-genes database use cases.....	131
3.9.5.1. To corroborate or to question the hypothesis pointed out by user's microarray experiments .....	131
3.9.5.2. To study the role of marker genes in the user's expression data in other microarrays .....	138
3.9.5.3. To assign biological meaning to the sample clusters obtained by statistical methods from the user's expression data .....	140
3.9.5.4. To identify or define the phenotype of different subsets of individuals under investigation .....	142
3.9.5.5. To compare the user's results with other assays in the same species .....	142
3.9.5.6. To compare the user's results with human assays .....	142
3.9.5.7. To broaden the analysis of the user's results with experiments in other species .....	142
3.9.5.8. To broaden the analysis of the user's experiments with other pathologies .....	145
3.9.5.9. To extrapolate prognosis information from other analogue studies .....	146
3.9.5.10. To compare the user's experiments with experiments in the same pathology, but in different tissues .....	146
3.9.5.11. To search for drugs whose effect cause the transition between the phenotypes studied in the user's expression data.....	148

3.9.5.12. To discover undesirable side-effects of a treatment studied in the user's microarray .....	149
3.9.6. Comparison of online tools to search microarrays of interest from marker genes .....	150
3.9.6.1. Interface analysis: operability, usability and functionality .....	150
3.9.6.2. Statistical analysis of the results provided by different tools .....	152
<b>3.10. STUDYING THE TRANSITION BETWEEN PHENOTYPES USING NON-LINEAR EXPRESSION RELATIONSHIPS TO SOLVE PARADOXES IN CANCER MARKER GENES .....</b>	<b>155</b>
3.10.1. Procedure that can be followed to solve different paradoxes .....	156
3.10.1.1. Classifying the expression-relationships by typology .....	156
3.10.1.2. Applying sample clusters to non-linear expression relationships .....	157
3.10.1.3. On-line tools .....	157
3.10.2. How our procedure to solve paradoxes works .....	158
3.10.3. Why our procedure to study phenotypic changes in cancer works well? .....	160
3.10.4. How do non-linear expression relationships help us to solve paradoxes or to reconcile previous studies with contradictory but correct results? .....	161
<b>4. Solving marker-gene paradoxes in cancer progression. The glucocorticoids case: Results .....</b>	<b>165</b>
4.1. Expression-relationship typology reveals the role of the genes in phenotypic changes .....	168
4.2. Revealing the phenotype of the sample clusters from the phenotypic changes described by expression relationships .....	171
4.3. The glucocorticoids' effect on different tumour tissues .....	173
<b>5. Solving marker-gene paradoxes in cancer progression. The glucocorticoids case: Discussion.....</b>	<b>177</b>
<b>6. Conclusions.....</b>	<b>199</b>
<b>7. Publications.....</b>	<b>203</b>
<b>8. Bibliography .....</b>	<b>205</b>





# Abbreviations Table

HC: Hierarchical Clustering  
SOTA: Self-organizing Tree Algorithm  
SOM: Self-Organizing Map  
PAM: Partitioning Around Medoids  
PC: Principal Components  
MDS: Multidimensional scaling  
PCOP: Principal Curve of Oriented Poings  
POP: Principal Oriented Point  
GTV: Global Total Variance  
RV: Residual Total Variance  
f: Uncorrelation factor provided by PCOP  
MST: Minimum Spanning Tree  
Clique: Complete Subgraph  
GEO: Gene Expression Omnibus  
GDS: Gene expression Data Set  
GO: Gene Ontology  
PubMed: Medical Publications  
OMIM: Online Mendelian Inheritance in Man  
InterPro: Interaction proteins  
KEGG: Kyoto Encyclopedia of Genes and Genomes  
NCBI: National Center for Biotechnology Information  
DB: DataBase  
NGS: Next Generation Sequencing  
Rseq: RNA Sequencing  
Cmap: Connectivity map  
A\_matrix: Survival matrix  
T\_matrix: Expression matrix  
AT\_matrix: Correlation-between-Suvival-and-Expression matrix  
2D: Two Dimensions Space  
ENaC: Epithelial Na<sup>+</sup> Channel  
CACNB1: Voltage-dependent calcium channel  
DCA: Dichloroacetate  
ROS: Reactive Oxygen Species  
SNC: Suprachiasmatic Nuclei  
SDPT: Spatial Dependent Proliferation Tissues  
TDPT: Temporal Dependent Proliferation Tissues



O	I
<b>INTRODUCTION</b>	
Bkg.	Background
<p>To solve the glucocorticoid paradox we use the holistic perspective provided by expression data, but we need the adequate tools to extract the proper information. Non-linear expression relationships and external databases help us to describe the interdependence between phenotypes.</p>	



The limited capacity of the human mind to consider, at a time, several interacting factors and the complexity of the living things, makes it difficult to understand the biological processes in all their depth. Today we are already well aware that understanding the function of a gene involves knowing the complex environment in which this gene is expressed. But, this great complexity of the biological systems, often impedes to answer to specific biological questions from experiments, because not having the appropriate tools, even if these answers are present in the analyzed data. Therefore, the first step is to have the right tools to analyse the data generated. Using current techniques for obtaining massive data on gene expression we can obtain an overall behaviour of the cell in our data. But we also need the right high-throughput tools that allow us to perform a holistic analysis of these data. These tools should allow us to obtain both a global and local view of the behaviours of the genes expressed in the cell by an automatic and unsupervised method to identify the patterns and relationships among genes. This powerful relational map among genes can guide the study of this complexity, while remaining independent of a researcher's cognitive bias. In such way, it is more difficult for the researcher to force a relationship that really does not exist in the data, or ignore a strong relationship that he/she would prefer not to be present. On the other hand, even a perfect clustering method is meaningless if it not allows us to infer biological significance to every cluster. Therefore it is necessary to have tools that allow us to cross information stored in different databases with relevant biological information to enrich our data and provide us biological sense. There is also the need to facilitate understanding of the information by the user displaying the information properly. The tools described in this manuscript have been designed to give answers to specific biological questions, and they are not sophisticated approaches designed merely using theoretical principles far from the needs of researchers. We have therefore chosen to present our approach in an illustrative form with a particular problem difficult to address, in order to emphasize not only the theoretical aspect of the work but also how to use these methodologies in a pragmatic way. The question addressed is the glucocorticoids administration in cancer patients and it is currently considered as paradoxical in the literature. The scientific community have not yet found the reason for the dual behaviour observed, but this is an ideal question to study with our tools, taking advantage from the holistic view that our approach provides. Sometimes paradoxes arise from failing to see the problem in its entirety. These tools have been developed with the ability to create a global representation of reality that is necessary to answer these type of elusive questions. Throughout the manuscript, we are going to show the methods we have developed to get the biological meaning from clouds of points in a multidimensional universe. We are going to show how we have analyzed the data to extract patterns and how these have been crossed with biological databases. We must also admit that even though we have been assisted by the tools developed, it has not been easy to find the answer to the question addressed, because we were studying the cell, and the hidden reason of this dual behaviour was not in this holon of study. The response has emerged in a higher order of complexity different from that initially studied; the tissue. Thus, to find answers, and test the correct operability of the tools, we also had to consult numerous scientific literature. Nevertheless, without the distilled information provided by our tools, it would have been seriously difficult, not to say impossible, in answering this question.

## 1.1. The glucocorticoid paradox

Paradoxes are common in cancer research. Often, the same gene is reported as a marker of tumour progression as well as a marker of tumour suppression. The research topic of the present work is the convenience of glucocorticoids administration in cancer therapy, and this topic remains as an open question precisely due to these paradoxes. Stressor agents like dexamethasone have shown positive and negative effects on tumour treatment. On the one hand, glucocorticoids are commonly used as adjuvant treatment for side-effects [Philips RS. et al 2010] and have anti-proliferative activity in several tumours [Philips RS. et al 2010; Lu YS. et al 2006; Meijer E. and Sonneveld P. 2009; Chen YX. et al 2006] while, on the other hand, the proliferative effect of synthetic steroids with predominantly glucocorticoid activity has been reported in several studies [Herr I. and Pfitzenmaier J. 2006], some of them involving the spread of cancer [Mattern J. et al 2007].

We have reconciled these apparent incongruities with our findings, facilitated by the development of specific bioinformatics tools. In our current work we present both the procedure and the practice trying to solve these paradoxes.

It is commonly accepted that the immunosuppressive activity of glucocorticoids and other stressors can increase cancer incidence [Reiche EM. et al 2005], but their causal relationship seems to be more complex than this. We have found that the pro- and anti-tumoural effects of stressor agents like glucocorticoids depend on the tumour-proliferation type of the tissue. In the first type of tumour proliferation we have found, the new cells can be immediately stressed after proliferation without stopping tumour proliferation. In these tissues, stressor agents would increase tumour proliferation progressively. In the second type of tumour proliferation we have found, cell stress (and function) is incompatible with tumour proliferation, then, during tumorigenesis the new cells remain unstressed and undifferentiated. In these tissues, stressor agents would stop tumour proliferation to then stress the cells.

As a key outcome of the present work, we will show how we deal with this dual behaviour of glucocorticoids, and how we try to solve this paradox.

## 1.2. What is the hidden reason for this dual behaviour of glucocorticoids?

As detailed further on, the underlying principle seems to be that tumours remain part of the properties of the healthy tissue. These conserved proprieties would be clearer in primary tumours and diluted in metastases. For example, as melanomas retain in many cases, the ability to produce melanin, or parathyroid tumours retain their ability to produce calcitonin, even if it is away from its original location in the neck. These conservative proprieties involve not only functional proprieties but also structural ones. Even in a much altered way, the crypts structure is conserved in colon cancers. It implies that the relationship between cell stress and cell proliferation is also conserved in these tumours.

Hence, the dual tumour proliferation seems to be closely related to the physiology of the original tissue. In the tissues where the proliferation and function are performed in different areas (like colon or thyroid), the new cells are immediately incorporated from the proliferation areas to the functional ones, where they can be stressed (SDPT). On these tissues the stressor agents increase tumour proliferation progressively. In the tissues that can't be functional during tissue remodelling (TDPT), the new cells remains non-functional until the remodelling of the tissue is finished, the whole area is checked, and can be stressed again (like glia or bone). On these last tissues, stressor agents stop the tumour proliferation to stress the cells.

Thus, the reason for the existence of two types of tumour proliferation seems to be that the type of proliferation pattern is an important property of the healthy tissues that tend to remain, at least in part, when these tissues become tumourous.

These findings have direct medical implications, because they entail that treatments can induce counterproductive effects if they are not applied properly. Concretely, stressors like dexamethasone should be applied only when tumours appear in the tissues that cannot be stressed during tumour proliferation. Furthermore, the possibility that secondary tumours with stressed proliferation can get worse should be considered, and the probability of it spreading to tissues of this type should be taken into account too.

### **1.3. To search for the hidden reason for glucocorticoids' dual behaviour we need tools to study dual behaviours in cancer findings**

The bioinformatics tools developed by us help to solve the cited paradoxes by means of the analysis of expression matrices with large sample series. These expression matrices contain the expression level of thousands of genes under hundreds of experimental conditions and can be obtained using Microarray technology or next-generation sequencing (NGS) among others [Barrett T. Et al 2013]. The developed tools are very versatile and lead to crossing the expression data with GO, PubMed, and other Biomedical DB. Using these tools, we have studied the multiple behaviours of some genes related to the glucocorticoid signal in tumours. And we have found the hidden reason for this glucocorticoids' opposite-effect on tumour progression. We clearly contrast our findings with the current bibliography.

This correct identification and characterisation of the mentioned two types of tumour proliferation can facilitate the adjusting of cancer treatments correctly to prevent future malignancy, recurrences and metastasis. Actually, agents like those cited are administered indiscriminately as adjuvant to treat the undesirable side-effects of chemotherapy, radiotherapy, surgery or cancer itself.

To achieve our objectives, we first need to identify the phenotypes hidden in the expression data, second, we need to identify the phenotypic changes, and third, to study the paradoxes linked to these phenotypic changes.



## 1.4. Studying phenotypes from expression data using sample clustering

One of the most habitual expression-data analyses is clustering, both gene clustering and sample clustering. A habitual use of gene clustering is the study of coexpressed genes that, in consequence, can take part in the same cellular process [ Eisen MB. et al 1998]. On the other hand, sample clustering is usually used in large-sample-series analysis to detect samples that could represent the same phenotype. [ Van Laere S. et al 2007; Bertucci F. et al 2005; Hongjuan Z. et al 2009; Liu Y. et al 2014; Lehmann BD. et al 2011; Bullinger L. et al 2004; Hwang T. et al 2012; Verhaak RG. et al 2010; Koestler DC. et al 2010; Braun R. et al 2011; Müller FJ. et al 2008; Diana L. et al 2012; Danza K. et al 2014; Tanic M. et al 2013; Miller DJ. et al 2008; Golub TR. et al 1999; Liang Y. et al 2006]

In the second case, the phenotypes represented by the different sample clusters affect the expression of the matrix genes differently enough to be detected by clustering methods. Thus, a large enough number of genes vary their expression between two sample clusters to a sufficient extent so that clustering methods detect these two sample clusters. For this reason, different sample clusters from the expression matrix separate phenotypes, determined by the expression level of the genes. Note however, that it is not easy to determine whether clusters are meaningful, reflecting true structure in the data or random aggregation [Valentini G. 2006].

Overall, we can consider that the samples of a sample cluster can represent the same phenotype in the analysed expression data because the genes are similarly expressed in all of the samples grouped in the cluster. As the genes are similarly expressed, the phenotype of the cluster's samples could be the same. It is true that this must always be regarded with caution, mainly for two reasons: (i) Not all of the genes are considered in the expression matrix. The more genes contained by the expression matrix, the more probability that the phenotype would be the same for the samples of a same sample cluster. (ii) Different experimental conditions can produce different *final phenotypes* although the genes would be equally expressed in these experiments. Despite these limitations, sample-cluster analysis is usually used for studying common genic responses to different sample conditions [Qiu Q. et al 2013]. For example, to test if two different experimental conditions give rise to the same phenotype [Golub TR. et al 1999] or to search for new applications of old treatments [Chen R. et al 2008] .

Another habitual analysis of the expression data is the search for genes differentially expressed among sample clusters (over-expressed in one sample cluster but under-expressed in other). Among other uses, the genes differentially expressed shared between different microarrays can be used to compare phenotypes and sample clusters between different microarrays [Qiu Q. et al 2013; Chen R. et al 2008; Yu Y. et al 2009; Vázquez M. et al 2010; Feng C. et al 2009; Huerta M. et al 2014; Lamb J. et al 2006; Lamb J. et al 2007]. Expression matrices like microarray data favour this kind of investigations since they allow for a holistic vision of the different phenotypes providing the gene expression of each one. In the research published in Science [Lamb J. et al 2006] and Nature [Lamb J. et al 2007], public microarray data from very diverse experiments are used by Cmap tool to search for possible therapies for new drugs. Cmap and others [Chen R. et al 2008; Huerta M. et al

2014] study the effect of drugs using genes differentially expressed in public microarrays to establish a correspondence between the effect of the drug and the microarray experiments. However, these experiments do not test the drug of study. The relation between the effect of the drug and the data from microarrays is established because the target genes of the drug are affected in the microarray experiments and if the affected genes are the same and in the same way, it is considered that we are dealing with the same phenotype.

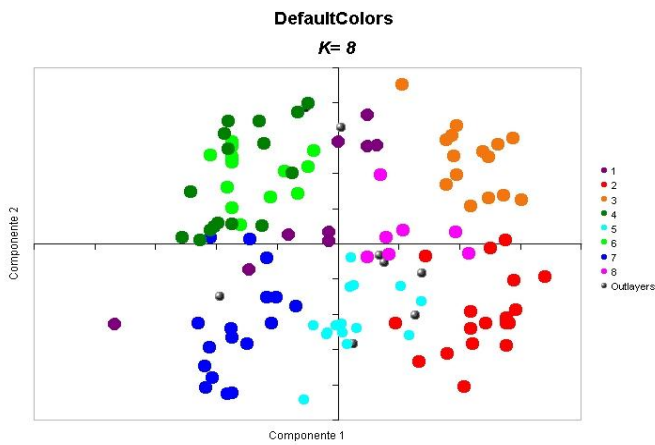
In the cases mentioned (Cmap and others), the genes differently expressed are compared because of the difficulty to analyse together the data coming from different microarrays using clustering methods (obtaining the sample clusters that would represent the same phenotype in the different microarrays)[Warnat P. et al 2005; Larsson et al 2006; Andreopolus et al 2009; Taminau J. et al 2014]. However, the purpose of this differently-expressed-genes comparison remains the same as sample clustering; to study if different experimental conditions give rise to the same phenotype.

The GEO gene-expression-matrices database allows one to study both sample clusters based on experimental criteria and sample clusters from statistical origin [Barrett T. et al 2013]. These sample clusters based on experimental criteria are guided by the interest of the researchers that have performed the experiment. Comparing both types of sample clusters (statistical and researchers' criteria) is also very useful for the researchers. It allows for checking if similar experimental conditions, which should trigger the same phenotypes in the cell, maintain a similarity at the expression level. Or it allows for checking the opposite, despite being sample conditions considered similar by the researchers, their gene expression is different therefore, the phenotypes would also be different [ Bullinger L. et al 2004; Diana L. et al 2012].

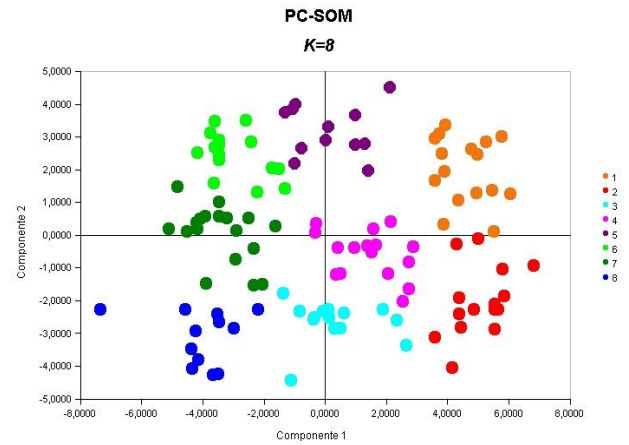
In the figures below (Figures 1.a, 1.b, 1.c, 1.d) the sample clusters resulting from the same expression data but after applying different clustering methods can be observed. The figures show how the samples in the border between neighbour clusters are located in one or another sample cluster by the different clustering methods (the samples of each cluster are coloured differently).

The figures show the samples of the expression data used along the next sections after applying Principal components to the data (the point cloud are the matrix samples, x and y axes are the two Principal components). In Figure 1.a, the sample clusters obtained by repeated bisection are shown with the sample clusters coloured as they will appear in the next sections. The green and blue sample clusters, the yellow and red sample clusters, and the pink sample cluster are shown in this figure. In Figure 1.b the sample clusters obtained after applying the SOM clustering method are coloured [Yin L. et al 2006]. In Figure 1.c the sample clusters obtained after applying the SOTA clustering method are coloured [Yin L. et al 2006]. In Figure 1.d the sample clusters obtained after applying the PAM clustering method are shown [Yin L. et al 2006]. As it can be observed, the different methods find similar sample clusters. In the figures, the samples on the border between sample clusters can be clearly seen, and how the different clustering methods place these samples in one cluster or another.

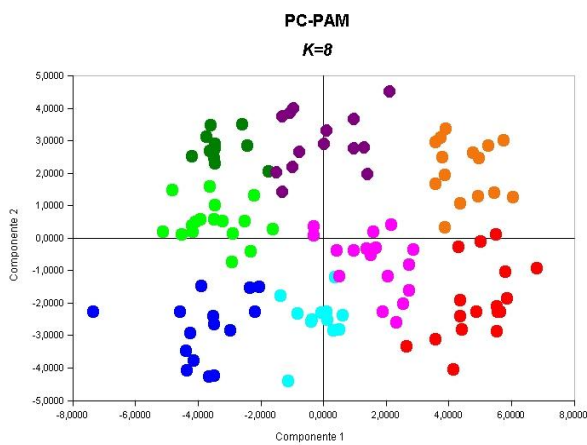
1.a)



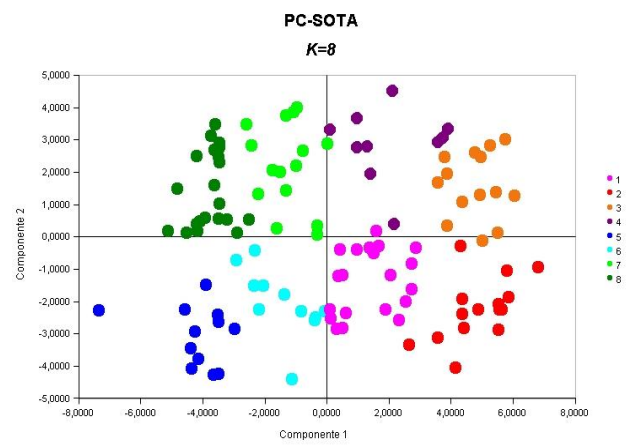
1.b)



1.c)



1.d)



**Figure 1. Sample clusters obtained by different methods comparison.** In the figures there can be seen how the samples in the frontier between sample clusters can determine the differences among the sample clusters obtained by different clustering methods from the same expression data. The samples are shown after applying Principal Components to the expression data (x and y axes are the two principal components and the data cloud the matrix samples).

If the phenotypes hidden in the expression data are clearly differentiated at expression level, the different clustering methods will obtain the same sample clusters. When the phenotypes are less detectable at the expression level, the disparity among the results provided by the different clustering methods, increases.

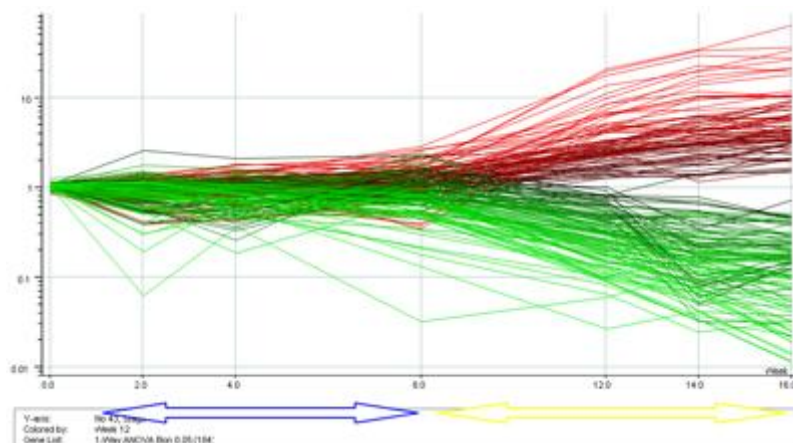
We largely use these sample clusters in our analysis, to try to identify the different phenotypes, the different phenotypic changes, and finally the role of glucocorticoids in the transition between phenotypes.

## 1.5. Studying phenotypic changes from expression data

It is not easy to define what a phenotype is, but it is even more difficult to define the processes that carry out one phenotype instead of another. Since our approach expects to be standard, one definition of phenotype as generalist as possible, could be: A phenotype is the result from the expression of an organism's genes as well as the influence of environmental factors and the interactions between both.

In our work, we refer to “phenotype” as the cell state resulting from the experiments performed. As we have previously commented, when clustering methods methods form the sample clustering, they detect the experimental conditions that have had a similar effect on gene expression, so these different experimental conditions could have the same cell state or phenotype as a result.

Let's now see a specific example about the phenotypic changes that can be detected by the tools developed by us. In the case of plants, such as blackberry, where phenotypic plasticity phenomena are most studied, they exhibit some homology with the kind of phenotypic changes that our tools are able to detect. Sometimes it is not easy to find a marker that allows us to establish the change in phenotype. For example, although it is easy to find flowering marker genes, it is not easy to separate two distinct phenotypes such as the winter phenotype, whose function is to resist the inclement weather, and the spring phenotype, in which the plant is more active and focuses on growing and storing energy. But even if we don't have clear markers between the winter and spring phenotypes, we can use microarray time-series data, where it becomes clear that the transit between winter and spring affects a large number of genes, whose expressions will increase or decrease compared with the period previous to the winter solstice [Mazzitelli L. et al 2007]. Thus, this phenotypic change is seen in **the change in slope**, which can be more or less sharp, in a large number of genes at the same time.



**Figure 2. Changes in gene expression around dormancy break.** Red = up-regulated, green = down-regulated. Blue arrow denotes winter period, yellow denotes spring. The phenotypic change is seen in the change in slope [Mazzitelli L. et al 2007].

In this example we observe how a phenotypic change occurs.

In addition:

1. The phenotypic change is not due to a trigger gene.
2. The phenotypic change is detected thanks to the use of time series.

With the methodology we have developed, we can detect phenotypic changes without using trigger genes, or time series. Note that most of the expression matrices don't use time series. Since we don't need time series, we can detect multiple phenotypic changes occurring in the expression data. But then, how do we detect the cited "change in the slope"?

The first point of the previous ones implies the need of tools that may consider phenotypic changes in subtle expression-level variations but with a huge number of genes involved. The second point implies that, whereas using time series the samples are sorted, making it easier to detect the moment when the phenotypic change occurs (Figure 2), by using non-time series we need to sort the samples by a hidden variable to detect the point when the phenotypic change occurs. We tackled both problems with the approach presented, sorting the samples by the expression dependences between gene pairs, and considering a multitude of gene pairs. The phenotype change is then found by the change in the slope of the expression dependencies.

As we mentioned, it is not easy to provide a precise definition of phenotype [Crusio WE. 2002], and even less so, a definition of the elements in which a phenotype can be decomposed, and it is not our intention with the present work to open a debate on this subject [Mahner M. et al 1997]. Furthermore, the expression matrices that can be analysed have an absolutely diverse origin. Depending on the expression data analysed, we could be talking about transversal and concurrent effects of different drugs, common or altered features between cancer phases, or the distinctive processes between the winter and summer-like phenotypes in plants, and so on. There are as many examples, as there are gene-expression matrices for the researchers to analyse. In conclusion, who decides if two sample clusters actually represent different phenotypes, is the researcher who analyses the gene-expression data. She/he must be the one who defines what he/she understands as different phenotypes and justifies why.

## 1.6. Using non-linear expression relationships to study phenotypes and phenotypic changes

Although the approach presented is valid to analyse time series, its true potential is analysing data that don't come from time series. As time series are ordered temporally, they already show us the point in time where the phenotypic change occurs. Thus, their study by means of non-linear expression relationships is not needed. As will be explained in the next sections, the tools presented have certain limitations with regard to the input data, but these limitations have nothing to do neither with the kind of experiments performed nor with the data origin.

We have published several works that highlight the relevance of the analysis of non-linear expression relationships from expression data [Cedano J. et al 2008; Cedano J. et al 2007; Huerta M. et al 2014; Huerta M. et al 2014; Huerta M. et al 2009; Huerta M. Et al 2008]. The different tools we use for analysing the expression data here are also explained in these papers. In the last cited work [Huerta M. et al 2008] the difference between considering non-linear expression relationships with respect to considering only the linear expression relationships can be seen very clearly.

As is shown in the mentioned paper [Huerta M. et al 2008], non-linear expression relationships allow for detecting new hubs in gene networks because they allow researchers to relate genes by complex expression relationships and to discover new relationships that otherwise would not be possible to detect. Other research teams have demonstrated the importance of non-linear expression relationships for the study of gene regulation [Guo X. et al 2014]. It is extremely important to detect non-linear dependences in gene regulatory networks inference because non-linear regulatory relationships are common in biology [Brunel H. et al 2010]. Furthermore, as we have shown in these works, non-linear expression relationships can be used as a starting-point to characterize complex regulations between genes in other analyses besides Gene Regulatory Networks [Huerta M. et al 2014; Huerta M. et al 2014]. Non-linear expression relationships can also be used for sample clustering. So far, this sample clustering has had two different approaches. A non-linear interpretation of dimensionality reduction [Cannistraci CV. et al 2010], and a clustering of the samples corresponding to each phenotype from the phenotypic changes described by non-linear expression relationships [Cedano J. et al 2008], as we stated in our work. The key point of that approach is the study of the phenotypes involved in the phenotypic change. In other words, the phenotype previous to phenotypic change and the following phenotype after the change. In the present work, the study of the transition between phenotypes is used to solve paradoxes such as the contradictory effect of glucocorticoids in tumour progression depending on the tissue affected by the tumour. This study has been also been published by us in another paper [Huerta M. et al 2014].

Our methodologies are based upon the following five properties of non-linear expression relationships: First, each fluctuation in a correlated expression relationship implies a phenotypic change. Second, multiple non-linear expression relationships are involved in the same phenotypic change. Third, if two genes maintain an expression relationship with a certain type of curve, the genes coexpressed with these two genes maintain expression relationships of the same type of curve between them. Fourth, if there exist an expression dependence between two sets of coexpressed genes, this expression dependence must be non-linear. Fifth, the curve type of the non-linear expression relationships will describe the role of the genes in the phenotypic change. The first point allows us to detect the phenotypic changes from the non-linear expression relationships that are highly correlated. The second point allows us to group the non-linearly highly-correlated expression relationships by the phenotypic change they describe. The third point allows us to expand the analysis from these non-linearly highly-correlated genes to the genes coexpressed with them. The fourth point allows us to study the activation/deactivation relationships among sets of coexpressed genes. The fifth point allows us to know the activation/deactivation relationship between the phenotypes involved in the phenotypic change based on the morphology of the curve.

The study of phenotypic changes is mainly possible thanks to the first property previously cited: “Each fluctuation in a correlated expression relationship implies a phenotype change”, so the samples at either side of the fluctuation point will represent different phenotypes. The fifth point, the analysis of the curve type, helps us to find dual behaviours and paradoxes in expression relationships.

It is important to note that the presented approach based on non-linear expression relationships does not open a new, totally unknown line of research. The procedure followed to study the double effect of glucocorticoids is not really so different from the methods usually used in expression-data analysis:

The methods we use for clustering the samples and detect the different tumour phenotypes are the clustering methods most commonly used in the calculation of sample clusters [Giancarlo R. and Utro F. 2011]. The same goes for the integrity methods to find which clustering method and k parameter provide the best sample clusters for each expression matrix [Giancarlo R. and Utro F. 2011].

The correlation degree between gene expressions is commonly used in the analysis of coexpressed genes to consider that this coexpression has a biological meaning [Eisen MB. et al 1998]. This biological significance increases when the number of samples of the expression matrix increases, and it increases even more if the samples cover a range of phenotypes as wide as possible. Also, it is of common use to look for those genes with a differentiated expression between the different sample clusters to study both phenotypes and phenotypic changes. Based on this differentiated expression, it is considered that these genes mark the presence of the phenotype corresponding to the sample cluster with their expression [Huerta M. et al 2014].

In our case we take it a step further, and we require not only a difference of expression between the phenotypes represented by each sample cluster, but that this expression variation implies a non-linear expression relationship with another gene. More concretely, a non-linear expression relationship resulting from the switch from one phenotype to another one (that is, the step from one sample cluster to another one). Besides this, we combine this requirement with the previous one of high correlation in order to consider the expression relationships as valid, that is, to consider it has a biological meaning and it is not a product of chance. The next step is to study whether these genes are related to glucocorticoids, to see whether this behaviour related to glucocorticoids takes us from one phenotype to another, and the implications that this fact could have. However the approach presented has several bottlenecks that must be solved.

Interoperability is one of the major problems in this kind of analyses that cross information from different sources and databases. As data is collected from heterogeneous resources in our analysis, the types of data would be different and it would be practically impossible to standardize them. Analyses need to be interoperable so that data could be analyzed and summarized correctly [Mohammadi A. et al 2011].

Another problem of this kind of analysis is that they need in order to be successful, a specialist who can make objective conclusions from the outputs that are created. In other words, our procedure is from being automatic.

Quality of data is the most important challenge faced in case of data mining. This is why the main handicap we have is the fact of using microarray data. When the analysis of gene expression is reduced to a general statistic, and the error in the measurement of gene expression can be balanced out (since we study only the general behaviour of genes), the fact of working with microarray data does not present a big problem. **The problem gets worsens when we wish to analyse the behaviour of concrete genes.** Although the filtration of genes and expression data can be statistically correct, as explained before, the fact of focusing on a few genes is still a problem if the expression levels are obtained using microarray technology or similar high-throughput procedures [Abdullah-Sayani A. et al 2006].

This is a problem shared by the multiple studies about gene expression that use microarrays or similar technologies to obtain the samples. Nevertheless, this problem is not exclusive of these technologies, and it also affects the credibility of works based on experiments that don't measure expression data [Swamidass SJ. 2011; Norton SM. et al 2001].

In the case of our study about the glucocorticoid paradox, however, the expression data used are especially robust. We use the expression data of Scherf U. et al. about tumour cells [Scherf U. et al 2000]. In the study of Scherf U., the data obtained by microarray technology are crossed with experimentally determined mortality indices. In this way, the expression level of a gene for a certain drug comes from the average of the samples of different tissues that have had the same mortality index under that drug. Then, genes that do not show a correlation between the mortality index and their expression level for each drug are discarded. This provides us with a dramatically larger reliability, plus a wider range of well characterised phenotypes (in which the different drugs produce a high mortality). This fact is truly appreciated, since these data about gene expression, so well elaborated, allow us to make analyses like the one described in the present work with reliability.

An especially remarkable point of the procedure followed, to obtain the gene-expression matrix, is that the resultant matrix does not only allow us to obtain the cellular behaviour, but also the tissue behaviour. In this way, we can use these data to study the effect of glucocorticoids on different tumour tissues.

It is true that a large part of the results is obtained thanks to the methodology applied to analyse the expression data. This methodology allows us to study the dual behaviours and phenotypic transitions that otherwise would be impossible to study, or at least much more difficult. But it is also true that the results obtained are largely due to using an appropriate expression array, which really differentiates the phenotypes, without largely repeated samples, and with extreme phenotypes and transition phenotypes between them. Without well characterised phenotypes, it is impossible to study the phenotypic changes and even less to extend the analysis using data mining and data crossing.



This is why high-quality public databases in the field of gene expression are so important. They are not only necessary advances in technology so biotechnologists can perform their experiments. Nor are they only necessary good, powerful tools for data analysis and data crossing like the ones presented in this work. We also need powerful public databases, with reliable data, that are useful not only to cross information, but also to analyse them in depth. In the case of gene expression this lack appears in all its splendour. The lack of reliable data is partly due to the nature of the technology to obtain gene expression in a massive way, but mainly because of the structure and definition of the gene-expression databases. Databases like NCBI's GEO content data from hundreds of thousands of microarrays, but the problem for the database, and its users, is that these data were developed to satisfy particular needs of the studies that certain researchers wished to publish. Not to satisfy the posterior use of that information in the database. For this reason, there are very rare and specific data sets in NCBI's GEO, rarely generalist and even more rarely, redundant, since no one will create a microarray about a previously published experiment. This redundancy is necessary to get reliable data, as well as the different phenotypes would be described with the only objective being that they appear well described, and not because they are linked to the publication of a concrete and specific experiment. As with the human genome and the genomes of different organisms, molecular-biology researchers also need standardised available information with respect to gene expression.

To some extent, the data generated by the work of Scherf et al. do accomplish these requirements and then allow the researchers to use them in order to study the expression of genes in different phenotypes of tumour development. There is a need to make a bigger scientific effort in this direction. But while waiting for the international community to come to an agreement, we must develop strategies and tools that lead us to extract as much useful information as possible from the large amounts of data generated.

O	2
<b>OBJECTIVES</b>	
Obj.	
What we aim for.	



## General Objectives

Gaining insight into the global behaviour of the eukaryote cell taking advantage of high-throughput technologies, concretely transcriptomics. Use this holistic perspective of the cell behaviour to solve different paradoxes in cancer findings and other pathologies.

## Specific Objectives

1. Develop methodologies necessary to achieve the objectives of this work.
2. Analyse the complex expression dependencies between genes, sets of coexpressed genes, and networks of them in relation to cell processes.
3. Decomposition of sample clusters in the combination of multiple cell processes that leads to each phenotype. Compare phenotypes by the cell processes that they have in common.
4. Compare the phenotypes and phenotypic changes described by expression datasets available in biomedical databases. Search for phenotypes and phenotypic changes similar to the ones studied in our expression data.
5. Analyse the transitions between the phenotypes. Detect extreme and transition phenotypes and classify the complex transitions between them by their biological meaning.
6. Gain insight into the glucocorticoid paradox in cancer treatment using the above analysis of expression data.



O	3
<b>METHODS</b>	
Mth.	The Methods and tools developed and the data analysed
<p>The public data reused to solve the glucocorticoid paradox. The methodologies and tools developed by us to study phenotypes and phenotype changes hidden in expression data.</p>	



### 3.1. The expression data analysed

The appropriate data to be analysed by our methods and tools must come from large sample series (expression matrices with a high number of sample conditions). This large sample series will not consist of repetitions of the same sample condition. On the contrary, it must include the highest number of different sample conditions. A sample series with few experiments or with repetitions of the same experiment, will not allow the detection of coexpressed genes, and even less the detection of complex expression relationships. Finally, if the phenotypes are not well described by the experiments, it will be extremely difficult to study the phenotypic changes, but it will be even more difficult to solve the paradoxes with respect to these phenotypic changes. It should be noted that de-noise, normalization, and similar procedures should be considered before using our tools.

The expression data analysed in the next sections correspond to the AT\_matrix. This expression matrix is the correlation between survival (A\_matrix) and expression (T\_matrix) data. The A\_matrix contains the growth inhibitory activities of 118 compounds tested on 60 tumour cell lines. This compound set includes most of the drugs that are currently in clinical use for cancer treatment. The microarray data (T\_matrix) reflect the level of expression of 1376 genes, plus 40 individually assessed targets (proteins), and 40 other targets in the previous 60 tumour cell lines. The AT\_matrix links both matrices using the 60 tumour cell lines as a sample space to generate a correlation matrix of 1416 rows (genes and targets) by 118 columns (substances) [Scherf U. et al 2000].

AT\_matrix, unlike most of the expression matrices, shows the expression levels necessary for drug action, not the resulting levels of drug action. These expression data were chosen because they cover the phenotypic changes shared by many different tumoural human tissues. The drugs used have a range of action wide enough to expose the main pathways involved in the maintenance of cellular homeostasis, cell-division regulation, tissue-remodelling type or any other important cellular feature. This makes it possible to reveal the different phenotypes that we could find in tumour cells of different tissues. These 118 drugs are a subset of drugs coming from a previous, more extensive study that included 1400 different substances. This subset of drugs has been chosen because they have a strong effect on some phenotypes but have only a slight effect on others [Scherf U. et al 2000]. Each drug selected has the desired effect on tumour cell-lines with certain levels of expression for a certain gene set, while for those tumour cell-lines in which the level of expression of this gene set is the opposite, the drug loses its effect. Thus, the data give us the opportunity to link the phenotypes where the drug acts with gene expression levels. Genes are filtered using the same criteria: only those genes showing expression variations between the phenotypes where a drug set acts, with respect to the phenotypes where the drug set does not act, are considered. The initial 9703 gene spots are reduced to a subset of 1376 genes in the final matrix.

These analysed data (AT\_matrix) are highly pre-processed. For details about the source of the data, normalization procedures, housekeeping-genes and other non-significant-genes filtration see [Scherf U. et al 2000]. The final data (AT matrix) ready to be analysed by our system has been added as supplementary material in the web, including the accession number of the microarray genes. The accession numbers of all the genes of the raw data (T\_Matrix) are also accessible at the



Gene Expression Omnibus (GEO) repository. These original data can be displayed in the GEO DataSet browser using the GDS ID of the microarray (GDS1761):

<http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS1761>

Note that dexamethasone or any other glucocorticoid drug has not been used in the microarray experiments. The glucocorticoid activity is obtained by analysing the role of the genes linked to the glucocorticoid activity in the phenotypic changes described by the expression data.

In the Figure 3 the correlation can be seen between the gene expression level (y axis) and the cell death after applying the drug (x axis). The sample space is all of the 60 tumour cell lines. Each tumour sample has got a gene level of expression before applying the drug and a cell death index after applying the drug. AT\_matrix reflects for each drug (column) and each gene (row), the level of expression of the gene in which the drug acts.

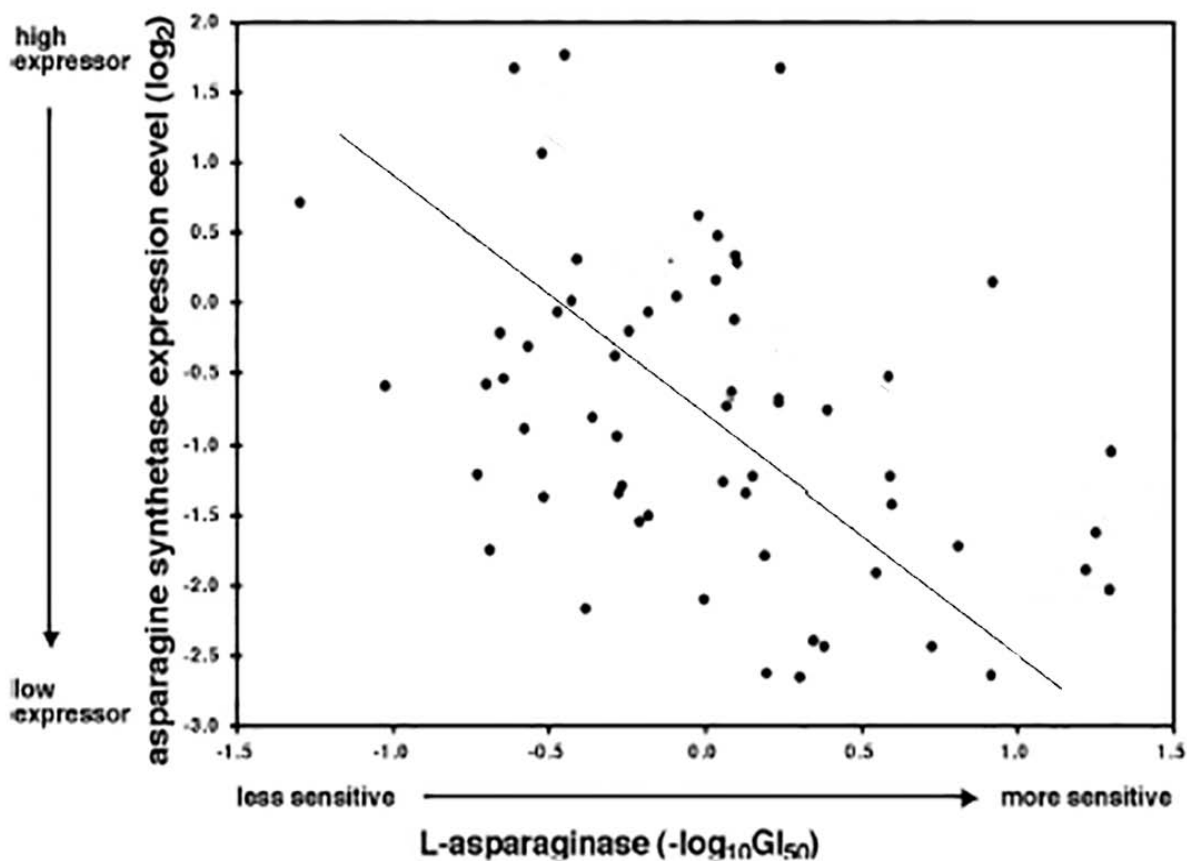


Figure 3. Relationship between ASNS gene expression and the action of L-asparaginase drug on the different tumour cell lines (the sample space) [Scherf U. et al 2000]. This relationship shows an inverse dependence between ASNS gene expression and the phenotypes chemosensitives to the drug.

As can be observed, the data analysed are not like usual microarray data. Our objective in studying these unusual data is to study the phenotypes in common among different types of tissue. The final data (AT\_matrix) show the effect of drugs on phenotypes, ignoring whether phenotypes belong to one or another tumour tissue (in T\_matrix, two samples of the same tissue can represent two different phenotypes, and two samples of different tissues can represent the same phenotype). This implies looking for the correlation between gene-expression levels and the effect of the drug, where the samples of the different tissues become the data cloud with its corresponding level of expression and drug effectiveness. Thus, in the final matrix (AT\_matrix) we have the phenotypes in which the different drugs act. These phenotypes where the drug acts have occurred only in certain tumour tissues.

The main attribute of these data is precisely that they give us the opportunity to compare the phenotypes of different tumour tissues. The value of AT\_matrix samples shows that for a certain level of expression of the genes in certain tumour tissues, the drug is having the desired effect, while for tissues in which the level of expression of the genes is different, the drug loses its effect, showing a linear correlation between level of expression and drug effect. For this reason, different sample clusters from AT\_matrix separate not only phenotypes, determined by the gene-expression levels, but also tissues in which these phenotypes appear.

Time is not a crucial factor in order to obtain our expression data because our microarray does not study the phenotype after the administration of drugs. The drugs are used in our expression experiments to obtain the phenotype where the drug acts (in contrast to those in which the drug does not act).

Of course, it is important to study the effect of the glucocorticoids doses as well as the effect of time, but those would be different studies from the current one. However, these new studies should consider the two types of tumour proliferation detected in our study, since they would affect their results. One of the most interesting aspects of bioinformatics studies like the present one, is that they allow the researcher to obtain models from re-used data thus, minimizing the number of experiments to be performed.



## 3.2. Non-linear expression relationships vs lineal expression relationships

Current biological research, specially transcriptomics, generates large amounts of data, which researchers need to analyse in order to understand the cell as a whole. The power of DNA microarray technology, among others, lies in its ability to provide data of a high number of gene expressions throughout different sample conditions. The objective of the present method is to analyse, relate, and show these expression data in a way that will permit researchers to reach a holistic view of the cell. Expression-matrices analysis challenges the traditional hypothesis-driven method of investigation and shifts the emphasis towards a hypothesis-generation paradigm. For achieving this new paradigm, the analysis procedure must be able to allow the researcher to have both a global vision of cell genetic networks and to focus on more specific aspects of subsets of genes. A global scenery alone does not permit the analysis of the expression connections among  $n$  non-directly correlated genes, nor the genes which relate them. To facilitate the examination of these local behaviours of interest for the researcher without losing the holistic perspective constitutes the main objective of this approach.

In this way, we move away from the genetic and regulatory networks towards the hypothesis-generation paradigm. Our methodology does not try to build regulatory pathways, but rather to gain insight into the relationships among genes. In the zoom-in operation, the system provides a set of biologically relevant genes for every query-set introduced by the researcher, then, presents the detailed expression dependence among all of them in an easy, visual way.

Current approaches for finding regulatory networks from the gene-expression data could be grouped into six categories according to their formalisms: (i) approaches based on linear models [D'Haeseleer P. et al 2000] (ii) approaches based on the description of the regulatory effects using sets of logical rules [Heidtke KR. and Schulze-Kremer S 1998; Thieffry D. and Thomas R. 1998]; (iii) approaches based on a set of deterministic differential equations characterising the regulation patterns of the network [Mestl T. et al 1995; Goryanin I. et al 1999; Becskei A. and Serrano L. 2000; Vohradsky J. 2001; Wessels LF. et al 2001; Vu TT. and Vohradsky J. 2002]; (iv) fully kinetic models which also include stochastic components at the molecular level [McAdams HH. and Arkin A. 1997; Wong P. et al 1997; Arkin A. et al 1998; Kastner J. et al 2002; Rosenfeld N. et al 2002; Swain PS. et al 2002], (v) probabilistic networks [Guet CC. et al 2002; Mao L. and Resat H. 2004], and (vi) neural networks [Weaver DC. et al 1999]. The combination of some of these approaches, as well as the inclusion of certain aspects of stochasticity, has also been utilized [Davidson EH. et al 2002]. The main problems of these approaches are related to the determinism of the boolean networks, which, though they facilitate visualisation, at the same time they hinder the complexity of the system. This reductionism could produce an incorrect model (especially when the data input does not contain all of the genes of the cell), but above all, in all of these approaches, it is impossible to disclose, recover and analyse the sub-layer information not depicted by the model. There are also some continuous models, which try to overcome the linearity limitations of the boolean and probabilistic network approaches, but they have serious visualisation interface problems. Our focussing approach, in addition to working on non-linear relationships among gene

expressions, facilitates an easy and strongly focussed visual interpretation. Finally, it is robust towards background noise and depicts all possible information of interest for the researcher at each moment (despite the problems previously commented on regulatory and genetic networks).

Our method starts by creating a minimum spanning-tree by linking the genes according to its best correlation, linear or non-linear. Then, for each zoom-in operation from a set of query genes provided by the researcher, and using the previously generated tree, the selection process automatically selects a desirable set of high-correlated genes which connect the query genes among them. This selection procedure is extremely sensitive to the provided set of query genes. For example, a large number of query genes or a smaller expression correlation among the query genes will yield a larger number of selected genes. Conversely, a smaller number of query genes or a greater correlation among their expression will yield a smaller number of selected genes. In this way, adding or subtracting a query gene can dramatically modify the number, and the genes that will be provided. The objective of this selection process is to provide the expression-behaviour pattern of the final set, including the query genes and the selected ones supplied by the system, as a way of understanding the initial query-genes behaviour.

The Bayesian, boolean and continuous models mentioned above usually work with temporal-series data. This kind of data is useful for modelling systems such as the cell cycle in order to detect regulatory genes or analyse the dynamics of the genetic networks. However, the temporal series can only study synchronic cellular events. This kind of data is not useful to provide a realistic behaviour of the genes. Our method does not require data such as temporal-series. The expression behaviour of the set of genes is obtained from all of the cellular conditions considered in the expression matrix where these conditions can be achieved as a response to a different stimulus (i.e., chemotherapy, temperature, radiation, starvation, etc). These different conditions constitute the sample space. From this sample space, the system obtains discretised states represented by coordinated-expression levels on local tendencies. The sum of these local tendencies constitutes the intra-set behaviour pattern of the analysed genes in expression terms, being particular to each different set.

The mathematics behind this system makes use of the Principal Curves of Oriented-Points calculation (PCOP) [Delicado P. 2001; Delicado P. and Huerta M. 2003], both in the construction of the gene tree (obtaining the pairwise correlations) used for the selection algorithm, and in the analysis of the intra-set expression behaviour of the resulting set of genes (obtaining the inner pattern of the expression behaviour of the set). The Principal Curves are a non-linear and non-hypothesis-driven analysis technique. The PCOP method uses a hidden variable as a dependent variable for ordering the sample data. In this way, for example, we do not need the time variable of temporal-series data yet.

### 3.2.1. The steps followed by our procedure

#### 1. Pre-process:

**1.a.** Obtaining the correlation degree between each pair of genes using the PCOP calculation.

**1.b.** Building the minimum-spanning tree among the expression-matrix genes using the previously calculated pair wise-correlations.

#### 2. Zoom-in operation:

**2.a.** Automatic selection of the genes that connect the query genes through the minimum-spanning tree calculated in the pre-process. Query genes are provided by the researcher in each new query.

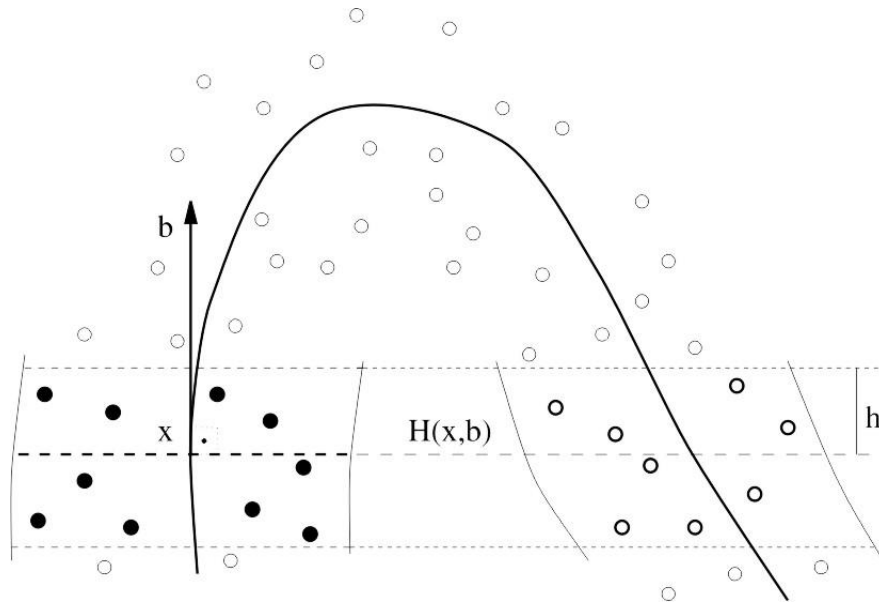
**2.b.** Obtaining the intra-set behaviour pattern of the gene set generated by the selection algorithm (using the PCOP calculation). This inner pattern relates the expression fluctuations of the selected genes plus the query ones (among all of them).

These steps involve capabilities such as: Gaining insight into relationships (known or hidden) among different pathways, analysing how the genes could be affected when a query gene is in a particular level of expression, or modelling and simulating the whole cell physiology (systems and integrative biology), among others.

### 3.2.2. The Principal Curves of Oriented-Points (PCOP) calculation

Principal curves have been devised to extract non-linear patterns from data in an automated, accurate, and reproducible way. It is a powerful method to extract behaviour patterns from different entities in multivariate-data analysis. The Principal Curves method describes the relationships among independent variables by a continuous curve passing through a discrete cloud of points such as, individual experimental measurements from DNA microarrays, or expression matrices from any other origin. PCOP calculation reduces the noise from signal processing, which is a major drawback in array technology, thus preserving the information. Principal components and principal curves do not use dependent variables, a difference with respect to other regression and estimation models. If the sampling space is ellipsoidal, the linear relations among the independent variables would suffice to describe the behaviour of these variables. In other cases, we need to use principal curves because the existent relationship is more complex.

The first approach to the Principal Curves [Hastie T. and Stuetzle W. 1989] defines them as smooth curves passing through the middle of a multi-dimensional data set. The problem with this first calculation of the Principal Curves is that if the distribution is not isotropic, it is difficult or impossible to find the principal curve. There are many approaches to implement Principal Curves [Kegl B. et al 2000; Chang K. and Ghosh J. 2001; Sandilya S. and Kulkarni SR. 2002]. Sandilya and Kulkarni being the first who realized a comparative analysis of Principal-Curves methods, including the PCOP. In addition, other approaches based on neural networks have followed the same objective as Principal-Curves methods [Mulier F. and Cherkassky V. 1995]. Indeed, there are some approaches that fused the two previous ones [Dong D. and McAvoy TJ. 1996].



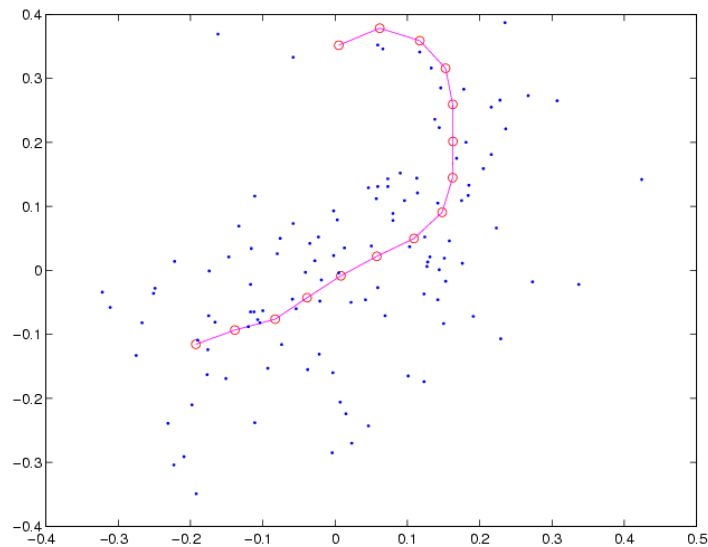
**Figure 4. A theoretical example of PCOP analysis.** The local areas are limited by bandwidth 'h' in two sides of the local area and adjusted to the shape of the data-point cloud in its other two sides. Using the samples clustered in this local area (black spots in the figure), and calculating the principal component for the local area.  $x$  will be validated, or not, as a new Principal Oriented Point(POP), and  $b$  as a tangent vector of the Principal Curve of Oriented Points(PCOP) if variance is minimum [Delicado P. and Huerta M. 2003].

In the present methodology, the PCOP as defined by [Delicado P. 2001] have been used. He defines PCOP on the generalisation, at the local level, of the next principal-component property: for a normal multivariable distribution  $X$ , if  $X$  is projected over a hyper plane, the total variance of the projection is minimized when the hyper plane is orthogonal to the first principal component. In this way, the Principal Component is found for each local area, and the principal-component vector will be the tangent vectors of the Principal Curve. Each Principal Component at local level conforms a Principal Oriented Point (POP), and the curve which visits all of the POP, conforms the Principal Curve of Oriented Points (PCOP). In a previous work [Delicado P. and Huerta M. 2003], it was described how to accurately define these local areas in the global space and the samples belonging to each one (Figure 4). One of the main advantages of PCOP over other Principal-Curves methods is its generalisation of the principal-components variance. As a result, the PCOP method provides a very good measure of the data dispersion around the curve [Sandilya S. and Kulkarni SR. 2002]. PCOP method also has the capability of defining higher-order curves. Both are very significant for our methodology.

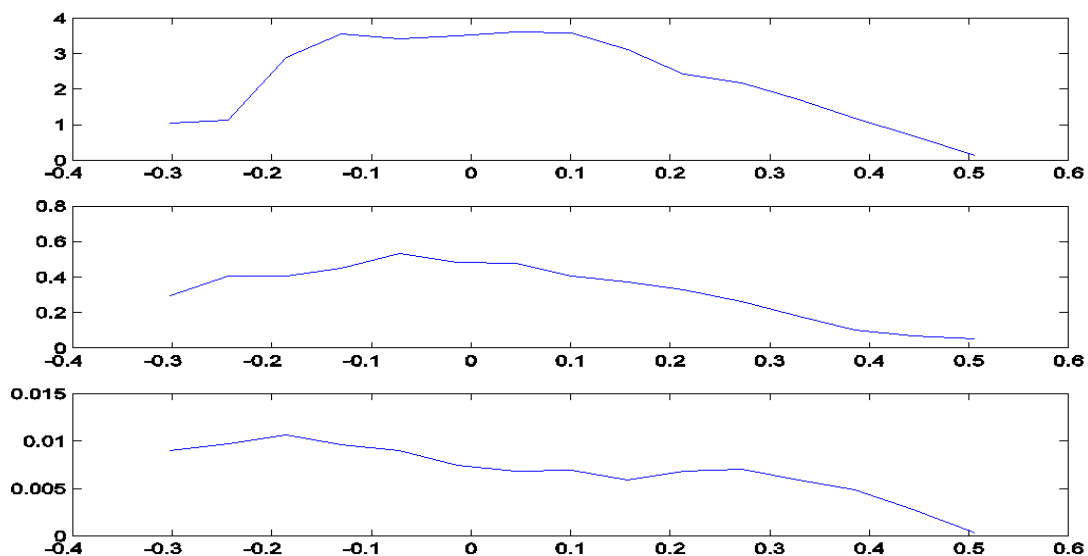
Now we shall describe how to apply the PCOP analysis to the expression data. For all of the following examples, the previously described experimental data from the National Cancer Institute (NCI, USA) (Scherf U. et al 2000) will be used. As an example of PCOP application to this kind of data, we show the Principal-Curve calculation for describing the expression relationship between two of the expression-matrix genes: DNA topoisomerase II beta and 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase (PFKFB). Each gene is an independent variable in a multi-dimensional space. DNA topoisomerase II beta is involved in processes such as chromosome

condensation, chromatic separation, and the relief of torsional stress that occurs during DNA transcription and replication. PFKFB is a key activator of glycolysis. Cancer cells maintain a high glycolytic rate. Figure 5 shows the PCOP obtained. The Principal Curve shows a non-linear relationship between the expressions of the two genes. The density and variance of the cloud of points along the Principal Curve can be also observed in this figure. In [Delicado P. and Huerta M. 2003] the recursive algorithm to calculate the PCOP for a multi-dimensional space is described, and the program is available on the web.

5.a)



5.b)



**Figure 5.** (5.a) Principal Curve of Oriented Points (PCOP) of DNA topoisomerase II beta (abscissas in figure 5.a) and 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase (PFKFB) (ordinates in figure 5.a) calculated from the gene-expression data. The expression patterns are extracted from the expression at\_matrix reported by [Scherf U. et al 2000]. (5.b) The three frames show, respectively: the density (top), span (middle) and residual variance (bottom) around the curve.



The correlation degree of this relationship will be provided by the values of the Generalized Total Variance (GTV) and the Residual Variance (RV). After calculating the PCOP between all pairs of genes, the values of Residual Variance and Generalized Total Variance provided by the PCOP calculations are the parameters used to filter out the insignificant relations between genes. The Variance explained by the curve permits one to know if the Principal Curve is able to follow the sample-cloud tendency, and goes up when the sample cloud has a regular behaviour being well identified by the Principal Curve. Residual Variance (RV) describes the degree of dispersion of the sample cloud around the Principal Curve; in other words, the residual variance is the variance not explained by the Principal Curve. The Generalized Total Variance (GTV) is the sum of these two dispersion parameters, the variance explained by the curve, plus the variance not explained by the curve. The uncorrelation factor ( $f$ ) is RV divided by GTV [Delicado P. 2001]. The lower the value of  $f$ , the higher the correlation degree between gene expressions. The  $f$  value is used to know how many, not exclusively linearly, sampled gene-expressions are related.

### 3.2.3. The minimum-spanning tree

A minimum-spanning tree is a tree built from a set of points which verify the following property: the sum of all edge-weights will be equal to, or lower than the sum of edge-weights of any other tree from the same input data. Based on the PCOP calculation between all the matrix genes, a minimum-spanning tree is built among all the matrix genes where the tree edges represent expression relationships between pairs of genes with its corresponding  $f$  value. The edges of the minimum-spanning tree maintain a decreasing-order based on the  $f$  value of each expression relationships.

The expression relationships of the tree edges can be of three basic types:

- Positively coexpressed genes.
- Negatively coexpressed genes.
- Non-linearly correlated genes.

A relationship of Type 1 or 2 involves genes with a linear expression relationship. A relationship of Type 3 corresponds to correlated relationships like that of Types 1 and 2, but in this case non-linear.

In Figure 6 a gene graph obtained using  $R^2$  (6.a) can be compared with a graph using the  $f$  value provided by PCOP calculation (6.b). We use the subset of matrix genes shown in Table I to build the gene network. In this subset we have mixed different genes of the at\_matrix [Scherf U. et al 2000], where some of these genes take part in the regulation and control of the G1 phase of the cell cycle, and some not.

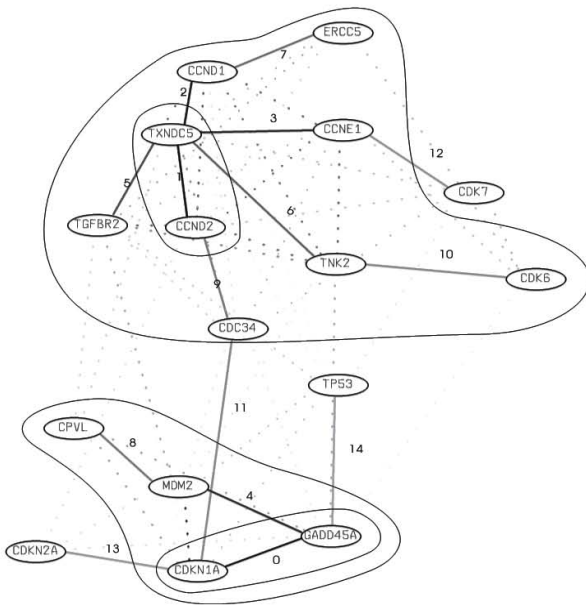
Entrez Gene
CDK7, cyclin-dependent kinase 7
CDK6, cyclin-dependent kinase 6
CCND1, cyclin D1
CCND2, cyclin D2
CCNE1, cyclin E1
TGFBR2, transforming growth factor, beta receptor II (70/80kDa)
TXNDC5, thioredoxin domain containing 5 (EndoPDI)
CPVL, carboxypeptidase, vitellogenic-like
CDKN2A, cyclin-dependent kinase inhibitor 2A (melanoma, p16)
CDKN1A, cyclin-dependent kinase inhibitor 1A (p21, Cip1)
GADD45A, growth arrest and DNA-damage-inducible, alpha
ERCC5, excision repair cross-complementing 5
CDC34, cell division cycle 34
TP53, tumour protein p53 (Li-Fraumeni syndrome)
MDM2, transformed 3T3 cell double minute 2, p53 binding protein (mouse)
TNK2, tyrosine kinase, non-receptor, 2 (ACK)

**Table I. The genes from at\_matrix analysed in the examples of this section.** They have been selected from a matrix of 1416 genes followed in a microarray study on the action of 118 anti-tumour substances (Scherf U. et al 2000). This microarray data is the sample input set used in the examples reported in the text.

In the two graphs, the gene location in the 2D plane depends on its correlation with the rest of the genes using a force-directed layout algorithm with three parameters: attraction, repulsion and gravity (aiSee-Graph-Layout Software).

In both graphs the best correlations are the same (the linear expression relationships), but some genes, like TP53, vary when non-linear expression relationships are considered. Considering non-linear expression relationships, four genes are linked to the minimum-spanning tree through the TP53 gene (Figure 6.b); however, considering only linear correlations, TP53 would have only a marginal role, being linked to the minimum-spanning tree by the worst correlation of the tree (Figure 6). The reason is that TP53 links genes of the G1 phase and some genes of error treatment by means of non-linear relationships. It can be observed meticulously in the detailed expression analysis.

6.a)

Pearson  $R^2$ 

6.b)

f value

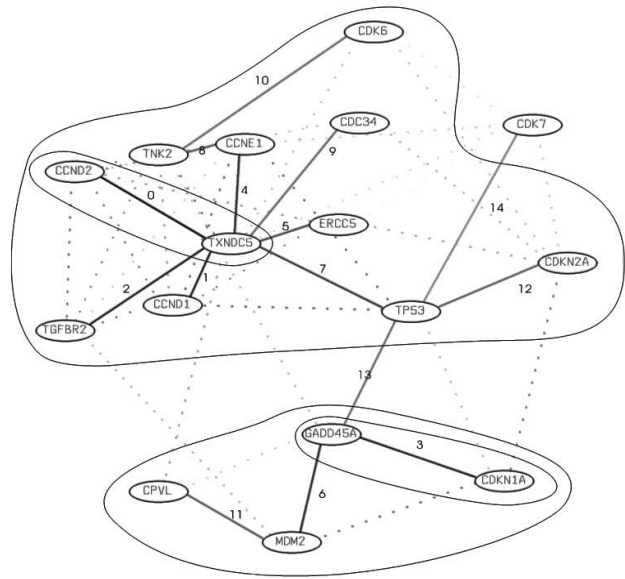


Figure 6. Cluster hierarchy and minimum-spanning tree (mst) for the genes of Table I using a  $R^2$  (Pearson product-moment correlation squared) at left side and the f value (PCPOP) at right side. The edges of the mst maintain a decreasing-order relationship based on the f value or  $R^2$  Pearson value of each one. Their corresponding f value can be observed in the ruler of Figure 7. In the b) graph, the relationships labelled as 0 and 3 represent the minimum intra-cluster pair, the relationship labelled as 13 represents the minimum inter-cluster pair. The nodes are located in the 2D plane using a force-directed layout algorithm. The expression-dependence degree is proportional to the grey scale of the edges.

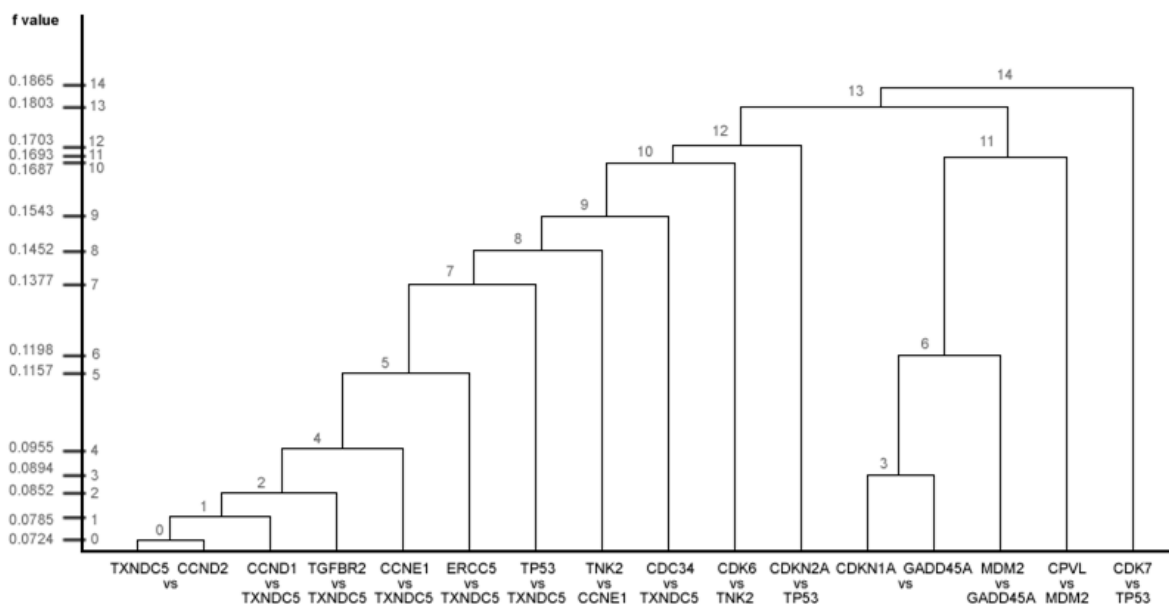
### 3.2.4. The gene-selection algorithm

Using only a graph or network it is difficult to notice the specific behaviour among the network genes. To examine this behaviour in detail, the zoom-in operation is necessary. In the zoom-in operation, the researcher begins introducing the genes of interest. Assuming that the query genes introduced by the researcher have a certain level of correlation as a set (correlation level which the researcher does not wish to lose), the first step of the zoom-in operation is to select the maximum number of genes which connect the genes of the query set, but conserving the correlation level of the query set in the new query-plus-selected-genes set. Notice that each gene is a variable in a multi-dimensional space when we calculate the PCOP of the gene set and its corresponding f value.

#### 3.2.4.1. Hierarchical clustering

We will use hierarchical clustering to expound the benefits of our selection algorithm. In hierarchical clustering, the genes are grouped by a likelihood function applied to their expression patterns. To build a hierarchical cluster, in each step, individual genes or a set of genes are added to

a gene or gene cluster previously added, following the best correlation. As a result, a dendrogram is obtained [Alon U. et al 1999; Eisen MB. et al 1999]. The hierarchical clustering built from the genes of Table I using the  $f$  value and single linkage [Hartigan JA. 1981] is shown in Figure 7. Note that in this way we are not grouping the genes by similarity in expression, but rather by their  $f$  value. Therefore, clusters will contain genes whose relationships show a low  $f$  value, even if these relationships are of different types (linear or non-linear). Using single linkage, we can guarantee that genes which belong to different clusters are actually independent.



**Figure 7. Hierarchical clustering for the genes of Table I using the  $f$  value (PCOP) and single linkage.** Under the names of each gene added to the tree appears the other gene of the relationship used to link the new gene to the tree (single linkage). Their relative position based on the  $f$  value is shown in the ruler at the left.

### 3.2.4.2. Gene selection following the minimum-spanning tree

We have described the properties of hierarchical clustering, but for selecting genes we only need the minimum-spanning tree implicit in it.

For the selection algorithm, we select the genes (nodes) of the minimum-spanning tree which connect the query genes conforming the minimum-spanning path between them.

When building a hierarchical clustering using a single linkage, we obtain a minimum-spanning tree where each edge of the tree represents the relationship used to add each new gene or gene cluster to the tree [Gower JC. and Ross GJ. 1969]. In this way, we can apply these clusters, their hierarchy and the properties of the hierarchical clustering to their corresponding minimum-spanning tree. Figure 6.b shows the minimum-spanning tree corresponding to the hierarchical clustering of Figure 7.

Due to these findings, we can formulate the following definitions with respect to the selection algorithm:

A set of genes of the tree linked by their best correlation will be considered as a cluster. These links between genes will be named intra-cluster relationships and link each gene with its most correlated gene.

A pair of genes whose intra-cluster relationship of each gene links the other gene of the pair, will be termed a minimum intra-cluster pair (the most correlated gene of each gene is the other gene of the pair). Each cluster will present one and only one minimum intra-cluster pair. There is a path between each gene of the cluster, and its minimum intra-cluster pair, following the intra-cluster relationships which link the members of the cluster. This implies that the minimum intra-cluster pair will be accessible from each gene of the cluster following the best correlation of each gene along the way. Note that from each gene, one, and only one intra-cluster relationship leaves, but some intra-cluster relationships can reach it. In Figure 6 two clusters and their respective minimum intra-cluster pair can be observed (represented in Graph b by the relationships with ranks 0 and 3).

In hierarchical clustering, each cluster is linked to the cluster-tree by means of the best relationship between one of its genes and the rest of the genes external to the cluster. This best external relationship of the cluster will be named its inter-cluster relationship. From each cluster one and only one inter-cluster relationship may leave, but some inter-cluster relationships can reach it.

A pair of clusters whose inter-cluster relationship links the other member of the pair will be termed a minimum inter-cluster pair. In Figure 6 the minimum inter-cluster pair formed by the two clusters in the picture can be observed (linked in Graph b by the inter-cluster relationship with rank 13).

Now, to select the genes, we only need to follow the best correlations among genes and clusters (intra-cluster and inter-cluster relationship) towards the minimum inter-cluster and intra-cluster pairs from the query genes until arriving to the meeting point. In this way, we shall follow the minimum-spanning path selecting the genes (nodes) which connect the query genes.

Focussing on the algorithm description, beginning from the query genes, it will take the best intra-cluster relationship in each step. If the different pathways do not cross because the query genes do not belong to the same cluster, the algorithm will search for the meeting cluster following the inter-cluster relationships, taking the best inter-cluster relationship in each step too. For each cluster visited, the algorithm will repeat the process to select the genes visited to cross the cluster. The algorithm will perform in the same way to deal with clusters of clusters. In this way, the algorithm will only need the best correlation of each gene and each cluster (intra-cluster and inter-cluster relationships) to establish the gene selection for all possible sets of query genes.

Now, there are two hot-points to mention about the minimum-spanning tree properties:

The intra-cluster relationship that leaves a gene always represents a better correlation than the intra-cluster relationships that reaches the gene. The inter-cluster relationship that leaves a cluster always represents a better correlation than the inter-cluster relationships that reaches the cluster. Therefore, the pathway to cross the cluster to link two inter-cluster relationships (which always crosses the minimum intra-cluster pair) usually represents a better correlation in the cluster that reaches the inter-cluster relationship than in the cluster from which it leaves. And the path between the

minimum intra-cluster pair and the gene from which the inter-cluster relationship leaves, usually represents a better correlation than the paths between the genes that reach an inter-cluster relationship and the minimum intra-cluster pair.

These two properties permit scientists to reach the selection objectives initially expected. First, the set of selected genes will tend to maximize the correlation among themselves and the query genes rather than another possible set of selected genes with the same number of genes. Second, the system will tend to select the maximum number of genes to connect the different query genes without losing the initial correlation degree existing among the query genes.

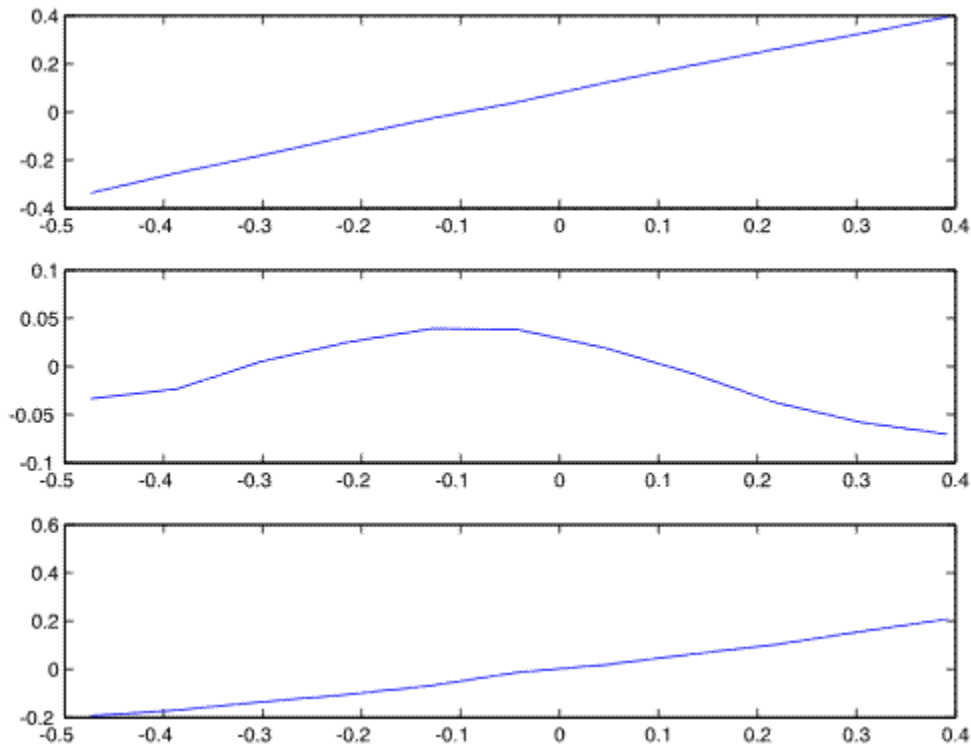
Let us consider two practical examples to illustrate the zoom-in operation and its gene-selection process. The hierarchical clustering and minimum-spanning tree needed for the examples-selection process are depicted in Figures 7 and 6.b. The first objective is to study what type of connection exist between the level of expression of genes cyclin E1 (CNNE1) and TP53. In the second example, the interest is to analyse what relation exists between cell-cycle regulators CDK6 and cyclin E1 (CNNE1) with TNK2 (which is involved in the tyrosine-phosphorylation signal-transduction pathway). In the first example, with the objectives being cyclin E1 and TP53, the gene-selection algorithm returns the thioredoxin-related protein endothelial protein disulphide isomerase gene, TXNDC5 (see Figure 6.b), but for the second example, none are selected. The next step is to calculate the intra-set behaviour pattern of the new gene set generated.

It's important to remark that the results of the selection algorithm are not the same, if one considers only the linear expression relationships with respect to also consider non-linear expression ones. If we try to obtain the selected genes from the query genes CDKN2A (P16) and CDKN1A (P21), basing ourselves only on linear correlations (Figure 6.a), no genes are selected. But if we consider the non-linear correlations (Figure 6.b), TP53 and GADD45A are selected, reflecting that TP53 is a pivotal element to determine the function of this sub-set of genes. For instance, the p53/p21 pathway is responsible for glycolated Collagen II-induced apoptosis, and p53/p21 coordinated activation with CDKN2A (P16) induces cell senescence [Chen J. et al 2006].

### **3.2.5. Obtaining the intra-set behaviour pattern of the generated gene subset**

Once the genes are selected, the following step is to search the dependence among the expression levels of all of them, plus the query genes. To represent this intra-set behaviour, we shall use the parametric representation of the PCOP calculated for the expression of all of these genes.

Let  $n$  be the number of the selected genes, plus the query ones. The description of the relative behaviour of each one of these genes with respect to the others, are the  $n$  equations of the parametric representation of the PCOP calculated for these  $n$  genes. The curve parameter is defined by the POPs found (Figure 8). The POP with a 0 value in this range represents the POP in the relationship where there are approximately the same number of samples on its left and right sides.



**Figure 8. Expression relationship among the selected genes TXNDC5 (top panel), TP53 (middle panel) and cyclin E1 (CNNE1) (bottom panel).** The panels show the gene-expression levels for the discretised states (POPs) of the relationship (PCOP). Abscissas correspond to the expression levels of each gene and ordinates the discretised states (POPs). The POP with a 0 value in ordinates points out the point of the PCOP with a greater sample density.

Taking up the first example, we need to obtain the inner pattern of the expression dependence among the objective genes cyclin E1 (CNNE1) and TP53, and the selected gene TXNDC5. We chose p53 because: (i) its relationship with other proteins is well described in the literature, (ii) **its expression relationships with other gene expressions are complex and non linear**, (iii) it is directly related to the development of tumours and its expression is altered in a high number of tumours. This last point is relevant from a technical standpoint, as it would be useful to test if the data processing is resistant to this kind of alterations. Figure 8 shows the behaviour among the three genes. Let us contrast these results with the current knowledge of cell-cycle events. Cyclin E1 is a regulatory subunit of the cdc2-related protein kinase CDK2, which is activated shortly before S-phase entry. Lower levels of cyclin E1 imply lower cell-division rates, whereas higher levels of cyclin E1 precede higher rates of cell division [Hinchcliffe EH. et al 1999]. High levels of TP53 induce either apoptosis (in the presence of appropriate mutations) or, alternatively, switch on the mechanisms of DNA repair. At low levels of TP53, less apoptosis is produced and mutations can accumulate more easily. It is known that rapidly dividing cells show a higher mutation rate, whereas slowly dividing cells show lower rates [Bielas JH. and Heddle JA. 2003]. It has been reported that constitutive cyclin E1 over-expression, in both immortalized rat embryo fibroblasts and human breast epithelial cells, results in chromosomal instability [Spruck CH. et al 1999; Tissier F. et al 2004]. A slight overproduction (just 5% more is enough) of cyclin E1 has been associated with the malignant phenotype and is strongly correlated with tumour size [Tissier F. et al 2004]. Other

authors have also reported the association of the above genes with cell division and apoptosis [Knoblach B. et al 2003; Sullivan DC. et al 2003]. In the light of the foregoing, we can find interesting new observations about the genes in relation to tumour evolution: Cyclin E1 (CNNE1) up-regulation leads to higher rates of cell division [Hinchcliffe EH. et al 1999]. Basal expression levels of cyclin E1 are related to over-expression of P53. When the expression level of cyclin E1 is very high, the cell loses the protective effect of TP53 (the TP53 level drops dramatically) and mutations can appear. At this moment, apoptosis is not only avoided by TP53 clearance, but also by the protective effect of thioredoxin on the protein-folding pathways, especially in the hypoxia conditions usually found in tumours (TXNDC5 is over-expressed). Therefore, we can follow the progression of a tumour, finding the same results as those reported experimentally by other authors [Knoblach B. et al 2003; Sullivan DC. et al 2003], but working at high-throughput. For example, we can easily explain several well-known facts jointly, such as: (i) low/medium rates of cell division imply low mutation levels, (ii) high rates of cell division imply the existence of more mutations without apoptosis and (iii) how cyclin E1 and TP53 expression levels are related to the acceleration / restraint of the G1 cell-cycle phase. Note that the selection of TXNDC5 by our method neither represents the belonging to the same activation pathway of TP53 and cyclin E1 nor to the interaction of these proteins. The basic reason for its selection is the adaptive response of the cell (TXNDC5) in order to survive in a low oxygen environment due to the cell growing at a high-division rate (cyclin E1). For this reason, our methodology is not directly comparable with the current methodologies which do not intend to represent these holistic and non-lineal gene relationships.

In the second example, we study the CDK6, cyclin E1 and TNK2 genes. Figure 9 shows the corresponding intra-set behaviour in expression terms. CDK6 and cyclin E1 genes show mutual-exclusion expression with respect to the TNK2 gene. When TNK2 is expressed above the control level, CDK6 and cyclin E1 levels are fixed around their minimum expression. When CDK6 and cyclin E1 are over-expressed, TNK2 is at its minimum level. The 0 value in the curve parameter (abscissa axes in Figure 9) is positioned when all of the genes are in basal expression. The location of this 0 value shows us that cyclin E1 and CDK6 over-expression is more usual than TNK2 over-expression for the analyzed data. Note that the intra-set behaviour pattern obtained depends directly on the concrete genes that are being compared. If the worst correlated genes are removed from the set, the pattern approximation of the new set will upgrade.



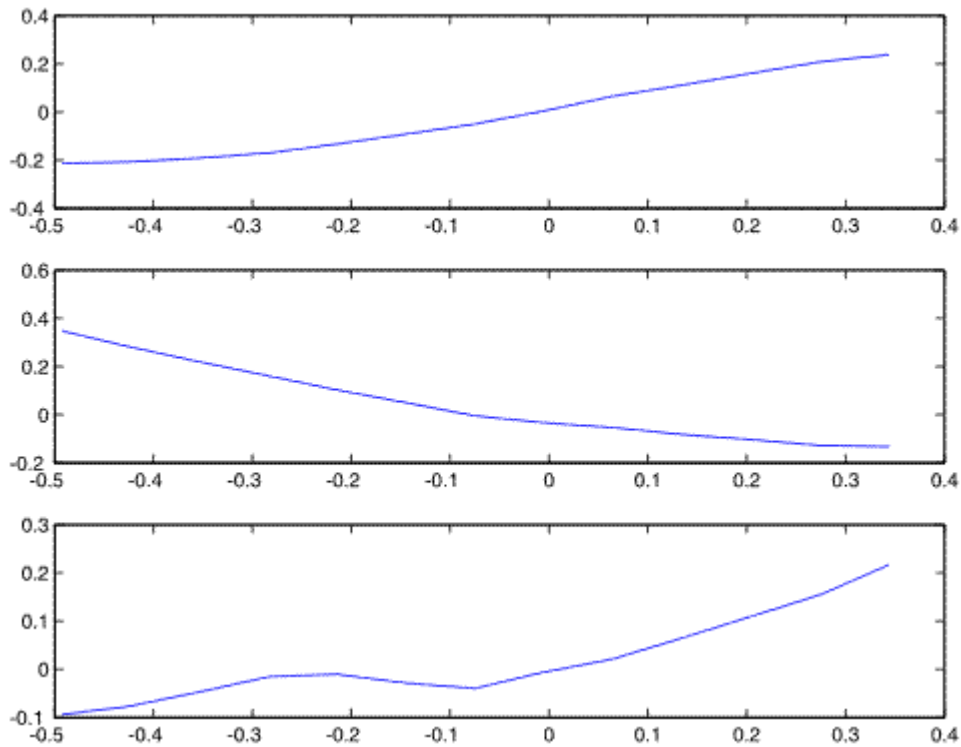


Figure 9. Expression relationship (PCOP) of the selected genes: cyclin E1 (CNNE1) (top panel), TNK2 (middle panel) and CDK6 (bottom panel). The panels show the gene-expression levels for the discretized states(POPs of the PCOP). Abscissas correspond to the expression levels of each gene and ordinates to the discretised states(POPs).

In the light of the foregoing, we can conclude:

Our methodology takes into account the great complexity of the biological systems but, at the same time, can adjust the operational scope dynamically without modifying the working data. In this manner, for each concrete study, the zoom-in operation automatically ignores the genes which can only generate distortion in the analysis of the relevant ones.

The non-predetermined interpretation of information is present throughout the whole procedure. PCOP method uses a hidden variable, which can be temporal or not, to sort the expression data.

The kind of information obtained by our methodology. The mRNA turnover in the cell is controlled by hundreds of different transcription factors (in their different activation states), by chromatin methylation changes, by histone modification (acetylation, phosphorylation, methylation, and recruitment of histone-binding protein) or by modulating mRNA half-life. That complex regulation, especially of the key genes, plus the noise mainly affecting the low-expressed genes [Tu Y. et al 2002], like transcription factors, severely complicate the reconstruction of the activation pathways. That gives meaning to our approach beyond the activation pathways or GO functional annotation. For instance, if we try to see the connection between the two genes seen previously, cyclin E1 and TP53, using the GO tool [Ashburner M. et al 2000] their functions/biological processes are only indirectly related. In the activation or metabolic pathways, the dependence of their respective

expression levels is not shown and they do not appear related to the TXNDC5 gene, and as has been shown previously, it plays an important role in understanding the relations between cyclin E1 and TP53 genes.

All the findings presented in this section have been included in the works [Delicado P. and Huerta M. 2003; Huerta M. et al 2008].



### 3.3. Clustering the samples along the PCOP

The suitable data for our type of analysis can be provided by: (i) temporal series, useful to study synchronous cellular events, and (ii) serial analysis of gene-expression samples under different conditions (i.e., chemotherapy, temperature, radiation, starvation, etc.) which are more useful for studying asynchronous events.

The progressive increase of sample-series [Hall PA. et al 2005] motivates a more thorough analysis of expression relationships and gene dependencies throughout these large series, trying to rescue global gene behaviours and phenotypes. The GEO database [Barret T. et al 2003] facilitates the study of the microarray experiments grouped into predefined subsets, introduced by the microarray authors, and the search for differentially expressed genes. Nevertheless, if the researcher wants to understand the microarray-experiments effect on expression relationships and elucidate cell states and phenotypes, he/she needs a more versatile approach.

#### 3.3.1. Defining sample classes

The sample-classes definition can be made in three different ways using our system:

- Clustering the samples from a part of a gene-expression range.
- Clustering the samples from a part of a gene-expression relationship.
- Clustering the experiments based on previous knowledge.

Sample clusters based on previous knowledge comes from a) medical/biological origin, provided for example, by the microarray developers, or b) obtained by statistical method such as k-nearest neighbour [Ripley B. 1999], Self-organizing Maps [Kohonen T. 1999], Principal Components, Biclustering [Wu CJ. and Kasif S. 2005], Locally Linear Embedding [Chao S. and Lihui C. 2005] and so on.

#### 3.3.2. Clustering the sample from a non-linear expression relationship

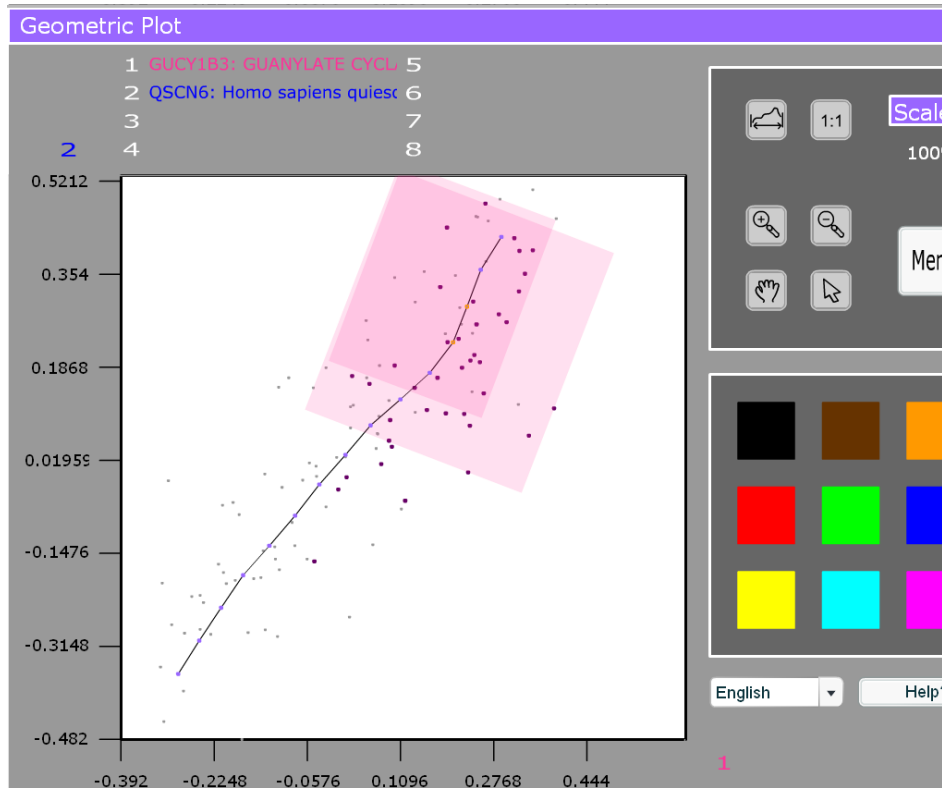
Using the clustering the samples along the PCOP tool, the user can select the different discretised states (the POPs) obtained in the PCOP calculation. With this action, the samples belonging to the local area of the selected POPs are clustered. In this way, the user can separate the sample data for their belonging to the different local behaviours of the expression relationship. So, this new clustering approach permits us to differentiate the samples belonging to a continuous dataset on the basis of a non explicit reason (or hidden-variable role) that guides the local tendencies. Identifying the samples that contribute to each local tendency, we can arrive at the understanding of the surrounding reason or hidden variables effect.

The advantage of defining the sample clusters from an expression relationship instead of an expression range, is that we are selecting the samples by the tendency maintained by the relationship among two or more genes. As a result, our clustering from an expression relationship more accurately defines the boundaries of the phenotype we wish to associate with the sample class [Cedano J. et al 2008]. In contrast to using a statistical method that clusters the sample conditions using all of the matrix genes (or some of them in Biclustering [Tanay A. et al 2002]), clustering from an expression relationship, only the marker genes of the processes in which the researcher is interested are used (the clustering will be based on the tendency of the expression relationship in which the researcher is interested).

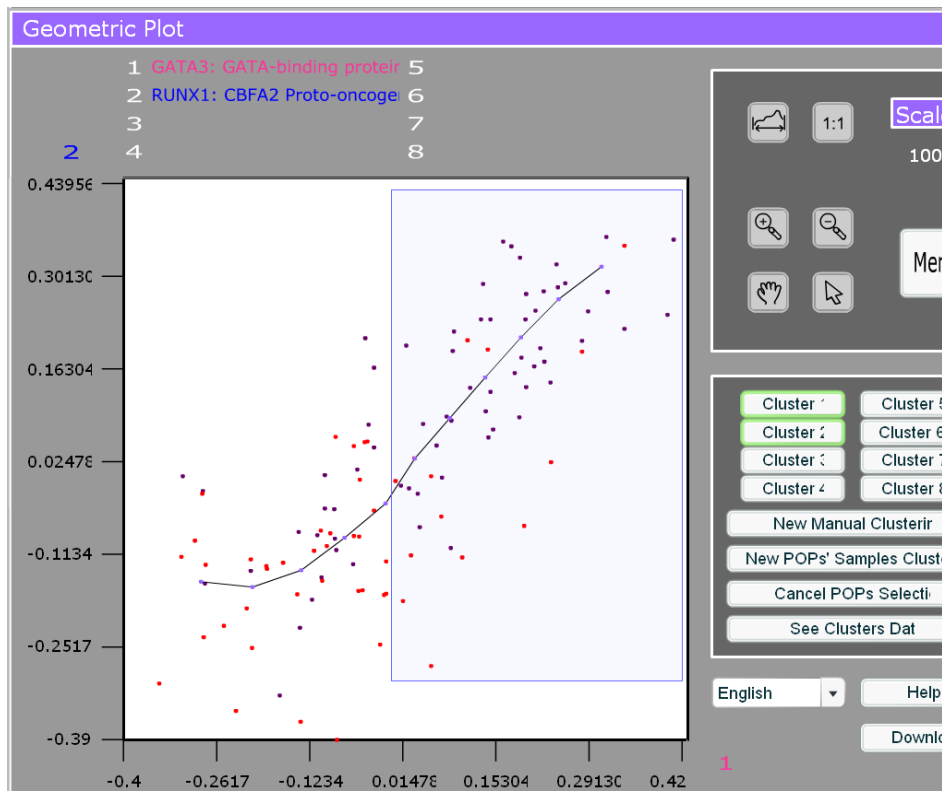
### **3.3.3. Colouring the sample clusters on a gene-expression relationship**

The graphical interface facilitates the visualisation of the defined sample classes on gene relationships by colouring the samples (as is shown in Figure 10) with the intention of studying their effect on expression-relationships. The study of this effect is especially relevant in non-linear expression relationships in order to understand the biological sense of the slope changes, possibly due to a phenotypic change.

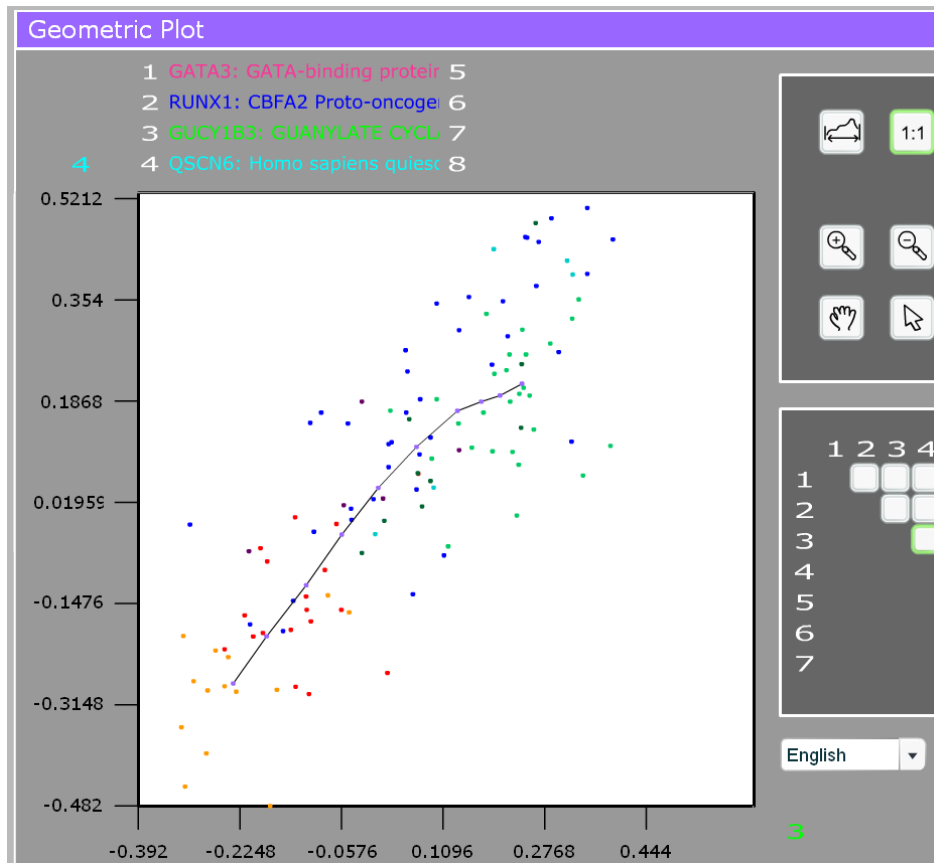
10.a)



10.b)



10.c)



**Figure 10. Sample class definition using the web interface.** Example 1: In the SGC(GUCY1B3) and Q6(Q6SN6) expression relationship (10.a;  $f=0.08$ ), the samples of the two extremes of the SGC / Q6 inner pattern are clustered into two classes by clicking on the POPs. Next, these sample classes are coloured in the GATA3 and AML1(RUNX1) expression relationship (10.b;  $f=0.10$ ). Example 2: In the GATA3 and AML1(RUNX1) comparison window (10.b), the samples with GATA3 over-expressed are clustered with the range-selection facility and coloured in the SGC and Q6 expression relationship (10.a). Example 3: In the GATA3, AML1(RUNX1), SGC and Q6 relationship (10.c;  $f=0.28$ ), the plot of the projection over the SGC and Q6 plane is shown. The coloured classes shown have been obtained using a clustering method. Each sample class represents a different phenotype involving most of the genes of the expression data. The phenotypes identified by sample clusters coincide with the two proliferation phenotypes described in the previous examples, the differentiated and the undifferentiated proliferation phenotypes.

### 3.3.4. Gene search based on the arrangement of the sample classes along the gene-expression ranges

In the 'sample-class-distribution' search, the genes can be searched by some of the defined classes being up-regulated or down-regulated with respect to the basal value, or by being disjointed, over-expressed or under-expressed, with respect to some of the other classes. All possible combinations of the requirements are allowed where different combinations will supply different gene-sets.

### **3.3.5. Basic analysis procedure**

The analysis begins in the researcher's genes of interest, usually marker genes of a specific disease, cell state or function. Once the pattern analysis of the query genes has been performed, the graphical interface will show their expression-relationship, the inner pattern (PCOP) and their fluctuations.

To find the expression dependence of these initial sets of correlated genes with other genes either linearly or non-linearly correlated with the first ones, the user should proceed as follows: first, he/she must cluster the samples belonging to the different local behaviours of the query-genes expression relationship (by clicking on the POPs along the relationship's inner-pattern in the plot interface); second, applying the 'classes-distribution' search tool for a certain distribution of the samples clusters, the genes that follow the required distribution in their expression are obtained. Finally, the researcher can now perform the pattern analysis of the genes provided by the search and observe, in the graphical interface, the effect of each sample-class on the expression relationship (with the samples of each one of the classes coloured, as in Figure 10). This procedure will show the non-continuous expression dependence among the genes provided by the search and the initial gene-set in the manner specified in the search. If the sample-classes distribution required in the search is changed, the genes provided and their non-continuous dependence, with respect to the initial ones, will vary too.

### **3.3.6. Contextualisation consulting GEO database**

In order to get complementary information from other public expression data, the GEO database could be consulted. These queries attempt to know if the genes supplied by the 'sample-class-distribution' search tool (genes that follow the user-defined sample-class distribution in their expression) are differentially expression genes in the GEO datasets. To achieve this information, the web application queries the microarray GEO Profiles [Barret T. et al 2003]. The GEO datasets where the query gene displays significant expression differences among the GEO predefined sample classes are obtained. In this way, the queried gene can be considered, for example, as a marker of osmotic stress in a microarray to analyse cellular stress response, or as a marker of metastasis in a microarray to analyse disease states of cancer, etc. Then, these query-gene properties can be also assigned to the user's sample clusters.

### **3.3.7. Sample-classes definition use cases**

Let us describe some real use cases and the relevance of the new knowledge supplied. The three ways to define the sample classes and a basic analysis procedure have been used.

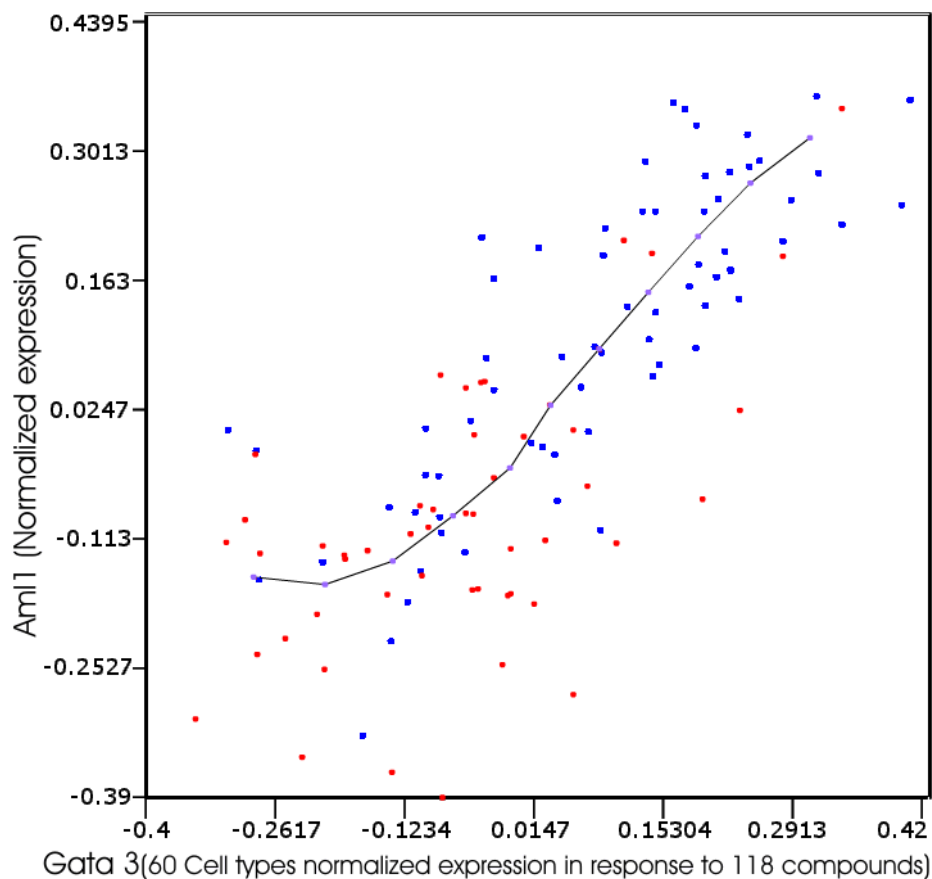


### 3.3.7.1. Case 1: Defining the sample classes from a non-linear expression relationship

We wish to relate the Soluble Guanylate Cyclase Beta1 3 (SGC) and Quiescin Q6 (Q6) genes (our genes of interest). The SGC is under-expressed in cellular stress [Roelofs J. and Van Haastert PJ. 2002], whereas Q6 is over-expressed in the last phases of tissue remodelling [Thorpe C. et al 2002]. Additionally, we would like to relate this pair of genes to GATA-binding protein 3 (GATA3) and acute myeloid leukaemia 1 (AML1, RUNX1). GATA3 is involved in growth control and the maintenance of the differentiated state in epithelial cells [Usary J. et al 2004]. The impairment of the AML1 function deregulates the pathways leading to cellular proliferation and differentiation [Michaud J. et al 2003]. The two gene pairs show an expression correlation (uncorrelation factors of: SGC vs Q6 = 0.08; GATA3 vs AML1 = 0.1). Their inner pattern can be visualised in the graphical interface, as is shown in Figure 10. The problem is that these two pairs of genes are neither linearly nor non-linearly correlated among all of them (uncorrelation factor of SGC vs Q6 vs GATA3 vs AML1 = 0.28). But perhaps they are maintaining a non-continuous expression dependence, and we cannot discern it with the analysis for continuous data-clouds. We can try to find this non-continuous relationship using the sample classes' definition from one expression relationship and colouring the sample clusters in the other expression relationship.

To perform this, two clusters are built from the SGC and Q6 expression relationship by selecting the POPs located at the two extremes of its inner-pattern, one with the cellular-stress samples, and the other with the tissue-remodelling samples (Figure 10, SGC and Q6 relationship).

Now, the defined classes are applied to the GATA3 and AML1 relationship (painting the samples of the two clusters with red and blue colours respectively in Figure 11). As can be observed in Figure 11, almost all of the samples corresponding to cellular stress (red) appear with an under-expression of GATA3 and AML1, indicating that both genes are not over-expressed in cellular stress. However, the tissue-remodelling-class samples (blue) appear along the GATA3 and AML1 relationship as being over and under-expressed, and indicate that some of these tissue-remodelling conditions are affected by the GATA3 and AML1 differentiation ways, while some others are not. This points out that the over-expression of GATA3 and AML1 implies an over-expression of SGC and Q6, but not the opposite. Note that this “uni-directional” relationship is impossible to detect by pattern-analysis methods.

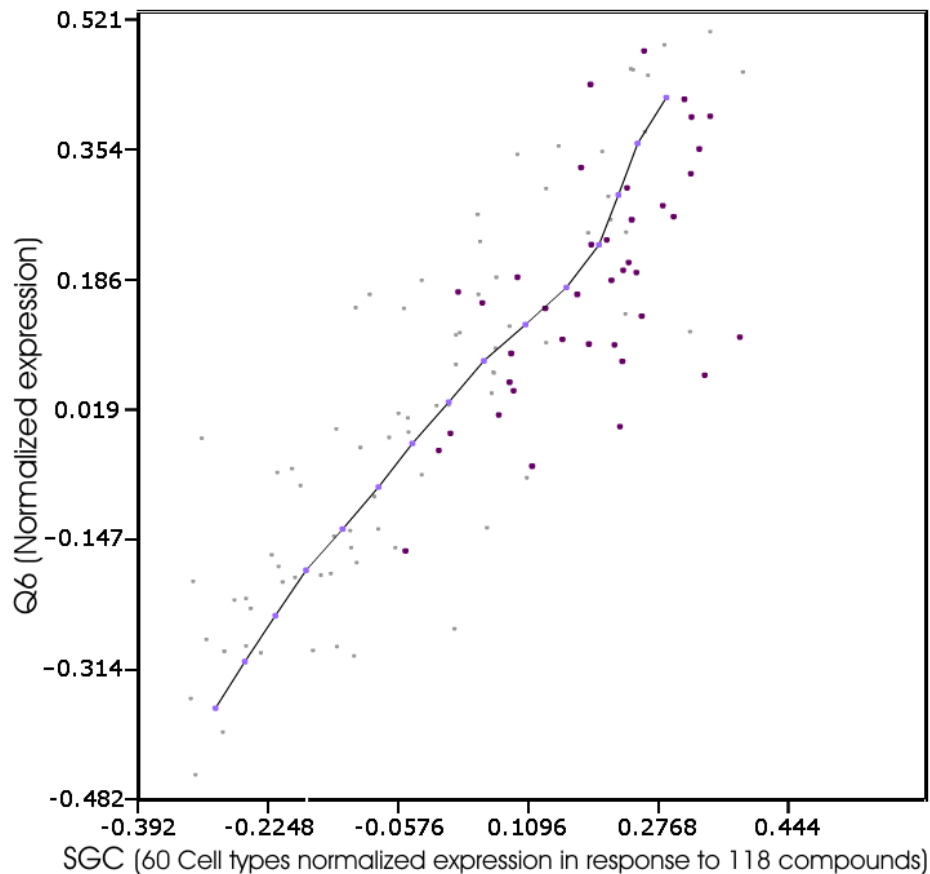


**Figure 11. Example 1: Effect of the two sample classes obtained from both extremes of Q6 and SGC expression relationship, on GATA3 and AML1 expression relationship.** In the picture, the GATA3(x axis) and AML1(y axis) expression relationship is shown, where the data-cloud shows the expression-matrix experiments and the PCOP, connecting the POPs, describes the inner pattern of the expression relationship. The samples under-expressed in the Q6 and SGC relationship are coloured in red in the GATA3 and AML1 expression relationship, and the samples over-expressed in the Q6 and SGC relationship are coloured in blue. This points out that the under-expression of Q6 and SGC implies an under-expression of GATA3 and AML1, but that an over-expression of Q6 and SGC does not always imply an over-expression of GATA3 and AML1. In this way, cell stress would imply the GATA3 and AML1 differentiation, but tissue remodelling would imply GATA3 and AML1 differentiation only in some conditions but not in others. The data set is the at\_matrix of [Scherf U. et al 2000].

### 3.3.7.2. Case 2: Defining the sample classes by means of selecting gene-expression ranges

As seen above, there are some experiments involved in tissue remodelling, but not in the GATA3 differentiation processes. It would be interesting to study them. For this purpose, those samples where GATA3 is over-expressed were clustered (using the graphical interface) to define a new sample class (Figure 10, GATA3 and AML1 relationship).

This class of samples is coloured in the SGC and Q6 relationship, as shown in Figure 12. As we can see, the differentiation induced by GATA3 is independent of the tissue-remodelling level achieved by the SGC and Q6 relationship.



**Figure 12. Example 2: Effect of the over-expressed samples from GATA3 and AML1 expression relationship on SGC and Q6 expression dependence.** The figure shows the SGC(x axis) and Q6(y axis) relationship, where the data-cloud shows the matrix experiments. The data set is the at\_matrix of [Scherf U. et al 2000]. The PCOP, connecting the POPs, describes the inner pattern of the expression relationship. The samples of the class representing the over-expression of GATA3 and AML1 are coloured. In this way, the four gene expressions are related, SGC and Q6 linearly, and GATA3 and AML1 maintaining a complex non-continuous relationship with the other two. In this way, we can observe the tissue-remodelling conditions with GATA3 and AML1 differentiation. Looking at Figure 10.c it can be seen how these tissue remodelling with GATA3 and AML1 differentiation samples coincide with one of the four sample clusters obtained by means of sample clustering (the green sample cluster). This fact points out that these samples represent a phenotype that affect most of the expression-matrix genes and that we have two tissue-remodelling phenotypes, one with GATA3 and AML1 differentiation and another without this. As it will be seen in the next sections the differentiated proliferation will be also stressed, and the undifferentiated proliferation will be unstressed.

### 3.3.7.3. Case 3: Defining the sample classes by means of classifying the experiments using previous knowledge

Previous knowledge can arise basically from two different origins: a biological/clinical origin or a statistical one. In our case the microarray experiments are grouped by Principal Components. Other methods like bi-clustering [Getz G. and Domany E. 2003; Tanay A. et al 2002] or Locally Linear Embedding [Roweis ST. and Saul LK. 2000; Roweis ST. and Saul LK. 2000] can be used with better accuracy to define the sample classes. In defining the classes by means of the matrix-experiments similarity or correlation, the genes are similarly expressed under the sample conditions

of a class but differently expressed with respect to the others. Thus, we can establish the hypothesis that each sample class represents a different phenotype.

The effect of the defined sample classes on the expression relationship among genes of interest can now be observed. Colouring these sample classes on the expression relationship among the previous four genes (Figure 10, GATA3, RUNX, SGC and Q6 relationship), and remembering the observations of the above examples, we can observe four different phenotypes: cell-stress phenotypes (yellow and red) and tissue remodelling phenotypes (green and blue), the last tissue-remodelling phenotypes being divided into the phenotype implicating GATA3 and AML1 differentiation (green), and the phenotype not implicating this differentiation (blue).

#### 3.3.7.4. Marker-gene search based on the arrangement of the sample classes along their expression range

With the sample classes obtained in the last example (Figure 10, GATA3, RUNX, SGC and Q6 relationship) and using the ‘sample-class-distribution’ gene-search tool, it is interesting to search the genes which mark the ‘transition’ from the tissue remodelling without GATA3 differentiation phenotype (blue), to the tissue remodelling with GATA3 and AML1 differentiation phenotype (green). For this purpose, a ‘class-distribution’ gene search like the next one can be performed: the blue and green classes over-expressed with respect to the basal value; the rest of the classes under-expressed with respect to the basal value; and the green-class samples more over-expressed than the blue-class samples.

Furthermore, from the supplied genes, we can identify marker genes of specific cell processes or pathologies by means of the queries made against the GEO Profiles. Thus, the analysis can be focused on the relevant genes for the user's biomedical interest.

In the light of the foregoing, we can conclude:

The presented tools strength resides in: (i) the flexibility of the sample classes’ definition, due to the non-linear pattern analysis of gene expressions and the sample clustering along the inner pattern, combined with (ii) the high-throughput approach of microarray technology, which, by means of the ‘class-distribution’ gene search and the gene-correlation table, leads the researcher to expand his/her analysis. As a result, our application can help to relate gene expressions when their relationships are non-continuous and cannot be found using linear or non-linear analytical methods.

The flexibility of the tool leads to the combination of the three ways to define and redefine the sample classes. For example, using the “sample class definition from a non-linear expression relationship” procedure, the sample classes can be clustered from two different gene sets apparently uncorrelated between them. Therefore, the user could search the genes that are partially related to both sets of genes in a specific manner, for instance, being correlated with one gene set in the under-expression, but with the other set in the over-expression. Or redefine the sample classes defined initially using previous knowledge by their effect on gene-expression relationships (performing sub-

classes from the original ones to study sub-processes in which the genes of interest are involved). In the cases of study previously analysed, a gene set linked with stress/proliferation and another gene-set linked with undifferentiation/differentiation have been related finding several sub-phenotypes. These sub-phenotypes will be key in the final analysis of the presented work.

All the findings presented in this section have been included in the works [Cedano J. et al 2007; Cedano J. et al 2008]

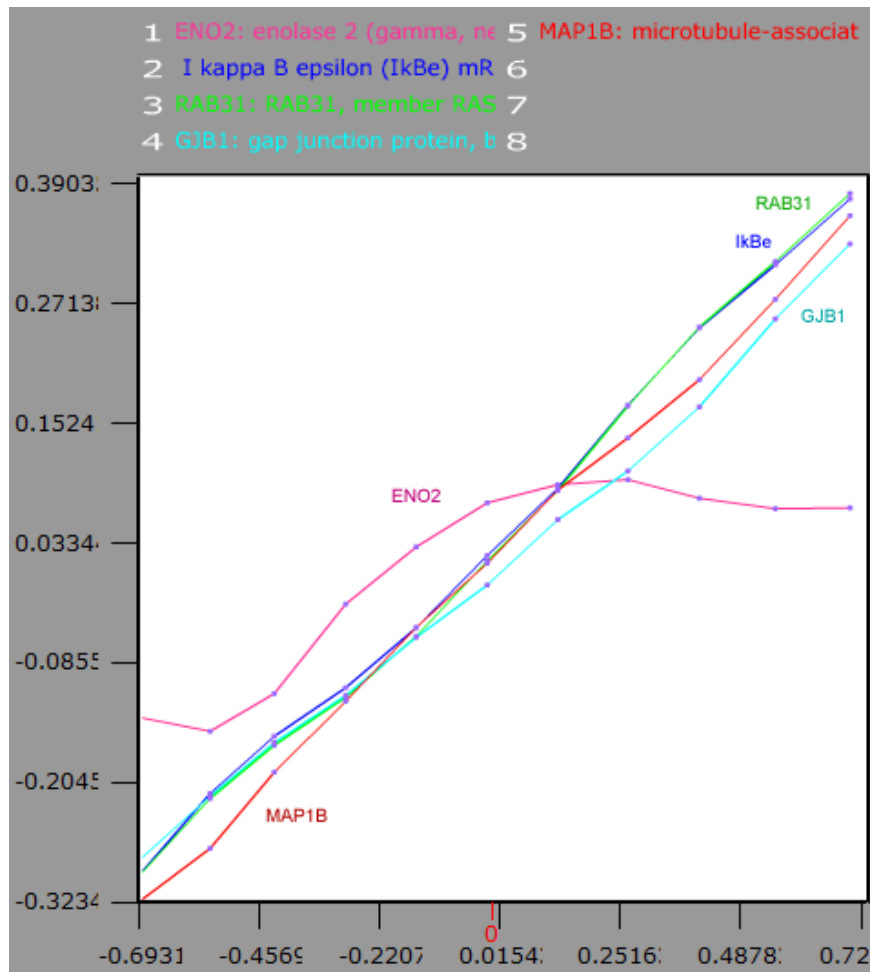
### 3.4. PCOPGene-Net : The interactive gene network.

There are several relevant web applications for microarray analysis, i.e., GEO [Barret T. et al 2005], BIOREL [Antonov AV. et al 2006], ArrayExpress [Parkinson H. et al 2005], MicroGen [Burgarella S. et al 2005] and GEPAS [Tarraga J. et al 2008]. Currently, most tools try to extract biological information from such high-throughput expression data combining information from coexpressed genes [Frickey T. and Weiller G. 2007] as well as additional annotations extracted from Gene Ontology (ADGO) [Nam D. et al 2006], phylogenetic information (CLANS)[Frickey T. and Lupas A. 2004] or pathway data (MAPMAN) [Thimm O. et al 2004].

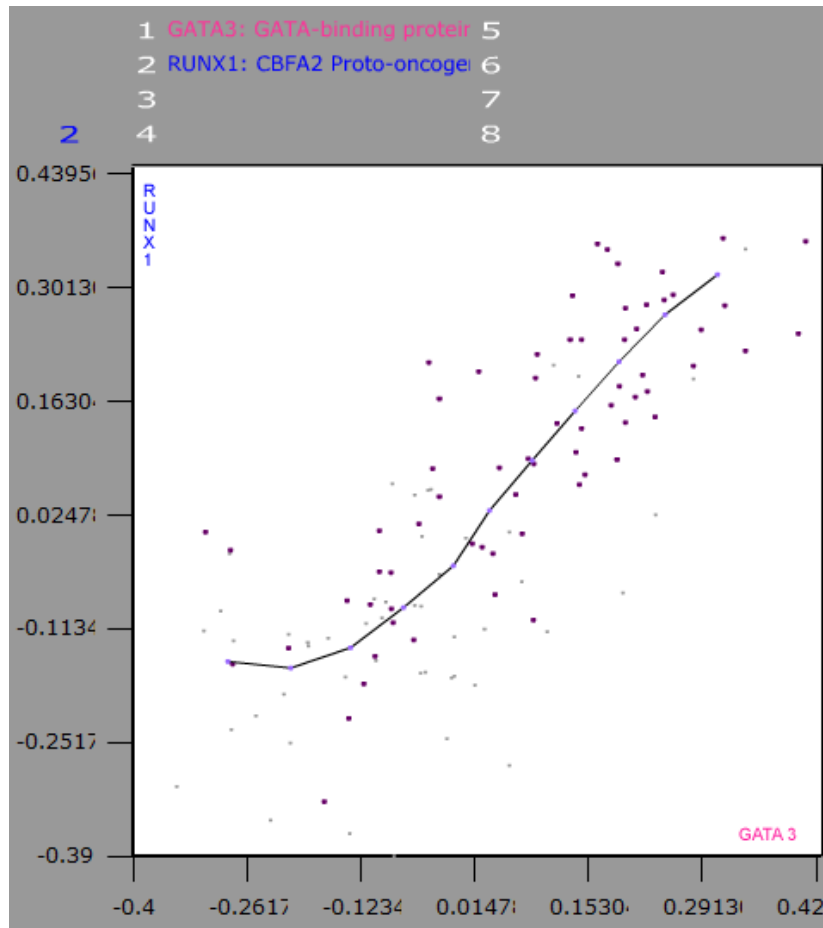
The large number of microarray genes involves genes belonging to very different processes and functions, thus leading to a holistic perspective. Our strategy to achieve this holistic perspective is to facilitate a progressive analysis leading the researcher from known marker genes, and progressively widen his/her scope analysis towards the holistic perspective. This progressive analysis is based on the navigation through the expression data based on: researcher interests, the linear, non-linear and non-continuous relationships among gene expressions, and the links among genes supplied by external biomedical databases.

### 3.4.1. The detailed view of the gene-expression relationship

In the same interface where the PCOP is shown (see detailed view, Figures 13 and 14), the samples can be clustered to define the sample classes clicking on the POPs. Previously-defined sample classes can be coloured simultaneously in the same interface to study their influence on gene-expression relationships and their fluctuations (Figure 14).



**Figure 13. The Detailed View (parametric plot) of a expression relationship.** In the display shows the dependence relationship among the selected genes. The ordinate axis indicates the expression level while the abscissa indicates the parameter of the function that describes the relationship. The lines represent the expression level of the compared genes for each point of the relationship. Selecting any point of the relationship (the POPs), the samples belonging to this relationship stretch are clustered to define the sample classes. In the display, it is shown that ENO2 has an under-expression phase and an over-expression phase, with respect to the rest of the analysed genes.



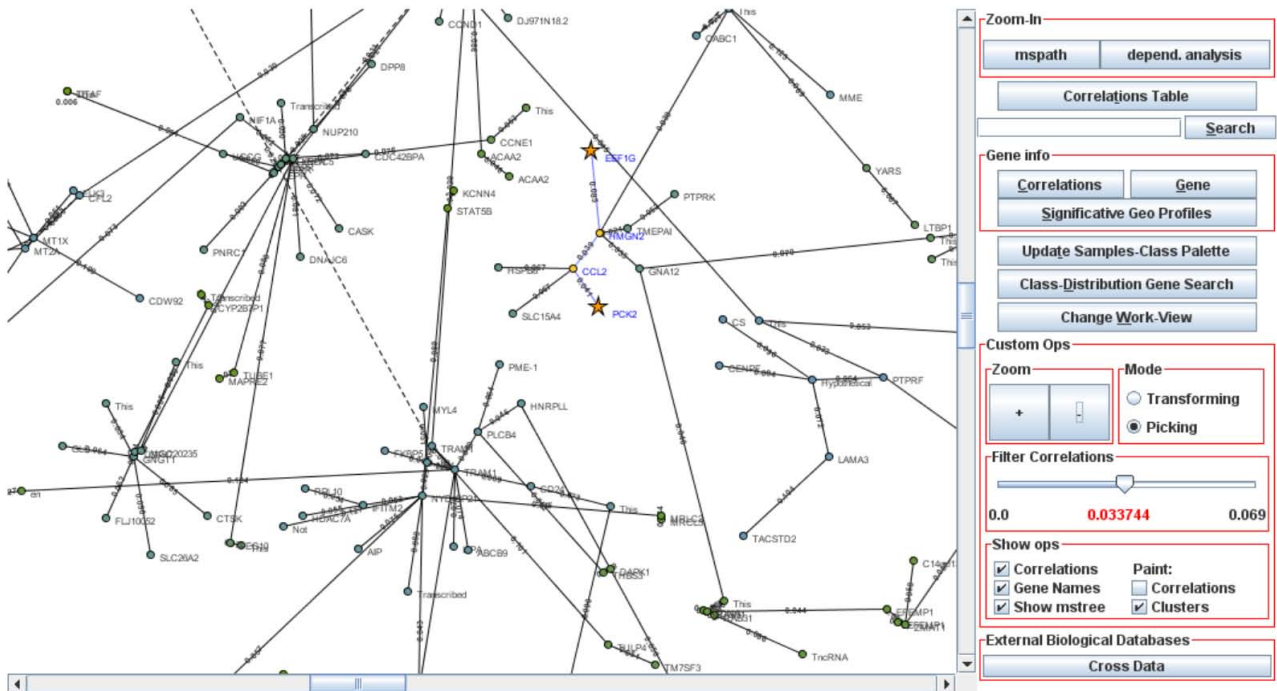
**Figure 14. The Detailed View (geometric plot) of a expression relationship.** The expressions of two genes are being compared in the display. Each axis represents the expression level of each gene, the data cloud represents the values of the expression-matrix experiments [Scherf U. et al 2000] for the compared genes, and the line represents the dependence between the gene expressions. The samples belonging to different sample classes previously defined are painted with different colours.

### 3.4.2. The gene network, gene clusters, minimum-spanning tree, and graph layout

The global-network view provides: The graph of continuous relationships (for all of the gene-expression range) among gene expressions and filtered by correlation, the minimum-spanning tree that links all matrix genes, the gene clusters with the nodes coloured by the average correlation degree among the cluster genes, the genes most correlated with a selected gene, and the correlation degree of the most correlated genes of a selected gene with the rest of the genes.

The Java JUNG libraries for analysis and visualization of network data by web [O'Madadhain J. et al 2005] have been used to mount the interactive graph of the global vision interface (Figure 15). On the graph layout, the genes are placed in the 2D space based on the correlation degree of each gene with its neighbours, grouping the genes in clusters and facilitating the showing of the minimum-spanning path among any set of any expression-matrix genes selected by the researcher. As it was previously described, the minimum-spanning path is necessary for the gene-selection process of the zoom-in operation.





**Figure 15. The Global-Network View.** Interactive gene network showing the expression-matrix-genes interdependence in expression terms. All of the operations of the PCOPGene-Net are launched from this interface.

### 3.4.3. Tool in use

The web application is composed of two main graphical interfaces: The global network view (Figure 15), and the detailed view of the expression relationship (Figures 13 and 14). The final web application provides four basic operations:

1. The gene-network view gives a global vision of the interdependence among the gene expressions of the analysed matrix.
2. At a given moment, the researcher wants to focus his/her research on a concrete process, or on apparently unconnected processes, to know what genes are involved and to study their expressions dependence in detail. This task is provided by the zoom-in operation.
3. However, a lot of genes are not continuously correlated for all of their expression ranges. To analyse these kinds of non-continuous relationships, the non-continuous analysis is provided.
4. There are many more links among genes besides the expression relationships. Remote biomedical data-bases will be consulted to search these gene links not based on gene expression.

The four operations are illustrated in Figure 16.

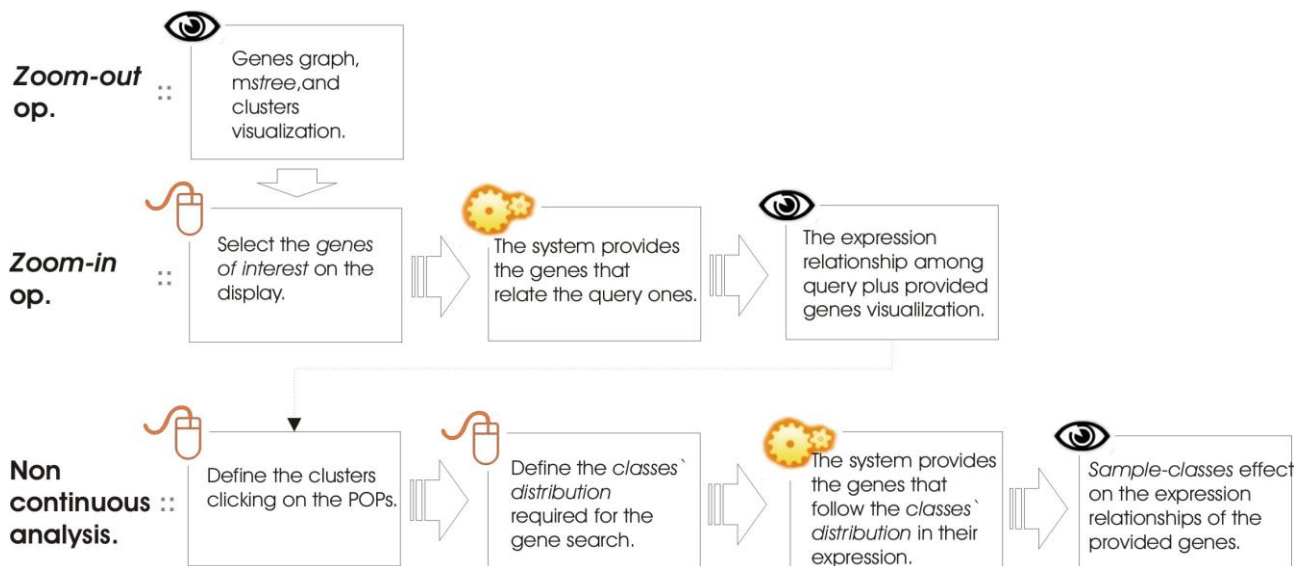


Figure 16. Basic-analysis procedure using the PCOPGene tools: Zoom-out operation, Zoom-in operation and the non-continuous analysis of gene expressions.

### 3.4.4. Biomedical-database remote access

The purpose of accessing remote biomedical databases is to supply complementary information beyond expression relationships. The expression analysis provides marker genes in expression terms. However, the researcher cannot know much more about these genes at first. By accessing the remote databases, the researcher can contextualise their results with new biological and medical information; for instance, observing their location in KEGG maps [Wixon J. and Kell D. 2000], their proteins' interaction [Alfarano C. et al 2005; Mishra GR. et al 2006; Breitkreutz BJ. et al 2008], PubMed [Motschall E. and Falck-Ytter Y. 2005] papers that talk about their connection, etc.

Furthermore, there are gene relationships that cannot be found in expression terms. Nevertheless, these relationships can be found accessing the remote biomedical databases (for instance, interaction at the protein level, activation by means of phosphorylation, proteolytic inactivation, etc.). In this way, the researcher can extend his/her analysis to new genes related with the current genes of interest, but not exclusively in expression terms. This gene search can also be oriented to search the genes linked to a biomedical-database topic of interest (for instance, a pubmed topic like glucocorticoid provides all the matrix genes that appear in papers related with glucocorticoids). This new gene search can be combined with the correlation analysis limiting the search to the genes of a gene-cluster or the most correlated genes of a query gene.

### 3.4.5. Assigning attributes to the sample classes

Our main objective in the use of the presented tools is the holistic study of cell behaviour taking advantage of the high-throughput potential of expression matrices with large sample series. The sample class's definition and the attributes assignment to sample classes are the key points to

achieve this. In this way, the phenotypes could be characterised, trying to associate them with a concrete class or subclass.

The attributes should give biological meaning to the sample classes, helping to identify and describe the cell state in its widest definition, which each class represents. The assignment of attributes comes from the four operations previously described. From the non-continuous analysis comes the definition of sample classes. From the zoom-in operation and non-continuous analysis comes the identification of marker genes of each sample class, and from the access to remote databases comes the attributes assigned to each class on the basis of the marker genes of each one. The procedure works as follows.

#### **3.4.6. Obtaining marker genes from continuous analysis**

The dependence relationship among gene expressions can be observed in continuous analysis (Figures 13 and 14). Each fluctuation of this dependence can be selected as a new sample class. Then, the compared genes become marker genes of the new sample class, giving new attributes to it. Note that the compared genes can be marker genes of apparently unconnected processes. This fact helps to describe the phenotype associated to the sample classes from a holistic point of view.

#### **3.4.7. Obtaining marker genes from non-continuous analysis**

The genes supplied by the search will be the marker genes for the sample classes used as search parameters. These marker genes will give new attributes to the sample classes.

#### **3.4.8. Defining sample subclasses and obtaining more marker genes**

When a subclass is defined from a sample class selecting only a part of the class samples, the subclass has the attributes of the original sample class plus the new and differentiated attributes. It is performed when a previously defined sample class is observed in a new expression relationship and the researcher wishes to define a new sample class from this relationship, while considering its intersection with the original sample classes (Figure 14). Thus, the genes compared in the new expression relationship will be marker genes of the new sample subclass but not of the original one.

#### **3.4.9. Obtaining attributes by accessing remote databases**

The attributes can come from two kinds of databases: Those about gene-expression and those not about gene-expression.

When searching a class marker gene in the GEO microarray datasets [Barrett T. et al 2005], the researcher obtains the microarrays for which this gene is also a marker gene. The attributes of these microarray sample series will be new attributes of the sample class.

Searching a class gene-marker in biomedical databases [Barrett T. et al 2005; Wixon J. and Kell D. 2000; Alfarano C. et al 2005; Mishra GR. et al 2006; Breitkreutz BJ. et al 2008; Motschall E. and Falck-Ytter Y. 2005; The Gene Ontology project 2008; Hamosh A. et al 2005; Wheeler DL. et al 2008; Wheeler DL. 2007], the researcher will obtain attributes that will give biological and medical significance to sample classes and help describe their phenotype. Notice that in this way the information will not be supplied for each marker gene separately but rather for all of the marker genes as a whole, defining the cell state of the sample class.

All of the operations are highly flexible and complementary using the results of one as an input of the next one, thus making the progressive analysis possible. The purpose of this is to obtain the sample classes and attributes that characterise these classes, and in a roundabout way, characterise the phenotype that these sample classes represent.

All the findings presented in this section have been included in the work [Huerta M. et al 2009].



### 3.5. The biological significance of the different expression-relationship curve types

Once the PCOP is calculated for the expression relationships between all the gene pairs, the non-linear expression relationships are detected and classified by curve type. The non-linearity and the type of curve of the relationship is calculated from the relative position of the POPs. Specifically the POPs that constitute the curvature points of the PCOP. Our system filters the expression relationships of each typology by its correlation degree. There are two correlation thresholds to filter the expression relationships, one which is more restrictive (demanding higher correlation) and the other one less restrictive. Working with the less restrictive threshold, the variations of the basic typologies are considered as the same typology (because being less correlated they have a higher variance). The correlation threshold is sensitive to the number of expression-matrix genes. A big number of genes imply a more restrictive threshold. The main reason is: The more genes implicated in the analysis, the more expression relationships with high correlation can be detected.

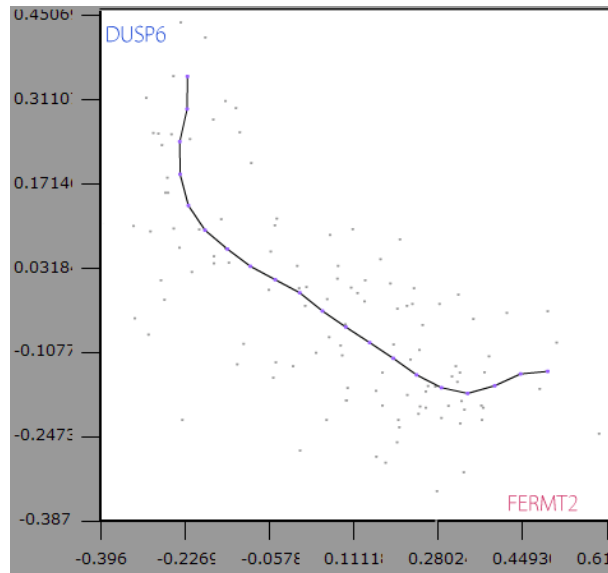
Next, the basic types of non-linear expression dependences are explained and their biological significance is illustrated with real examples. Each one of the shown topologies has the positive and negative variation. In most of the cases, both are explained. There are other typologies of non-linear expression relationships but basically they are variations of the already shown typologies.

#### **$y = e^x$ typology**

This is the expression-relationship typology followed by enhancer and trigger genes. One of the genes must be over-expressed to make it possible for the other gene to over-express.

#### **$y = e^{-x}$ typology**

The negative version of the logarithmic curve type implies an expression relationship which correlates two mutual excluding genes. One of the genes must be under-expressed to make it possible for the other gene to over-express.

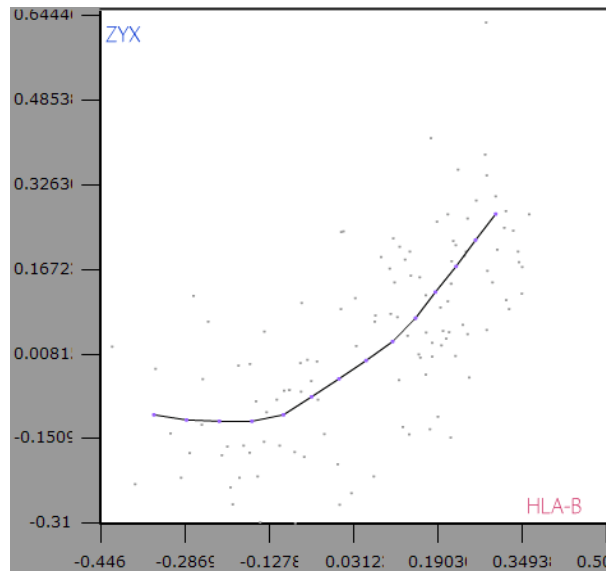


**Figure 17.** FERMT2(x axis) and DUSP6(y axis) expression relationship of the type  $y = e^{-x}$ . The different sample conditions of the expression matrix (the sample series) constitute the data cloud. The FERMT2 expression is linked to tumour proliferation, whereas DUSP6 stops tumour proliferation.

The DUSP6 and FERMT2 genes maintain this type of relationship between their expressions (Figure 17). The FERMT2 expression is linked to tumour proliferation, whereas DUSP6 stops the tumour proliferation. Thus, this non-linear relationship divides the sample space in two phenotypes, the cell-stress phenotype (which stops the tumour proliferation) and the tumour proliferation phenotype (which stops the cell-stress).

The HLA genes (HLA-A, HLA-B, HLA-H, plus RAB14) constitute a cluster of coexpressed genes. Furthermore, the genes of this cluster maintain expression relationships of type  $y = e^x$  and  $y = e^{-x}$  with other gene clusters. Thus, the HLA cluster maintains a non-linear expression dependence of activation or deactivation with respect to other clusters of coexpressed genes. The HLA genes mark the cell making the proliferation possible, the inflammation of the tissue and the activation of the immune system. Then, the HLA-genes activity facilitates the expression of the genes involved in inflammatory processes ( gene clusters linked by a  $y = e^x$  relationship with the HLA cluster) and prevents the expression of genes involved in antagonistic processes (gene clusters linked by a  $y = e^{-x}$  relationship with respect the HLA cluster).

For example, the HLA-B and ZYX genes maintain a relationship of type  $y = e^x$  between their expressions. The HLA genes are indicative of cell maturation marking the cells to be recognised by the immune system. ZYX participates in the tissue restructuring facilitating the cell adhesion during the tumour proliferation. Then, looking at their relationship in the figure 18, we can observe that it is necessary that the HLA genes mark the cell before the proliferation. As a result, these cell marks make the tumour destruction by the immune system difficult. Once the cell is marked and the proliferation starts, the ZYX expression level defines the level of cell adhesion and tissue restructuring of the tumour.



**Figure 18. HLA-B(x axis) and ZYX(y axis) expression relationship of the type  $y = e^x$ .** The different sample conditions of the expression matrix (the sample series) constitute the data cloud. HLA-B marks the cells to be recognised by the immune system. The ZYX participates in the tissue restructuring facilitating the cell adhesion.

### $y = -x^2$ typology

When this kind of expression relationship is maintained between two genes, the first gene over-expression involves a unique phenotype but its under-expression involves two different phenotypes, one for the second gene under-expressed and the other for the second gene over-expressed. This kind of relationship can be observed, for instance, when the expression of a gene is regulated by two hetero co-activators, and these co-activators are inversely coexpressed (the absence of any co-activator implies the under-expression of the gene). This type of expression relationship is useful to study contradictions in expression behaviour.

CCNE1 and TP53 maintain this kind of relationship in their expressions. The CCNE1 expression implies a high proliferation and TP53 expression implies a high control of mutations. As we have seen before and in the figure 19, at low levels of CCNE1 there are low levels of TP53, because a low proliferation does not need a high error control. When the CCNE1-expression increases, the mutations control needs to be increased too, but when the proliferation grows too much, like in tumour proliferation, the TP53-expression decreases again and mutation control falls.



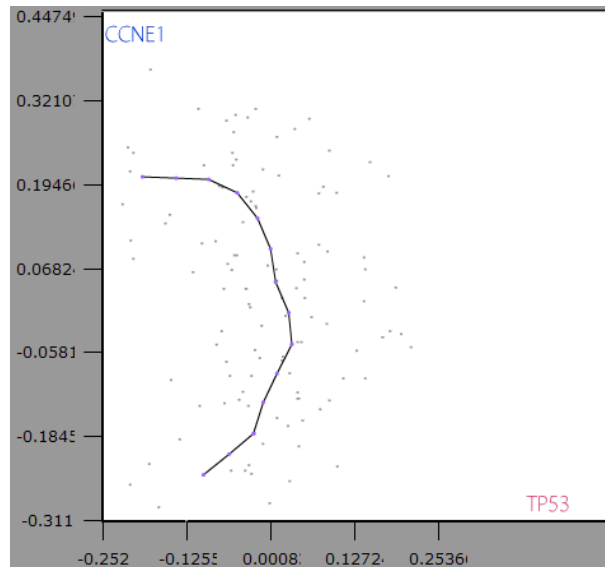


Figure 19. TP53(x axis) and CCNE1(y axis) expression relationship of the type  $y = -x^2$ . The different sample conditions of the microarray (the sample series) constitute the data cloud. The CCNE1 expression implies a high proliferation and TP53 expression implies a high control of mutations.

MAKP1/DUSP1 and FKBP5 maintain this  $y = -x^2$  typology in their expression relationship too (Figure 20). MAKP1/DUSP1 inhibit S10, facilitating the metilation and inhibiting "the exploration" (a high control of gene expression). FKBP5 has a key role in the histone liberation. The histone liberation is connected to function loss and "exploratory phases" of the tumourgenesis (gene expression out of control). Looking at the relationship between the two genes, we can see that at low levels of DUSP1, there are medium levels of FKBP5, being the exploratory phase low too. At high levels of DUSP1, the expression levels of FKBP5 are very low too, being the exploratory phase totally inhibited. Therefore, exploratory phase is really done then when FKBP5 raise its maximum expression, being DUSP1 at basal values at this point.

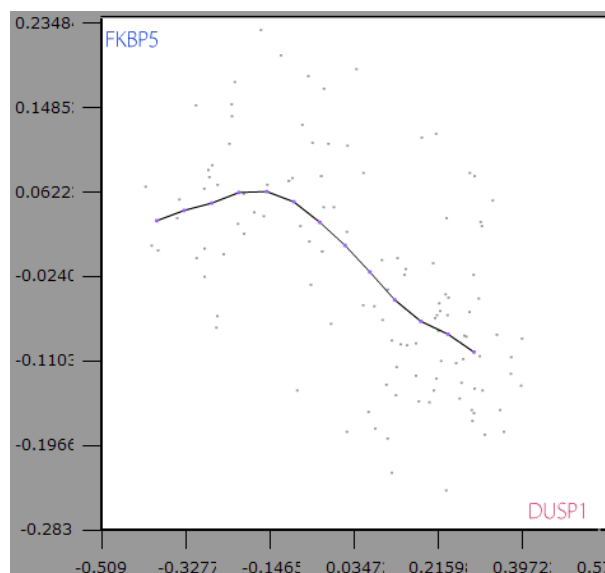
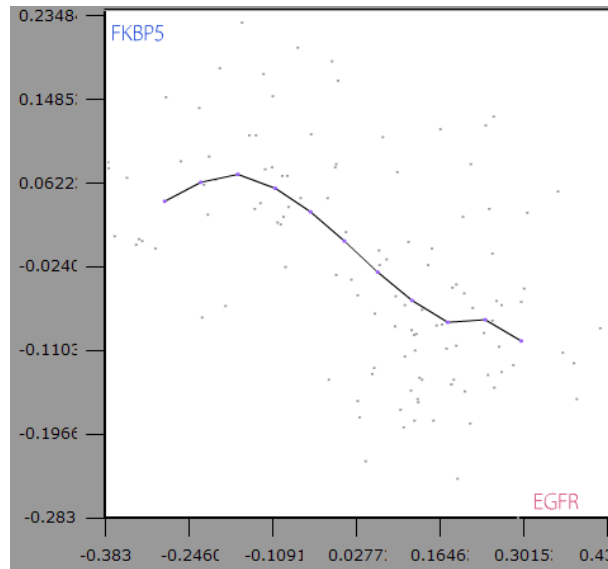


Figure 20. DUSP1(x axis) and FKBP5(y axis) expression relationship of the type  $y = -x^2$ . The different sample conditions of the expression matrix (the sample series) constitute the data cloud. MAKP1/DUSP1 facilitates the metilation. FKBP5 facilitates the histone liberation.



**Figure 21. EGFR(x axis) and FKBP5(y axis) expression relationship. It shows a  $y = -x^2$  curve type.** The different sample conditions of the expression matrix (the sample series) constitute the data cloud. EGFR is linked to tumour proliferation. FKBP5 is linked to tumour "exploratory phases".

EGFR maintains with FKBP5 (Figure 21) the same curve type in their relationship than with MAKP1/DUSP1. As EGFR is linked with tumour proliferation, this relationship can help us to understand the two FKBP5 under-expressed phenotypes. Looking at the EGFR and FKBP5 relationship, we can distinguish two phenotypes for low levels of FKBP5: One with high proliferation and other with low proliferation. It implies that there is no gene exploration in the proliferation phenotype and there is exploration only in part of the cell-stress phenotype, possibly in the more acute stress phases because FKBP5 over-expression is a marker of high cell-stress level.

These contradictions are present in the bibliography of FKBP5, for this reason, the detection of this curve type is so interesting when trying to understand these paradoxes. This curve type allows us to differentiate different phenotypes were the gene is equally expressed. FKBP5 is expressed during the tumour-growth phases in cancers like colon cancer, but under-expressed during the tumour growth of other cancers like colorectal cancer. Thus, if colorectal cancers achieve the "exploratory phase" during a growth phase, it stops the proliferation. But if colon cancers achieve the exploratory phase, it promotes the proliferation, maybe no immediately but near in the future. As shown later, the possible reason of this dual behaviour is that colon cancer combines the cell stress and its "exploratory phase" with the tumour proliferation, in the same tissue, whereas colorectal cancers need to separate these two phases because the rectus is a structurally more complex tissue. As we will see later, there are two clearly differentiated tumour-proliferation phenotypes.

### **$y = x^2$ typology**

In the inverse version of the previous typology, it shows the two phenotypes in the first gene over-expression instead of its under-expression. Now the gene could be regulated by two hetero co-inhibitors (instead of two co-activators), which are inversely coexpressed. This type of expression relationship is also useful to clarify contradictions in expression behaviour.

### $y = x^3$ typology

This is the typology of the expression-relationship between two genes basically coexpressed but with a non-linear coexpression. The fluctuations in the expression dependence indicate the pass from one phenotype to another. It can be more clearly observed when sample clusters obtained by means of clustering methods are coloured over the relationships (this procedure will be described in detail in the next sections). This curve type could also imply a switch expression dependence, but in contrast to the  $y = e^x$  relationship, now the gene starts its over-expression when the other gene achieves a certain level of expression but not its over-expression. The figure 13 illustrates this type of relationship between the gene ENO2 and the genes MAP1B, GJB1, ikBe, and RAB31. As was commented before, the fluctuation points out the phenotype change.

### $y^2 + x^2 = 1$ typology

In this case, the expression relationship covers the four combinations of the expression increase and decrease of both genes. This typology implies a double relationship, usually of type  $y = x^2$  and  $y = -x^2$  or  $y = e^x$  and  $y = e^{-x}$ . Detecting this curve type we can study these dual behaviours of the gene relationships (and solve paradoxes if this duality is unknown).

CITED2 and NEDD4L maintain this kind of relationship between their expressions (Figure 22). CITED2 expression indicates a proliferation increase. NEDD4L is related with ion-channels inhibition turning the cell refractory to activity signals. Accordingly, this relationship implies: First a phenotype with low proliferation and ion channels active and second, two different ways for proliferation increase: One with active ion channels (and  $y = -x^2$  typology) and the other with inhibited ion channels (and  $y = x^2$  typology).

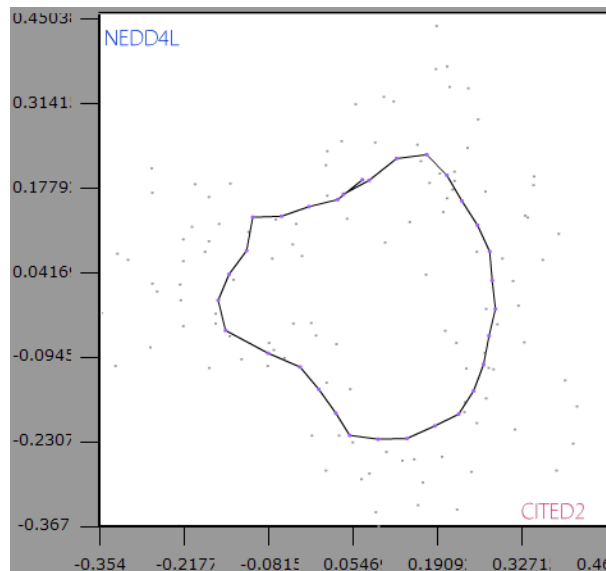


Figure 22. CITED2(x axis) and NEDD4L(y axis) expression relationship. It shows a  $y^2 + x^2 = 1$  curve type. The different sample conditions of the expression matrix (the sample series) constitute the data cloud. The CITED2 expression indicates a proliferation increase. NEDD4L is related with turning the cell refractory to activity signals.

As we can see, this expression-dependence type helps us to well discriminate phenotypes because the relationship covers all expression combinations between the two genes. In practice, each relationship of this typology has its own interpretation.

All the findings presented in this section have been included in the work [Huerta M. et al 2014].

There are other kinds of expression-relationship typologies detectable by the system, but they are in most cases variations of the mentioned ones. The web-tool to study the curve types is available at: <http://platypus.uab.es/GCinC>. The aforementioned expression data are available for the shared user. You can create your own user and make the shared-user data available for your user. All the previously-analysed expression relationships are available in our server for the Shared User in the directory of the `at_matrix` expression matrix named "Examples of different non-linear expression-relationship typologies".



### 3.6. Non-linear expression relationships between sets of coexpressed genes

Organisms have evolved to vary internal and external cell environments by carefully controlling the abundance and activity of these proteins to suit their conditions. To simplify this task, genes whose products function together are often under common regulatory control. This regulatory control is such that these genes are co-ordinately expressed under the appropriate conditions. The experimental observation that a set of genes is coexpressed frequently implies that the genes share a biological function and are under common regulatory control [Eisen MB. et al 1998]. These regulators that govern the expression of sets of coexpressed genes that carry out the appropriate cell functions are also regulated and synchronized among them. Nevertheless, the regulation and synchronization among the regulatory mechanisms is much more complex. The regulatory mechanisms are not directly regulated by the other regulatory mechanisms, but by the coexpressed genes product of their activation cascade. In this way, a regulatory mechanism indirectly regulates a different regulatory mechanism by means its activation cascade. These coexpressed genes switch from the inhibition to the allowance of the regulatory mechanism depending on whether they reach or lose certain expression levels. Furthermore, these regulatory mechanisms are multiregulated. The final activation or deactivation of a regulatory process will depend on internal and external factors to the cell, not only on the expression level of a set of coexpressed genes. For this reason, to study the synchronization and regulation among the regulatory mechanisms based on gene expression is really difficult. Furthermore, many proteins have multiple roles in the cell, and act with distinct sets of cooperating proteins to fulfil each role. The genes that synthesise these proteins are therefore coexpressed with different sets of genes, each one governed by a distinct regulatory mechanism. The experimental conditions will determine the regulatory mechanism activated in each condition, and thus, the set of coexpressed genes that will be activated. An increased number of different experimental conditions for the same genes will provide: less genes in each set of coexpressed genes, more different sets, and higher alternation of the activation and deactivation of the coexpressed-genes. But irrespective of the conditions, how can we study the effect of their activation and deactivation on the rest of the sets of coexpressed genes?

Microarray technology, as well as the new techniques of Next Generation Sequencing (NGS), allows us to obtain large size gene-expression matrices [Barret T. et al 2013]. A usual analysis of these data is the use of clustering methods to obtain the sets of coexpressed genes. It is of vital importance to detect these clusters of coexpressed genes, among other reasons because as mentioned before, these clusters of coexpressed genes carry out the different cellular functions [Eisen MB. et al 1998].

There are powerful coexpression-analysis tools for this purpose [Van Dam S. et al 2012; Stuart JM. et al 2003]. As expression arrays allow simultaneous analyses of thousands of genes, we can study genes responsible for very diverse cellular functions. Therefore, it makes it easier for the researcher to understand the cellular phenotype in the performed experiments from a holistic point of view. That is, involving the largest number of cellular processes possible. Without this holistic point of

view, it is very difficult to deal with the multiple functions, phenotypes or states of the living beings, in which a large amount of genes are collaborating. This holistic point of view can be useful to characterize phenotypes previously unknown, for instance, the description of the “fish fever” in zebrafish [Boltana S. et al 2013]. Nevertheless, even though clustering methods allow us to obtain coexpressed genes and thus differentiate diverse cellular processes, clustering methods explain very little of the expression relationships between the different sets of coexpressed genes. As a consequence, the researcher is constrained to study each one of the coexpressed-gene sets individually, losing much of the potential that technologies for obtaining gene-expression matrices offer.

Current statistical technologies allow us to study inclusion relationships between clusters of coexpressed genes. That is, to study which clusters of coexpressed genes would be more correlated and which others would be more uncorrelated [Thalamuthu A. et al 2016]. But they do not go much beyond that. The main obstacle to the tools that attempt to study the regulation of coexpressed genes is that the low number of copies of the regulatory genes impedes the correct capture of their expression by technologies obtaining gene expression. Furthermore, the coexpression of genes is strongly linked to the chromatin structure, especially in multicellular organisms. This chromatin structure depends on a complex network of cellular signalling and post-transcriptional modifications of proteins (phosphorylation, acetylation, ubiquitination,...). So, even having detected the regulatory genes, we would have an incomplete puzzle. All of this makes it enormously difficult for the tools to be able to obtain information about the regulation among the sets of coexpressed genes and the processes they carry out. So, without these regulatory elements, how can we describe a regulatory network among the processes performed by the sets of coexpressed genes? Well then, it could be based on the dependences between the gene expressions product of this complex regulation. Furthermore, these final genes present expression ranges wide enough so that the fluctuations of their expression dependences can be analysed.

With the developed tools, we expect to detect the complex expression dependences between the different sets of coexpressed genes, synthesise them and make them easier for the researcher to interpret. With this purpose we will provide the researcher with networks that show the expression dependences between all the coexpressed-gene sets of the network. In this way, the researcher will be able to study the alternation or synchronism among all these sets of coexpressed genes.

Our methodology is based upon the following three principles: First, the interdependence between sets of coexpressed genes cannot be described by linear expression relationships. Second, if two genes maintain an expression relationship with a certain type of curve, the genes coexpressed with these two genes maintain expression relationships of the same type between them. Third, the curve type of the inter-group expression relationships will describe the dependence of activation and deactivation between these sets of coexpressed genes. Thus, the strategy proposed here is focused on the detection of non-linear expression relationships between sets of coexpressed genes.

Activation and deactivation dependences between sets of coexpressed genes can be very complex. Some of the most common ones are: those in which a set of coexpressed genes acts as a trigger of another set of coexpressed genes; the case of antagonist processes, where the coexpressed-gene set that carries out each process needs to be totally deactivated so the other set can express; or sets of coexpressed genes that activate or deactivate another set of coexpressed genes when losing their

basal values of expression. In any case, the system does not anticipate any type of expression relationship. Since the system is able to recognize curves of very different shapes, it can process unknown activation and deactivation relationships as reliably as when processing the best known relationships.

The ultimate goal of our approach is that researchers are able to know the networks of processes hidden in their experimental data, as well as the activation and deactivation relationships between all of these processes. Furthermore, if the researcher is particularly interested in specific genes, the system will allow him/her to study the way the expression of a gene activates and deactivates different processes.

### **3.6.1. Genes coexpressed with a pair of genes with a non linear expression relationship between them**

All the non-linear expression relationships are detected from the expression array. These non-linear expression relationships are classified by the type of curve. The curvature points of the PCOP are used to identify and classify the non-linear expression relationships. Curvature points are those POPs in the PCOP in which a change in slope occurs. The detection of curvature points in expression relationships identifies the non-linear expression relationships. The type of curve is described by the function of the curve:  $y = e^x$ ,  $y = -e^x$ ,  $y = -e^{-x}$ ,  $y = x^2$ ,  $y = -x^2$ ,  $y = x^3$ ,  $I = x^2 + y^2$ , ... The genes coexpressed with each gene are also detected (none curvature points are detected in the PCOP of two coexpressed genes). This will allow us to study the non-linear expression relationships between a gene of interest and different sets of coexpressed genes.

The correlation degree provided by the PCOP calculation is what guarantees us that the linear expression relationships (coexpressed genes) as well as the non-linear expression relationships (inter-group expression relationships) are not a product of chance and have a biological meaning. For this reason, we require a high correlation degree for the linear expression relationships as well as for the non-linear expression relationships. We are also restrictive in the classification of the expression relationships as linear expression relationships, and the consequent consideration of two genes as coexpressed genes. Even a small curvature in the relationship of two coexpressed genes can cause a diversity in the typology of the expression relationships of these two genes with the genes of another set of coexpressed genes. More concretely, being A and B two coexpressed genes whose expression relationship has a small curvature, and being C a set of coexpressed genes that maintain non-linear expression relationships with A and B. The expression relationships of gene A with set C may describe a different typology with respect to the expression relationships of gene B with set C.

### **3.6.2. Cliques of non-linear expression relationships between genes**

A clique in an undirected graph is a subset of its vertices such that every two vertices in the subset are connected by an edge. If we consider a graph of all the non-linear expression relationships with



a high correlation, we obtain its cliques. The genes of a clique must be at least three and they must maintain non-linear expression relationships between all the genes of the clique. These cliques will not yet relate sets of coexpressed genes, but rather, genes individually. Nevertheless, as the cliques are relating several genes in a network of non-linear expression relationships, the cliques will be the seed to relate the sets of coexpressed genes between all of them.

### **3.6.3. Pairs of isomorphic and linear Cliques of non linear expression relationships**

Once the cliques are detected, they are grouped in clique pairs by relating genes that belong to the same set of coexpressed genes.

The cliques that will form each clique pair will meet two conditions:

- Each one of the genes of a clique will be coexpressed with a different gene of the other clique, forming pairs of coexpressed genes.
- The type of curve that relates each two pairs of coexpressed genes will be the same in both cliques of the clique pair.

This provides us with pairs of cliques. Each gene of a clique will be coexpressed with a different gene of the other clique forming pairs of coexpressed genes. Then, the two expression relationships that relate the two coexpressed genes with different pairs of coexpressed genes will maintain the same type of curve for the two genes of the pair.

### **3.6.4. Cliques of isomorphic and linear cliques of non-linear expression relationships between genes**

Previously, we obtained the non-linear expression relationships between pairs of coexpressed genes. Now, we obtain sets of coexpressed genes that maintain non-linear expression relationships between all of the sets by means of grouping these pairs of coexpressed genes non-linearly related, into sets. If we consider a graph where: the vertices are the cliques of non-linear relationships between genes, and the edges link linear isomorphic cliques, now we will calculate the cliques of this new graph obtaining the cliques of cliques. Thereby we obtain the skeleton of the sets of coexpressed genes and the non-linear expression relationships between these skeletons. The different networks among sets of coexpressed genes are constituted from these relationships between the skeletons of the sets of coexpressed genes.

The genes of the skeleton of a set of coexpressed genes are those genes of the set that maintain a highly-correlated non-linear expression relationship with the skeleton of the other sets of the network. The genes of the set of coexpressed genes that are not part of the skeleton will be coexpressed with the genes of the skeleton. These genes coexpressed with the skeleton will maintain the same type of non-linear expression relationship with the other sets of coexpressed

genes as the genes of the skeleton. In this way, the genes of the gene set are coexpressed among them and their expression relationships with the genes of another gene set maintain the same curve type. Deformations of the  $y=x$  relationship between the genes of the gene set will produce a distortion of the expected curve between gene sets. A higher variance in the coexpression with respect to the skeleton genes will also imply a higher variability in the expected type of curve.

The higher the number of genes of the skeleton, the more representative the process carried out by the coexpressed-gene set. Thanks to the second condition of the linear-isomorphic-cliques definition we can make sure that the relationships between the genes of the skeleton of different sets of coexpressed genes maintain the same type of curve for all the genes of the skeleton. As aforementioned, it can be extended to all the genes of the gene set.

The correlation degree needed to consider an expression relationship coexpressed enough will depend on the number of genes of the expression array. It is useful for small expression matrices, because non-linear expression relationships between sets of coexpressed genes can be detected, although these expression relationships have high entropy. The aim is to always detect enough non-linear expression relationships to be able to find the skeletons that relate the sets of coexpressed genes.

The threshold to consider an expression relationship as linear or non-linear will also depend on the number of genes, (this threshold) being more restrictive for the linear relationships in matrices with less genes. Thus, large sets of coexpressed genes with very sharp curves between them will be formed for large expression matrices, whereas smaller sets of coexpressed genes, as well as more subtle non-linear expression relationships between the sets, will be considered for small matrices.

The expression relationships have been filtered by the uncorrelation factor provided by the PCOP calculation to be considered correlated enough [Delicado P. and Huerta M. 2003]. The threshold formula is:

$$0.12 \times \frac{1600}{num\ genes} - \left( \frac{num\ genes}{40000} \right)^{18}$$

The threshold formula for the curvature to consider whether the relationships are linear or non-linear is:

$$160 - \left( \frac{\frac{15.0}{20000} + \frac{14.0}{18400}}{2 \times num\ genes} \right)$$

However, a formula that depends only on the number of genes is not enough to guarantee the quality of the analysis extracted from the data. Because the data come from very different experiments of different nature. The diverse nature of the experiments is a qualitative variable that cannot be quantified with a formula. For this reason the correlation threshold calculation uses an on-line correction: From the expression relationships already analysed and the number of relationships

pending to be analysed, the system makes an estimation of the number of relationships that would finally pass the threshold. From this estimation, the system automatically modifies the threshold.

A higher number of genes in the expression array increase the number of coexpressed genes and non-linear expression relationships, which facilitates finding skeletons. But in any case, the number of expression relationships with high correlation, as well as the number of linear expression relationships with respect to the non-linear ones, will always depend on the nature of the experiments of the sample series.

### 3.6.5. Tool's output interfaces

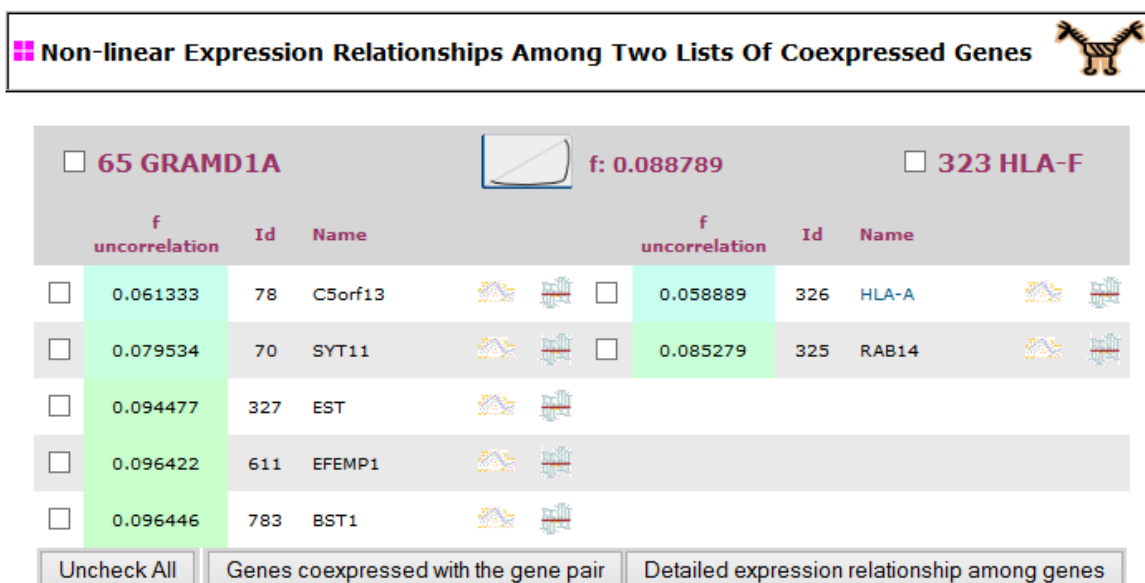
The system allows us to study sets of coexpressed genes that maintain non-linear expression relationships among them, as well as to study the non-linear expression relationships that a concrete gene of interest maintains with different sets of coexpressed genes. There are 4573 non-linear expression relationships that have been found, plus 20269 pairs of coexpressed genes (all highly correlated) from the expression matrix of 1416 genes used in the examples (*at\_matrix*).

#### 3.6.5.1. Studying the complex expression relationships between a target gene and sets of coexpressed genes

The study of the expression relationships between sets of coexpressed genes can start from the researcher's gene of interest. All the non-linear expression relationships that a gene of interest maintains with different sets of coexpressed genes will be shown. These relationships will be shown classified by curve type, because each curve type implies a different activation/deactivation relationship. Only the non-linear expression relationships that maintain a sufficient correlation degree will be shown.

We will study the activation and deactivation relationship between our gene of interest and different sets of coexpressed genes starting from these high-correlated non-linear expression relationships. Two lists of coexpressed genes will be shown in a new view for each highly-correlated non-linear expression relationship of the gene of interest (Figure 23) when clicking on them. The first list will show the genes coexpressed with the gene of interest. The second one will show the genes coexpressed with the gene that maintains the highly-correlated expression relationship with the gene of interest. In this way, the user can study the expression relationships between the two sets of coexpressed genes. The first list of coexpressed genes is ordered by their correlation degree with the gene of interest, the second list is ordered by their correlation degree with respect to the gene non-linearly related to the gene of interest. The user can select genes from both lists of coexpressed genes to study their expression relationship in detail using the *expression-relationship detailed view* [Cedano J. et al 2008; Huerta M. et al 2009] (Figure 24).

An icon shows the type of non-linear relationship between the two main genes and, by extension, between the two sets of coexpressed genes. The curve type is very important, since it determines the role of the genes in each expression dependence.



**Figure 23. Non-linear Expression relationships between coexpressed gene sets.** This view shows at the top, a non-linear expression relationship where a researcher's gene of interest participates. The gene of interest is displayed on the top of the view on the left side. The column on the left side displays the genes coexpressed with the gene of interest, while the column on the right displays a set of coexpressed genes that maintain a non-linear expression relationship with the gene of interest. The coexpressed genes are ordered by their correlation degree with their respective gene at the top (the f value obtained by the PCOP calculation). The icon shows the curve type of the non-linear expression relationship. Each curve type implies different expression dependence: mutual exclusion, trigger, double trigger... All the expression relationships relating genes from the two sets of coexpressed genes should be of the type shown by the icon. By selecting genes from the two sets, their expression relationship can be studied in detail in a new interface (Figure 24).

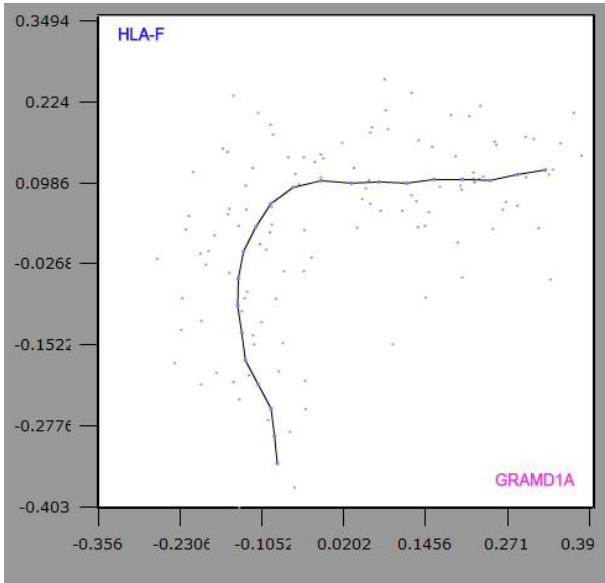
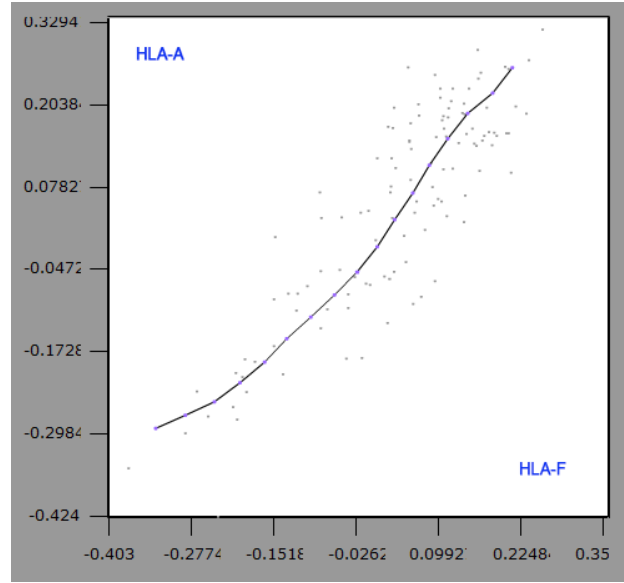
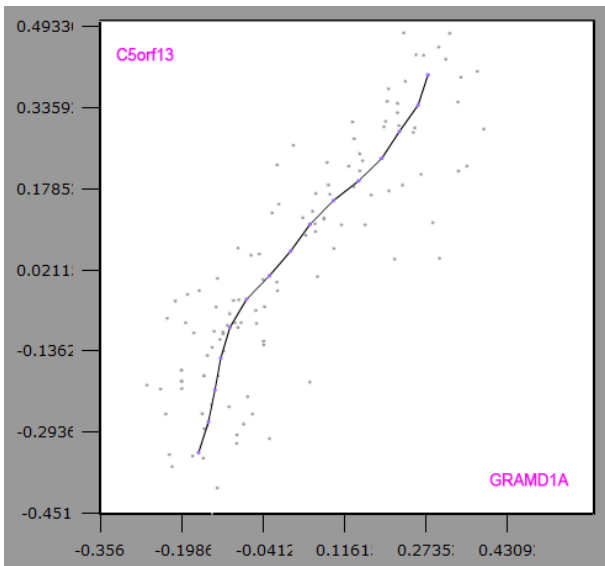
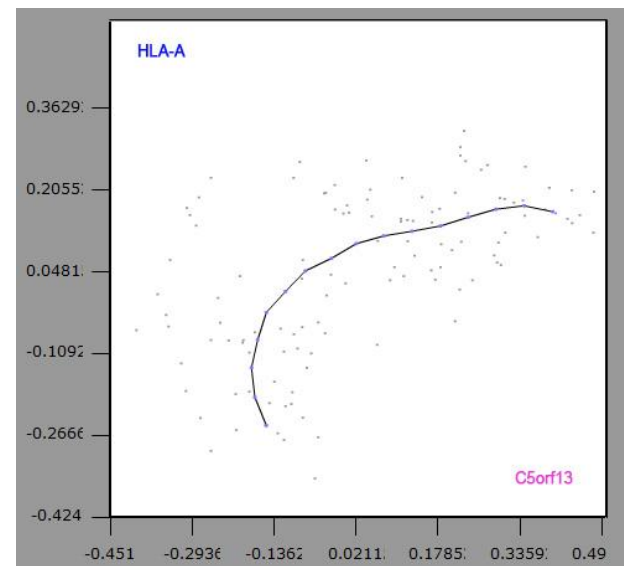
### 3.6.5.2. The curve type indicates the type of activation and deactivation relationship between sets of coexpressed genes

The system obtains the inner pattern of the expression relationship for any type of curve and classifies it. The only requirement is that the data cloud must be continuous. Remembering the meaning of each curve type,  $y = e^x$  relationship will provide a switch activation relationship between a set of coexpressed genes and the other set. In other words, the first set of coexpressed genes must over-express, so that the second set starts to express. A  $y = e^{-x}$  relationship indicates a mutual-exclusion dependence. That is, one of the two sets of genes must be deactivated so that the other set of coexpressed genes expresses. Note that these types of relationships are different from the positive and inverse coexpression relationships. This difference is precisely what allows us to detect different sets of coexpressed genes as well as the complex expression dependences between them.

A  $y=x^2$  relationship, would indicate a deactivation of the second set of coexpressed genes by the over-expression as well as the under-expression of the first set. A  $y=-x^2$  relationship would indicate an activation of the second set of genes, for the over-expression as well as the under-expression of the first set. Whereas in the relationships of type  $|e^x|$ , the over-expression of the set of coexpressed genes affects the other set. In the relationships of type  $|x^2|$ , the over-expression as well as the under-expression of the genes has an inhibitory or activatory effect on the other set of coexpressed genes.

Other relationships, such as those of type  $y = x^3$  or  $I=x^2+y^2$ , will indicate other complex expression dependences between different sets of coexpressed genes.

As pointed out in the introduction, one of the principles of our analysis is: The type of curve between two genes is also maintained between the genes coexpressed with each gene of the pair. Let's see an example: HLA genes are indicative of cell maturation marking the cell so it is recognised by the immune system [Parkes M. et al 2013]. HLA genes mark the cell making possible the inflammation of the tissue and the activation of the immune system. GRAMD1A is a not well known membrane receptor that inhibits programmed cell death and it is linked to disease resistance [Parkes M. et al 2013]. HLA-F and GRAMD1A maintain a non-linear expression relationship of type  $y=e^x$  (Figure 24.a). This points out that, possibly, the function associated with GRAMD1A can only be performed once the cell is marked by HLA genes. HLA-F and HLA-A are coexpressed genes (Figure 24.b), and GRAMD1A is coexpressed with NREP (Figure 24.c). Thus, HLA-A and NREP will also maintain a non-linear expression relationship of type  $y=e^x$  (Figure 24.d). C5orf13 (NREP) expression is linked to hypertrophic scar [Tan J. et al 2010]. This points out that hypertrophic scar, and the function associated with NREP, can only be performed once the cell is marked by HLA genes [McCarty SM. et al 2010]. Even though the relationship of these genes with hypertrophic scar was already known [Tan J. et al 2010; McCarty SM. et al 2010], there was no knowledge about how it was regulated.

**24.a)****24.b)****24.c)****24.d)**

**Figure 24.** Two sets of coexpressed genes will maintain non-linear relationships of the same type between the two sets. Four expression relationships are shown. The sample conditions of the expression matrix (the sample series) constitute the data cloud. The PCOP describes the expression-relationship inner pattern. The second and third plots (b,c) show coexpressed genes. HLA-A and HLA-F are coexpressed genes (b), and GRAMD1A and NREP(C5orf13) are also coexpressed genes(c). The first and last plots (a,d) show non-linear expression relationships of  $y=e^x$  type, a switch-activation relationship. Since HLA-F and GRAMD1A maintain a non-linear expression relationship of type  $y=e^x$  (a), and HLA-F is coexpressed with HLA-A (b), and GRAMD1A is coexpressed with NREP (c), therefore HLA-A and NREP (C5orf13) maintain a non-linear expression relationship of the same type (d). HLA-A and GRAMD1A would also maintain a non-linear relationship of type  $y=e^x$ , and HLA-F and NREP would maintain a non-linear relationship of the same type. This is the key-point of our approach: all the non-linear expression relationships that relate genes from two sets of coexpressed genes will maintain the same type of curve.

Hypertrophic scarring (HS) is a result of increased fibrogenesis, which is thought to be caused by an exaggerated inflammatory response [Lawrence JW. et al 2012; Van Loey NE. and Van Son MJ. 2003; Liu S. et al 2014]. There is a clear association between specific HLA alleles and cutaneous fibrosis. Specific examples of cutaneous fibrosis include hypertrophic scars (HS) among others [McCarty SM. et al 2010].

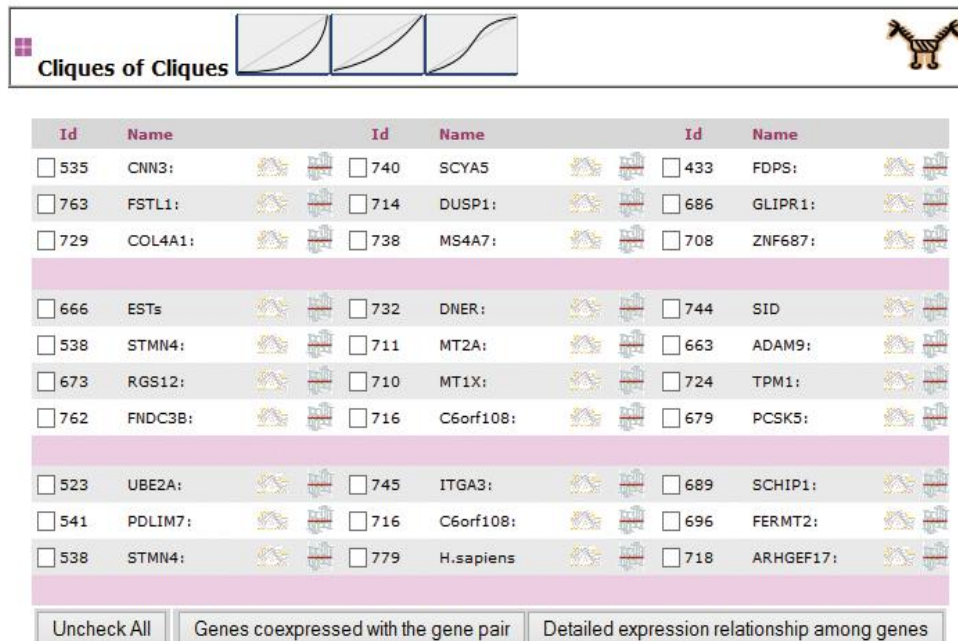
The relation of HLA and hypertrophic scar is already documented, but using our tool, we found that this relation is mediated by NREP, because HLA genes must be over-expressed to activate NREP (a gene directly linked to HS [Tan J. et al 2010]).

In this way, it is valuable that even though the technologies to obtain gene-expression arrays do not capture regulatory genes because of the low variability in their gene expression, these technologies do allow studying the regulation between processes through the genes that perform these processes (coexpressed genes that result from the activation cascade started by regulatory genes). This is because these final genes do maintain wide enough expression ranges, which allows our high-throughput tools to analyse the expression dependence between the sets of coexpressed genes.

### 3.6.5.3. Studying the complex expression relationships between sets of coexpressed genes

The different networks of non-linear expression relationships between sets of coexpressed genes are classified by the number of sets related and the curve types of the relationships between the sets. The number of sets related and the curve types between the sets will define the network type. Once a network type is selected in the main view, the networks that maintain this pattern in their inter-sets expression relationships are displayed.

In the view that shows the networks of a network type (Figure 25), the genes that belong to the skeleton of each set of coexpressed genes are displayed in a different column. That is, each column displays the genes of the skeleton of a coexpressed-gene set related by the network. By selecting genes from the different skeletons, the expression relationships among them can be studied in detail (Figure 24). Starting from the genes of the skeleton, the expression relationships between the rest of coexpressed genes of the different sets can also be studied. By selecting one gene from two different skeletons (columns), the genes coexpressed with each one of the two genes will be shown (in the way shown in Figure 23). The genes of the two sets should maintain the same type of non-linear expression relationship between them. These listed genes can also be selected to study their non-linear expression relationships in detail (Figure 24).



**Figure 25.** This view shows networks of concrete types of non-linear expression relationships between sets of coexpressed genes. The icons at the top show the curve-type pattern of the networks listed. The networks will always form a complete graph. The pink line separates the networks found for the curve-type pattern. The columns contain the genes of the skeleton of each set of coexpressed genes. The genes of the different skeletons can be selected to study their expression relationship in detail [Cedano J. et al 2008; Huerta M. et al 2009] (Figure 24). The genes of the different skeletons can be selected to study the expression relationships between the rest of coexpressed genes of the two sets, opening the view of Figure 23 for the two skeleton genes.

In this way, it can be studied whether the genes coexpressed with the skeleton maintain the type of curve, or whether it is distorted or lost. In the new window, the genes coexpressed with the gene of each skeleton appear ordered by their correlation degree with the gene of the skeleton. The higher correlation between a coexpressed gene and the gene of its skeleton, the lower variations of the curve types between this gene and the genes of the other set with respect to the curve type between the genes of both skeletons.

In Figure 24 we can see the key-point of our approach: all the non-linear expression relationships that relate genes from two different sets of coexpressed genes will describe the same type of curve. Since each gene set carries out a different cellular process, using our tool we can describe the relationships between these independent cellular processes.

In summary, to respond to diverse and frequently changing conditions, cells must precisely mediate the synthesis and function of the proteins in the cell. This is controlled in part by the overall genomic expression program that results from the combined action of different regulatory factors, each of which responds to specific extra- and intra-cellular signals. These regulators govern the expression of sets of coexpressed genes that perform the appropriate cell functions. The variations in the expression of these coexpressed genes can be captured by high-throughput technologies to obtain gene-expression arrays. In this way, the researcher is able to know which processes are carried out in the conditions he/she wishes to study, by knowing the different genes coexpressed in



them. But, what if the researcher wishes to know which relations have those different processes between them? How do we know how these processes are activating or deactivating and activating again among them? If the researcher suspects that certain target genes can be a therapeutic target, how can he/she know the effect of their expression on the rest of the processes, which this target gene does not belong, since it expresses with a different set of coexpressed genes? To know this could have several implications, from discovering unknown side effects to finding new ways to manipulate the expression of this gene.

As we present in the introduction, the expression dependences between sets of coexpressed genes, as well as the expression dependences between the processes which these sets of coexpressed genes carry out, would never be linear. This is why new tools like the presented one are necessary.

The methodologies and tools presented in this section have been included in the work [Huerta et al 2014].

**Availability:** expression relationships between coexpressed gene sets web-tool:  
<http://platypus.uab.es/nlnet>

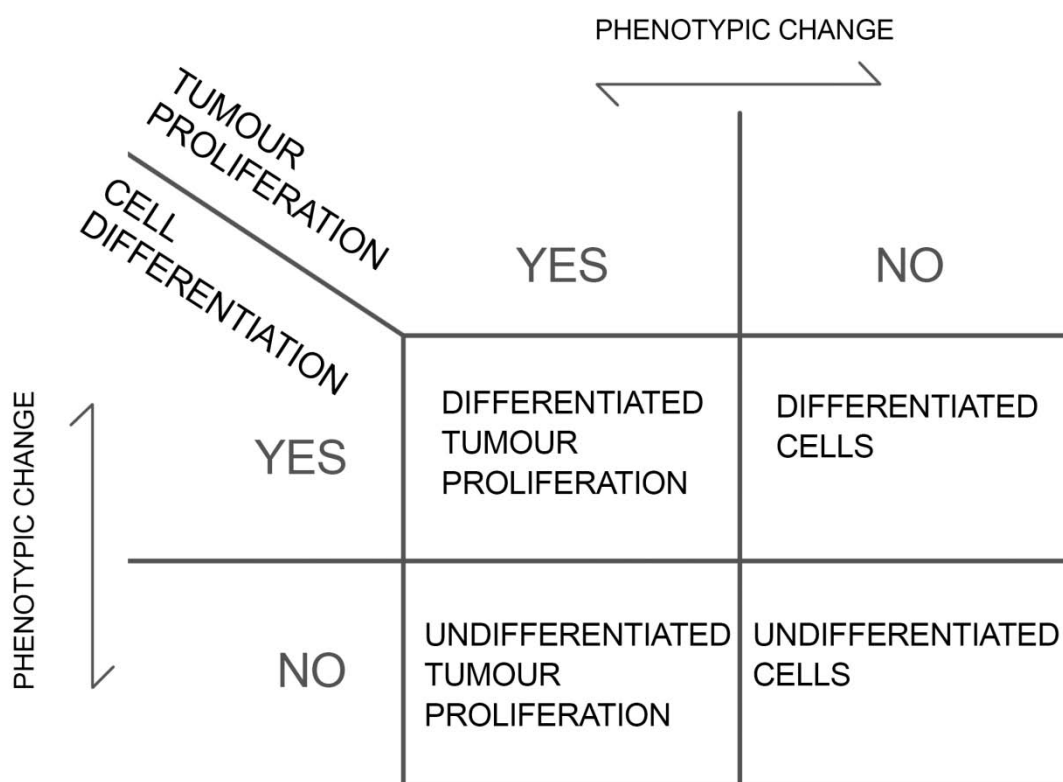
### 3.7. Deconstructing sample clusters in multiple concurrent phenotypic changes

Large sample series allow us to study expression relationships to detect fluctuations in these relationships and allow us to cluster samples based on the tendency of gene-expression dependences, as is shown in previous sections and papers [Cedano J. et al 2007]. These changes in expression dependences are especially relevant because they are associated with a modification of the phenotype of the cells studied [Huerta M. et al 2014]. Then, using the appropriate tools, it is possible to integrate the expression matrices (gene-expression levels under specific conditions) into the context of biological knowledge.

When a phenotypic change is due to a change or suppression of a single gene, as it is common in monogenic diseases, the level of expression of this single gene could be enough to establish the healthy or pathological phenotype of the tissue. But the world of the monogenic characters and diseases is a very narrow and relatively well-known field of study. What is much more difficult is to infer the pool of genes involved in the maintenance of a multi-genetic disease, in its variations, and in the different disease progressions. Normally, these different phenotypes are a product of a combination of cell processes carried out by different sets of genes. These cell processes are transverse cell processes, because the same process occurs in different and even antagonistic conditions and can be combined with each other. Besides, these multiple transverse cell processes take place in a concurrent way, overlapping under certain conditions but not under others.

Let us consider two cellular processes, such as cell differentiation and tumour proliferation (Figure 26). As they are transverse cell processes, then we can find up to four different combinations with the presence or absence of each process (they can appear concurrently, only one appears, or none). These different combinations will determine the final phenotype of the cell. For instance, we are going to observe tumour proliferation in differentiated cells, but also tumour proliferation in undifferentiated cells, thus, tumour proliferation and cell differentiation are transverse cell processes. Our system helps to detect all of the combinations of transverse cell processes from the expression data, plus the genes most involved in the appearance and disappearance of each transverse cell process. The study of these concurrent transverse cell processes has direct medical applications. For instance, besides using retrovirals to treat viral infections in fish, it is possible to promote temperature-dependent processes that induce a fever phenotype that increases fish survival under viral challenge [Boltana S. et al 2013]. The next step would be to study the underlying mechanisms that cause this, to understand how the phenotypic changes work and to be able to control them. But how can we understand the underlying mechanisms of these phenotypic changes? To achieve this purpose, it is not enough to detect when a phenotypic change is linked to a strong variation in gene-expression levels, but also when the variations are subtle. For instance, a change in the diet of bees induces modifications in the methylation pattern of genes that leads to a metabolic acceleration and increased growth. These phenotypic changes are the product of relatively subtle variations in terms of expression levels but extended to a large number of genes [Barchuk AR. et al 2007; Kucharski R. et al 2008].

To mention one example, the presented decomposition of sample clusters in a combination of transverse phenotypes has a direct application in the study of the effect of the drugs. Unlike principal component or clustering methods, the concise analysis of our tools allows researchers to separate genes by their link with the different effects of the drugs, where each effect would correspond to a different transverse phenotype. Then, the effects of the drugs are transverse to different treatments, and in each treatment different effects appear concurrently. Clustering methods group the treatments with the same effects and thus treatments that produce a similar final phenotype [Ling Y. and Kelemen A. 2006]. Two treatments could be linked to the same transverse phenotype (effect) in one transverse phenotypic change (passing from one effect to another) but to a different transverse phenotype in another transverse phenotypic change (because these treatments share some effects but not others). Thus, not only can the shared effects of the drugs be studied, but also the effects that are different for similar drugs.



**Figure 26. Cell processes are transverse between themselves whether they are concurrent sometimes yes and sometimes not.** Two transverse cell processes and four transverse phenotypes are shown in the figure: differentiated cells, undifferentiated cells, and tumour proliferation or no tumour proliferation. The two phenotypes from the first transverse cell process are transverse to the two phenotypes from the second transverse cell process. In other words, the two first phenotypes are transverse to the two last phenotypes. However, neither the two first phenotypes nor the two last phenotypes are transverse between themselves and a phenotypic change is needed to pass from one phenotype to the other. The combination of the cited transverse phenotypes occurring concurrently in the cell, results in four different final phenotypes of the cell. Note that all of the combinations between transverse phenotypes are not always possible as in this case.

When analysing expression data with large sample series using sample clustering, usually the researcher wishes to extract the maximum information about the sample clusters obtained [Ling Y. and Kelemen A. 2006]. Furthermore, this further information will depend on the researcher's interests and the data analysed in each case. Clustering methods obtain sample clusters using statistics, but these methods do not say anything about the phenotypes these sample clusters could represent. There are several methods to obtain further information about the researcher's sample clusters. One powerful method is to use tools to compare these sample clusters with the sample clusters from other microarrays [Huerta M. et al 2014] as will be seen in next sections. Nevertheless, the most common method is to directly study the differentially-expressed genes between sample clusters [Huerta M. et al 2009] as has been seen in previous sections. In this method, from the genes that over-express in one sample cluster but not in another, the researcher tries to know what phenotype represents each sample cluster. Nonetheless, there can be hundreds of genes differentially expressed between one sample cluster and another one. Moreover, even though the genes coexpress for these sample clusters, they can actually be part of independent cellular processes that take place concurrently for those sample clusters. The most common procedure of attempting to distinguish among those independent, although concurrent, cellular processes that occur in each sample cluster, consists of performing a combination of sample and gene clustering. The results can be seen in a Heatmap. The Heatmap displays the genes that, in addition to expressing differently in each sample cluster, they are coexpressed for all of the matrix samples. These different sets of coexpressed genes would carry out different processes that occur concurrently for a given sample cluster but not for another. This fact differentiates one sample cluster from another. In this way, the analysis is much easier. We just need to study these cellular processes because the combination of the mentioned processes occurring concurrently in each sample cluster is what will define the hypothetical phenotype of the sample cluster. Now, instead of an arbitrary list of hundreds of genes differentially expressed, we have several sets of coexpressed genes that carry out independent and transverse cellular processes [Eisen MB. et al 1998]. The set of coexpressed genes of some of these processes are also differentially expressed. Nevertheless, the use of Heatmaps and hierarchical clusterings of samples and genes is not always intuitive. This way of studying the phenotypes of the sample clusters has two main limitations: First, it provides a static view, and second, and more important, we need great differences in the expression levels of a set of coexpressed genes between one cluster and another in order to distinguish the sample clusters where this process is occurring and the sample clusters where it is not. This problem is especially serious because most of the coexpressed genes maintain continuous expression relationships, that is, not discontinuous. The problem is even worse if large sample series are used. The continuity in the expression relationships between coexpressed genes can be better appreciated as the size of the sample series increases, because the transition or intermediate phenotypes are also considered, not only the extreme phenotypes.

A complementary approach to the differentially-expressed-genes analysis queries GO database to classify the differentially-expressed genes trying to find genes involved in the same transverse process based on the literature [Cedano J. et al 2007; Boltana S. et al 2013; Mohammadi A. et al 2011]. The main problem of this approach is that it can only work with the genes described in GO, ignoring the rest of differentially-expressed genes. And the second one, is that these tools really do not detect the transverse processes occurring in the expression data, because they work with the

categories previously defined in GO. All of this, in addition to the aforementioned problems of the differentially-expressed-gene search.

It is not easy to identify when a change in the cellular state has occurred only by observing discontinuities in the level of expression of the genes, and is even worse if the change is subtle. Many times, agglomerative or partitive clustering methods cannot differentiate the phenotypes because the changes of phenotype do not create the necessary discontinuity in the sample distribution to be easily detected, especially when it affects only a subset of genes [Thalamuthu A. et al 2006]. To address this problem, we choose to detect the phenotypic changes using the changes in slope of the expression relationships.

Our methodology is based upon the following four properties of the non-linear expression relationships: First, each fluctuation in a correlated expression relationship implies a phenotypic change. Second, multiple non-linear expression relationships are involved in the same phenotypic change. Third, if two genes maintain an expression relationship with a certain type of curve, the genes coexpressed with these two genes maintain expression relationships of the same type of curve between them. Fourth, the curve type of the non-linear expression relationships will describe the role of the genes in the phenotypic change. The first point allows us to detect the phenotypic changes from the non-linear expression relationships that are highly correlated. The second point allows us to group the non-linear high-correlated expression relationships by the phenotypic change they describe. The third point allows us to expand the analysis from these non-linearly highly-correlated genes to the genes coexpressed with them. The fourth point allows us to know the activation/deactivation relationship between the phenotypes of the phenotypic change based on the morphology of the curve.

As was seen in previous sections, non-linear expression dependences can be used for sample clustering. This sample clustering has had two different approaches, a non-linear interpretation of dimensionality reduction [Cannistraci CV. et al 2010] and a clustering of the samples corresponding to each phenotype from the phenotypic changes described by the non-linear expression relationship [Cedano J. et al 2008]. This second approach is possible thanks to the first property previously cited: "Each fluctuation in a correlated expression relationship implies a phenotype change", so the samples at either side of the fluctuation point will represent different phenotypes. Nevertheless, the second property is what makes this approach especially relevant: "Multiple non-linear expression relationships are involved in the same phenotypic change". However, as the previous section reflects, until now this approach has been limited to single relationships between pairs of genes. The next step will be then, to exploit this second principle to obtain the multiple phenotypic changes that occur concurrently in the expression data and to study how their transverse phenotypes combine to result in the final phenotypes. This is so because this strategy will allow us to show the sample clusters as the intersection of multiple phenotypic changes that occur concurrently in the cell.

Identifying the phenotypes based on the changes in slope of the expression relationships allows us to detect both abrupt changes in expression levels plus more subtle phenotypic changes. In turn, the system detects those phenotypic changes that are repeated in a large number of gene-expression relationships as well as those that only occur in certain pairs of genes. As was mentioned before,

one of the limitations of the clustering methods is that some changes do not affect all genes of the matrix, but rather a subset of them. In these situations, common clustering methods struggle to detect sample clusters [Thalamuthu A. et al 2006], and clustering methods specialized in clustering samples when only a subset of genes is involved in the changes [Csardi G. et al 2010] still need much more discontinuity in the sample distribution to detect the phenotypic changes.

In this way, the transverse phenotypes of the concurrent phenotypic changes obtained could be considered as quite basic states or transverse to the main cellular states depending on whether those phenotypic changes are recurrent in multiple pair-of-genes relationships (basic states) or if they only occur in a few pairs of genes (transverse to the main cellular states). Using our system, it is possible to detect the multiple combinations of these transverse phenotypes and establish a hierarchy among the concurrent phenotypic changes. This hierarchy would highlight the ones that affect the cell to a greater extent and the ones that affect the cell to a lesser extent. Besides this, the genes involved in each phenotypic change and the role they play in the transition will be obtained.

As we stated in a previous section, an especially relevant point in the study of non-linear expression relationships is that the type of curve of the relationship determines the role of each gene in the expression relationship [Huerta M. et al 2014]. That is, their activation/deactivation role. This fact can be used to study the phenotypes of the phenotypic change, in other words, the phenotype previous and next to each phenotypic change. The aim of studying the "gene role in phenotypic changes" here is exactly the same as in differentially-expressed-genes analysis: the activation or deactivation of the phenotype. However, what the curve type describes to us is complex activation relationships (between the genes that take part in the phenotypic change). For this reason, our application provides the non-linear expression relationships that participate in each concurrent phenotypic change classified by curve type. The researcher will obtain the biological meaning of each concurrent phenotypic change from the gene Paris involved in the phenotypic change and the type of curve of their expression relationships.

Below, the new approach and method are detailed. After that, the web application that follows this new approach, as well as a comparison with the results obtained by classical methods for sample clustering can be observed. In these examples, one can observe how a sample cluster obtained by classical methods is decomposed.

As the presented methodology intends to be a standard method valid for expression data from any origin, the terminology used in the following sections to describe the analysis pipeline attempts to be as generic as possible, to also be valid for expression data from any origin.

### **3.7.1. Automatic detection of sudden changes in expression relationships**

The curvature points of each expression relationship are obtained from the POPs that describe the inner pattern of the relationship (the PCOP). Curvature points are those POPs of the PCOP in which a change in slope occurs. The curvature point will be the POP among a set of nearby POPs that

maximizes the change in slope, with respect to the previous and the next curvature point [Huerta M. et al 2014].

As seen in previous sections, the samples belonging to each POP local area can be selected [Cedano J. et al 2007; Cedano J. et al 2008] and, in this way, the samples belonging to each tendency between two curvature points are clustered. These sample subspaces will represent the phenotypes involved in the phenotypic change linked to the sudden change that the expression relationship describes.

Only the highly-correlated expression relationships are considered in the analysis [Huerta M. et al 2014]. The correlation degree provided by PCOP calculation is what guarantees us that the non-linear expression relationships are not a product of chance and have a biological meaning. The correlation degree will also affect the criterion to consider an expression relationship as linear or non-linear: Expression relationships that are less correlated will need more curvature to be considered as non-linear [Huerta M. et al 2014].

In this way, all the highly-correlated non-linear expression relationships are detected from the expression data. Furthermore, the sample subspaces between two consecutive curvature points are obtained for each non-linear expression relationship detected. In the next step, these sample subspaces will be compared in order to detect the expression relationships involved in the same phenotypic change.

### **3.7.2. Automatic detection of phenotypic changes**

The sample subspaces obtained from each expression relationship are compared with the subspaces obtained from the rest. Then, the gene relationships that divide the sample space in the same sample subspaces are clustered together, because these expression relationships describe the same phenotypic change. This is the key point of our approach. The samples belonging to each phenotype of a phenotypic change are averaged from all of the non-linear expression relationships involved in the same phenotypic change. Note also that the gene relationships of the same group will perform the same phenotypic changes but each gene will play a different role in the phenotypic change. As we stated in a previous section, the type of curve of the non-linear expression relationships will describe the role of each gene in the phenotypic change [Huerta M. et al 2014].

As was mentioned earlier, curvature points are those points in the expression relationship in which a change in slope occurs. Depending on the number of pairs of genes that share the same sample subspaces on both sides of the curvature point, the phenotypes of the phenotypic change, that is, both subspaces, can be considered as quite basic states (a high number of gene pairs) or transverse to main cellular states (a small number of gene pairs). These transverse phenotypes take place in a concurrent way, overlapping under certain conditions but not under others. In other words, these transverse phenotypes will take place concurrently with different phenotypes of the main phenotypic change.

One of the most delicate points of the procedure followed is determining the phenotype to which the samples in the limit of the phenotypic change (the samples close to the curvature point) belong. The information provided by one expression relationship is not enough to decide whether these bordering samples belong to one phenotype or another. This point is solved only when all of the expression relationships that represent the same phenotypic change are considered. The samples will be assigned to the phenotype in which they appear more often considering all expression relationships of a certain phenotypic change.

Obtaining the transverse phenotypes from all of the expression relationships for a certain phenotypic change provides us the accumulated error (as can be seen in Figure 27). This accumulated error explains the number of samples that has been assigned to a transverse phenotype incorrectly for each expression relationship of a phenotypic change, divided by the total of relationships of the phenotypic change. The smaller the accumulated error is, the more clearly differentiated the phenotypes described in the phenotypic change are. On the contrary, a bigger error will indicate a bigger number of samples in the border area between the two phenotypes. The bigger the number of expression relationships involved in the phenotypic change, the more reliable the assignment of samples to each transverse phenotype.

### **3.7.3. The showing of the intersections between transverse phenotypes**

Once the transverse phenotypes of each concurrent phenotypic change are obtained, the intersection between the phenotypes of the different phenotypic changes will be displayed via web-browser. The phenotypic changes detected are concurrent for all of the samples of the expression matrix. The intersection between the phenotypes of the different phenotypic changes results in the final phenotypes that can be recognized by clustering methods. Depending on whether a sample describes the previous or subsequent phenotype of each concurrent phenotypic change will condition the final phenotype to which this sample belongs. It is the combination of transverse phenotypes occurring in a sample for all the phenotypic changes detected that will determine the final phenotype of the sample.

The web interface allows ordering the experimental conditions so that the intersection between the transverse phenotypes can be studied (Figure 27). By grouping the samples by the phenotypes of a transverse phenotypic change, the intersection of these phenotypes with the rest of concurrent phenotypic changes will be shown. The priority given to the different phenotypic changes will establish the final order of the sample conditions. In this way, when prioritizing the phenotypic changes of interest for the researcher from all of those detected, their intersection with the rest of the concurrent phenotypic changes will be shown. By default, the highest priority will be assigned to the phenotypic changes that involve the highest number of genes in the transition, for being considered as the main phenotypic changes of the cell, and the least priority will be assigned to those that just affect a small number of genes, for considering that they represent the most specific processes and are transverse to the first ones. In this way, the hierarchy of phenotypic changes that take place concurrently in the analysed expression data is shown (Figure 27).



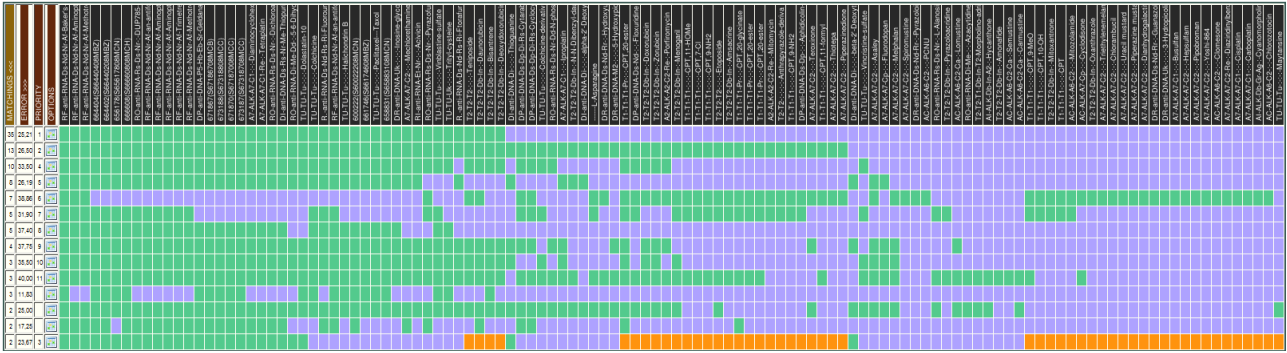
### **3.7.4. Crossing the sample clusters obtained by common clustering methods with the transverse-phenotypes intersection**

SOM, SOTA, PAM and Repeated Bisection clustering methods [Yin L. et al 2006] have been used in the examples shown in the next section. PC and MDS dimensionality reduction methods have also been used, which improves the results of the clustering. Silhouette and Dunn integrity methods have been used to find the k parameter value that provides the best clustering for each combination of clustering and dimensionality reduction methods. Those cited are the most used procedures in gene-expression sample clustering [Yin L. et al 2006]. The results of these clustering methods (plus dimensionality reduction and integrity) are automatically calculated in our server for the user's expression data. In this way, the researcher can study the correspondence between the sample clusters obtained and the multiple concurrent phenotypic changes.

### **3.7.5. Tool's output interfaces**

#### **3.7.5.1. The phenotypic-changes hierarchy view**

The web application provides two types of lists: the list of transverse and concurrent phenotypic changes found for the analysed expression data and the list of the main gene relationships involved in each transverse and concurrent phenotypic change. In the list of phenotypic changes (Figure 27), the sample sub-spaces of the transverse phenotypes involved in each phenotypic change are shown. The concurrent phenotypic changes are the rows, the samples are the columns, and the samples of each transverse phenotype are coloured differently. As a same sample (column) belongs to a different transverse phenotype in each phenotypic change (row), the transverse-phenotype intersections are shown as a hierarchy of sample subspaces (where the subspaces are progressively subdivided by the phenotype intersections). The samples belonging to each phenotype intersection represent a new sample subspace that involves the simultaneity of both phenotypes.



**Figure 27.** The transverse-phenotypic-changes list view shows the transverse phenotypic changes detected by our system. The rows correspond to each concurrent phenotypic change detected by our system, and the columns correspond to each sample. All cells of the same column represent the same sample, and these cells are coloured based on the phenotype the sample belongs to in each transverse phenotypic change. Then, looking at the column of the sample, the phenotype which this sample belongs to in each one of the detected transverse phenotypic changes, can be seen. All of these transverse phenotypic changes take place concurrently for that sample; therefore, the columns will show us the intersection of transverse phenotypes for these samples. The phenotypic changes are sorted by the number of expression relationships involved in each phenotypic change. The two first rows describe the main transverse phenotypic changes. 35 non-linear expression relationships are detected for carrying out the first phenotypic change, and 13 for carrying out the second one in the list. There exists a phenotype intersection between the second phenotype of the first phenotypic change and the first phenotype of the second phenotypic change. Those samples in the intersection of phenotypes will inherit the characteristics of both phenotypes. The priority column shows the criterion followed to sort the concurrent phenotypic changes to study their phenotypes intersection (see phenotype-intersection section for more details). The error column shows the mismatch among the non-linear expression relationships that describe the same phenotypic change to constitute its sample subspaces. The last concurrent and transverse phenotypic change in the list describes a transition among three phenotypes.

The main transverse phenotypic changes are those that involve the highest number of genes in the phenotype transition, whereas the transverse phenotypic changes involving a small number of genes are those that least affect the cell. In the hierarchy of phenotypic changes, the phenotypic changes involving the highest number of non-linear expression relationships are shown first (Figure 27). In this way, the intersection of these main phenotypic changes with the rest of the transverse phenotypic changes is shown. Nevertheless, if the researcher is interested in studying the intersection of a phenotypic change of interest with the rest, he/she can increase the priority of this one. Note that, as every concurrent phenotypic change occurs in the totality of the samples, for each sample we can study the intersection between phenotypes from any pair of phenotypic changes detected. Likewise, in each sample (column) the totality of concurrent phenotypic changes occur. In order to avoid any confusion, it is important to clarify that our analysis-pipeline term “concurrent phenotypic changes”, does not mean phenotypic changes occurring simultaneously, but rather either the previous or the next phenotype of a phenotypic change occurs simultaneously with the previous or the next phenotype of any other phenotypic change. The samples are grouped in the table by the intersection of phenotypes between the phenotypic changes detected. The samples are grouped by the phenotype to which each sample belongs in the first transverse-phenotypic-change (1st row by default). Then, the samples from a same transverse phenotype are sub-divided by belonging to a different phenotype in the next phenotypic change with less priority, and so on. This continues until all of the samples are ordered. This arrangement of the columns facilitates observing the transverse-

phenotype intersections, and the transverse-phenotypes intersection is what will define the cell state of the sample.

The search for transverse and concurrent phenotypic changes can be parameterized by the correlation degree and curvature degree required for the expression relationships involved in the transition. A sharper curvature will imply a more clear transition between phenotypes.

### 3.7.5.2. The gene-relationships list view

In the gene-relationships list view, the gene relationships involved in each transverse and concurrent phenotypic change are shown (Figure 28). Only the expression relationships with a sufficient correlation will be listed. The expression relationships listed are classified by type of curve, and a curve-type icon points out the type of the expression relationship. Different curve types imply a different role of the genes in the phenotypic changes, for example, a  $y=e^x$  expression relationship represents a switch gene that activates the other gene (resulting in a new phenotype due to the gene activation) [Huerta M. et al 2014]. More information about the role of the genes in each type of curve can be obtained from the previous sections. The detailed view of each gene-expression relationship [Cedano J. et al 2008] can be launched from the list of gene relationships. In this detailed view, the expression dependence between the genes can be seen in a 2D plot (one dimension per gene). In this view, the samples belonging to the phenotypes involved in each phenotype transition are coloured differently. Thus, the link between the expression-relationship suddenly changes and the transition between phenotypes can be observed (Figure 29). Remember that multiple non-linear expression relationships are involved in the same phenotypic change, contributing to the phenotypic transition in different ways. The relative location of the samples, further from or closer to the curvature point in the 2D plot, shows us whether the sample clearly represents the phenotype or if it is a transition state between two phenotypes. In addition to showing the phenotypes involved in the phenotypic change, this 2D plot shows the sample clusters obtained by common clustering methods [Yin L. et al 2006] (Figure 29). This allows the researcher to study the correspondence between sample clusters and phenotypic changes.

The correspondence between the sample clusters obtained by means each clustering plus dimensionality-reduction methods and its best k value, with respect to the intersection of transverse phenotypes obtained by our application has been studied for the expression data used as example (at\_matrix). This correspondence is shown in the tables below. The table rows represent the clustering methods, and the columns represent the phenotype intersections. Each cell of the table shows the percentage of sample matching between the sample clusters and each phenotype intersection. The "Coincidence" column of these tables shows the percentage of correspondence between each clustering method and the phenotype intersections. This correspondence between clustering methods and phenotype intersections is calculated from the samples in common between each cluster from the clustering method and the phenotype intersection that best matches with this sample cluster.

As can be seen in Table II, our system can recognize the sample clusters obtained by common clustering methods (which affect all of the matrix genes)[Yin L. et al 2006]. But, at the same time, it detects the more subtle (and transverse) phenotypic changes that only affect a subset of genes, and that are not detectable by common clustering methods. As stated earlier, the sample clusters obtained by common clustering methods precisely represent several intersections of these multiple transverse phenotypes detected. In this way our procedure is able to decompose these sample clusters in a phenotypes intersection. Note that only the intersections shown in the hierarchy of transverse phenotypes (Figure 27) are detectable in the expression matrix studied (some transverse-phenotype intersections not occur in the expression data).

	Id	Gene-1	$f_i$ uncorrelation	Id	Gene-2	Graphical interface
	326	HCG4P6	0.070317	645	AKT3	
	337	ENO2	0.074276	645	AKT3	
	337	ENO2	0.080728	943	IFI30	
	325	RAB14	0.082282	650	RBMS1	
	325	RAB14	0.083830	608	TMEPAI	
	327	This term is	0.084097	607	CCL2	
	337	ENO2	0.085224	685	PLAUR	
	324	HLA-B	0.088765	952	NET1	
	337	ENO2	0.088959	617	RAB31	
	325	RAB14	0.089919	609	HMG2	
	432	ANX8	0.091318	924	ASNS	
	337	ENO2	0.091791	938	RREB1	
	337	ENO2	0.092488	615	RAB31	
	432	ANX8	0.094323	781	CITED2	
	369	RDX	0.095766	703	TAX1BP3	
	319	DUSP14	0.097222	571	NCDN	
	337	ENO2	0.097405	647	XTP3TPA	
	337	ENO2	0.097780	614	This term is	
	328	GYPE	0.097812	613	GNA12	
	364	This term is	0.099688	767	PARP3	
	325	RAB14	0.099742	780	MGP	
	337	ENO2	0.099788	622	GJB1	

**Number of merges: 22**      [Update your clusters file](#)      [Apply your clusters file](#)

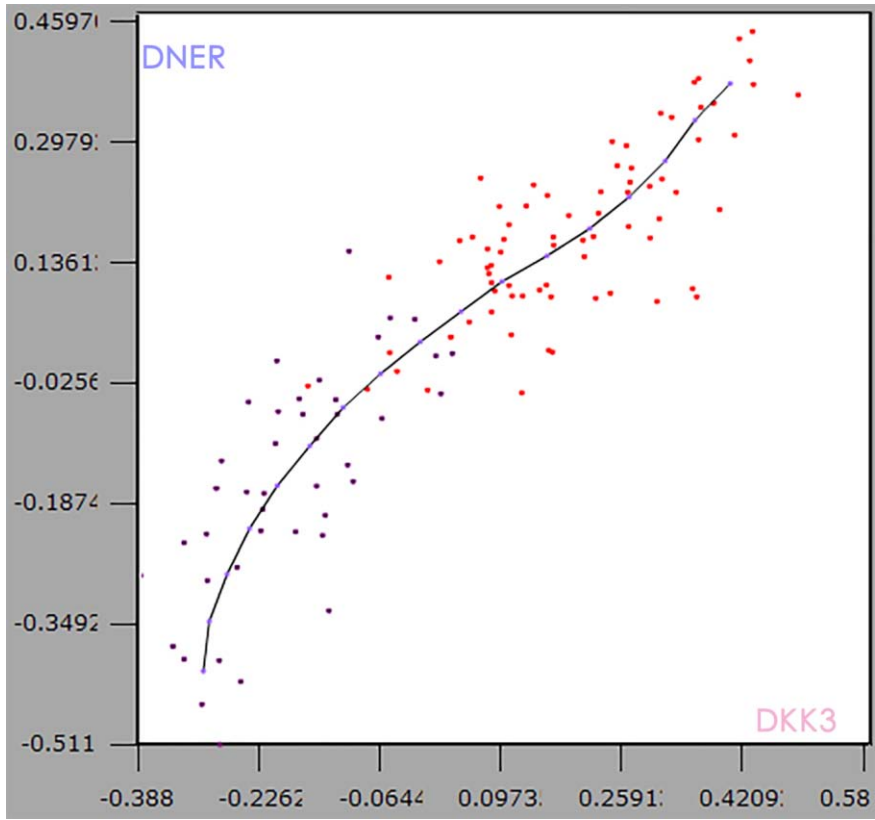
Figure 28. The gene relationships involved in each concurrent phenotypic change are shown in the gene-relationships list view. The expression relationships listed are ordered by correlation degree. They are also classified by curve type. An icon at the right of the expression relationship describes its typology. The detailed view of each gene-expression relationship (Figures. 29-30) can be launched by clicking on this icon. A button at the bottom of the list allows one to switch from the classical-clustering-methods sample clusters to the phenotypic-change transverse phenotypes in order to be shown in the detailed view (Figures. 29-30). The other button at the bottom opens a menu to choose the clustering method the researcher wishes to use to cluster the samples. This allows the researcher to relate these sample clusters with the transverse-phenotypic change.

Method	K	Ph. Intersection (1,1)	Ph. Intersection (1,2)	Ph. Intersection (2,2)	Coincidence
Mds Som	4	1 <sub>65,38%</sub> 2 <sub>100%</sub> 3 <sub>7,89%</sub> 4 <sub>3,03%</sub>	1 <sub>30,77%</sub> 3 <sub>65,79%</sub> 4 <sub>3,03%</sub>	1 <sub>3,85%</sub> 3 <sub>26,32%</sub> 4 <sub>93,94%</sub>	79,66%
Som	4	1 <sub>100%</sub> 2 <sub>64%</sub> 3 <sub>3,13%</sub> 4 <sub>10%</sub>	2 <sub>32%</sub> 4 <sub>65%</sub>	2 <sub>4%</sub> 3 <sub>96,87%</sub> 4 <sub>25%</sub>	79,66%
Mds Pam	4	1 <sub>3,03%</sub> 2 <sub>95,65%</sub> 3 <sub>83,33%</sub> 4 <sub>9,09%</sub>	1 <sub>12,12%</sub> 2 <sub>4,35%</sub> 3 <sub>16,67%</sub> 4 <sub>59,09%</sub>	1 <sub>84,85%</sub> 4 <sub>31,82%</sub>	77,12%
Pam	4	1 <sub>100%</sub> 2 <sub>4,88%</sub> 3 <sub>100%</sub> 4 <sub>9,75%</sub>	2 <sub>21,95%</sub> 4 <sub>60,98%</sub>	2 <sub>73,17%</sub> 4 <sub>27,27%</sub>	77,12%
Mds Sota	5	1 <sub>95,45%</sub> 2 <sub>100%</sub> 5 <sub>38,89%</sub>	4 <sub>72,73%</sub> 5 <sub>55,56%</sub>	1 <sub>4,55%</sub> 3 <sub>100%</sub> 4 <sub>24,24%</sub> 5 <sub>5,56%</sub>	84,75%
Sota	5	1 <sub>88,46%</sub> 2 <sub>100%</sub> 5 <sub>35,71%</sub>	1 <sub>7,69%</sub> 4 <sub>70,59%</sub> 5 <sub>57,14%</sub>	1 <sub>3,86%</sub> 3 <sub>100%</sub> 4 <sub>29,41%</sub> 5 <sub>7,14%</sub>	83,90%
Pc Sota	4	1 <sub>57,58%</sub> 2 <sub>95,83%</sub>	1 <sub>21,21%</sub> 2 <sub>4,27%</sub> 4 <sub>74,29%</sub>	1 <sub>21,21%</sub> 3 <sub>100%</sub> 4 <sub>25,71%</sub>	79,66%
Pc Pam	5	1 <sub>95%</sub> 2 <sub>100%</sub> 5 <sub>23,33%</sub>	1 <sub>5%</sub> 4 <sub>71,87%</sub> 5 <sub>33,33%</sub>	3 <sub>100%</sub> 4 <sub>28,13%</sub> 5 <sub>43,33%</sub>	77,12%
Pc Som	4	1 <sub>95,24%</sub> 2 <sub>93,75%</sub> 4 <sub>2,78%</sub> 5 <sub>24%</sub>	1 <sub>4,76%</sub> 2 <sub>6,25%</sub> 4 <sub>72,22%</sub> 5 <sub>24%</sub>	3 <sub>100%</sub> 4 <sub>25%</sub> 5 <sub>52%</sub>	79,66%
Pam	9	2 <sub>20%</sub> 4 <sub>100%</sub> 5 <sub>100%</sub> 8 <sub>50%</sub>	1 <sub>7,69%</sub> 3 <sub>89,47%</sub> 7 <sub>38,46%</sub> 8 <sub>50%</sub>	1 <sub>92,31%</sub> 2 <sub>80%</sub> 3 <sub>10,53%</sub> 6 <sub>100%</sub> 7 <sub>61,54%</sub>	86,24%
Pc Som	9	1 <sub>42,86%</sub> 3 <sub>85,71%</sub> 4 <sub>100%</sub> 5 <sub>21,43%</sub>	1 <sub>42,86%</sub> 3 <sub>14,28%</sub> 5 <sub>78,57%</sub> 6 <sub>70%</sub> 8 <sub>10%</sub>	1 <sub>14,28%</sub> 2 <sub>100%</sub> 6 <sub>30%</sub> 7 <sub>100%</sub> 8 <sub>90%</sub>	81,82%
Som	9	1 <sub>16,67%</sub> 2 <sub>100%</sub> 3 <sub>10%</sub> 4 <sub>100%</sub> 8 <sub>57,14%</sub>	1 <sub>16,67%</sub> 3 <sub>60%</sub> 5 <sub>54,55%</sub> 6 <sub>73,91%</sub> 8 <sub>42,86%</sub>	1 <sub>66,67%</sub> 3 <sub>30%</sub> 5 <sub>45,45%</sub> 6 <sub>26,09%</sub> 7 <sub>100%</sub>	80,36%
Mds Pam	7	1 <sub>25%</sub> 2 <sub>5,26%</sub> 4 <sub>100%</sub> 5 <sub>100%</sub> 7 <sub>18,18%</sub>	1 <sub>58,33%</sub> 3 <sub>84,21%</sub> 7 <sub>50%</sub>	1 <sub>16,67%</sub> 2 <sub>94,74%</sub> 3 <sub>15,79%</sub> 6 <sub>100%</sub> 7 <sub>31,82%</sub>	83,05%
Mds Sota	7	1 <sub>27,27%</sub> 4 <sub>95,45%</sub> 5 <sub>100%</sub> 7 <sub>57,14%</sub>	1 <sub>63,64%</sub> 3 <sub>72,73%</sub> 7 <sub>42,86%</sub>	1 <sub>9,09%</sub> 2 <sub>100%</sub> 3 <sub>27,27%</sub> 4 <sub>4,55%</sub> 6 <sub>100%</sub>	85,59%

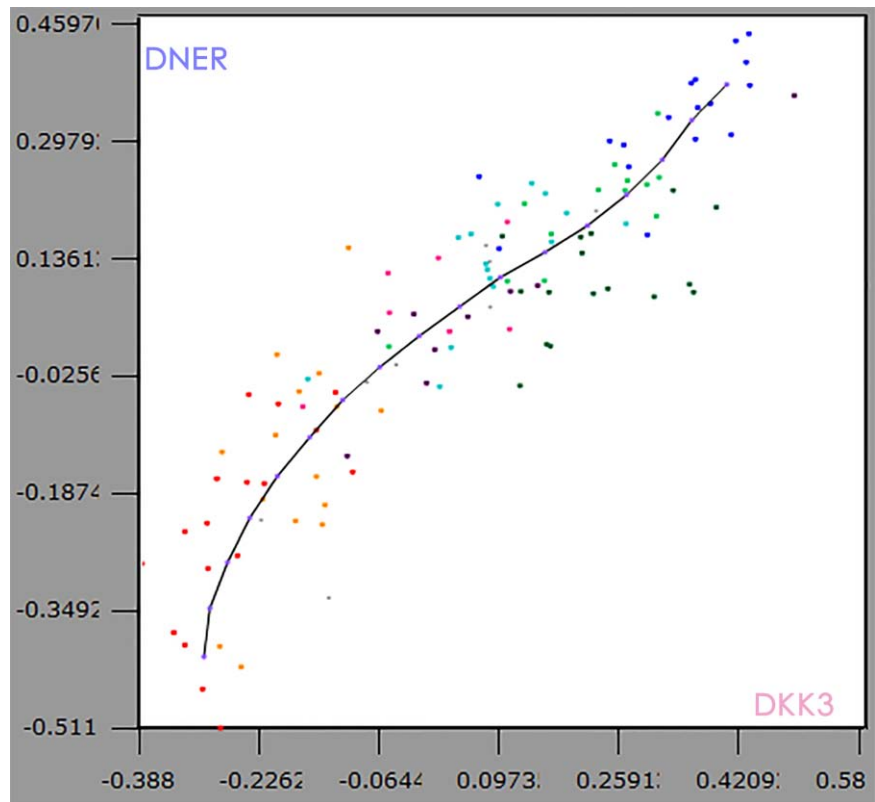
**Table II. Correspondence between the sample clusters obtained by clustering methods commonly used in expression analysis and the transverse phenotypic changes found by our system.** The rows represent the common clustering methods. The columns represent the transverse-phenotype intersection between the two first phenotypic changes shown in Figure 27. These transverse phenotypic changes describe the two main phenotypic changes inside the expression data (35 and 13 non-linear expression relationships are detected for the first and second transitions respectively). The first column represents the samples that belong to the intersection between the first phenotypes of each one of the two transverse phenotypic changes, the third column represents the samples belonging to the intersection between the second phenotypes of the two transverse phenotypic changes, and the second column represents the samples belonging to the intersection between the second phenotype of the first phenotypic change and the first phenotype of the second phenotypic change. The intersection between these transverse phenotypes can be seen in Figure 27. The cells show the correspondence between the clusters obtained by common clustering methods (rows) and the phenotype intersections (columns). It can be seen that there exists a correspondence between these sample clusters and the intersection of the transverse phenotypes found. The "Coincidence" column shows the coincidence between the sample clusters from clustering methods and the phenotypes intersection. The matching between the sample clusters obtained by SOTA K=5 and by SOM k=4 is 82%, which is less than the matching between SOTA K=5 and the phenotype intersections (see "coincidence" column). Silhouette and Dunn Integrity methods were used to choose the best k for each clustering method. In this way, our application recognizes the sample clusters obtained by common clustering methods but, furthermore, it decomposes these clusters into a hierarchy of phenotypic changes that shows the intersection of transverse phenotypes that constitutes the sample clusters.

In Figure 29 and Table II, the correspondence between the transverse-phenotypes intersection and the sample clusters obtained by common clustering methods [Yin L. et al 2006] can be observed. The correspondence between the sample clusters obtained by means of SOTA K=5 and SOM k=4 is 82%, similar to the correspondence (72%-86%) between the sample clusters obtained by means these clustering methods and the intersection of transverse phenotypes (for more details, see the "Coincidence" column of Table II). As can be seen by examining Table II and the plots of Figure 29 in depth, the sample clusters can be described as the intersection of transverse phenotypic changes.

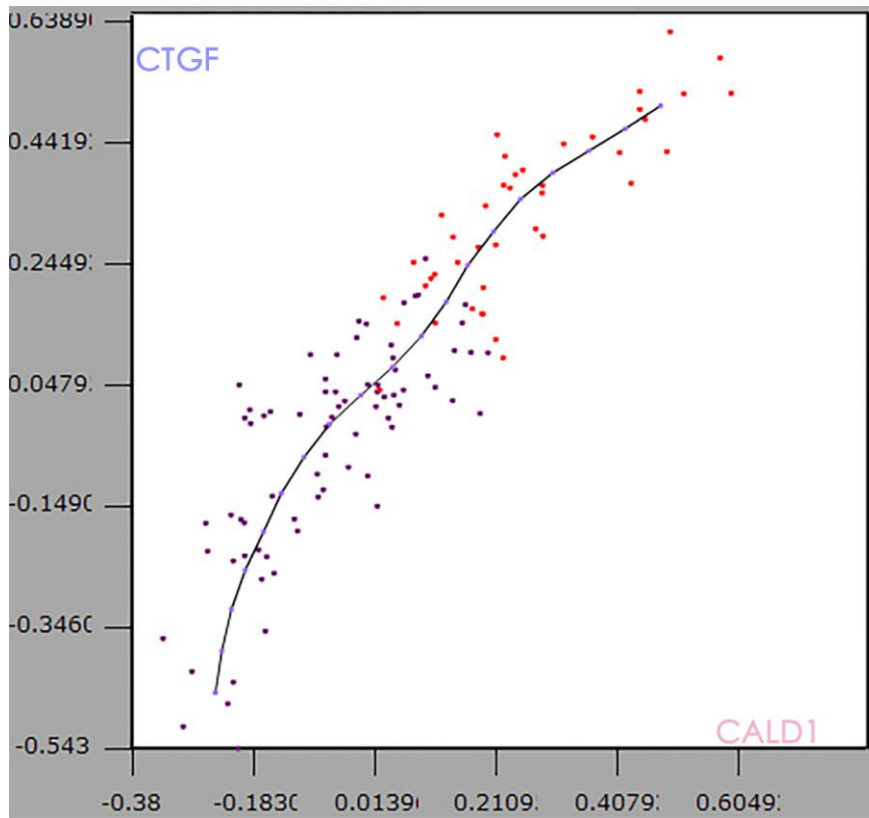
29.a)



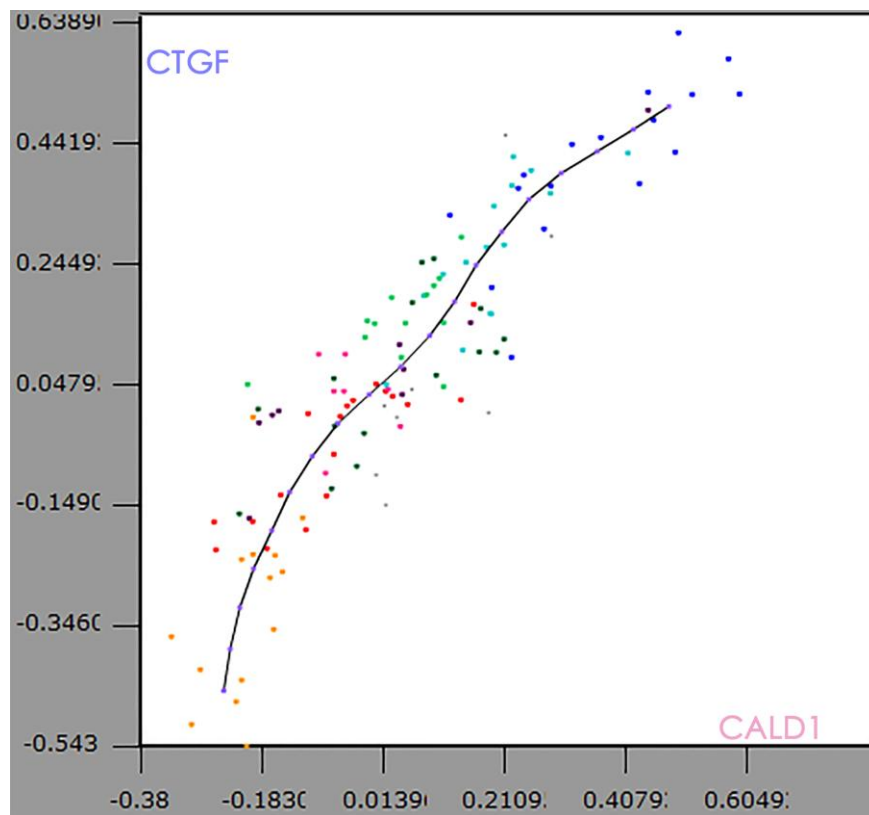
29.b)



29.c)



29.d)

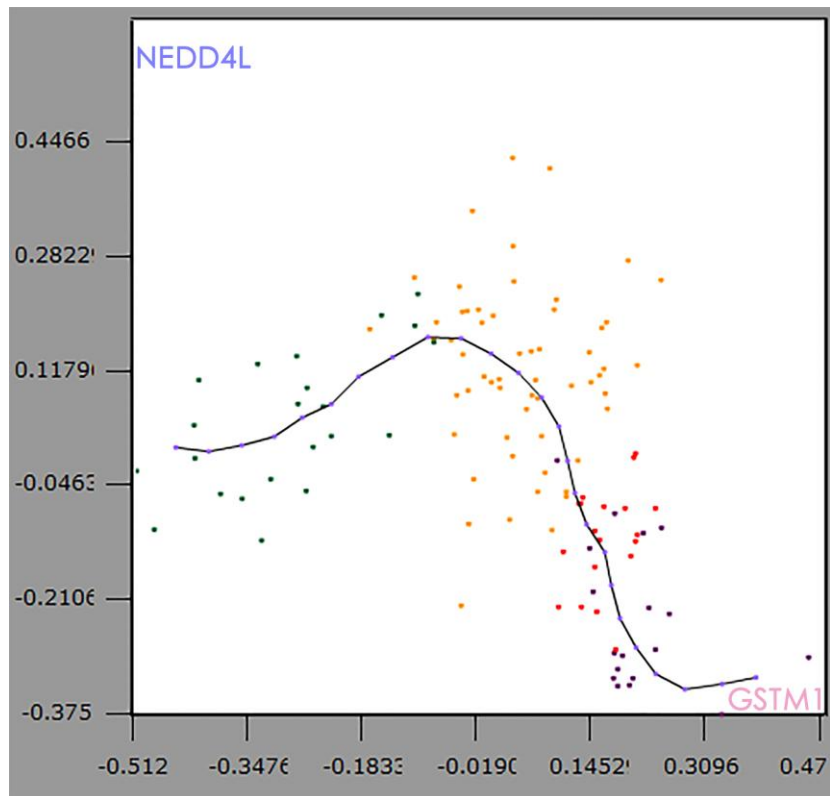


**Figure 29.** The detailed view of two expression relationships involved in the phenotypic changes analysed in Table II. DKK3 and DNER genes in (a) and (b), and CALD1 and CTGF genes in (c) and (d). The axes represent the expression level of the genes, and the point cloud represents the samples. The curve crossing the data cloud is the PCOP that describes the inner pattern of the expression relationship. The samples are coloured based on the cluster or transverse phenotype they belong to. The DKK3 and DNER expression relationship performs the first transverse phenotypic change listed in the phenotypic-change hierarchy shown in Figure 27. The CALD1 and CTGF expression relationship carries out the second transverse phenotypic change in the hierarchy. In (a), the two phenotypes obtained from the first phenotypic change listed in Figure 27 are shown. In (c), the two phenotypes obtained from the second phenotypic change listed in Figure 27 are shown. These transverse phenotypic changes describe the two main phenotypic changes detected, which are the ones used in Table II. In (b) and (d), the sample clusters obtained by Repeated-Bisection clustering are shown. Looking at the relationship between DKK3 and DNER genes ((a) and (b)), yellow and red clusters from Repeated-Bisection, in (b), correspond to the first phenotype of the phenotypic change, in (a), and blue, green and pink clusters from Repeated-Bisection, in (b), correspond to the second phenotype in (a). Looking at the relationship between CALD1 and CTGF genes ((c) and (d)), yellow, red, green and pink clusters from Repeated-Bisection, in (d), correspond to the first phenotype of the phenotypic change, in (c), and blue clusters from Repeated-Bisection, in (d), correspond to its second phenotype, in (c). In this way, it can be observed how sample clusters (pink and green clusters) are the product of the intersection of transverse phenotypes.

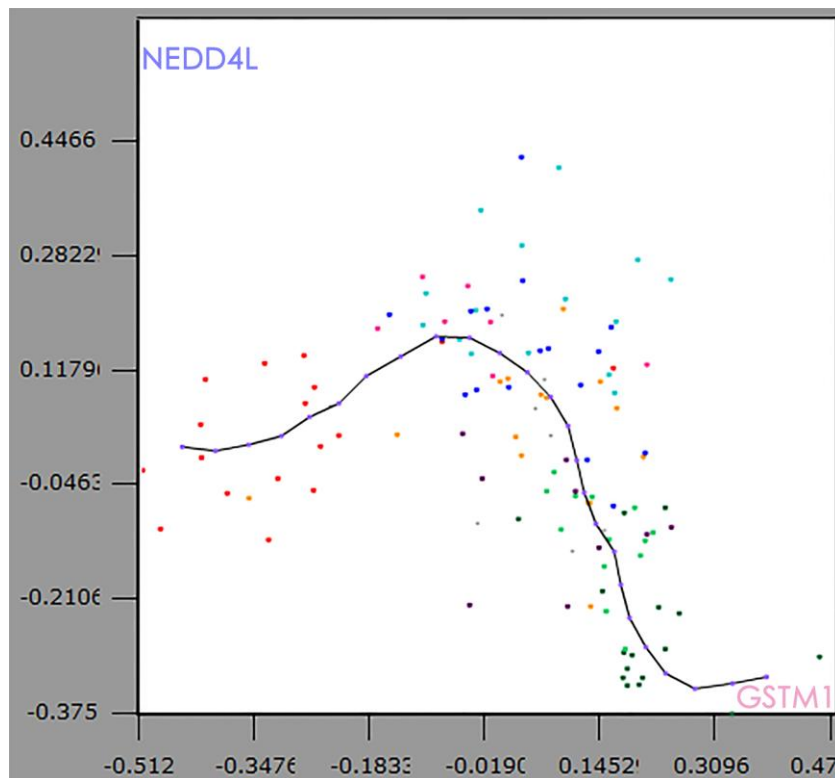
Looking at Figures 29.a, 29.b, 29.c and 29.d, it can be observed how the concurrent phenotypic changes are transverse among them and how the intersection between transverse phenotypes constitutes new sample clusters (i.e. green and pink clusters in Figures 29.b and 29.d). These green and pink clusters (obtained by common clustering methods) overlap with two transverse phenotypes, one from each phenotypic change. This indicates that the green and pink clusters inherit the characteristics from both transverse phenotypes (each one from a different phenotypic change). However, these transverse phenotypes involve more sample clusters and these sample clusters are different from one phenotype and another (red and yellow clusters and blue clusters respectively). This indicates that the final phenotypes of the pink and green clusters will share some characteristics with the red and yellow clusters, and some others with the blue clusters. In other words, the two transverse phenotypes to which the pink and green clusters belong, will define the characteristics they share and the characteristics they do not share with the rest of the clusters (the pink and green clusters will share the characteristics of one transverse phenotype with the blue clusters, and the characteristics of the other transverse phenotype with the red and yellow clusters). The researcher finds out the characteristics of the phenotypes by studying the expression relationships (genes and curve type) linked to each transverse phenotypic change.



30.a)



30.b)



**Figure 30.** The detailed view of the expression relationship involved in the phenotypic changes analysed in Table III. The expression relationship is involved in three phenotypic changes. In Figure 30.a the four transverse phenotypes obtained are shown. In Figure 30.b, the sample clusters obtained by Repeated-Bisection clustering method are shown. The figures show the correspondence between the clusters obtained by the clustering method and the three phenotypic changes described by the curve.

Note that the system can detect very subtle changes in expression relationships, indicating a phenotypic change (as in Figure 29). Note also that each transverse phenotypic change is obtained from several non-linear expression relationships. As more non-linear expression relationships are involved in the same phenotypic change, the more accurate is the assignation of samples to each phenotype and the more "global" are the transverse phenotypes. This also implies a bigger correspondence between the sample clusters obtained by common clustering methods, the transverse phenotypes, and their intersections. Now, looking at Figure 30, a gene pair involved in three phenotypic transitions allows us to accurately study the transitions between several sample clusters. As can be seen in Table III, there is also a correspondence between the sample clusters obtained by common clustering methods [Yin L. et al 2006] and the four transverse phenotypes found in this expression relationship.

In this figure, it can be observed how the expression relationship guides the transition from some clusters to others.

The most common methods in gene-expression sample clustering have been used in the previous examples. Nevertheless, there exist more specific clustering methods [Andreopoulos B. et al 2009] that have not been compared in the examples. The results of our tool have not been compared either with the clusters from clustering methods that only work with part of the samples, like Bi-clustering [Csardi G. et al 2010], since, in our approach, all of the concurrent and transverse phenotypic changes consider the complete sample series. Neither have we analysed here the results of decomposing sample clusters from a non-statistical origin based on previous criteria, like the samples origin or experimental criteria used in the GEO database [Barret T. et al 2013]. However, regardless of the clustering method used, if they obtain the same or similar sample clusters, the results of their decomposition would be identical. If the final phenotypes of the expression data are clearly differentiated, the different clustering methods will obtain the same sample clusters, so the results, when crossing them with our tool, would be the same. When the phenotypes are less detectable, the disparity among the results provided by the different clustering methods and the probability of non-correspondence increases.

If a high number of concurrent phenotypic changes are considered in the deconstruction of the sample clusters, not only will the number of samples in the intersection between the transverse phenotypes be reduced, but also their coincidence with the sample clusters obtained by clustering methods. This is because of the limitations of clustering methods [Thalamuthu A. et al 2006] and also because of the accumulated error when assigning the samples to each transverse phenotype. This limits us when we try to infinitely decompose the sample clusters obtained by clustering methods, and the phenotypes they could represent, in a combination of multiple concurrent phenotypic changes.

Method	K	Phenotype 1	Phenotype 2	Phenotype 3	Phenotype 4	Coincidence
Mds Som	4	3 <sub>42,11%</sub> 4 <sub>6,06%</sub>	1 <sub>15,38%</sub> 3 <sub>31,58%</sub> 4 <sub>9,09%</sub>	1 <sub>76,92%</sub> 2 <sub>9,58%</sub> 3 <sub>23,68%</sub> 4 <sub>81,82%</sub>	1 <sub>7,69%</sub> 2 <sub>90,48%</sub> 3 <sub>2,63%</sub> 4 <sub>3,03%</sub>	69,49%
Som	4	3 <sub>3,13%</sub> 4 <sub>42,5%</sub>	2 <sub>16%</sub> 3 <sub>9,38%</sub> 4 <sub>30%</sub>	1 <sub>9,52%</sub> 2 <sub>76%</sub> 3 <sub>84,38%</sub> 4 <sub>25%</sub>	1 <sub>90,48%</sub> 2 <sub>8%</sub> 3 <sub>3,13%</sub> 4 <sub>2,5%</sub>	69,49%
Mds Pam	4	4 <sub>40,91%</sub>	1 <sub>3,03%</sub> 3 <sub>22,22%</sub> 4 <sub>31,82%</sub>	1 <sub>90,91%</sub> 2 <sub>17,39%</sub> 3 <sub>72,22%</sub> 4 <sub>25%</sub>	1 <sub>6,06%</sub> 2 <sub>82,61%</sub> 3 <sub>5,56%</sub> 4 <sub>2,27%</sub>	67,8%
Pam	4	4 <sub>43,9%</sub>	1 <sub>25%</sub> 2 <sub>2,44%</sub> 4 <sub>34,15%</sub>	1 <sub>62,5%</sub> 2 <sub>90,24%</sub> 3 <sub>12,5%</sub>	1 <sub>12,5%</sub> 2 <sub>7,32%</sub> 3 <sub>90%</sub>	70,34%
Mds Sota	5	3 <sub>3,23%</sub> 4 <sub>45,45%</sub> 5 <sub>11,11%</sub>	2 <sub>28,57%</sub> 3 <sub>9,68%</sub> 4 <sub>33,33%</sub> 5 <sub>5,56%</sub>	1 <sub>9,09%</sub> 2 <sub>64,29%</sub> 3 <sub>83,87%</sub> 4 <sub>21,21%</sub> 5 <sub>77,78%</sub>	1 <sub>90,91%</sub> 2 <sub>7,14%</sub> 3 <sub>3,23%</sub> 5 <sub>5,56%</sub>	71,19%
Sota	5	4 <sub>47,06%</sub> 5 <sub>14,29%</sub>	2 <sub>28,57%</sub> 3 <sub>10%</sub> 4 <sub>32,35%</sub> 5 <sub>7,14%</sub>	1 <sub>19,23%</sub> 2 <sub>64,29%</sub> 3 <sub>86,67%</sub> 4 <sub>20,59%</sub> 5 <sub>78,57%</sub>	1 <sub>80,77%</sub> 2 <sub>7,14%</sub> 3 <sub>3,33%</sub>	70,34%
Pc Sota	4	2 <sub>4,17%</sub> 3 <sub>3,85%</sub> 4 <sub>45,71%</sub>	2 <sub>20,83%</sub> 3 <sub>1,54%</sub> 4 <sub>31,43%</sub>	1 <sub>54,55%</sub> 2 <sub>50%</sub> 3 <sub>76,92%</sub> 4 <sub>22,86%</sub>	1 <sub>45,45%</sub> 2 <sub>25%</sub> 3 <sub>7,69%</sub>	55,93%
Pc Pam	5	2 <sub>6,25%</sub> 3 <sub>5%</sub> 4 <sub>50%</sub>	1 <sub>5%</sub> 2 <sub>25%</sub> 3 <sub>15%</sub> 4 <sub>31,25%</sub> 5 <sub>3%</sub>	1 <sub>20%</sub> 2 <sub>56,25%</sub> 3 <sub>75%</sub> 4 <sub>18,75%</sub> 5 <sub>80%</sub>	1 <sub>75%</sub> 2 <sub>12,5%</sub> 3 <sub>5%</sub> 5 <sub>16,67%</sub>	66,95%
Pc Som	4	2 <sub>6,25%</sub> 3 <sub>5%</sub> 4 <sub>44,44%</sub>	1 <sub>4,76%</sub> 2 <sub>18,75%</sub> 3 <sub>15%</sub> 4 <sub>33,33%</sub>	1 <sub>23,81%</sub> 2 <sub>62,5%</sub> 3 <sub>75%</sub> 4 <sub>22,22%</sub> 5 <sub>80%</sub>	1 <sub>71,43%</sub> 2 <sub>12,5%</sub> 3 <sub>5%</sub> 5 <sub>20%</sub>	64,41%
Pam	9	3 <sub>73,68%</sub> 6 <sub>9,09%</sub> 7 <sub>7,69%</sub> 8 <sub>20%</sub>	3 <sub>15,79%</sub> 5 <sub>26,67%</sub> 6 <sub>18,18%</sub> 7 <sub>61,54%</sub> 8 <sub>20%</sub>	1 <sub>92,31%</sub> 2 <sub>100%</sub> 3 <sub>10,53%</sub> 4 <sub>5,56%</sub> 5 <sub>60%</sub> 6 <sub>63,64%</sub> 7 <sub>30,77%</sub> 8 <sub>50%</sub>	1 <sub>7,69%</sub> 4 <sub>94,44%</sub> 5 <sub>13,33%</sub> 6 <sub>9,09%</sub> 8 <sub>10%</sub>	75,23%
Pc Som	9	5 <sub>21,43%</sub> 6 <sub>65%</sub> 7 <sub>8,33%</sub> 8 <sub>10%</sub>	4 <sub>23,08%</sub> 5 <sub>42,86%</sub> 6 <sub>25%</sub> 7 <sub>8,33%</sub> 8 <sub>30%</sub>	1 <sub>85,71%</sub> 2 <sub>92,31%</sub> 3 <sub>14,29%</sub> 4 <sub>61,54%</sub> 5 <sub>35,71%</sub> 6 <sub>10%</sub> 7 <sub>75%</sub> 8 <sub>60%</sub>	1 <sub>14,29%</sub> 2 <sub>7,69%</sub> 3 <sub>85,71%</sub> 4 <sub>15,38%</sub> 7 <sub>8,33%</sub>	70,91%
Som	9	5 <sub>72,33%</sub> 6 <sub>34,78%</sub> 8 <sub>28,57%</sub>	4 <sub>25%</sub> 5 <sub>27,27%</sub> 6 <sub>39,13%</sub> 7 <sub>10%</sub> 8 <sub>14,29%</sub>	1 <sub>100%</sub> 2 <sub>5,88%</sub> 3 <sub>80%</sub> 4 <sub>58,33%</sub> 6 <sub>26,09%</sub> 7 <sub>85%</sub> 8 <sub>57,14%</sub>	2 <sub>94,12%</sub> 3 <sub>20%</sub> 4 <sub>16,67%</sub> 7 <sub>5%</sub>	73,32%
Mds Pam	7	3 <sub>68,42%</sub> 6 <sub>8,33%</sub> 7 <sub>18,18%</sub>	3 <sub>15,79%</sub> 5 <sub>30,77%</sub> 6 <sub>25%</sub> 7 <sub>40,91%</sub>	1 <sub>91,67%</sub> 2 <sub>100%</sub> 3 <sub>15,79%</sub> 4 <sub>14,29%</sub> 5 <sub>53,85%</sub> 6 <sub>58,33%</sub> 7 <sub>36,36%</sub>	1 <sub>8,33%</sub> 4 <sub>85,71%</sub> 5 <sub>15,38%</sub> 6 <sub>8,33%</sub> 7 <sub>4,55%</sub>	71,19%
Mds Sota	7	3 <sub>45,45%</sub> 6 <sub>8,33%</sub> 7 <sub>28,57%</sub>	3 <sub>33,33%</sub> 5 <sub>28,57%</sub> 6 <sub>25%</sub> 7 <sub>14,29%</sub>	1 <sub>90,91%</sub> 2 <sub>100%</sub> 3 <sub>21,21%</sub> 4 <sub>9,09%</sub> 5 <sub>64,29%</sub> 6 <sub>58,33%</sub> 7 <sub>57,14%</sub>	1 <sub>9,09%</sub> 4 <sub>90,91%</sub> 5 <sub>7,14%</sub> 6 <sub>8,33%</sub>	71,19%

**Table III. Correspondence between the sample clusters obtained by common clustering methods and the transverse phenotypes found by our system from expression relationships that involve three phenotypic changes.** As in Table II, the results show a correspondence between the sample clusters obtained by the clustering methods and the four transverse phenotypes identified by our system. Looking at the *expression relationship detailed view* shown in Figure 3O, we can study the transition between one cluster and another one and the expression level of the genes responsible for the transitions.

Independent cell processes can participate together in the same concurrent phenotypic change. For this reason, not all the cellular processes that participate in a phenotype are identified by our system as an independent cell process. These processes take place together for the analysed sample series, and the genes that carry these processes out appear coexpressed. This is because the concurrent phenotypic changes detected will depend on the experiments done, even though the analysed genes would be the same. This would mean that, for the same genes, a sample series could obtain more concurrent phenotypic changes than another one, together with a bigger amount of sample clusters. All concurrent processes that do not cause a different phenotype for the analysed data will be grouped in the same phenotypic change. Only the concurrent cell processes that differentiate one final phenotype from another one will be considered separately. In other words, these independent cellular processes must be transverse among them for the analysed expression data. It should be remembered that the final phenotypes are the result of different combinations of concurrent transverse cell processes, where different combinations of these transverse cell processes result in the different final phenotypes. The system specifically aims to detect these combinations, and for this purpose it groups together all of the concurrent cell processes that, although independent, appear coexpressed in the analysed expression data.

For each transverse and concurrent phenotypic change are provided the genes most involved, but these genes could be involved in the cause or in the effect of the phenotypic change. In fact, the second one is more usual. Even though the technologies to obtain gene-expression arrays do not capture the regulatory genes that cause the phenotypic changes because of the low variability in their gene expression, these technologies do allow for studying the genes that result from the activation cascade started by these regulatory genes [Wu B. 2007]. Our analysis is possible because these final genes do maintain wide-enough expression ranges, which allows high-throughput tools to analyse their complex expression dependences.

By means of PCOP calculation, we can detect phenotypic changes without using time series that sort the data temporally. In contrast, we have a static vision of the phenotypic changes. We do not know which of the two phenotypes, one at one side and the other at the other side of the phenotypic change, occurs first or later in time. Nevertheless, this static vision is enough for the analysis we intend to do, since we are just interested in knowing which one of the two phenotypes involved in each phenotypic change occurs in each sample cluster.

The use of non-linear expression relationships in sample clustering opens up an enormous potential beyond the study of gene regulation [Guo X. et al 2014; Cannistraci CV. et al 2010]. Common clustering methods [Yin L. et al 2006; Andreopoulos B. et al 2009] can detect the main phenotypes well enough, phenotypes that affect most of the genes of the expression matrix, but these methods are not so precise in the detection of secondary phenotypes, which only affect the expression of part of the genes, and much less in the detection of the transversality of these secondary phenotypes with the main ones. Local clustering methods also have this flaw. Our procedure allows one to study the following: the combination of phenotypic changes that result in each final phenotype; which phenotypic changes are common in different final phenotypes; and which phenotypic changes turn them into different final phenotypes. It also allows for studying the specific role of the genes in the final phenotypes for participating in one phenotypic change or another.

Considering all the expression relationships of a certain phenotypic change and assigning the samples to the phenotype in which they appear most often, we solve the problems explained previously, according to which a discontinuity in the sample space is needed to detect sample clusters and differentially expressed genes; but to achieve this, we need multiple non-linear expression relationships for a same phenotypic change to assign the samples to their phenotype accurately, especially with the samples on the border between phenotypes.

We move away from the approach followed by differentially-expressed-genes analysis, Heatmaps and double clustering of samples and genes, due to the limitation of these methods, because of the need of discontinuity in the expression data is over-come.

Our approach provides a new kind of information in the representation of the hierarchy of concurrent and transverse phenotypes. Unlike the dendrogram representation of hierarchical clustering, our representation shows the intersection between the transverse phenotypes. It not only shows clusters that are the product of the union of sample clusters, but also clusters that are product of the intersection of two or more sample clusters(our transverse phenotypes). This intersection between transverse phenotypes, because the phenotypes take place concurrently in the same samples, represents the final phenotype of these samples.

Our approach also denotes the difference, with respect to the methods that query the GO database to classify the differentially-expressed genes for involvement in the same processes [Boltana S. et al 2013; Mohammadi A. et al 2011]. Our approach does not start from previous premises; it is purely statistical, detecting the transverse processes that occur in our experiments from the data, and whose different combinations make it possible for us to obtain different sample clusters from our expression data.

The main obstacle in the search for multiple concurrent phenotypic changes is mainly due to the data source: non-linear expression relationships need to be detected for a number of gene-pairs big enough to be useful. Large enough sample series and wide enough gene-expression ranges facilitate the detection of non-linear expression relationships. If one wishes to perform a brief study of an expression matrix with large sample series, it is better to use only common clustering methods and Heatmaps, from the methods mentioned in this paper to more advanced clustering methods with more precision obtaining sample clusters [Andreopoulos B. et al 2009], but if you need a more accurate analysis, to study the differences among the sample clusters obtained by these clustering methods, and to understand the cellular processes behind each one, the tools developed here allow for the obtainment of very relevant information for the researcher that can enrich the classical analyses.

As previously stated, it is very important to study the underlying mechanisms that carry out the different phenotypes that can be detected by clustering methods. The difference among these final phenotypes is rarely caused by a change in a single gene. Indeed, the different final phenotypes will be the product of a combination of cell processes carried out by different sets of genes. They are transverse cell processes that overlap (are concurrent) in certain phenotypes but not in others. This is why new tools like the one presented here are necessary.

All the findings presented in this section have been included in the work [Huerta M. et al 2015].

**Availability:** sample-cluster-deconstruction web-tool: <http://ibb.uab.es/phscl>



### 3.8. Crossing Clusters: Studying the relationships among the sample clusters obtained by clustering methods from expression data

Microarray analysis, RSeq analysis, etc allow the study of large sample series, which allows one to describe multiple phenotypes. Clustering methods are used to describe these multiple phenotypes by grouping the sample conditions that cause a similar effect on the genes. But conventional expression analysis does not go much further in the study of the phenotypes detectable in expression experiments. For instance, they do not analyse the relationships between the sample clusters in-depth, or how the transitions between the phenotypes that these clusters represent could happen.

Clustering methods, like hierarchical clustering, establish inclusion relationships between clusters; nevertheless, they are not able to detect any other relationship. With the approach presented next, we intend to help researchers to study the transition between the phenotypes, as well as to discover other types of more complex relationships between them.

To study these interdependencies between phenotypes, we can rely again on the study of gene-expression relationships. As aforementioned, differentially expressed genes show us how we can change from a sample cluster to a new one depending on expression variations of individual genes. Non-linear expression relationships show us how we can change from a sample cluster to a new sample cluster or to a different one depending on expression variations of combinations of gene pairs, instead of individual genes. Now, what we intend with this tool is to show all of those transitions between sample clusters described by non-linear expression relationships ordered and categorised, in order to study these transitions and understand their biological meaning.

Our methodology is based on the fact that each fluctuation in a correlated expression relationship implies a phenotypic change [Huerta M. et al 2014]. Thus, the proposed strategy here is focused on the detection of non-linear expression relationships, their crossing with the sample clusters obtained by clustering methods, and to classify the different arrangements of sample clusters along the phenotypic changes.

The analysis of differentially expressed genes allows us to study the biological meaning of the different sample clusters from the genes that over-express and under-express in each sample cluster [Huerta M. et al 2014]. Studying the step from under-expression to over-expression of these genes allows us to study the transition between phenotypes. Nevertheless, this is limited to an activation/deactivation transition. Furthermore, we need these genes to be known enough in the literature to obtain biological meaning about the sample clusters. With our approach, complex dependences between sample clusters are detected. Besides, biological meaning of the phenotypes can be obtained although the genes are unknown by the literature. This is because, as aforementioned, the curve type of the non-linear expression relationship already describes the dependence among phenotypes. For instance, a  $y = e^x$  relationship will provide a switching dependency, that is, a certain phenotype must occur so that the other phenotype can occur. A  $y = e^{-x}$  relationship describes a mutual exclusion relationship, that is, for one phenotype to occur the other



one must not occur.  $y=x^2$  or  $I= x^2 + y^2$  relationships indicate two different ways of the same cell process or feedback effects [Huerta M. et al 2014].

In addition, clustering methods need to deal with intermediate phenotypes due to the increasing number of samples in sample series. Our analysis also interrelates these transition phenotypes. Sample clusters closer to the inflection point of the curve will represent the intermediate phenotypes between the more extreme phenotypes (represented by the sample clusters far away from the inflection point).

### **3.8.1. Detection of curvature points and non-linear expression relationships**

Curvature points are those POPs in the PCOP in which a change in slope occurs. The detection of curvature points identifies the non-linear expression relationships [Huerta M. et al 2014]. All the non-linear expression relationships inside the expression data are detected.

### **3.8.2. Arrangement of sample clusters along the PCOP**

The samples of each cluster are bound to their closer POP. This binding of the samples to the POPs provides a map of the cluster arrangement along the expression relationship.

### **3.8.3. Classification of the non-linear expression relationships that cross each sample-cluster arrangement**

As aforementioned, if two different clusters are located one on either side of the curvature point, it is considered that this expression relationship describes the phenotypic change between those two sample clusters. So, we group the expression relationships by whether they place the same clusters at each side of the phenotypic change, and also by the order of these clusters at each side of the phenotypic change (which allows us to detect transition phenotypes). Different groups of expression relationships show dependence between different sample clusters, and those grouping a larger number of relationships would describe the main and more global phenotypic changes (which affect the expression of a larger number of the expression-matrix genes).

### **3.8.4. Calculation of the sample-cluster arrangement by different clustering methods**

For each uploaded expression matrix, the server calculates the most common clustering plus dimension-reduction analyses applied to expression matrices: HC, SOTA, SOM and PAM, combined with PC and MDS. The parameters are adjusted using the most common integrity analysis: Silhouette and Dunn. For the sample clusters obtained by means of each clustering method

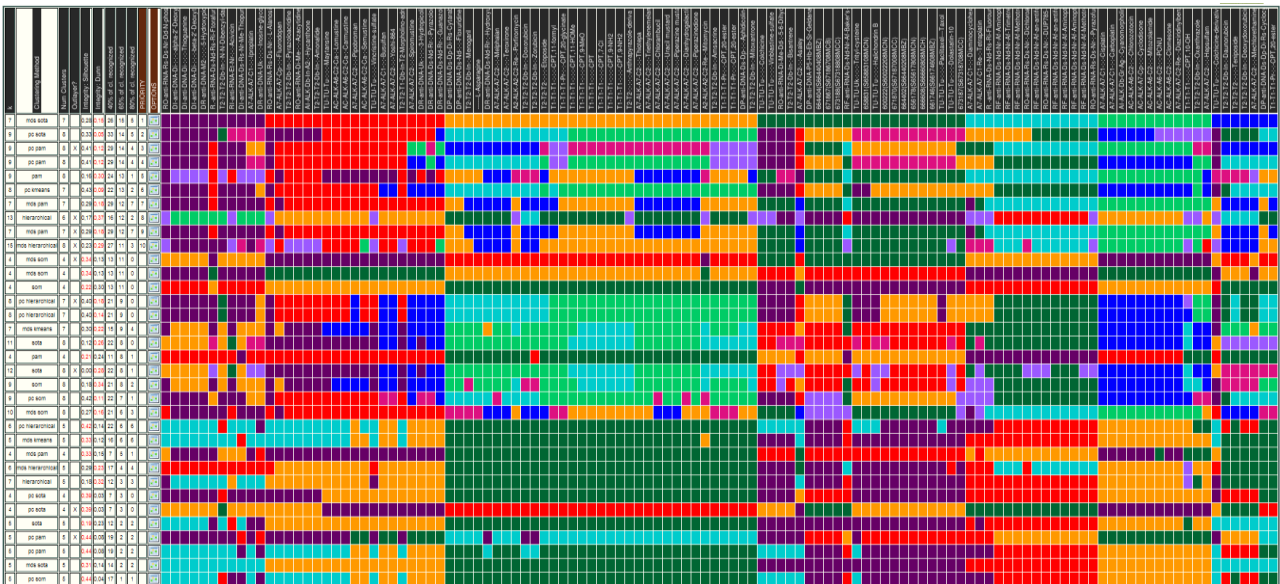
and its best parameter value, the non-linear expression relationships that cross these sample-clusters are grouped based on the phenotype changes among the sample clusters.

### 3.8.5. Tool's output interfaces

#### 3.8.5.1. List of sample-cluster arrangements view

This allows one to compare the results obtained by different clustering methods and their respective k values. The list is represented as a table where each row displays the sample clusters resulting from applying a clustering plus dimension-reduction method and a k value. The columns represent the matrix samples. In each cell of the table, the samples that belong to each different cluster are coloured with a different colour. In this way the sample clusters obtained by means of each clustering method and k value can be identified and compared with the results from the rest of methods and k values (Figure 31).

A different table will be displayed for each clustering plus dimension-reduction method. For each table, the difference between one row and another one is the k value used in the same clustering method. In this way the sample clusters resulting from the use of different k values in the same clustering method can be compared among them. An extra table with one row per method shows the clustering results from each methodology and its k value with best integrity. In this way the best results from each clustering method can be compared among them.



**Figure 31. List of sample-cluster arrangements view.** In the display it can be seen the best results obtained from the different clustering methods applied to the expression data (at\_matrix). Integrity methods are used to evaluate the sample clusters and choose the best k value for each clustering method. The columns represent the matrix samples and the rows the clustering methods (one method per row). The colour of the samples represents the cluster they belong after applying each clustering method.

### 3.8.5.2. List of non-linear expression relationships crossing the sample clusters view

Once the clustering method whose result we wish to study is chosen, the non-linear expression relationships are listed, grouped and classified into three categories: First, by whether these expression relationships place the same sample clusters at one side and the other of the curvature point; second, by whether the sample clusters maintain the same order along the expression relationship; and third, by their cluster intersection degree (the last two categories lead to study transition phenotypes).

For each expression relationship, a graph shows the curvature point and the cluster arrangement along the expression relationship (Figure 32). This allows one to immediately observe how the clusters are distributed at either side of the curvature point and the degree of belonging to each phenotype. An icon shows the curve type. As aforementioned, this describes the type of relationship between the sample clusters [Huerta M. et al 2014].

	Id	Gene-1	$f_1$ uncorrelation	Id	Gene-2	Graphical interface	Clusters on PCOP
	579	COL6A1	0.074533	688	GLIPR1		
	650	RBMS1	0.071312	953	HISPPD2A		
	617	RAB31	0.073138	788	GNAS		
	325	RAB14	0.082282	650	RBMS1		
	325	RAB14	0.082769	647	XTP3TPA		
	324	HLA-B	0.074741	675	TRIM3		

**Figure 32. List of expression relationships crossing the sample clusters view.** For a given sample-clusters arrangement, non-linear expression relationships that separate the sample clusters are listed and grouped into three categories: First, those that separate the same sample clusters at one and the other side of the curvature point (yellow), second, if the sample clusters at the same side of the curvature point maintain the same order (orange) and third, if they also maintain the same degree of cluster intersection (red). The last two categories are used to study the transition phenotypes. The graph on the right shows the location of the curvature point and the arrangement of the sample clusters along the expression relationship. This allows one to observe how the clusters are distributed at one and the other side of the phenotypic change and the degree of belonging to each phenotype. An icon shows the type of non-linear expression relationship, which allows researchers to know the type of dependence between the sample clusters.

By clicking on the curve icon, the 2D interface shows the gene expression relationship in detail (one dimension per gene), with the distribution of the samples of each cluster along the relationship.

This tool can be a good complement for expression-matrix analysis, especially if the researcher wishes to go beyond the study of differentially expressed genes. Now she/he can study complex transitions among sample clusters plus the transition sample clusters.

All the findings presented in this section have been included in the work [Huerta M. et al 2015].

**Availability:** Crossing-clusters tool: <http://wallace.uab.es/crscl>



### 3.9. MGDB: comparing differentially expressed genes from different microarrays to compare phenotypes and sample clusters from different microarrays

Since 2002, the Nature journals, among others, have announced that authors were thereafter required to deposit microarray data in public repositories like GEO [Barrett T. et al 2011] or ArrayExpress [Parkinson H. et al 2007] so that anyone could freely access and critically evaluate the data discussed in manuscripts [Nature 2002]. But how could all of these data help your particular research? How could all of this information enrich your expression data analysis and the phenotypes you have found in your data?

To date, GEO archives approximately 20,000 studies comprising 500,000 samples, 33 billion individual measurements, for over 1,300 organisms, submitted by 8,000 laboratories from around the world, and supporting data for over 10,000 published manuscripts [Barrett T. et al 2011].

Samples within GEO datasets are further grouped and classified into subsets according to the experimental variables under examination in each study, for instance tissue or strain. So, any gene of the microarray will be a marker gene of the microarray if its expression displays a significant effect in relation to subsets, that is, if the expression values pass a threshold of statistical difference between any experimental-variable subset and another [Barrett T. et al 2011].

The experimental variables of GEO datasets are based on the sample origin, such as species, specimen, strain, individual, tissue, development stage, cell type, cell line, etc. on individual features like age or gender, on pharmacological experimentation such as agent, dose, protocol, etc. on the disease genesis such as genotype/variation, disease state, infection, shock, stress, etc. or on other experimental conditions like temperature or time.

How can we establish a correspondence between these sample subsets perfectly classified by experimental variable and the sample clusters obtained by statistical methods from our expression data under study? We can do this by verifying that differentially expressed genes are the same in both microarrays, similar to cMAP procedure [Lamb J. et al 2006], because this can imply that the subsets of both expression matrices describe the same phenotypes.

So, although the experimental-variable subsets of GEO microarrays are defined by microarray developers, and the experimental variables are subjected to the hypothesis that the researchers wanted to investigate, these microarray data can be reused for investigations around the world even when the hypotheses of these investigations are completely different from the microarray original hypotheses.

Our tool for crossing marker genes can be used for different purposes:

- To study the role of marker genes of the user's microarray in other microarrays.
- To assign biological meaning to the sample clusters of the user's microarray. This can include:
  - To compare the user's experiments with experiments for the same pathology but in different tissues, in different species, or directly different pathologies.
  - To search for drugs whose effect causes the transition between the phenotypes studied in the user's microarray.
  - To study the phenotype of different subsets of individuals under investigation.
  - To search undesirable side-effects of a treatment studied in the user's microarray.

The marker-gene database tool is composed by the elements set out below.

### **3.9.1. Database of Microarray marker genes**

The samples of GEO microarray datasets are classified by experimental variable, such as treatment, protocol, disease state, patients condition, tissue, etc. So, our database of microarray marker genes contains the genes with statistically significant differences in their expression between any experimental variable subset and another for each GEO microarray. The aim of the marker-genes database is to use the marker genes (differentially-expressed genes) of each microarray to search for matching microarrays with respect to the user's microarray. The system obtains these marker genes directly from GEO. The database of marker genes is updated monthly with the marker genes of the new datasets from GEO.

### **3.9.2. Marker-gene search in the user's expression matrix**

The definition of sample clusters and the marker-gene search in the user's microarray is completely versatile. Our system provides the sample clusters calculated by common clustering methods (HC, SOM, SOTA, etc.) or allows the researcher to define the sample clusters based on his/her hypothesis. The marker genes can be searched for by having some of the clusters being up-regulated or down-regulated with respect to the basal value, or by being over-expressed or under-expressed with respect to other clusters. All possible combinations are allowed and different combinations will supply different sets of marker genes for the user's expression matrix [Huerta M. et al 2009]. This procedure permits the researcher to delimit the search of matching microarrays to a specific phenotypic change.

### **3.9.3. Crossing the user's microarray marker genes with the database of microarray marker genes**

By comparing the marker genes of the user's sample clusters with the marker genes of the database, the system returns all microarrays with common marker genes with respect to the specific search for the user's microarray. Then, the correspondence between the sample clusters of the user's expression matrix and the sample subsets of the matching-microarrays can be elucidated using the graphical interface (Figure 33). A correspondence can be established between two sample clusters from different microarrays if the common marker genes over-expressed in the sample cluster from one microarray are over-expressed in the sample cluster from the other microarray.

### **3.9.4. Tool's output interfaces**

When our application crosses the marker genes of the user's microarray sample clusters with the database of microarray marker genes, two types of lists are provided.

#### **3.9.4.1. List-of-microarrays view**

The list of matching microarrays is ordered by the number of marker genes in common between the user's microarray specific search and the matching microarrays. The list of microarrays can be filtered by the experimental variable, like “agent”, “dose”, or “time”, in order to delimit the matching microarrays listed to a concrete scope, or by keywords like “breast cancer”. The common marker genes between the user's expression matrix and each matching microarray can be analyzed in the list-of-marker-genes view.

#### **3.9.4.2. List-of-marker-genes view**

The common marker genes between the two microarrays are listed. The sample clusters arrangement along the gene expression is shown for each marker gene and both expression matrices. In this way, the researcher can quickly establish a correspondence between the sample clusters of his/her own microarray and the subsets of the public microarray. The more matching microarrays found, the more attributes that can be assigned to the user's microarray sample clusters. Note that the user's sample clusters could have been calculated by statistical methods and thus, their biological significance is unknown.



**Matching marker genes GDS3003: House dust mite extract effect on a bronchial epithelial cell line**



Gene Name	matching-gds cluster distribution	Dist	user-gds cluster distribution
<a href="#">NNMT: nicotinamide N-methyltransferase</a>		0.318600	
<a href="#">NNMT: nicotinamide N-methyltransferase</a>		0.318600	
<a href="#">MT1X: metallothionein 1X</a>		0.292924	
<a href="#">MT1X: metallothionein 1X</a>		0.292924	
<a href="#">ASNS: asparagine synthetase</a>		0.259311	
<a href="#">MT2A: metallothionein 2A</a>		0.245549	
<a href="#">TFPI2: tissue factor pathway inhibitor 2</a>		0.236698	
<a href="#">TFPI2: tissue factor pathway inhibitor 2</a>		0.236698	
<a href="#">TGFB2: transforming growth factor, beta 2</a>		0.223802	
<a href="#">HIF1A: hypoxia-inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor)</a>		0.211458	

**Figure. 33. The list of marker genes view.** The common differentially-expressed genes between a matching microarray from GEO database and the user's expression data are listed (each row is a different marker gene). The arrangement of the sample clusters along the gene expression is shown for each gene and the sample clusters of the two microarrays (in two bar charts per row). Comparing the two bar charts of each marker gene the researcher can establish the correspondences between the sample clusters of his/her expression matrix and the subsets of the GEO microarray (the sample clusters on the same side in the two bar charts should represent the same phenotype). The right side of the bar charts represents the gene over-expression with respect to their left side. Dist value shows the difference in expression among the user's microarray sample clusters specified in the marker-gene search. The common gene markers are sorted by this Dist value.

As a result, the user can enrich his/her expression data analysis, improve his/her future experimental design and check the hypotheses generated from the data in the ways cited previously and detailed below.

This work has been published in [Huerta M. et al 2014].

**Availability:** Marker-gene database tool: <http://ibb.uab.es/MGDB>

### 3.9.5. Marker-genes database use cases

Samples within GEO datasets are grouped and classified into subsets according to the experimental variables under examination in each study, for instance ‘tissue’ or ‘strain’. How could we establish a correspondence between these sample subsets previously classified by the experimental variables with respect to the sample clusters obtained by statistical methods from the expression matrix we are studying? The user's matrix genes will be marker genes of a public microarray if their expressions display a significant effect in relation to the public microarray subsets. Thus, by verifying that marker genes are the same in both expression matrices we could establish a correspondence between them, because both microarrays could describe the same phenotypic changes performed by the same marker genes.

So, GEO and other expression-data repositories could be reused for subsequent investigations around the world even when their hypotheses under investigation are absolutely different from the originals. It is the aim of our Database of Microarray Marker Genes.

Let's see different research strategies or use cases to take advantage of the MGDB. The use cases are ordered from more generic to more specific.

1. To corroborate or to question the hypothesis pointed out by user's microarray experiments.
2. To study the role of marker genes in the user's expression data in other microarrays.
3. To assign biological meaning to the sample clusters obtained by statistical methods from the user's microarray data.
4. To identify or define the phenotype of different subsets of individuals under investigation.
5. To compare the user's results with other assays in the same species.
6. To compare the user's results with human assays.
7. To broaden the analysis of the user's results with experiments in other species.
8. To broaden the analysis of the user's experiments with other pathologies.
9. To extrapolate prognosis information from other analogue studies.
10. To compare the user's experiments with experiments in the same pathology, but in different tissues.
11. To search for drugs whose effect cause the transition between the phenotypes studied in the user's expression data.
12. To discover undesirable side-effects of a treatment studied in the user's expression data.

#### 3.9.5.1. To corroborate or to question the hypothesis pointed out by user's microarray experiments

In that example, we wish to corroborate if our microarray for studying cancer really differs between cell stress and tumour proliferation. Our microarray under study contains the expression data provided by the National Cancer Institute (USA) [Scherf U. et al 2000] aforementioned (section 2.1).

Our system incorporates the calculation of the most common global-clustering methods to cluster the microarray samples. These sample clusters represent the phenotypes that imply a higher number of coexpressed genes. We have used the Repeated-Bisection analysis to cluster the matrix samples.

In our expression-data analysis we suspect that we have identified the phenotypes linked to two sets of sample clusters, one set (red and yellow clusters) corresponding to stressed phenotype and the other one (green and blue clusters) corresponding to proliferative phenotype. We have identified the sample clusters corresponding to the stressed phenotype because we recognize genes linked to stress that are over-expressed on these clusters, and we have identified the sample clusters corresponding to the proliferative phenotype because we know genes linked to tumour proliferation that are over-expressed on these clusters.

Now, we wish to corroborate our hypothesis comparing our data with several microarrays about stress and cell-proliferation using MGDB queries.

To perform this task, first we make a search of the genes differentially expressed between the sample clusters corresponding to the stressed phenotype and those corresponding to the proliferative phenotype. These differentially-expressed genes will be the marker genes of our study for the current microarray.

**Search genes comparing microarray-condition classes**

Choose the classes to be overexpressed (>) or infoexpressed (<) with respect to the basal value.

1

2

3

4

5

6

7

8

Choose the classes to be disjointed (#), over-expressed (>), or info-expressed (<) with respect to the others.

	1	2	3	4	5	6	7	8
1								
2								
3								
4			#					
5			#					
6			#					
7			#					
8								

alfa : 0.003 [Update your clusters file](#) [Launch Search](#)

**Figure 34. Marker-gene search from the user's expression matrix.** We want to identify those genes that differentially expressed the samples that correspond to the stressed phenotypes from those samples corresponding to a proliferative phenotype. In the menu of the search for marker genes, we require a different expression between red and yellow clusters versus blue and green clusters. These are the marker genes provided by our web-application.

Rank	Dist	Id	Name		
1	0.132317	776	TXNDC5: thioredoxin domain containing 5		
2	0.131321	777	H.sapiens OB-RGRP gene Chr.1 [485960, (EW), 5♣:AA040627, 3♣:AA040165]		
3	0.123048	779	H.sapiens OB-RGRP gene Chr.1 [470194, (E), 5♣:, 3♣:AA030058]		
4	0.121753	722	COL4A1 Collagen, type IV, alpha 1 Chr.13 [489467, (IEW), 5♣:AA054624, 3♣:AA054564]		
5	0.117941	778	H.sapiens OB-RGRP gene Chr.1 [365926, (E), 5♣:, 3♣:AA025885]		
6	0.117036	720	CYR61: cysteine-rich, angiogenic inducer, 61		
7	0.114373	719	ANLN: anillin, actin binding protein		
8	0.111154	775	H.sapiens OB-RGRP gene Chr.1 [265571, (EW), 5♣:N31358, 3♣:N21401]		
9	0.111081	767	PARP3: poly (ADP-ribose) polymerase family, member 3		
10	0.109756	696	FERMT2: fermitin family homolog 2 (Drosophila)		
11	0.108012	745	ITGA3: integrin, alpha 3 (antigen CD49C, alpha 3 subunit of VLA-3 receptor)		
12	0.107836	746	MYH9: myosin, heavy chain 9, non-muscle		
13	0.107084	687	GLIPR1: GLI pathogenesis-related 1 (glioma)		
14	0.106854	721	THBS1: thrombospondin 1		
15	0.106451	768	QSOX1: quiescin Q6 sulfhydryl oxidase 1		
16	0.104872	702	CFL2: cofilin 2 (muscle)		
17	0.104485	714	DUSP1: dual specificity phosphatase 1		
18	0.104225	788	GNAS: GNAS complex locus		
19	0.104210	688	GLIPR1: GLI pathogenesis-related 1 (glioma)		
20	0.104120	742	NNMT: nicotinamide N-methyltransferase		
21	0.103871	534	VGLL3: vestigial like 3 (Drosophila)		
22	0.103678	700	DKK3: dickkopf homolog 3 (Xenopus laevis)		
23	0.103572	728	MYLK: myosin light chain kinase		
24	0.102365	701	SID W 429623, Homo sapiens clone 24659 mRNA sequence [5♣:AA011634, 3♣:AA011635]		
25	0.101931	740	SCYAS Small inducible cytokine A5 (RANTES) Chr.11 [306743, (IW), 5♣:W24183, 3♣:N91767]		

Figure 35. Marker genes obtained as a result of the previous search from our expression data about cancer. The total number of marker genes found for this concrete search is 181. The full list is available in the supplementary material.

Now, we will search for matching microarrays querying the Marker-Genes Database. These matching microarrays will be GEO DataSets that have marker genes in common with our microarray query. So, if the marker genes of the matching microarrays are the same as in our microarray query, the phenotypic changes studied in these microarrays could be also the same than those described by our expression data and may help to confirm our hypothesis.

The more GEO GDSs confirm our hypothesis, the better our hunch will be corroborated. Let's see several examples of microarrays that confirm our hypothesis.

The matching GDS2307 about "Oxidatively modified LDL effect on retinal pigment epithelial cell line" identifies our stress clusters (red and yellow clusters) with the cell-stress phenotype described by these experiments, concretely with oxidative stress.

To perform the correspondence between the sample clusters from both expression matrices, the user only should see the sample clusters over-expressed in his/her expression matrix and the sample subsets over-expressed in the matching microarray (at the right side of the bar charts), because these sample clusters, although from different datasets, will represent the same phenotypes (Figure 33).

Likewise, for the sample clusters under-expressed (at the left side of the bar charts), the user can establish a correspondence between the sample clusters of his/her expression matrix and the sample subsets of the matching microarray. The green triangle in the bar charts points out the basal value. This separates the sample clusters over-expressed from the under-expressed ones. The length of each bar points out the range in expression terms of the sample cluster represented by this bar. The circle on the bars points out the point with higher cluster's samples density. The correspondence between a sample cluster from the user's expression matrix and a sample subset from the matching microarray may be carried out comparing their bars, or the high-density mark of each bar.

In the analyses performed next, the sample clusters are identified by the colour of the bars, both in the user's expression matrix (right column) and in the matching microarray (left column).

**Matching marker genes GDS2307: Oxidatively modified LDL effect on retinal pigment epithelial cell line**



Gene Name	matching-gds cluster distribution	Dist	user-gds cluster distribution
ITGA3: intera <sup>n</sup> , alpha 3 (antigen CD49C, alpha 3 subunit of VLA-3 receptor)		0.329017	
COL4A1 Collagen, type IV, alpha 1 Chr.13 [489467, (IEW), 5⚡:AA054624, 3⚡:AA054564]		0.313831	
COL4A1 Collagen, type IV, alpha 1 Chr.13 [489467, (IEW), 5⚡:AA054624, 3⚡:AA054564]		0.313831	
DUSP1: dual specificity phosphatase 1		0.305825	
MATN2: matrilin 2		0.304930	
CYR61: cysteine-rich, angiogenic inducer, 61		0.298079	
CYR61: cysteine-rich, angiogenic inducer, 61		0.298079	
MT1X: metallothionein 1X		0.292924	
MT1X: metallothionein 1X		0.292924	
THBS1: thrombospondin 1		0.274740	
THBS1: thrombospondin 1		0.274740	
THBS1: thrombospondin 1		0.274740	
MYLK: myosin light chain kinase		0.272782	
MYLK: myosin light chain kinase		0.272782	
COL4A2: collagen, type IV, alpha 2		0.268795	
COL4A2: collagen, type IV, alpha 2		0.268795	
PLK2: polo-like kinase 2 (Drosophila)		0.265910	
JUB: jub, ajuba homolog (Xenopus laevis)		0.257207	
AKR1B1: aldo-keto reductase family 1, member B1 (aldose reductase)		0.253591	
HLA-B: major histocompatibility complex, class I, B		0.249237	

**Figure 36.** GDS2307 about “Oxidatively modified LDL effect on retinal pigment epithelial cell line” is a matching microarray provided by our previous search. It identifies our stress clusters (red and yellow clusters) with the cell-stress phenotype described by their subsets, concretely with oxidative stress. The total number of common marker genes found for this matching GDS is 51. The full list can be compared in the supplementary material.

The matching GDS2307 about “Oxidatively modified LDL effect on retinal pigment epithelial cell line” identifies our stress clusters (red and yellow clusters in the second column) with the cell-stress phenotype, concretely with oxidative stress (yellow cluster in the first column). In GDS2307 lilac cluster is untreated, red cluster is LDL treatment and yellow cluster is ok-LDL treatment (Figure 36).

GDS2307 is part of a study that concluded that the oxidized low density lipoprotein induces transcriptional alterations in genes related to lipid metabolism, oxidative stress and apoptosis. These data support the hypothesis that proteins oxidation could be a trigger for initiating early events in degenerative diseases.

The matching GDS2090 about “Sphingosine 1-phosphate effect on glioblastoma cells” identifies our proliferation clusters (green and blue clusters) with a tumour-proliferation phenotype, even with invasion.

GDS2090 and the associated study allow us to connect the green and blue sample clusters with aggressive tumours, whose growth is modulated by SP1, which could trigger the tumour expansion by inducing the expression of metalloproteases (Figure 37).

**Matching marker genes GDS2090: Sphingosine 1-phosphate effect on glioblastoma cells** 

Gene Name	matching-gds cluster distribution	Dist	user-gds cluster distribution
<a href="#">FERMT2: fermitin family homolog 2 (Drosophila)</a>		0.315349	
<a href="#">GLIPR1: GLI pathogenesis-related 1 (glioma)</a>		0.309935	
<a href="#">DUSP1: dual specificity phosphatase 1</a>		0.305825	
<a href="#">CITED2: Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2</a>		0.302750	
<a href="#">CITED2: Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2</a>		0.302750	
<a href="#">CYR61: cysteine-rich, angiogenic inducer, 61</a>		0.298079	
<a href="#">MT1X: metallothionein 1X</a>		0.292924	
<a href="#">PLK2: polo-like kinase 2 (Drosophila)</a>		0.265910	
<a href="#">FOSL2: FOS-like antigen 2</a>		0.264952	
<a href="#">FOSL2: FOS-like antigen 2</a>		0.264952	
<a href="#">HLA-B: major histocompatibility complex, class I, B</a>		0.249237	
<a href="#">SERPINE1: serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1</a>		0.210577	
<a href="#">SERPINE1: serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1</a>		0.210577	
<a href="#">CAV1: caveolin 1, caveolae protein, 22kDa</a>		0.191832	
<a href="#">RCN1: reticulocalbin 1, EF-hand calcium binding domain</a>		0.149320	
<a href="#">RCN1: reticulocalbin 1, EF-hand calcium binding domain</a>		0.149320	
<a href="#">FLNA: filamin A, alpha (actin binding protein 280)</a>		0.144379	
<a href="#">ZYG: zyxin</a>		0.141069	
<a href="#">ZYG: zyxin</a>		0.141069	
<a href="#">IL6 Interleukin 6 (B cell stimulatory factor 2) Chr.7 [310406, (RW), 5q:W31016, 3q:N98591]</a>		0.130466	

20 genes matched & 98 genes mismatched

**Figure 37.** The matching GDS2090 about “Sphingosine 1-phosphate effect on glioblastoma cells” identify our proliferation clusters (green and blue clusters in the second column) with the tumour-proliferation phenotype, even with invasion (red cluster in the first column). In GDS2090 lilac subset is under the influence of epidermal growth factor (EGF), an important growth factor, and red subset is under SIP administration. The total number of common marker genes found for this matching GDS is 20. The full list can be compared in the supplementary material.

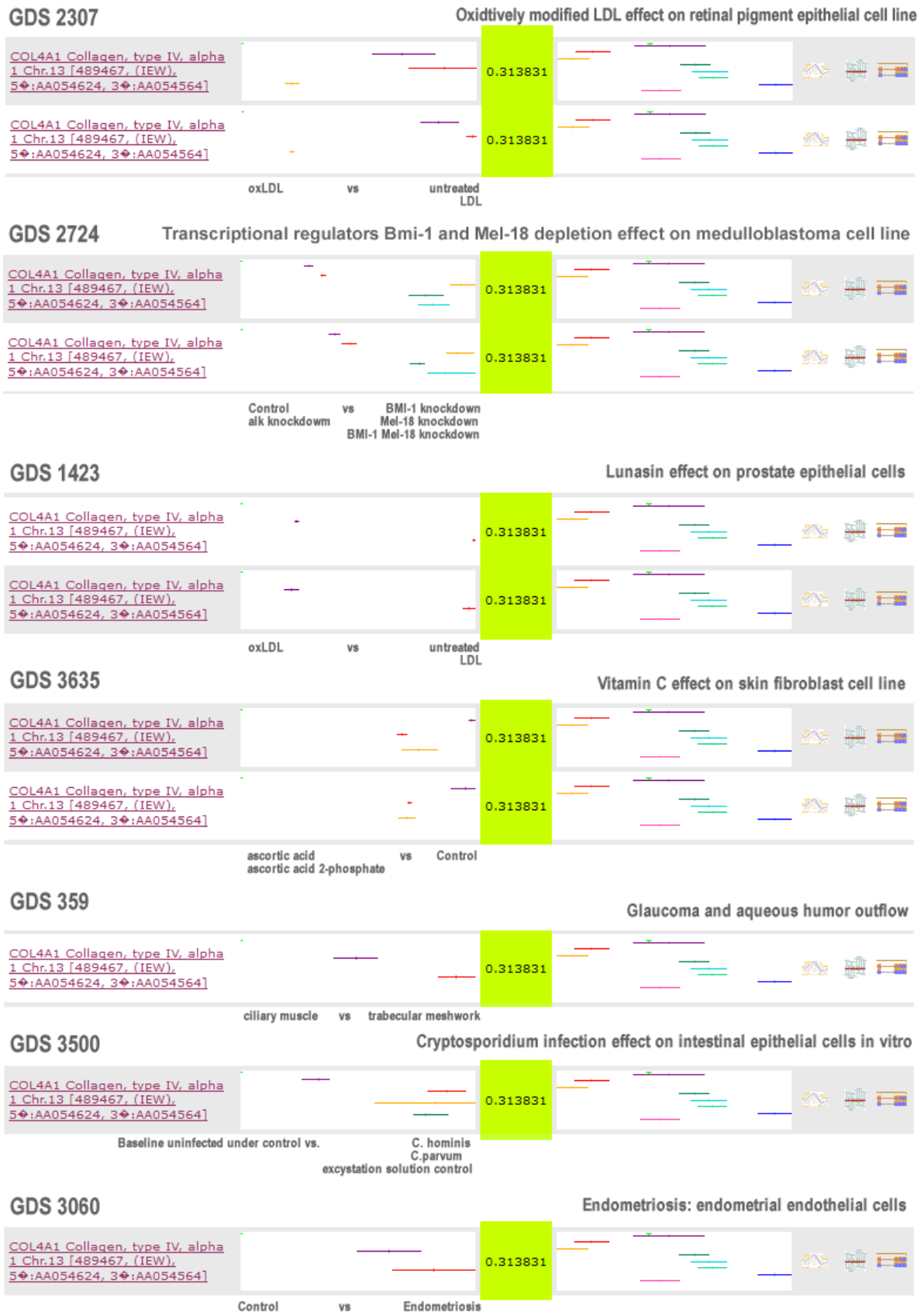


### 3.9.5.2. To study the role of marker genes in the user's expression data in other microarrays

From the above comparison between expression matrices we can observe the common marker genes between both microarrays and identify the role of the marker genes: they can be marker genes of cell stress or marker genes of tumour proliferation.

We are especially interested in one of the marker genes provided by the search in our expression matrix under study: Collagen IV, because it is known to be a gene linked to tumour proliferation, since it is necessary to rebuild the collagen bridges in the process of tissue remodelling.

We searched for COL4A1 in the matching microarrays and we found it in some of the matching GDSs that better discriminate among their sample subsets.



**Figure 38.** We wish to study the role of Collagen IV in matching GDSs because its link to tumour proliferation is well known. Collagen IV gene is a common marker gene in the comparison of some of the matching GDSs that better discriminate their sample subsets. The title of each GDS is shown on the right of the GDSID. The title of each sample subset is shown under the bar chart of each matching GDS.

One way to reach a profound understanding of the function of a gene in our phenotype under study is to exam those GDSs that match our microarray sample clusters and at the same time have this marker gene in common. These could supply important details on how this gene is implicated in cellular/tissular functions. As we examine more matching GDSs, we get to know more about our gene of interest.


The GDSs listed in the Figure 38 describe phenotypic changes involving tissue remodelling, and the Collagen IV gene appears to be involved in these phenotypic changes.

3.9.5.3. To assign biological meaning to the sample clusters obtained by statistical methods from the user's expression data

Looking at the previous comparison with the GDS2090 about "Sphingosine 1-phosphate effect on glioblastoma cells" of the case 1, we could assign attributes to the proliferative sample clusters (green and blue clusters) of our microarray like: invasion, migration, cell survival... (Figure 37) On the other hand, looking at the comparison with the GDS2307 about "Oxidatively modified LDL effect on retinal pigment epithelial cell line" of the case 1, we could assign attributes to the stress sample clusters (red and yellow clusters) of our microarray like: oxidative stress, apoptosis or future inflammation (Figure 36).

As we examine more matching GDSs we can assign more attributes to our clusters.

For instance, if we examine the comparison with the GDS2471 about "Serum amyloid A effect on endothelial cells: time course" we can assign new attributes to the proliferation sample clusters like matrix degradation (Figure 39).

**Matching marker genes GDS2471: Serum amyloid A effect on endothelial cells: time course** 

Gene Name	matching-gds cluster distribution	Dist	user-gds cluster distribution
<a href="#">NNMT: nicotinamide N-methyltransferase</a>		0.318600	
<a href="#">FERMT2: fermitin family homolog 2 (Drosophila)</a>		0.315349	
<a href="#">GLIPR1: GLI pathogenesis-related 1 (glioma)</a>		0.309935	
<a href="#">DKK3: dickkopf homolog 3 (Xenopus laevis)</a>		0.305348	
<a href="#">IL32: interleukin 32</a>		0.275290	
<a href="#">HLA-B: major histocompatibility complex, class I, B</a>		0.249237	
<a href="#">HLA-B: major histocompatibility complex, class I, B</a>		0.249237	
<a href="#">HLA-B: major histocompatibility complex, class I, B</a>		0.249237	
<a href="#">ALCAM: activated leukocyte cell adhesion molecule</a>		0.242438	
<a href="#">TFPI2: tissue factor pathway inhibitor 2</a>		0.236698	
<a href="#">EFEMP1: EGF-containing fibulin-like extracellular matrix protein 1</a>		0.235543	
<a href="#">TPM1: tropomyosin 1 (alpha)</a>		0.233620	
<a href="#">RBMS1: RNA binding motif, single stranded interacting protein 1</a>		0.228108	
<a href="#">PRSS23: protease, serine, 23</a>		0.219008	
<a href="#">PRICKLE1: prickle homolog 1 (Drosophila)</a>		0.200081	
<a href="#">PLOD2: procollagen-lysine, 2-oxoglutarate 5-dioxygenase 2</a>		0.196676	
<a href="#">PLOD2: procollagen-lysine, 2-oxoglutarate 5-dioxygenase 2</a>		0.196676	
<a href="#">CAV1: caveolin 1, caveolae protein, 22kDa</a>		0.191832	
<a href="#">UGCG: UDP-glucose ceramide glucosyltransferase</a>		0.175710	
<a href="#">ITGB1: integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12)</a>		0.168013	

**Figure 39.** In the comparison with GDS2471 about “Serum amyloid A effect on endothelial cells: time course” we could assign new attributes to the proliferation clusters (green and blue clusters in the second column) like matrix degradation (red cluster in the first column). In GDS2471 lilac cluster is untreated and red cluster is treated with Serum amyloid A. The total number of common marker genes found for this matching GDS is 26. The full list can be compared in the supplementary material.

#### 3.9.5.4. To identify or define the phenotype of different subsets of individuals under investigation

The process is exactly the same as the one followed in previous cases, and only the user's expression data to be analyzed is different. In the case that the samples of the user's expression data were patients, clustering would detect patients sharing the same phenotype. So, if we look for the marker genes which separate these patient groups and seek the GEO GDSs with the same marker genes, we could establish the correspondence between the subsets of the matching GDS and our patient groups. Then, we could assign the phenotypes of these subsets to the patient groups (as a starting hypothesis for further investigations).

The more matching GDSs assigning the same or similar phenotypes to the patient groups, the greater is the actual correspondence between patient groups and the phenotypes we try to relate.

#### 3.9.5.5. To compare the user's results with other assays in the same species

Our expression data and the matching GDSs compared in cases 1 and 3 are all human experiments. We could restrict the search for matching GDSs to a concrete taxonomy.

#### 3.9.5.6. To compare the user's results with human assays

The matching GDSs analyzed in cases 1 and 3 are all human experiments. But the purpose of this strategy is to search, for example, whether the results of our experiments in mice could be extrapolated to humans or if otherwise, the phenotypes pointed out by the same marker genes appear to be completely different in humans. This strategy can include physiological differences, for example, if we want to discover if there is a difference between the phenotypes mediated by genes that control the circadian rhythm in humans with respect to our experiments with mice (a nocturnal animal).

#### 3.9.5.7. To broaden the analysis of the user's results with experiments in other species

The purpose of this case of study is the opposite of the previous one. We want to know whether the phenotypic changes studied in the user's expression data (human) are also repeated in other species experiments.

This strategy is very interesting to detect different metabolic pathways that carry out the phenotypic change we are studying in our expression data. We achieve this goal by comparing our microarray

with animal-testing GDSs that study the effect of gene knockdown and similar procedures. If the phenotype resulting from the knockdown is the same as the phenotype observed in our expression data, then the metabolic pathway could be involved in our phenotypic change under study. This is especially useful if we are searching for a way to reverse the phenotypic change.

Focusing on our study about stress and tumour proliferation, we could examine the comparison of our microarray with GDS483 about "DACH1-responsive genes" in *mus musculus*. GDS483 forces an over-expression of DACH1 gene induced by ponasterone A. This comparison identifies the phenotype with higher DACH1 activity with our proliferative phenotype (green and blue clusters). This could indicate a possible metabolic pathway for our tumour-proliferation phenotype, especially if we consider that DACH1 is an important regulatory gene in development and was found to inhibit (TGF-beta)-induced apoptosis (Figure 40).

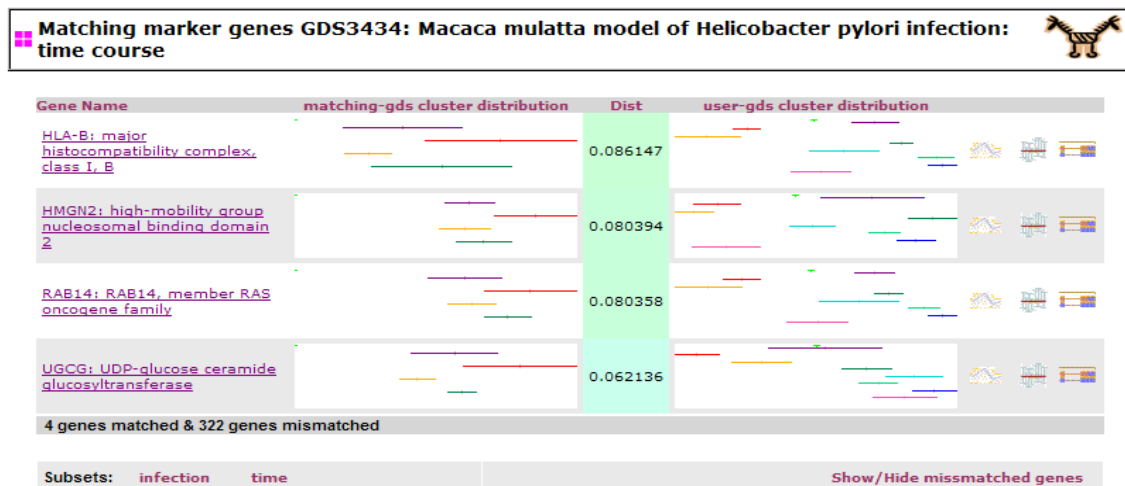
Matching marker genes GDS483: DACH1-responsive genes



Gene Name	matching-gds cluster distribution	Dist	user-gds cluster distribution
<a href="#">CYR61: cysteine-rich, angiogenic inducer, 61</a>		0.117036	
<a href="#">CYR61: cysteine-rich, angiogenic inducer, 61</a>		0.117036	
<a href="#">ITGA3: integrin, alpha 3 (antigen CD49C, alpha 3 subunit of VLA-3 receptor)</a>		0.108012	
<a href="#">DUSP1: dual specificity phosphatase 1</a>		0.104485	
<a href="#">DUSP1: dual specificity phosphatase 1</a>		0.104485	
<a href="#">MT1X: metallothionein 1X</a>		0.100860	
<a href="#">MT1X: metallothionein 1X</a>		0.100860	
<a href="#">RXRA: retinoid X receptor, alpha</a>		0.096497	
<a href="#">MT2A: metallothionein 2A</a>		0.094362	
<a href="#">AKR1B1: aldo-keto reductase family 1, member B1 (aldose reductase)</a>		0.078117	
<a href="#">PLOD2: procollagen-lysine, 2-oxoglutarate 5-dioxygenase 2</a>		0.074315	
<a href="#">HIF1A: hypoxia-inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor)</a>		0.073965	
<a href="#">SERPINE1: serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1</a>		0.071708	
<a href="#">SERPINE1: serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1</a>		0.071708	
<a href="#">UGCG: UDP-glucose ceramide glucosyltransferase</a>		0.062136	
<b>15 genes matched &amp; 213 genes mismatched</b>			
Subsets: <b>time</b>		<a href="#">Show/Hide mismatched genes</a>	

**Figure 40.** GDS483 forces an over-expression of DACH1 gene induced by ponasterone A. This comparison identifies the phenotype with higher DACH1 activity (red cluster in the first column) with our proliferative phenotype (green and blue clusters in the second column). This could indicate a possible metabolic way for our tumour-proliferative phenotype. In GDS483 lilac cluster is 0 hours, red cluster is 18 hours and yellow cluster is 36 hours. The total number of common marker genes found for this matching GDS is 15. The full list can be compared in the supplementary material.

Another type of experiment that could be carried out only on animals, but not on humans, is the response to virus or other lethal experimental conditions. In the comparison with GDS3434 entitled “Macaca mulatta model of Helicobacter pylori infection: time course” we can observe a correspondence between the response to H. pylori infection and our proliferative phenotype (green and blue clusters). However we could observe that the macaques with a PAI isogenic knockout attenuate the H. Pilyry host response, indicating that the antimicrobial response could be mediated in a cag-pathogenicity island-dependent manner. This shows a possible way to develop our tumour-proliferation phenotype, at least in gastric mucosa (Figure 41).



**Figure 41.** In the comparison with GDS3434 entitled “Macaca mulatta model of Helicobacter pylori infection: time course” we could observe a correspondence between the response to H. pylori infection (red cluster in the first column) and our proliferative phenotype (green and blue clusters in the second column). The cag-PAI isogenic knockout revert the H. pylori infection phenotype (lilac subset in the first column). In GDS3434 lilac cluster is H. pylori - PAI isogenic mutant from 1 to 13 weeks, the red cluster is H. pylori infection from 1 to 13 weeks, the yellow cluster is the mock-infected control group and the green subset is the pre-inoculation control group.

### 3.9.5.8. To broaden the analysis of the user’s experiments with other pathologies

The matching GDS2471 compared in case 3 (Figure 39) is part of a coronary artery disease experiment. This matching GDS allows us to compare the phenotypes of different pathologies (its coronary-artery disease and our cancer).

This pathology is not as usual as cancer but could be a clue for searching common treatments, for instance, by their effect on the cell-matrix degradation.

This strategy is very useful because in some cases, apparently very different pathologies share similar phenotypes, and this strategy could reveal the opportunity of applying a specific pathology treatment in other different pathologies.



### 3.9.5.9. To extrapolate prognosis information from other analogue studies

This strategy is focused on the comparison between the phenotypes studied in the user's expression data with matching GDSs linked to prognosis.

This is the most fruitless search of matching microarrays, because the researcher's searches for his/her phenotypes matching different phases of a disease or time series on a GEO GDS can hardly be expected to produce results, but it is also one of the most useful, if you are fortunate, and find this type of matching. If you find it, you not only can assign a stage of a disease to your phenotype, but you can also predict what will happen in the future (the incoming phenotype).

Comparison with microarray time series and forecasting could also serve for studying the order in which the phenotypic changes studied in your expression data occur. That is, the previous phenotype and the subsequently phenotype, plus the intermediate phenotypic stages.

In the example of case 1, GDS2090 about "Sphingosine 1-phosphate effect on glioblastoma cells" is contrasted with the stress and proliferative sample clusters of our expression data (Figure 37). In this GDS there are no future worst stages (like in our data), but the matching microarray could provide us information about the progression of the proliferative phenotype studied in our expression data, including invasion of surrounding tissues.

### 3.9.5.10. To compare the user's experiments with experiments in the same pathology, but in different tissues

We examine the comparison of our expression data with GDS3060 about "Endometriosis: endometrial endothelial cells" (Figure 42) and with GDS2090 about "Sphingosine 1-phosphate effect on glioblastoma cells" (Figure 37) which shows us that cancer behaviour maintains common features along different tissues. Because our proliferative sample clusters correspond, at least in part, with the proliferative phenotypes of GDS3060 (endometriosis) and GDS2090 ( glioblastoma). Note that the common marker genes point out the part of the phenotypes in common.

This strategy is interesting in pathologies like cancer where the pathology behaviour can be the same or can be different depending on the tissue affected.

**Matching marker genes GDS3060: Endometriosis: endometrial endothelial cells**



Gene Name	matching-gds cluster distribution	Dist	user-gds cluster distribution
<a href="#">FERMT2: fermitin family homolog 2 (Drosophila)</a>		0.315349	
<a href="#">COL4A1: Collagen, type IV, alpha 1 Chr.13 [489467, (FEW), 5⚡:AA054624, 3⚡:AA054564]</a>		0.313831	
<a href="#">DKK3: dickkopf homolog 3 (Xenopus laevis)</a>		0.305348	
<a href="#">MYLK: myosin light chain kinase</a>		0.272782	
<a href="#">MYLK: myosin light chain kinase</a>		0.272782	
<a href="#">TPM1: tropomyosin 1 (alpha)</a>		0.233620	
<a href="#">PRSS23: protease, serine, 23</a>		0.219008	
<a href="#">SERPINE1: serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1</a>		0.210577	
<a href="#">SERPINE1: serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1</a>		0.210577	
<a href="#">PLAUR: plasminogen activator, urokinase receptor</a>		0.205673	
<a href="#">PLAUR: plasminogen activator, urokinase receptor</a>		0.205673	
<a href="#">PLOD2: procollagen-lysine, 2-oxoglutarate 5-dioxygenase 2</a>		0.196676	
<a href="#">CAV1: caveolin 1, caveolae protein, 22kDa</a>		0.191832	
<a href="#">CAV1: caveolin 1, caveolae protein, 22kDa</a>		0.191832	
<a href="#">MMP2: matrix metalloproteinase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)</a>		0.191385	
<a href="#">ITGB1: integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12)</a>		0.168013	
<a href="#">PAPSS2: 3⚡-phosphoadenosine 5⚡-phosphosulfate synthase 2</a>		0.161128	
<a href="#">IL6: Interleukin 6 (B cell stimulatory factor 2) Chr.7 [310406, (RW), 5⚡:W31016, 3⚡:N98591]</a>		0.130466	
<b>18 genes matched &amp; 270 genes mismatched</b>			
Subsets: <b>disease state</b>		<a href="#">Show/Hide mismatched genes</a>	

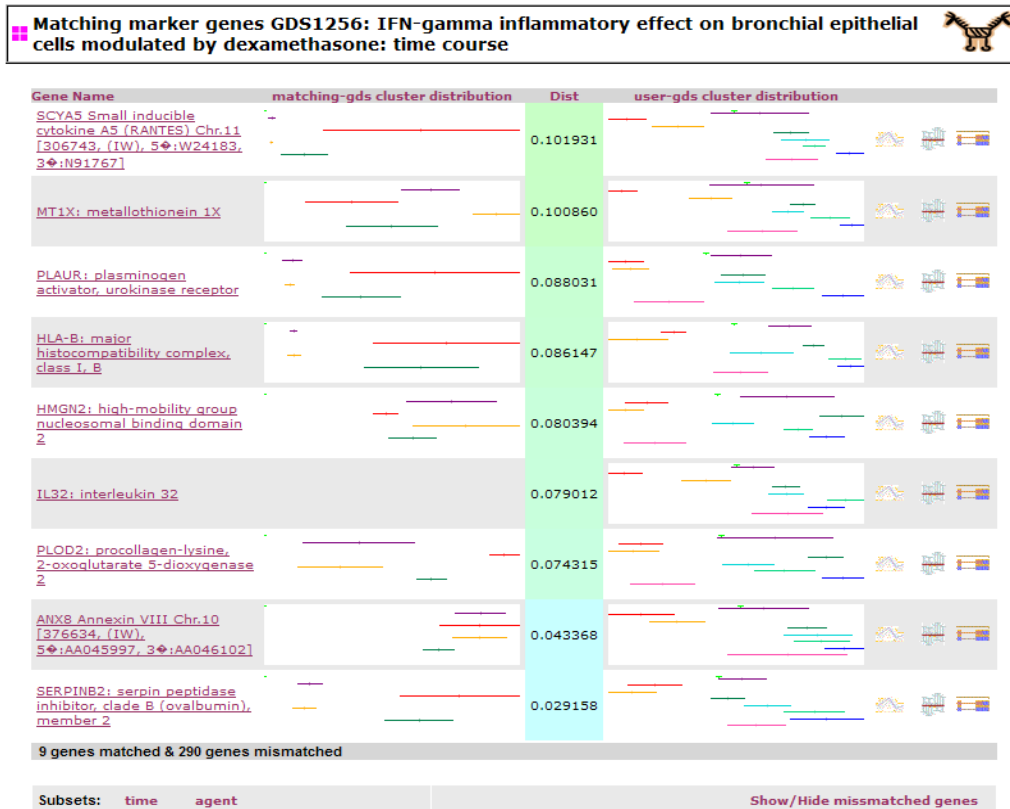
**Figure 42.** The comparison of our expression data with GDS3060 about “Endometriosis: endometrial endothelial cells” or with GDS2090 about “Sphingosine 1-phosphate effect on glioblastoma cells” (Figure 37 of case 1) shows us that cancer general behaviour remains common among different tissues. In both GDSs the proliferative phenotypes from two different tissues, correspond with our proliferative phenotype (green and blue clusters). In GDS3060 (shown in the figure) lilac cluster is the control group and red cluster is endometriosis group. The total number of common marker genes found for this matching GDS is 18. The full list can be compared in the supplementary material.

3.9.5.11. To search for drugs whose effect cause the transition between the phenotypes studied in the user's expression data

In our study, we are interested in finding drugs that revert the proliferative phenotype of cancer towards the stress phenotype.

Filtering the matching gds by agent subset type, we found GDS1256 entitled "IFN-gamma inflammatory effect on bronchial epithelial cells modulated by dexamethasone: time course" (Figure 43). In this microarray study, whereas the IFN-gamma would cause a phenotype similar to our proliferative phenotype, dexamethasone would partially reverse this process. In the comparison with our expression data, the red subset (IFN-gamma) in GDS1256 would correspond with our proliferative phenotype (Green and blue clusters) and the application of dexamethasone after IFN-gamma (green subset) could correspond with our stress phenotype (yellow and red clusters) or a similar phenotype (because dexamethasone plus IFN-gamma produce different effects on gene expression than dexamethasone alone). In either case, this comparison aims to investigate dexamethasone as a possible drug to force the phenotypic change we desire.

In the next sections we will disclose very interesting findings about the effect of dexamethasone on cancer by means of an in-depth analysis (also using bioinformatics tools).



**Figure 43.** Comparison with GDS1256: in the comparison of both microarrays the red subset(IFN-gamma) in GDS1256 would correspond with our proliferative phenotype (Green and blue clusters in the second column) and the application of dexamethasone after IFN-gamma (green subset in the first column) could correspond with our stress phenotype (yellow and red clusters in the second column). This comparison aims to investigate dexamethasone as a possible drug to force the phenotypic change we desire. In GDS1256 lilac subset is control group, yellow subset is dexamethasone, red subset is IFN-gamma and green subset is IFN-gamma plus dexamethasone.

### 3.9.5.12. To discover undesirable side-effects of a treatment studied in the user's microarray

The expression data used to illustrate the different cases of use are data about treatments in cancer. However, it is not the type of expression data for which this strategy n° 12 is specifically addressed. This strategy is especially useful when we have control versus a treatment, or control versus different doses of a treatment. In such cases, if our subsets of treated samples match with a pathologic phenotype in a GEO GDS, while our subsets of control or applying a lower dose match with an absence of pathology phenotype, we can then suspect possible secondary effects of our treatment.

Focusing on the expression data we are analyzing, we have previously seen in the last case of use that dexamethasone could interrupt the tumour proliferation to stress the tissue. Now we are interested in knowing if the post-drug administration phenotype could imply undesirable side effects.

Studying the matching GDSs with this purpose, we found the GDS2307 about "Oxidatively modified LDL effect on retinal pigment epithelial cell line" previously seen (Figure 36), and we realised that dexamethasone, when changes the cell from the proliferative phenotype to the stress phenotype, according to the GDS2307 comparison, could also imply some side-effects like: oxidative stress, apoptosis, future inflammations, future degenerative diseases, etc., as described by the GDS2307 study.

Now the researcher must consider if these side effects could be serious enough to study them in depth.

### **3.9.6. Comparison of online tools to search microarrays of interest from marker genes**

Several on-line tools that can be used to search for matching microarrays have been compared with our marker-genes database. To perform this comparison we once again apply the `at_matrix` used in the previous use cases. But now, we would like to find those microarrays which distinguish the two sample clusters that represent the tumoural proliferation in our expression data: blue clusters with respect to green clusters. The first step is to search for the marker genes which distinguish both phenotypes. We will use these marker genes in the different tools for searching for matching microarrays and compare the results.

Symbols of the marker genes have been inserted in each one of the online tools and then microarrays with the same marker genes have been obtained. i.e., these genes will be both marker genes of the phenotypes studied in our expression data and the phenotypes studied in the matching microarrays.

The analyzed online tools are: Marker Genes Database (MGDB), MicroArray Rank Query (MarQ) [Vazquez M. et al 2010], GEOGLE [Yu Y. et al 2009], Gene Change Browser (GeneChaser) [Chen R. et al 2008], EXALT [Qiu Q. et al 2013] and Gene Expression data Mining Toward Relevant Network Discovery (GEM-TREND) [Feng C. et al 2009].

#### **3.9.6.1. Interface analysis: operability, usability and functionality**

After comparing the different tools, the most appropriate ones for our study are determined: MGDB, MarQ and GEOGLE. MGDB turns out to be the most functional tool for our type of analysis and also the one with better usability and operability.

Amongst the different online tools, MGDB is the only one that allows one to search for marker genes by expression difference between sample clusters from the user's expression data. MGDB online tool provides then, a visual comparison between these sample clusters and the subsets of the microarrays with marker genes in common. Thus, it allows making a correspondence between the sample clusters of both microarrays for all the common marker genes. In addition to this, it allows

the search for microarrays with common marker genes for multiple species simultaneously. Furthermore, MGDB interface is very easy to use and the seek time for matching-microarray search is short. It also allows one to easily visualise the description of the microarrays with common marker genes.

MarQ does not allow making a visual comparison between the sample clusters of the microarrays with common marker genes. On the other hand, its interface is very easy to use and the seek time for matching microarray search is not too long. However, it seems to be difficult to visualize the descriptions of the obtained matching microarrays.

MarQ online tool has provided the largest number of microarrays with common marker genes (713 microarrays). The criteria used by MarQ to consider which marker genes discriminate between subsets in each GEO microarray are possibly more permissive than those used by MGDB (340 microarrays found) and GEOGLE (350 microarrays found).

GEOGLE online tool allowed to find a number of matching microarrays similar to MGDB, although a 59,43% (208 microarrays) of the matching microarrays found are not the same in both tools. Most of the matching microarrays found by GEOGLE have been also detected by MarQ, those also detected by MGDB are not so many (320 microarrays of MarQ in common with GEOGLE and 257 microarrays in common with MGDB). As well as MarQ, GEOGLE does not allow to make a visual comparison between the sample clusters of the microarrays with common marker genes. However, it does allow correctly visualizing the descriptions of the microarrays with marker genes in common. The seek time for matching-microarray search is not too long. Nevertheless the GEOGLE on-line tool demands the user to start a short session which is somewhat uncomfortable.

The less appropriate tools for our type of analysis are Genechaser, EXALT and GEM-TREND.

Genechaser is not functional to search for matching microarrays to study our phenotypes of interest. Genechaser requires microarrays to content all the marker genes in common to be considered matching microarrays. So, if just one of the marker genes in the user's expression data is not also a marker gene in the GDS, this GDS is considered as not valid. Genechaser is useful for making a matching-microarray search with only one marker gene or a few marker genes, where this marker gene is the focus of the investigation. Instead, it is not recommendable to compare different microarrays and their respective phenotypes.

EXALT and GEM-TREND have been excluded because they are not enough operational tools that are operational enough for our type of study. EXALT has problems connecting with its server and does not allow to make a search for microarrays with common marker genes. GEM-TREND has a difficult data-entry system for users. This system demands to insert the symbols of the marker genes in a document. If the text is not recognised by the tool, the search for matching microarrays is not done.

### 3.9.6.2. Statistical analysis of the results provided by different tools

The search for marker genes using MGDB allowed finding 56 marker genes. Some of these genes have been excluded (SID W 487351, SID 29828, SID W 346334, ESTs, SID W 115769, ESTsSID 116296, ESTsSID 361648, Glutathione S-Tranferase p-log Correlation factor, Thioredoxin Reductase mRNA-log Correlation factor and Raf-1-log) for incompatibility in the search for matching microarrays using the other tools, remaining 46 marker genes. Most of the 46 remaining marker genes are marker genes for the blue sample-clusters proliferation.

Marker genes marking the blue type proliferation:

STX3, NEDD4L, C5orf32, CD9, SLC25A24, TSPAN6, CDH17, PTK2, NDRG2, ARHGEF12, SH3D19, TMBIM1, LGALS3, FCHO2, CTDSPL, GALNT3, DECR1, ERMP1, SLC25A13, CTTN, LOC151162, TSC22D1, LTBR, DDAH1, KLF3, PERP, MFAP4, GULP1, EPS8, LGMN, NRBP2, SYK, FAM3C, DDEF2, RHOBTB3, AZGP1 e ITGA6.

Marker genes marking the green type proliferation:

DCTN4, AIF1, FADS2, NAP1L1, IRX3, TDG, SLC15A4, RUNX1 y CENPF.

Regarding the analysed species in the comparison of online tools, the fact that MGDB includes microarrays of other species and MarQ and GEOGLE do not, is not apparently relevant for the coincidence in the results of the different tools. MGDB provides 5 matching microarrays of non human species (GDS3267 *Canis lupus familiaris*, GDS2991 y GDS3623 *Danio rerio* y GDS1299 and GDS35 *Saccharomyces cerevisiae*) while the other 335 microarrays belong to *Homo sapiens*. The microarrays compared with MarQ as well as those compared with GEOGLE belong to *Homo sapiens*. (These tests have been performed before introducing the homologene option for the comparison among microarrays of different species in MGDB).

Genes which expressions separate green and blue type proliferation used in MarQ and GEOGLE:							46
STX3	NEDD4L	C5orf32	CD9	SLC25A24	TSPAN6	CDH17	
PTK2	NDRG2	ARHGEF12	SH3D19	TMBIM1	LGALS3	FCHO2	
CTDSPL	GALNT3	DECR1	ERMP1	SLC25A13	CTTN	LOC151162	
TSC22D1	LTBR	DDAH1	KLF3	PERP	MFAP4	GULP1	
EPS8	LGMN	NRBP2	SYK	FAM3C	DDEF2	RHOBTB3	
AZGP1	ITGA6	DCTN4	AIF1	FADS2	NAP1L1	IRX3	
TDG	SLC15A4	RUNX1	CENPF				

Table IV. Marker genes used in the search for matching microarrays in MarQ and GEOGLE.

Online tools				
	Coincidences			
	nº GDS	GDS MGDB	GDS MarQ	GDS Geogle
<b>MGDB</b>	<b>340</b>	–	<b>257</b>	<b>142</b>
<b>MarQ</b>	713	257	–	320
<b>Geogle</b>	350	142	320	–

Table V. Number of matching microarrays (with common marker genes) in common between the different online tools.

From the matching microarrays detected by MGDB (340 microarrays), a 41,76% (142 microarrays) have been detected by GEOGLE and a 75,59% (257 microarrays) have been detected by MarQ.

Coincidences of MGDB with respect to:	
MarQ-MGDB	<b>75,59%</b>
GEOGLE-MGDB	<b>41,76%</b>

Table VI. Percentage of matching microarrays (with common marker genes) found by MGDB and the rest of tools.

From the microarrays detected by MarQ (713 microarrays), a 36,04% (257 microarrays) have been also detected by MGDB and a 44,88% (320 microarrays) have been also detected by GEOGLE.

Coincidence of MarQ with respect to:	
MGDB-MarQ	<b>36,04%</b>
GEOGLE-MarQ	<b>44,88%</b>

Table VII. Percentage of matching microarrays (with common marker genes) found by MarQ coincident with the rest of tools.

Finally, from the microarrays detected by GEOGLE (350 microarrays), a 40,57% (142 microarrays) have been detected by MGDB and a 91,43% (320 microarrays) have been detected by MarQ.

Coincidence of Geogle with respect to:	
MGDB-GEOGLE	<b>40,57%</b>
MarQ-GEOGLE	<b>91,43%</b>

Table VIII. Percentage of matching microarrays (with common marker genes) found by GEOGLE coincident with the rest of tools.



A total of 142 matching microarrays have been detected in common by all the tools.

### 3.10. Studying the transition between phenotypes using non-linear expression relationships to solve paradoxes in cancer marker genes

A phenotype is a consequence of the interaction among genetic and environmental factors. There are several thousands of disorders caused by single genes, in these cases the variation in certain traits or cellular functions can be controlled by single-expression variations. But this is not the general rule. Usually, traits or cellular functions are controlled by multiple genes, and a combination of multiple traits or cellular functions lead to the final phenotype. As aforementioned, only by using specific tools to study the dependence among phenotypes is it possible to deal with this complexity. This is the case of our study about the dual effect of glucocorticoids. The dramatic increase of expression sample series motivates a more subtle analysis of gene dependences throughout these large series in order to infer phenotypes from gene behaviour. There are several analytical methods that perform a global clustering of the microarray samples, such as Self-Organizing Maps [Kohonen T. and Somervuo P. 2002], or which perform local clusterings by considering only a subset of coexpressed genes, such as Biclustering [Wu C.J. and Kasif S. 2002], and so on. All of them build sample clusters based on gene-expression levels. The methodology developed improves these clustering methods because it enriches the final study of the generated sample clusters and also directs this study to the researcher's objectives.

Initially, our system detects different phenotypic changes involving a limited number of relevant genes which are greatly affected by these phenotypic changes. This detection is not based upon the study of gene-expression levels; instead, it is based upon: (i) the full expression-dependence among genes, and (ii) expression-dependence fluctuations. Finally, the phenotypic changes found by our system are linked with the sample clusters obtained by the different classical clustering methods. This fact will facilitate the subsequent study of complex phenotype interdependences.

Our methodology to detect phenotypic changes is based upon the following two principles: First, since the different phenotypes are performed by different sets of coexpressed genes not linearly correlated among each other, the interdependence among these phenotypes cannot be described by a linear-expression relationship. Second, each fluctuation in a correlated expression relationship implies a phenotype change. Whereas the classical global and local clustering methods are focused on the detection of coexpressed genes, the proposed strategy is focused on the detection of non-linear expression relationships between sets of coexpressed genes, because these relationships describe the interdependence among the phenotypes.

Accordingly, we have incorporated a tool for the detection and classification of non-linear expression relationships in our web-application server for micorarray analysis [Cedano J. et al 2008; Huerta M. et al 2009; Cedano J. et al 2007], and on the other hand, we have incorporated the most common clustering methods applied to sample series [Yin L. et al 2006]. The clustering output data are represented together with the detected non-linear expression relationships with the aim to link both (Figure 45 of the next section). Furthermore, the system allows for crossing the results with GO, KEGG, PubMed, and other biomedical DB, to direct the final study to genes related to the researcher's objectives [Huerta M. et al 2009]. As a result, an accurate dissection of phenotype

interdependences can be performed, and this dissection can be linked to biomedical information and current bibliography to focus the data analysis on the researcher's area of interest. The conclusions that can be extracted from each work will depend in a great extent on the gene-expression matrix used as a base for these analyses and on how well this matrix represents the phenotypes that the researchers wish to study.

### 3.10.1. Procedure that can be followed to solve different paradoxes

The procedure followed to achieve our goals is: (i) to describe the different phenotypes, (ii) to study the glucocorticoid effect in each phenotype and the phenotype transitions promoted by glucocorticoid activity and, finally (iii) to discover if the difference in the effect of glucocorticoids in different tumour tissues lies in the phenotype that appears in each tumour tissue. As we finally confirm after our investigations. To achieve this, the expression-matrix genes related to glucocorticoids are identified. The links between glucocorticoid activity and the expression levels of these genes, and their variation, are obtained by the tools used for crossing databases. Then, the effect of applying glucocorticoids to each phenotype is studied from the non-linear expression relationships between the genes related to glucocorticoids, observing the variations in expression after the glucocorticoids supply pointed out by the literature. The details of the followed methodology are explained below.

#### 3.10.1.1. Classifying the expression-relationships by typology

The curvature points of each expression-relationship are obtained from the POPs that describe the relationship's inner pattern. In this manner, the expression-relationships are classified by typology based on these curvature points. This classification is a key point of the phenotype-interdependence analysis, given that each expression-relationship typology represents a concrete interdependence among phenotypes. Note that, in this way, for each non-regulatory gene we know its role in all phenotypic changes it is involved in.

For our study of the contradictory effect of glucocorticoid administration in cancer, we search for complex expression relationships that can help to understand this dual behaviour. If there would be no paradox, the relationship between the drug administration (or another source of interest) and the tumoural proliferation would be linear, as well as the expression relationship between the marker genes of drug activity and the marker genes of tumoural proliferation. But in the case of the administration of glucocorticoids, the relationship between the drug administration and the tumoural proliferation is non-linear, so the expression relationship between these genes should be non-linear too. This is why detecting these non-linear expression relationships helps us to solve the paradoxes.

As aforementioned, curve types like  $y^2+x^2=1$  or  $y=x^2$  could detect dual behaviours of genes, because the over-expression/under-expression of a gene could imply a different phenotype depending on the other gene's expression level. Therefore, these will be the expression-relationships in which we will focus our research.

A variation in the expression dependence implies a phenotypic change. However, from an expression relationship between a pair of genes we do not know if its phenotypic changes affect the expression of the rest of the matrix genes. To discover it, we can relate this expression dependence with the sample clusters, which describe the global phenotypes of the expression matrix.

### 3.10.1.2. Applying sample clusters to non-linear expression relationships

For each uploaded expression matrix, the server calculates the most common clustering analysis applied to expression data: HC, SOTA, SOM, PAM, combined with the most common dimension-reduction analysis: PC and multidimensional scaling. The clustering parameters are adjusted by evaluating the results by means of the most common integrity analysis: Silhouette and Dunn. The sample clusters are calculated for all matrix genes and represent the different phenotypes of the expression matrix.

The graphic interface (the aforementioned expression-relationship detailed view) facilitates the analysis of the distribution of the sample clusters throughout the expression relationships [Cedano J. et al 2007]. Thus, we can study the links between the global phenotypes and the phenotypic changes described by the non-linear expression relationships (as is shown in the figures below). What brings significance to the current study of non-linear expression relationships is the fact that non-linear expression relationships distribute the clusters along their phenotypic changes (at both sides of the inflection point) if these phenotypic changes are global phenotypic changes. This fact helps us to relate the matrix phenotypes among them (and to study the global phenotypic changes).

For our study of the contradictory effect of glucocorticoids administration in cancer, we apply the Repeated-Bisection method to cluster the samples.

### 3.10.1.3. On-line tools

Once the expression-matrix high-throughput analyses are performed, the on-line tools lead the researchers to focus their study in their topics of interest [CedanoJ. et al 2008; Huerta M. et al 2009]. To facilitate this, the system allows crossing the results of the highthroughput analysis with GO, KEGG, PubMed, and other Biomedical databases [CedanoJ. et al 2008; Huerta M. et al 2009].

In the current research, this facility is used to focus our study on genes related to the glucocorticoid signal in tumours and on tissue physiology (healthy and tumoural). Furthermore, this tool facility helps to contrast our findings with the current bibliography.

Once the expression relationships that involve genes related with the drug activity are provided, we can study the phenotypic changes resulting from the drug administration based on the effect of the drug on the expression level of these genes related with the drug. Note that these expression variations must be previously known, since they should be provided by the literature. The gene-

expression variations provided by the literature point out the direction of the phenotypic changes described by the expression relationship. Note that both starting phenotypes and resulting phenotypes after applying the drug must be well described in the expression data and must be detectable by clustering methods.

### 3.10.2. How our procedure to solve paradoxes works

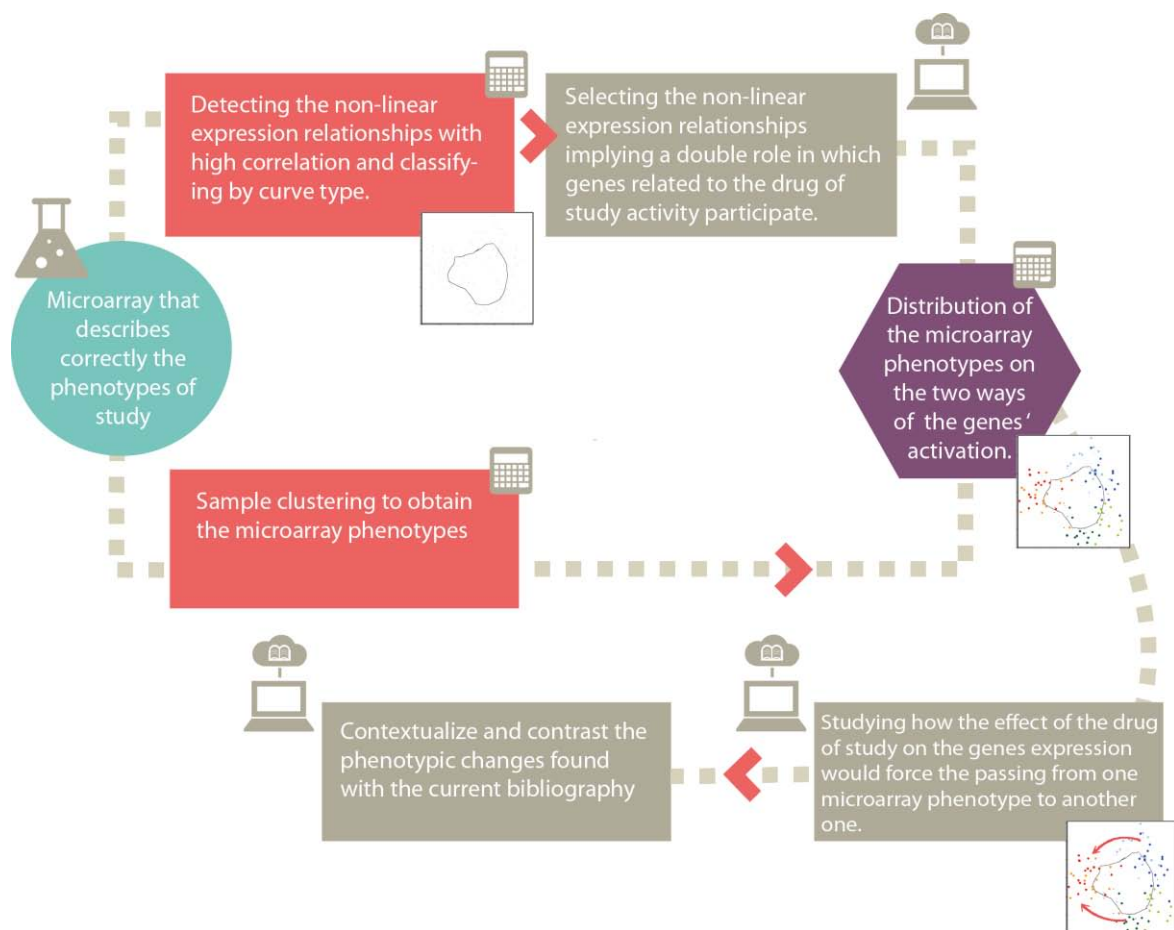
This is a bioinformatics study, and glucocorticoids have not been directly applied in the microarray experiments analyzed (the `at_matrix`). The expression data used in the analysis has been used because it describes phenotypic changes in common among multiple tumour tissues, and thus, it can help us to describe the role of the genes implicated in the glucocorticoid response in these phenotypic changes.

By means of this bioinformatics procedure we detect complex expression dependences that describe the different behaviours of the genes related to glucocorticoid response. Finally we have found that this different behaviour seems to be linked to the original healthy tissue of tumours and thus we solve the following question: why can glucocorticoids perform a different role in each of the two kinds of tumour tissues found? Because the genes implicated in glucocorticoid response have different levels of expression in the tumour tissues that can be stressed (SDTP) with respect to the tumour tissues that cannot be stressed (TDPT).

In more detail, the bioinformatics procedure followed by us is as follows (Figure 44).

The aim of our study is to describe the different tumour phenotypes, in which glucocorticoids could be applied, and thus discover if the difference in the effect of glucocorticoids depends on the phenotype of the different tumour tissues where the glucocorticoids are applied (as we finally confirm after our researches). The drugs used in our microarray experiments have a range of action wide enough to reveal the different phenotypes that are present in tumour cells of different tissues. **Note that AT\_matrix shows the expression levels necessary for drug action, not the resulting levels of drug action** (this has been emphasised in the analyzed-expression-data section). These 118 drugs are a subset of drugs coming from a previous, more extensive study that includes 1400 different anti-tumoral substances. The final subset of drugs has been chosen because they seem to be a good representation of the other drugs with similar effects or because their way of action is well known. All of these details are widely explained in the paper of Scherf et al [Scherf U. et al 2000].

The matrix genes related to glucocorticoids are identified to study the effect of glucocorticoids on the different starting phenotypes. Chemotherapeutic agents act on one or several therapeutic targets in the cell. In the case of glucocorticoids, their action is mediated by specific receptors, therefore, the cell-process activation or inhibition by glucocorticoids can only take place if these targets are expressed in the cell. In our study we have verified that some sample clusters that describe the different tumour phenotypes do not only express mature glucocorticoid receptors, but also express marker genes of concrete glucocorticoid activity.



**Figure 44. Analysis pipeline to study paradoxes and contradictory behaviours of drugs.** Two types of analyses are shown in the figure, one type by calculations on the expression matrix and another one by crossing previous results with remote databases. First, we need a gene-expression matrix with a wide range of sample conditions and well differentiated phenotypes. Otherwise, it will not be possible to study the transition from one phenotype to another. On the other hand, if the expression values are not robust enough the results will not be valid. The first step of the analysis clusters the samples, where each cluster will represent a different phenotype because they affect differently the expression of the matrix genes. In our case these clusters will detect different tumoural phenotypes. Simultaneously, the non-linear expression relationships with high correlation will be detected and classified by curve type. Curve types like  $y^2+x^2=1$  or  $y=x^2$  will detect dual behaviours of genes. Next, the expression relationships related with the paradox of study are detected by consulting remote databases, in our case those in which genes related to glucocorticoid activity participate. Then, these non-linear expression relationships are crossed with the sample clusters, so that we can observe how the variations in expression lead from one phenotype to another one. When an expression relationship shows a dual behaviour, the activation of a gene will point to one phenotype or to another one depending whether the other gene activates or deactivates. Since all the expression relationships we study are between genes related to glucocorticoids, studying the effect of the drug on the expression of these genes (whether it increases or decreases their expression) we are able to know how the drug forces the transition from one phenotype to another. We get the effect of the drug on the expression of each gene from the genes' state of art. Finally we access to the remote biomedical databases to contextualize the phenotypic changes found and finally find the hidden reason of our paradox. This procedure is useful if the paradox cause is two starting phenotypes (as in our case) or two final phenotypes after applying the drug. Note that both starting phenotypes and resulting phenotypes after applying the drug must be well described by the expression data and must be detectable by clustering methods.

The links between glucocorticoid activity and the variations of gene-expression levels are obtained by using the tools for crossing our results with biomedical databases. The effect of applying glucocorticoids to each tumoural phenotype has been studied from the phenotypes of the expression data, identified by the sample clusters, and the interdependencies between these phenotypes, described by the nonlinear-expression relationships of genes related to glucocorticoids. Once we know the link between glucocorticoids and the variations of gene-expression levels, we can follow the expression-relationship inner pattern to observe the transition from one sample cluster to another sample cluster. Meaning, from the starting phenotype to the phenotype after applying glucocorticoids. In our case of study we have two starting phenotypes; the two different tumour proliferation phenotypes.

### **3.10.3. Why our procedure to study phenotypic changes in cancer works well?**

To study the phenotypic changes, we do not use all the expression relationships. Rather, we use only non-linear expression relationships. The biological significance of the non-linear patterns of non-linear expression relationships is supported by the correlation degree obtained in the calculation of the PCOP, and by the correspondence between the global sample clusters and the slope changes of the curves (pointing out that these slope changes describe global phenotypic changes). The drugs of the expression matrix have been classified using global clustering criteria, considering the entire dataset. In the representation of the expression relationship between a pair of genes (local representation), the points belonging to each subpopulation of treatments found by clustering methods (global representation) are distributed along the curve in a consistent manner. For example, the same cluster of treatments will be neither distributed at the two ends of the curve nor cleaved along the curve. As a result, if the nonlinear-expression relationships describe the transitions among the different phenotypes found by clustering methods (sample clusters), these non-linear expression relationships are not a product of random chance. On the contrary, this fact implies a biological significance, because these relationships are product of phenotypic changes that imply a variation in the expression dependence. Note that each phenotype of these phenotypic changes is identified by a different sample cluster and that the sample clusters identify the phenotypes that affect the expression of most of the matrix genes. Therefore, both the analyzed data and the results obtained are robust and consistent.

This is an interesting way to make the complex world of gene expression understandable and to study the interdependence among sample clusters and phenotypes. Our tool allows researchers to have a very detailed view of how genes behave at the expression level and simultaneously, links the variations in the expression dependences with the sample-clustering analysis of the data (previously performed), and also with the literature. In this way, we can focus our analysis in concrete subjects and study the phenotypes resulting after a concrete variation of gene expression. The latter is particularly relevant if the phenotypes resulting after a gene-expression variation could be antagonist, from which the paradox stems.

### 3.10.4. How do non-linear expression relationships help us to solve paradoxes or to reconcile previous studies with contradictory but correct results?

The expression relationships between gene pairs work in 2 dimensions, one for each gene. One of the dimensions should represent the effect we wish to study (in our case, a marker gene of the presence or absence of tumoural proliferation). The other dimension should be a gene related to the paradox (for example, a target gene or membrane receptor of the drug). If we consider that the over-expression and the under-expression of these two genes provide us 4 combinations of their expressions (g1 down + g2 down, g1 up + g2 up, g1 up + g2 down, g1 down + g2 up), a linear expression relationship would cover 2 expression combinations for these two genes, while a non-linear expression relationship would cover 3 or 4 combinations depending on the typology of the curve. Each combination would represent a different phenotype, a phenotype that would involve these two genes, unless it would match with a sample cluster obtained by clustering methods, in this case, the phenotypes would imply most of the genes of the expression matrix. Given that each fluctuation of the curve would indicate a phenotypic change, the non-linear expression relationship would indicate how the transition from one phenotype to another one would be. If we study the effect of a concrete drug like glucocorticoids, the effect of this drug on the gene expression will indicate the direction that follows the expression relationship to pass from one phenotype to the other one (note that one of the genes is related with the drug of study and the other with the paradox). This makes the work much easier for researchers and makes possible the present study.

There are obviously more genes affected by the application of glucocorticoids beyond the genes directly involved in the glucocorticoid response. When a phenotypic change occurs due to the administration of glucocorticoids, the effect over gene expressions and the number of genes affected is big enough to be detected by clustering methods (this effect is large enough to group the samples of the resulting phenotype in a different cluster than the starting one). The glucocorticoid target genes contribution is pointing out that this phenotypic change occurs after the administration of glucocorticoids. Since the level of expression of these genes before and after the application of glucocorticoids correspond with the phenotype of different sample clusters. The rest of the matrix genes whose expression also vary from one sample cluster to another one, will also be affected indirectly by the effect of glucocorticoids, and will also be involved in the phenotypic change represented by the cited sample clusters.

In other words, a sufficiently large number of genes vary their expression between two sample clusters to a sufficient extent so that clustering methods detect these two sample clusters. How do we know that the effect of glucocorticoids will lead us from the phenotype that represents one sample cluster to the phenotype that represents the other sample cluster? Because using our tool we see that genes involved in the glucocorticoid response have relatively the same level of expression in one sample cluster as before glucocorticoid administration, and they have relatively the same level of expression in the other sample cluster as after glucocorticoid administration (the effect of glucocorticoids in these genes must be well known). From this fact we deduce that the effect of glucocorticoids could lead us from the phenotype represented by one sample cluster to the phenotype represented by the other sample cluster. As will be seen in the next section, analysing the complex expression relationships we see that the starting phenotype could correspond with a



tumoural proliferation with stress or to a tumoural proliferation without stress. Depending on the phenotype present in the tumour initially, the consequences of applying glucocorticoids will be different.

Remember that Dexamethasone has not been used as a primary drug treatment in the expression data. As shown before, the microarray data have been used to obtain the common phenotypes in several cancer types. This is a very complex task, and the described procedure has made it possible. Once the different phenotypes in common among several tumour tissues have been detected using the expression-analysis tools, the description of these phenotypes has been performed crossing these results with the biomedical databases.

Expression matrices can describe phenotypes involving very different aspects of the cell, because the expression of thousands of genes, which cover almost all cellular functions, can be observed for a same sample condition. Furthermore, the expression of these genes can be observed for a number of sample conditions large enough to describe a large number of phenotypes, which help us study the progress of pathologies and the eukaryotic cell performance.

Nowadays there is a lot of expression data describing a multitude of different sample conditions. It is necessary to study all of these data and to extract the maximum amount of knowledge from all this public information about gene expression. Nevertheless, powerful bioinformatics tools able to extract information and search for patterns are necessary to perform this task.

We are not the first researchers to re-use expression data for purposes different from the ones the microarray experiments were designed for. Expression matrices favour this kind of investigations since they allow a holistic vision of the different phenotypes providing the gene expression of each one. In the research published in *Science* [Lamb J. et al 2006] and *Nature* [Lamb J. et al 2007], public microarray data from very diverse experiments are used to search for possible therapies for new drugs also based on the gene expressions affected by the drug. In this research published in *Nature* and *Science* [Lamb J. et al 2006], they search for those microarrays that study similar phenotypic changes as the ones caused by the drug (comparing the genes affected in both cases). In our case, what we do is analyse an expression matrix with large sample series that properly describe the phenotypes we wish to study, in order to search for how the drug can cause the different phenotypic changes (studying the influence of genes affected by the drug). In this way, we can envisage whether the possible contradictions observed in cancer research, or others, can be solved by studying these phenotypic changes in depth. If so, the problem could be that we are considering several phenotypes that are actually different as the same phenotype (and so we can have different responses to the same drugs).

Based on this hypothesis and using the developed tools, we have worked until we have clearly described those different phenotypes as well as the multiple behaviours of genes linked to glucocorticoid response.

Note that the procedure of the aforementioned tools [Lamb J. et al 2007] is not useful for our study. As well as we do, Cmap, our GMDB, and others study the effect of drugs using marker genes like us. These drugs' marker genes may vary their expression in public microarrays to establish a

correspondence between the effect of the drug and the microarray experiments (although those experiments do not test the drug of study). This relation between the effect of the drug and the microarray experiments is established because the target genes of the drug are affected in the phenotypes described by the microarray experiments, and if the affected genes are the same and are affected in the same way, it is considered that we are dealing with the same phenotype. However, their approach cannot be used for our study because these bioinformatics methods are more oriented to observe side effects and possible new applications of the studied drug than to solve paradoxes or to design new models like us. That is, the first ones would be exploratory methods, whereas the second approach would need bibliographic confirmation so that the conclusions are consistent with the previous studies (and finally solve their paradoxes).

All the procedures presented in this section have been included in the work [Huerta M. et al 2014].

**Availability:** tools and supplementary data: <http://platypus.uab.es/GCinC>



O	4
<b>RESULTS</b>	
Rs.	Solving marker-gene paradoxes in Cancer progression. The glucocorticoids case.
Studying the phenotypic changes from a public expression data and accesing to remote biological databases, we try to solve the glucocorticoid paradox.	



We have performed the analysis of expression data corresponding to the genes from 60 different tumoural cell-lines in response to 118 drugs [Scherf U. et al 2000]. The final data analysed are a robust gene-expression matrix (described in the section 2.1). Our objective in analysing these data is to study the phenotypes in common among different types of tissue. These data show the phenotypes where each drug acts, ignoring whether these phenotypes belong to one tissue or to another (two samples of the same tissue can be in two different phenotypes, and different tissue samples can be in the same phenotype). For this purpose, the correlation between the level of expression of the genes and the drug's effect has been initially sought, gathering samples from all tumourous tissues. The samples of different tumour cell-lines become the data-cloud, with their corresponding level of expression and drug effectiveness. From these data we obtain the phenotypes, defined by the expression of the genes, in which the drugs act. Thus, in the final matrix we see that each drug will act in a concrete phenotype, which is common among different types of tissue (see the expression-matrix section for more details).

**Why does the expression matrix used (at\_matrix), excels in describing the tumour phenotypes?** These are the more relevant key points:

1. Those genes whose expression is relevant only in a small subtype of cells will not be included in the final matrix, since only the genes with expression levels highly correlated with the drug's effect will be selected regardless of the cell subtype. Then, we are only going to see those genes that would be expressed in representative cells with common features of many cell subtypes. Using this correlation, the samples of the final matrix represent the phenotypes in which the drug has its best performance.
2. If the absence or presence of a gene does not affect the rate of survival or death of cell subtypes, this gene is not selected. This implies that the subset of selected genes and finally included in at\_matrix is directly or indirectly implicated in relevant cellular functions. The oncological substances used in T\_matrix block the main paths essential for cell survival, such as protein synthesis, DNA synthesis, DNA repair, synthesis of microtubules, etc.
3. Only the genes with a differentiated expression between the cell subtypes where the drug acts with respect to the cell subtypes where the drug does not act, are included in the final matrix. This fact assures us that the final matrix genes can discriminate among different tumour phenotypes with their expressions. Note that another drug should act in the phenotype where the first one does not act. In this way the different tumour phenotypes are present in the final matrix.
4. Another interesting advantage of using the correlation between gene expression and survival rate, instead of using the rough-expression data, is that we can work with cell types that express genes in quite different ranges of expression. That is, this kind of data could lead us to compare different cell types even when some of them do not share the same range of expression.
5. Another interesting point is linked to laboratory conditions. Usually, cells present serious alterations in the levels of expression of some of their genes. This is usually a big problem, but as the data we use come from the correlation between the gene expressions of all cell lines with

respect to the survival for each substance, the problem is reduced (as it can be seen in Point 4.). The fact that some of these cell lines had some of their genes altered would not make it difficult to extract a general behaviour, because the rest of the considered cell lines are numerous enough to generate a strong correlation, if the gene is relevant enough to reach the phenotype where the drug has its best performance.

The procedure followed trying to solve the paradoxes about the double effect of glucocorticoids in cancer progression is shown in the Figure 44 in the previous section. This includes sample clustering, detection of non-linear expression relationships, classification by type of curve, access to remote databases, and the analysis of dual behaviours. Throughout this section we will outline the results we obtain from the different phases of the pipeline.

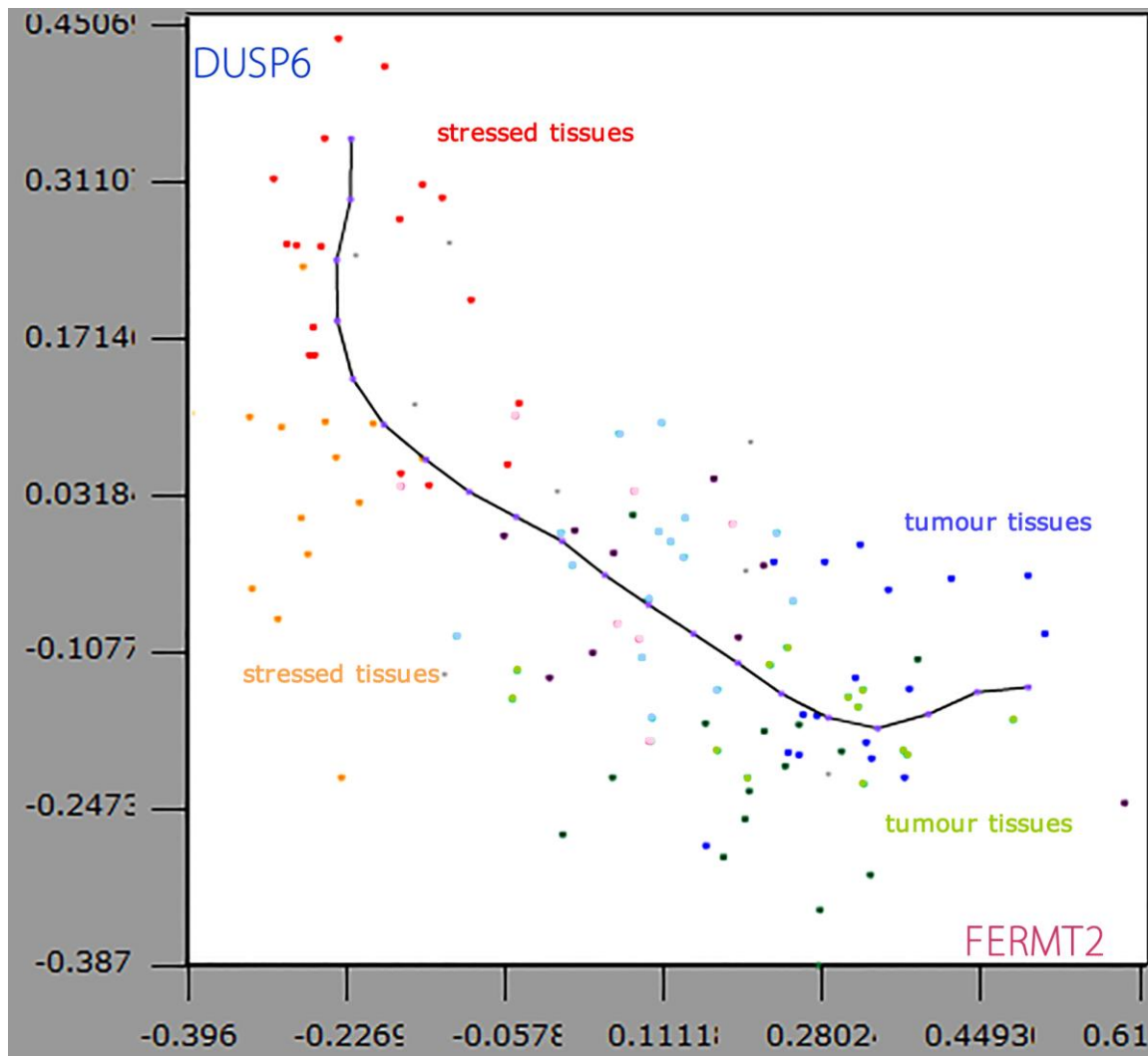
#### 4.1. Expression-relationship typology reveals the role of the genes in phenotypic changes

Whereas an expression-relationship fluctuation points to a phenotype change, the relationship typology indicates the kind of interdependence among the phenotypes. As aforementioned, our tools classify gene relationships by typology, because each different typology indicates a different kind of interdependence. So, these tools have been used to study the role of some genes related to the glucocorticoid signal in tumours, as this glucocorticoid signal does not always promote the same phenotypic changes.

In this process, for the 1416 genes of the final matrix, 11244 significant non-linear expression relationships classified into 11 different typologies have been detected. From them, 122 genes and 415 non-linear expression relationships were related to glucocorticoid activity. Below, some of the main relationship typologies are described together with the expression relationships that can help us to describe our findings.

**$y = e^{-x}$  typology:** This curve type describes two mutually excluding genes. One of the two genes must be under-expressed to make the other gene's over-expression possible.

The DUSP6 and FERMT2 genes maintain this type of relationship between their expressions (Figure 45). FERMT2 is linked to tumour proliferation [Gozgit JM. et al 2006], whereas DUSP6 has a tumour-suppressive function [Furukawa T. et al 2003; Okudela K. et al 2009]. Thus, this 'mutually excluding' relationship divides the sample-space into two phenotypes: the cell-stress phenotype (which stops tumour proliferation to perform the cell function) and the tumour proliferation phenotype (which stops cell functionality).



**Figure 45. FERMT2(x axis) and DUSP6(y axis) expression-relationship.** The different sample conditions of the expression matrix [Scherf U. et al 2000] constitute the data cloud. FERMT2 expression is linked to tumour proliferation, whereas DUSP6 stops tumour proliferation. Their non-linear expression-relationship divides the sample space into two phenotypes, the cell-stress phenotype (which stops tumour proliferation) and the tumour proliferation phenotype (which stops cell stress). The samples of each of the clusters obtained by the Repeated-Biclustering method are coloured differently. Cell stress phenotype corresponds to red and yellow clusters. Tumour-proliferation phenotype corresponds to blue and green clusters.

**$y = -x^2$  typology:** When this kind of expression-relationship is observed between two genes, the over-expression of the first gene involves only one phenotype, but this first gene under-expression involves two different phenotypes: in one phenotype the second gene is under-expressed, and in the other phenotype this second gene is over-expressed. This type of expression relationship is useful to study contradictions in expression behaviour. FKBP5, a key gene for our research, and EGFR maintain this typology in their expression relationship (Figure 46). EGFR is linked to tumour proliferation [Watanabe K. et al 1996], while FKBP5 is linked to chronic cell-stress [Billing AM. et al 2007]. FKBP5 expression is induced by the activation of steroid receptors, but it is involved in a negative-feedback, generating a high reduction of the glucocorticoid-receptor sensitivity when FKBP5 is over-expressed [Binder EB. 2009]. When looking at the relationship between the two genes (Figure 46), it can be seen that in the non-proliferative phenotype there are different levels of



cell stress marked by FKBP5 expression. In the chronic cell-stress phenotype, at maximum FKBP5-expression levels, the glucocorticoid-receptor sensitivity is fully inhibited.

**$y^2 + x^2 = 1$  typology:** This typology usually implies a dual relationship, normally of type  $y = \ln(x)$  and  $y = e^x$ , or  $y = \ln(-x)$  and  $y = e^{-x}$ , or  $y = x^2$  and  $y = -x^2$ . Searching for this type of expression relationship is very useful to solve paradoxes. NEDD4L, another key gene for our research topic, and CITED2 maintain this type of relationship in their expressions (Figure 47). CITED2 expression indicates an increase of the cell-growth rate [Tien ES. et al 2004], whereas NEDD4L is related to ion-channels inhibition by means of ubiquitination [Araki N. et al 2009] turning the cell refractory to activation signals. Therefore, this relationship implies, first, a phenotype with low proliferation and active ion channels, and second, two different ways for proliferation increase: one with active ion channels (and  $y = x^2$  typology) and the other one with inhibited ion channels (and  $y = -x^2$  typology).

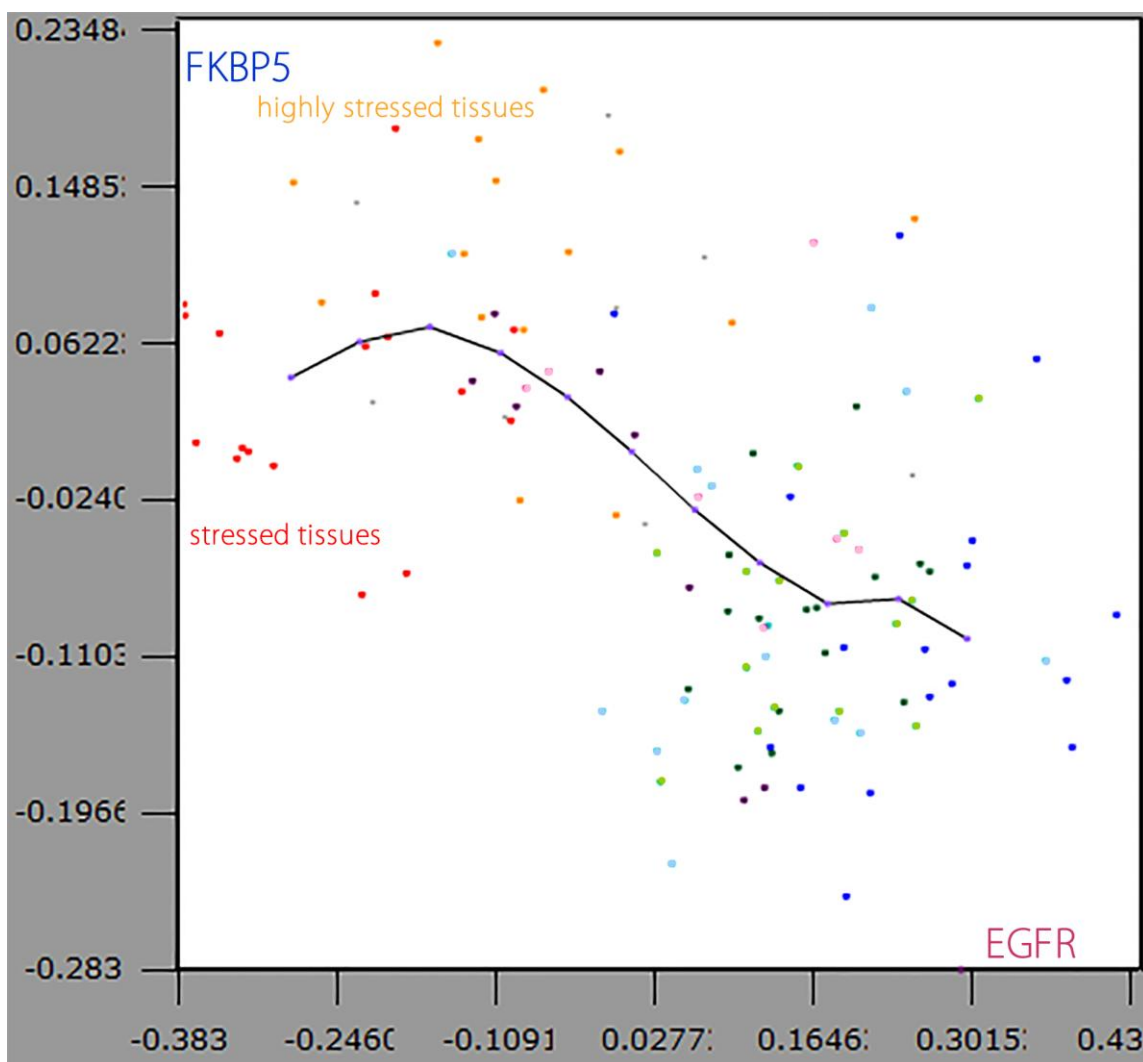
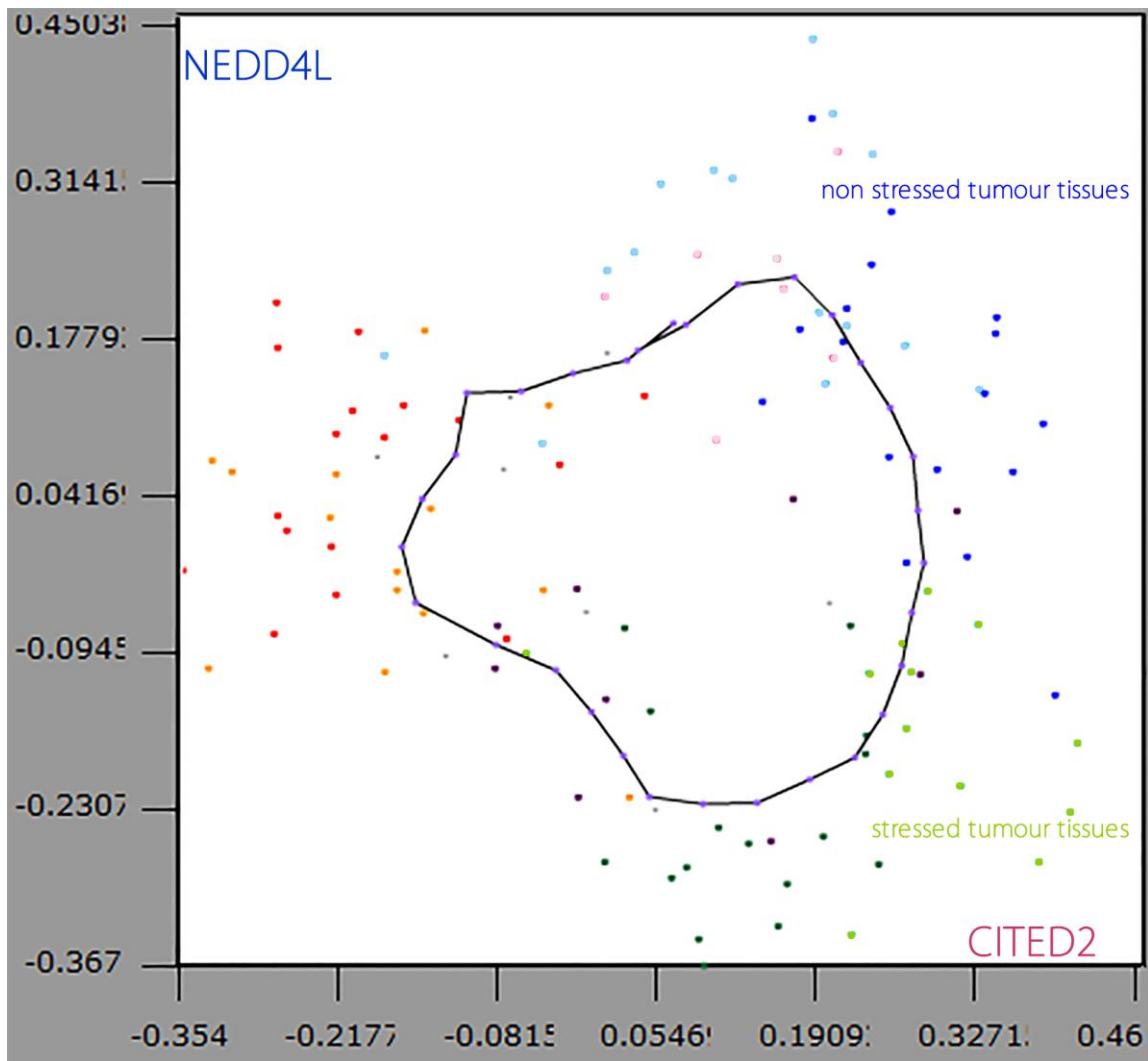


Figure 46. EGFR (x axis) and FKBP5 (y axis) expression-relationship. EGFR enhances tumour proliferation [Watanabe K. et al 1996]. FKBP5 stabilises the glucocorticoid signal response at its maximum levels [Binder EB. 2009]. The EGFR and FKBP5 expression-relationship shows us that the yellow samples represent the highly stressed phenotype within the cell-stress phenotype.



**Figure 47. CITED2 (x axis) and NEDD4L (y axis) expression-relationship.** CITED2 expression indicates a proliferation increase. NEDD4L is linked to ion-channel inhibition. Their relationship implies: first, a non-proliferative phenotype and, second, two different ways of tumour proliferation: one with active ion channels ( $y=x^2$ ) and the other with ion channels being inhibited ( $y=-x^2$ ). The samples of each of the clusters obtained by Repeated-Bisection are coloured differently in the display. The clusters are distributed in the following way: two in the non-proliferative phenotype (red and yellow), three in proliferation with ion channels being inhibited (pink and blue), and two in proliferation with active ion channels (green).

## 4.2. Revealing the phenotype of the sample clusters from the phenotypic changes described by expression relationships

The system takes the clusters obtained with global methods like SOM, HC or SOTA. The phenotypes represented by these sample clusters affect the gene expression of the matrix genes differently enough to be detected by clustering methods. The system also facilitates the analysis of the distribution of these clusters throughout the non-linear expression relationships (Figure 45).

This leads researchers to analyse the links between these clusters that affect all matrix genes and the phenotypic changes described by the expression relationships between a gene pair. As aforementioned, if the sample clusters are distributed along the curve in a consistent manner, these non-linear-expression relationships are not a product of random chance and describe a phenotypic change that affects most of the genes of the expression matrix. In these expression relationships, the change of phenotype implies a variation in the expression dependence. As each one of these phenotypes is described by different sample clusters, the expression relationships can be used to study the interdependence among these clusters. The sample clusters used for our analysis were obtained by using the Repeated-Bisection method. The results are as follows:

The FERMT2 and DUSP6 relationship (Figure 45) separates the sample-space into two phenotypes: the samples of cell-stress and non-proliferation (yellow and red clusters) and the samples of tumour proliferation (pink, blue and green clusters).

The EGFR and FKBP5 expression-relationship (Figure 46) shows that the yellow samples correspond with the chronic cell-stress phenotype.

The CITED2 and NEDD4L expression-relationship (Figure 47) helps us to differentiate between two proliferative phenotypes (green clusters vs blue clusters): first, separating the cell-stress phenotypes, with low proliferation and active ion channels (yellow and red clusters), and second, separating and differentiating the two different ways for proliferation increase: one with inhibited ion channels (pink and blue clusters) and the other one with active ion channels (green clusters).

Note that each sample cluster clearly represents a distinct phenotype; since each sample represents a different drug that acts in a concrete phenotype but not in others (see the expression-matrix section for more details).

From the analysed microarray which contains 1416 genes, 11244 significant nonlinear-expression relationships have been detected and classified in 11 different typologies. From them, the non-linear expression relationships of 122 genes were related to glucocorticoid activity. Non-linear gene-expression relationships can help us to describe the interdependence between phenotypes, so we select the non-linear gene-expression relationships performed by the genes linked to glucocorticoids, and in this way, we can study if the glucocorticoid activity can promote the change from one phenotype to another. In summary, from the results provided by the system (11244) we select the expression relationships that can better describe the phenotypes and phenotypic changes guided by glucocorticoid activity (122). The selection of these genes has been done using the bioinformatics tools previously described. These genes have been provided by crossing the genes of the expression matrix with the PubMed database and the query terms. The final genes have been chosen from the non-linear expression relationships of glucocorticoid-linked genes with known activity in tumoural tissues for their idoneity to explain the results of the analysis performed. These genes can be used to compare the phenotype of tumour tissues by their different levels of expression, but furthermore, since these genes are linked to glucocorticoid activity, they can also relate the tumour phenotypes with glucocorticoid activity. For instance, if glucocorticoid administration changes their levels of expression in a well known way (info provided by the

database crossing), these genes can help us to understand the phenotypic changes due to the glucocorticoid administration observing the change of phenotype when their expression levels vary.

Note that dual behaviours are searched by means of non-linear expression relationships study (in our case the expression relationships of glucocorticoids-linked genes) but **the phenotypes** (in our case the two ways of tumour proliferation) **are obtained by sample clustering and considering all the genes of the expression matrix**. Thus, the phenotypes found (in our case the two ways of tumour proliferation) imply almost all of the genes of the expression matrix. In other words, **there exists enough difference in expression terms between the stressed tumour tissues (SDPT) and the non-stressed tumour tissues (TDPT) and for a number of genes big enough to be detected by clustering methods**.

The non-linear expression relationships of glucocorticoid target genes with known activity in tumoural tissues detected by our system have been added to the supplementary material available in the web. The typology of each expression relationship has been also added to the supplementary material because the typology of the expression relationship reveals the gene role in the phenotypic changes. The bibliography that describes the link between the expression-relationship genes and the glucocorticoid activity has been also added to the supplementary material. With this supplementary data anyone can study the role of these genes with glucocorticoid activity in the two kinds of tumour proliferation described by the expression data analyzed.

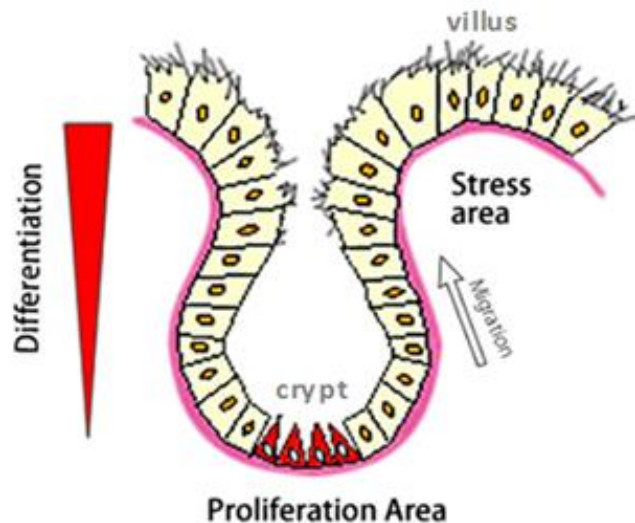
Not all of the non-linear expression relationships involving glucocorticoid-related genes provide new insights, because not all the genes of the gene pairs are known well enough, and because some of them provide us the same or similar information. Obviously, there are other expression relationships, not included in this section that provide relevant information about the two tumour-proliferation phenotypes found, as the ones seen in previous sections.

### 4.3. The glucocorticoids' effect on different tumour tissues

The two different ways of tumour proliferation (pink plus blue clusters vs green clusters) seem to be the key point of the different effects of glucocorticoids administration in cancer patients. As shown below, NEDD4L plays a crucial role both in differentiating the tumour progression of both proliferation ways, as in this glucocorticoid effect. Using our online tool to cross the expression analysis with remote biological databases [Huerta M. et al 2009], these two tumoural growth ways have been related to the physiology of healthy tissues, and we have found that structural properties of the healthy tissue determine whether the tumour proliferates in one way or another. As we see next, the underlying reason for this phenomenon seems to be that some structural properties in healthy tissues are conserved even when the tissues become tumourous, conditioning the way tumours proliferate.

NEDD4L plays a crucial role in phenotypic changes by marking membrane proteins for proteasomal and endosomal degradation. In this way, NEDD4L promotes ion-channel inhibition through Fe65 ubiquitination. Fe65 ubiquitination starts the degradation of ion channels like, for

example, ENaC (Epithelial Na<sup>+</sup> Channel) [Araki N. et al 2009] or CACNB1 (voltage-dependent calcium channel) [Persaud A. et al 2009]. As a result, one of the two tumour-proliferation types (pink and blue clusters) prevents the cell from entering into stress during tumour growth by means of NEDD4L expression, which turns the cell refractory to ion-channel signals. This NEDD4L activity grows when the level of proliferation is not yet too high (pink and light blue clusters) and falls when the proliferation is already uncontrolled. The glucocorticoids activate the ion channels again, partly by means of the glucocorticoid inhibition of NEDD4L [Snyder PM. et al 2004], and lead the cells to the cell-stress phenotype (red and yellow clusters). It happens in tumour tissues like ovarian cancer cells [Chen YX. et al 2006], osteosarcoma [Yamamoto T. et al 2002], or glioblastoma [Kaup B. et al 2001], stopping the tumour proliferation of these tissues (TDPT). During this tumour-proliferation type, a strong signal like glucocorticoids is needed to switch from the proliferative phenotype to the cell-stress phenotype again. During the other tumour-proliferation type (green clusters), the ion channels are not altered by NEDD4L. This fact facilitates the normal switch between proliferation and cell stress during this kind of tumour growth. This combination of cell stress and proliferation is common in tissues like colon, endometrium [Gargett CE. et al 2008], prostate or mammary gland [Phesse TJ. and Clarke AR. 2009]. We found using our tools and the bibliography, that the physiology of these tissues facilitates the cell stress and proliferation combination: in all of them the proliferation and function are performed in different areas of the tissue and the new cells are immediately incorporated from the proliferation areas to the functional areas (SDPT). In colon tissue, for example (Figure 48), the proliferation is performed in the crypts and the colon function (and cell stress) in the villus. In mouse colon normal tissue, the stress signals increase the proliferation rate of the crypts [Clarke RM. 1975], even inducing hyperplasia (cells/crypt rate) [Newmark HL. et al 1991]. An excessive physiologic stimulation by dihydrotestosterone may induce nodular prostatic hyperplasia and by estrogen it may induce endometrium hyperplasia [Jasonni VM. et al 2005], by thyroid-stimulating hormone it may induce the development of nodular goiters in the thyroid [Kumar V. et al 2009], and even a progesterone proliferative signal is required in breast cancer for a robust tumourgenesis induction by carcinogen or in mice Brcal-mediated mammary gland tumour models [Poole AJ. et al 2006]. In colon, aberrant crypt foci, adenoma and adenocarcinoma induction by carcinogenic agents are positively correlated with gastrin levels, a local stimulation signal of the colon promoted by the nervous system [Singh P. et al 2000]. Tumour-cell proliferation is retarded with an adrenalectomy, and this adrenalectomy retardation is reversed by the administration of hydrocortisone or synthetic steroids with a predominantly glucocorticoid activity [Tutton PJ. and Barkla DH. 1981]. In differentiated mammary glands, the mammary-specific stressor agent prolactin stress the cells and induces tumour growth. Cortisol achieve the same results, but to a lesser extent [Wennbo H. and Törnell J. 2000]. All of this show us that the effectiveness of stressor agents like glucocorticoids to stress these kind of tissues (stressed tumour tissues (SDPT)) and promote the proliferation has multiple regulations and intermediaries, which are different for each tissue, and that the relation between cell stress and induced proliferation is at least stress-level dependent. When high levels of cell stress are achieved, the glucocorticoid signal reception is finally disabled by FKBP5 in the way cited early [Binder EB. 2009]. And it is in this glucocorticoid-receptor absence when the worst survival is manifested in cancers like colon [Theocharis S. et al 2003] or non-small cell lung cancer, as well squamous cell carcinoma as alveolar adenocarcinoma [Lu YS. et al 2006].



**Figure 48.** The two proliferation ways found appears linked to the physiology of the tissues. The physiology of tissues like colon or endometrium makes possible the cell stress and proliferation combination because the proliferation and function are performed in different areas of the tissue (SDPT) and the new cells are immediately incorporated from the proliferation areas to the functional areas. In colon tissue, for example, the proliferation is performed in the crypts and the colon function (and cell stress) in the villus.

FKBP5 is a marker gene for chronic cell stress (Figure 46). However, FKBP5 has an opposite behaviour depending on tumour tissues: FKBP5 is expressed during tumour growth in colon cancer, whereas FKBP5 expression stops tumour growth of other cancers such as colorectal [Mukaide H. et al 2008]. In the same way, the glucocorticoid signal, which expresses FKBP5 [Billing AM. et al 2007], has the same, opposite effect: It induces the proliferation of (SPDT) tissues like colon [Tutton PJ. and Barkla DH. 1981], pancreas [Zhang C. et al 2006], hormone-sensitive prostate [Herr I. and Pfitzenmaier J. 2006], or at a lesser extent, squamous-Cell-lung Carcinoma [Herr I. et al 2003] and mammary gland [Sui M. et al 2006], but stops the proliferation of (TPDT) tissues like ovarian cancer cells [Chen YX. et al 2006], osteosarcoma [Yamamoto T. et al 2002], medulloblastoma [Heine VM. et al 2010] or glioblastoma [Kaup B. et al 2001]. The origin of these opposite behaviours seems to be the two kinds of tumour proliferation described before: Tissues like colon combine chronic cell stress with tumour proliferation (SDPT), whereas cancers like glioblastoma separate these two phases and, therefore, they cannot be stressed during tumour proliferation (TDPT). In this way, for glioblastoma, if the cells change their growth phenotype (blue cluster) to cell stress (red or yellow clusters), tumour proliferation stops. And for colon cancer, if the cells switch their phenotype (from green or red clusters) to a highly stressed phenotype (yellow cluster), the tumour becomes progressively more malignant (as described above). These tumoural features have direct medical implications, because they entail that treatments can induce counterproductive effects if they are not applied properly. Concretely, stressors like dexamethasone should be applied only when tumours appear in (TDPT) tissues that cannot be stressed during tumour proliferation (like glioblastomas), but even in this case it can become a problem: Although glioblastoma-type proliferation could be reduced, stressing agents could induce tumour proliferation

in other tissues with the colon-type proliferation (SDPT). This fact may be responsible for some metastasis increase after glucocorticoid supplementation during cancer therapy [Mattern J. et al 2007; Sherlock P. and Hartmann WH. 1962]. With an opposite treatment, applying beta-blockers that inhibit the systemic adrenalin-stress signal, the spread of cancer is reduced [Powe D. 2010; Sloan EK. et al 2010]. Metastasis spread is a serious issue and its underlying mechanisms must be studied in-depth.

All the findings presented in this section have been included in the work [Huerta M. et al 2014].

**Availability:** Supplementary data are available at: <http://platypus.uab.es/GCinC>

O	5
<b>DISCUSSION</b>	
Ds.	Solving marker-gene paradoxes in Cancer progression. The glucocorticoids case.
The hidden reason of glucocorticoids paradox seems to be linked to the physiology of the healthy tissues. This link seems to have a functional motivation.	





We can solve paradoxes or apparently incongruent experimental results by the acute dissection of the interdependencies among the phenotypes described by global sample clusters (obtained by SOTA, SOM, PAM, etc.). In the present study, NEDD4L and the other genes helped us to identify and characterize the two types of tumour proliferation. This correct identification and characterization of the two types of tumour proliferation may help to adjust the treatments correctly in cancer and other pathologies. It could also help in understanding and preventing future malignancy, recurrence and metastasis. Controlled randomized trials evaluating the impact of stressor agents on growth and metastasis in stressed (SDPT) and non-stressed (TDPT) tumour tissues need to be performed. Patient survival needs to be evaluated also. Currently, agents like those cited in this work are administered as adjuvant to treat undesirable side effects of radiotherapy, surgery or cancer itself [Philips RS. et al 2010].

New items around the glucocorticoid-administration paradox are constantly supplied to the medical discussion. This relationship between tumour growth and systemic hormones related with stress and produced in the adrenal glands is not restricted to glucocorticoids, a similar behaviour have been observed also in the case of the catecholamines. For instance, there is an interesting talk of the 7th European Breast Cancer Conference (2010) entitled: “Applying beta-blockers that inhibit the systemic adrenalin-stress signal, the spread of cancer is reduced” [Powe D. 2010; Sloan EK. et al 2010], in which the relation between hypertensive stressor blockage and the reduction of metastasis incidence in oncologic patients is described. Furthermore, not only do glucocorticoids seem to have this dual behaviour depending on the original healthy tissue. Other drugs that stress tissues, like chemotherapy, appear to have a different behaviour depending on the original tissue (SDPT or TDPT). In a work about therapeutic effects of chemotherapy it can be observed that its therapeutic effects could depend on whether the tumours can be stressed or cannot be stressed [Morgan G. et al 2004].

Table 1 - Impact of cytotoxic chemotherapy on 5-year survival in Australian adults

Malignancy	ICD-9	Number of cancers in people aged >20 years (1)	Absolute number of 5-year survivors due to chemotherapy (2)	Percentage 5-year survivors due to chemotherapy (3)
Head and neck	140-149, 160, 161	2486	63	2.5
Oesophagus	150	1003	54	4.8
Stomach	151	1904	13	0.7
Colon	153	7243	128	1.8
Rectum	154	4036	218	5.4
Pancreas	157	1728	0	0
Lung	162	7792	118	1.5
Soft tissue sarcoma	171	665	0	0
Melanoma of skin	172	7811	0	0
Breast	174	10661	164	1.5
Uterus	179+182	1399	0	0
Cervix	180	867	104	12
Ovary	183	1207	105	8.7
Prostate	185	9869	0	0
Testis	186	529	221	41.8
Bladder	188	2802	0	0
Kidney	189	2176	0	0
Brain	191	1116	55	4.9
Unknown primary	195-199	3161	0	0
Non-Hodgkin's	200+202	3145	331	10.5
Hodgkin's disease	201	341	122	35.8
Multiple myeloma	203	1023	0	0
Total		72903	1690	2.3

(1) Numbers from Ref. [21].

(2) Absolute numbers (see text).

(3) % for individual malignancy.

(4) Total for Australia 1998 = 80 864 people.

Table IX. [Morgan G. et al 2004].

As it can be observed in the Table IX, in (SDPT) tissues where stress and proliferation occur in different areas (stomach, colon, pancreas, melanoma, prostate, bladder) the survival after chemotherapy does not increase, whereas in (TDPT) tissues in which proliferation stops when stress occurs (Oesophagus, Rectum, Cervix, Ovary, Testis, Lymphoma, Hodgkin), chemotherapy gets an increase of survival.

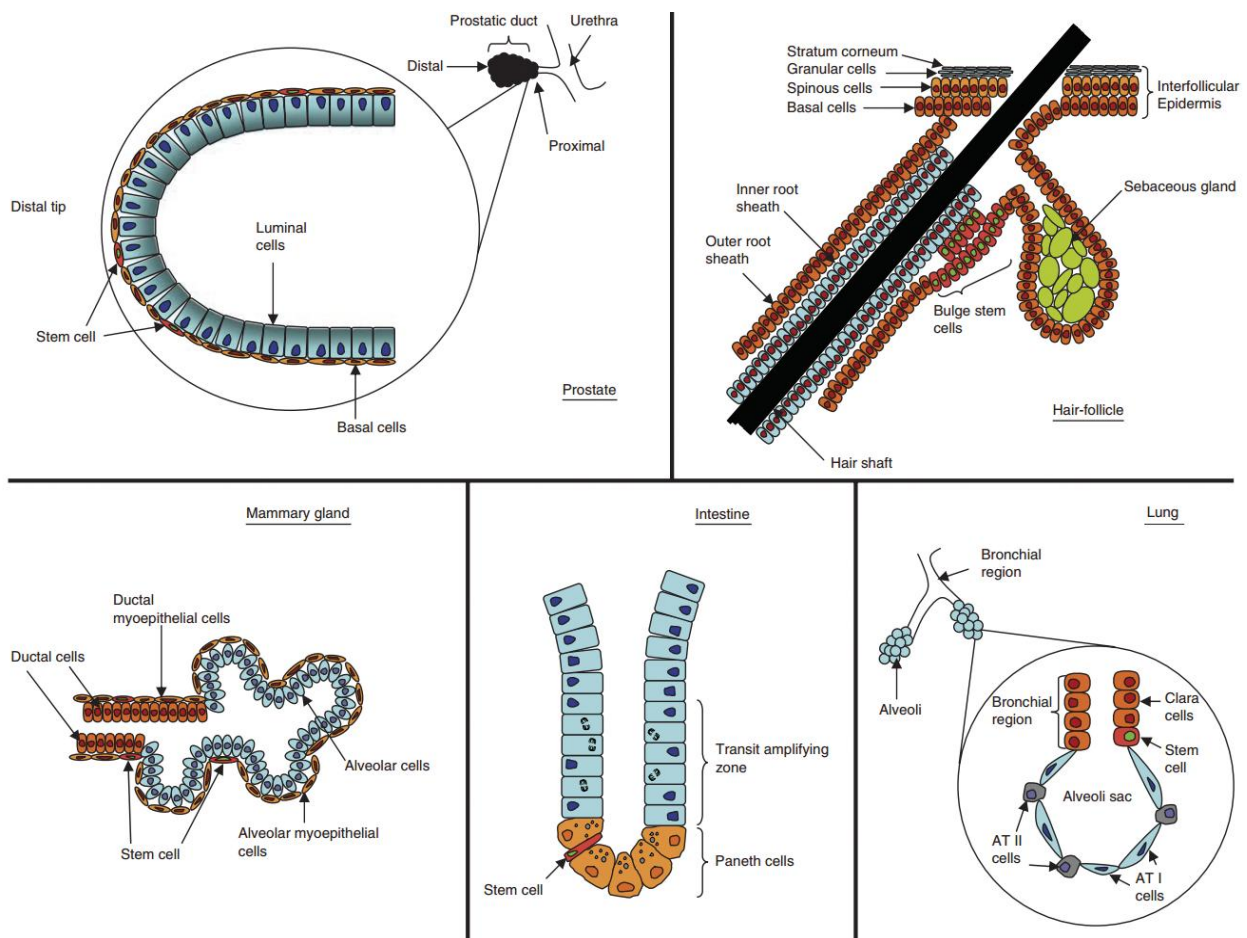
Without a doubt this is an issue that needs to be studied in depth, and future researches may start to consider it. We can finally reconcile these apparent incongruities with our findings; however, there are still many remaining questions to answer.

One of the great strengths of our analysis is that we find two major types of tumour proliferation, one (SDPT) in which the cell is actively dividing but which the main part of the tissue structural properties are preserved, the cells are differentiated (Figures 11-12) and very stressed (Figure 47), and another (TDPT) in which cell division is accompanied by a full tissue remodelling, inactivating the cells, and maintaining the cells unstressed until the tissue remodelling ends. Furthermore, these two types of tumoural cell divisions seem to be dependent on the structure of the original healthy tissue.

At this point new questions come up. For instance, **why is this different tumour proliferation dependent on the structure of the tissue?**

One of the possible reasons of this dependence can be that tissular function in one type of tissue is more dependent on the tissue structure than in the other type. The organogenesis of the lung (bronchi, bronchioles and alveoli) can serve to illustrate the idea. Proper lung function requires that the air comes from outside the body to the alveolar, so each and every one of the bronchi and bronchioles have to be interconnected in a dense network of pipes carefully subdivided to ensure that the subsequent bronchioles reach the correct air flow. All of this complies with the physical laws of gas dynamics, even responding effectively to multiple levels of demand for ventilation. To accomplish this, tissular function bronchioles are highly dependent on tissue structure and in lesser extent on cell function. Thickness, diameter and flexibility of the cartilage generated by the bronchial cell must change as branches into the lung toward the alveolar to properly fulfil the function of the tissue. However, the alveoli located at the end of this complex structure do not need to be sized depending on their relative position within the bronchial tree, as this is always the final element in this complex structure. In the case of alveolar cells of type I, their main function is to spread gases through it, between outside and inside the body, a function that is highly dependent on the cell and to a lesser extent on the tissue structure.

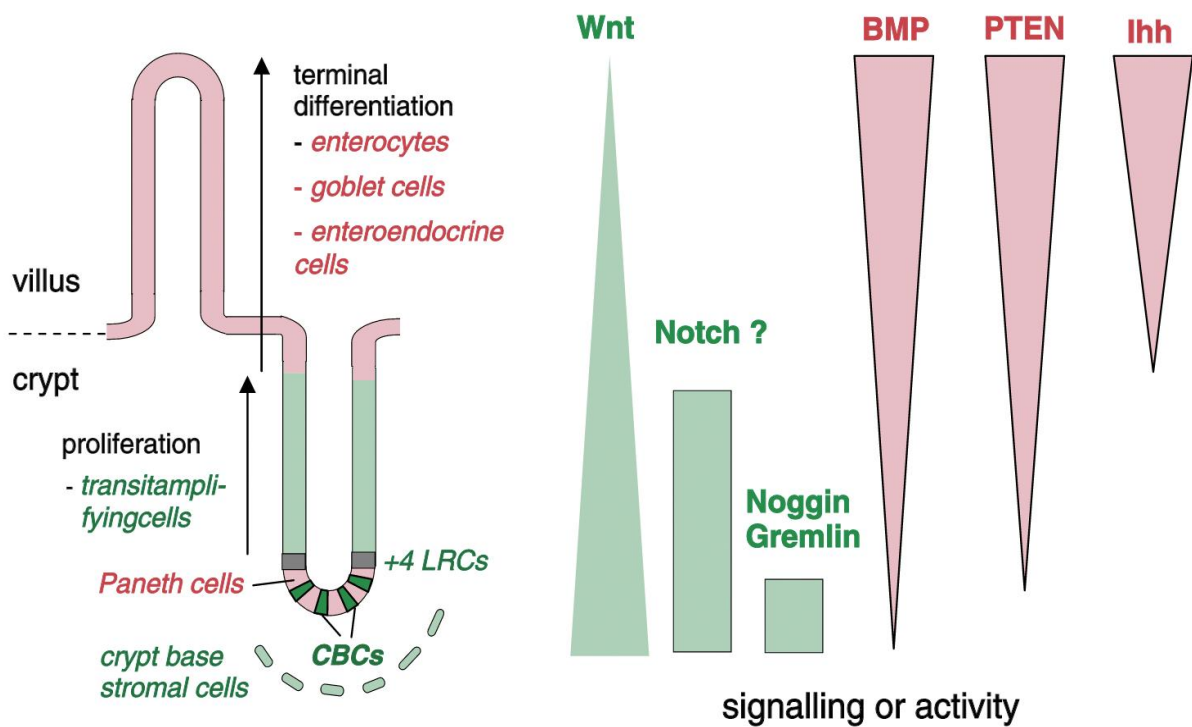
This fact would have practical implications in tissues growth. During bronchi remodelling, the tissue cannot develop its cellular function, thereby it keeps in isolation from the external stimuli, which did not happen in low-complexity tumours (SDPT). As a result, whereas one type of tissue (SDPT) can be highly active during tissue remodelling (low complexity tissues of Figure 49), the other type (TDPT) cannot develop its cellular function until the tissue remodelling ends so it keeps in isolation from the external stimuli during remodelling (high-complexity structure linked to tissue function). As we show in our work, this fact is directly related to the observed differences in the response to glucocorticoids.



**Figure 49. Location of stem cells in low-complexity tissues (SDPT) which main function is more dependent on cell than on tissue structure.** Prostate gland. The putative prostatic stem cells are located in the basal cells surrounding the columnar secretory cells of the distal prostatic duct. Hair-follicle. Skin stem cells are located under the sebaceous gland in a region known as the bulge. During rest periods, stem cells of the bulge region form the base of the hair-follicle. During the start of each new growth cycle stem cells located at the base of the bulge become active to form the highly proliferative new hair germ. The interfollicular epidermis is a stratified epithelium, containing unipotent progenitor cells and transit-amplifying cells located in the basal layer. Basal cells differentiate upward to form the spinous, granular, and stratum corneum layers of the epidermis. Mammary gland. The mammary gland consists of a branching network of ducts, terminating in alveolar buds. Mammary stem cells are thought to be located in the basal, myoepithelial layer, which tightly surrounds the ductal epithelial layer. The secretory alveolar cells are also surrounded by a looser association of myoepithelial cells. Intestinal crypt. The crypt stem cell had previously been located to position 4-6, just above the base of the crypt. Recent data now suggests the putative stem cells of the intestine (red) are narrow cells located between Paneth cells near the base of the crypt. Cells leaving the proliferation zone migrate upward towards the villus tip and differentiate into one of three cell types, enteroendocrine cells, goblet cells, and enterocytes, to form the villus. A fourth cell type, the Paneth cells, migrate downward to the crypt base. Lungs. The putative lung stem cells are located at the junction between the branching, bronchial region and the alveolar sac, and express markers from ATII cells and Clara cell. [Phese TJ. and Clarke AR. 2009].

According to our findings, the tissue information is quite relevant in relation to cancer prognosis and the effect of dexamethasone on it. As it is shown in our work, the two types of tumoural cell division found by our analysis seem to be dependent on the structure of the original healthy tissue (Figure 50). Therefore, another question to solve is: To what extent does this dependence exist? Does it exist also in metastasis?

We have already explained that in some tissues the proliferation and function are performed in different areas of the tissue and that the new cells are immediately incorporated from the proliferation areas to the functional ones (SDPT). In colon tissue for example, the proliferation is performed in the crypts and the colon function (or cell stress) in the villus [Phese T.J. and Clarke AR. 2009].



**Figure 50. Architecture and signalling in the small intestinal mucosa.** Terminally differentiated intestinal epithelial cells (red) are found in the villi; only Paneth cells are located in the crypt base. Green indicates location of putative intestinal stem cells (dark green) and the transit-amplifying cells. Crypt base columnar cells (CBCs) selectively express the Wnt target gene *Lgr5/GPR49*. Gradients of pathways (or components) that induce stemness or proliferation are indicated in green. Triggers of differentiation are shown in red. [Brabletz S. et al 2009].

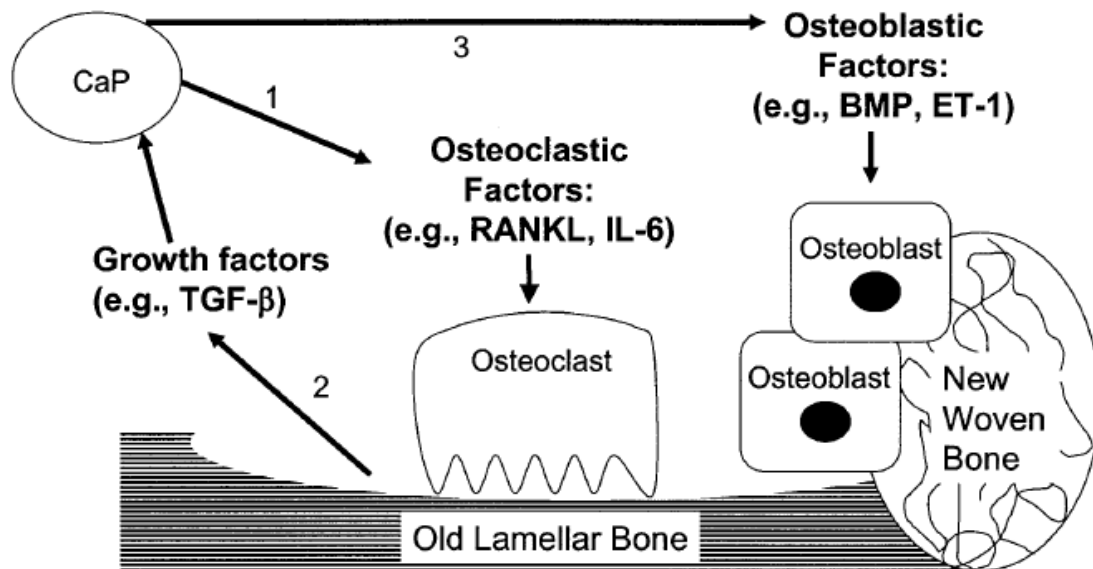


**Figure 51.** From osteolytic to osteoblastic tumour progression. [Keller E.T. and Brown J. 2004].

But the dependence on the tissue structure seems to go even further beyond this.

In the evolution of tumours in tissues with a function highly dependent on structure (TDPT), such as bone, it is common to observe two phases: first an osteoclastic phase followed by an osteoblastic one. In the first phase tissue resorption occurs with an abundance of osteoclasts. The second one is characterized by the proliferation of osteoblasts increasing the bone mass largely (Figure 51).

This trend is so strong in this type of tissue, that even metastasis bone tumours coming from prostate primary tumours have these two phases (Figure 52). In this case, it has been pointed out that the bone microenvironment plays a role in the establishment and progression of prostate-cancer bone metastasis [Keller E.T. and Brown J. 2004]. Notice that these are not similar tumours, but that the primary tumour comes from a tissue with stress and proliferation separated areas (SDPT) and the secondary tumour develops in a tissue that must alternate stress and proliferation (TDPT).



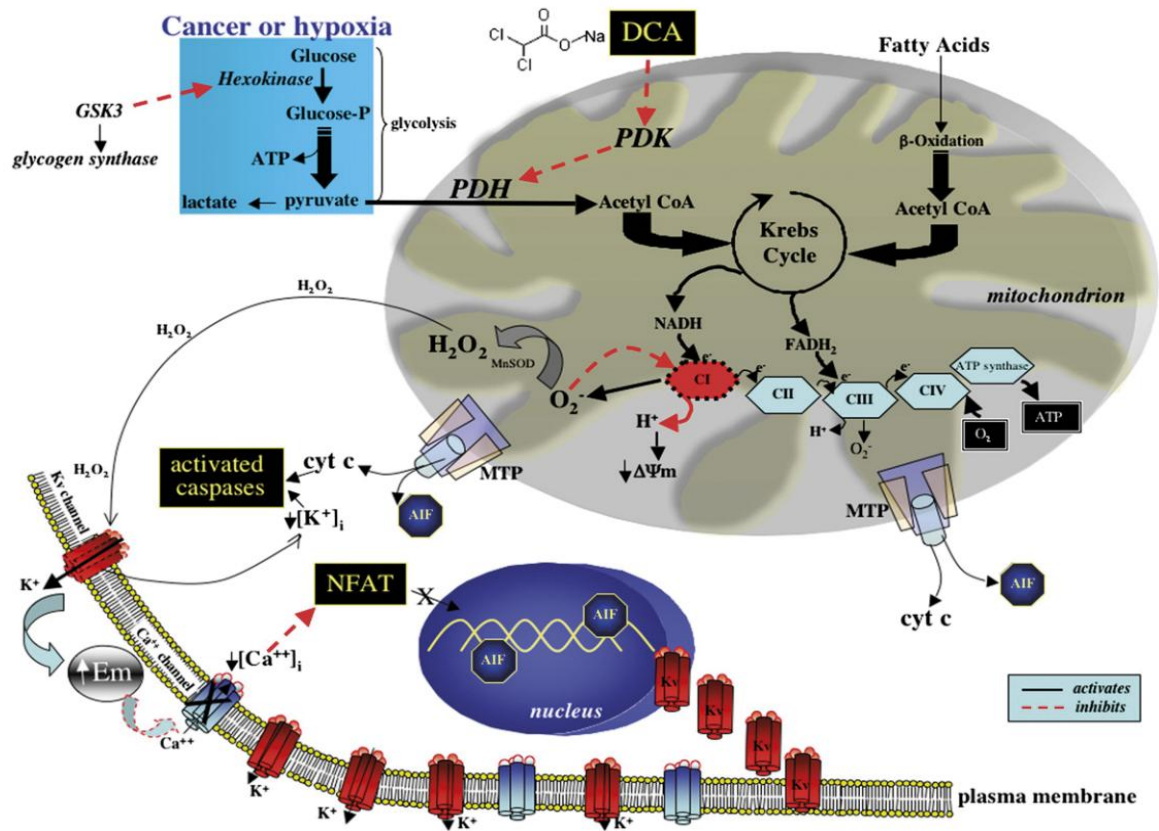
**Figure 52. Model for how prostate cancer induces bone remodelling.** The prostate cancer cells initially (i) induce osteoclastogenesis and resorption of mature lamellar bone. As the bone matrix is destroyed, it releases growth factors (ii) that induce prostate cancer cell's growth factors growth and alter their phenotype. The changing bone microenvironment, enhances the prostat cancer cells' production of osteoblastic factors (iii) resulting in production of woven bone. BMP, bone morphogenetic protein; CaP, prostate cancer cell; ET-1, endothelin-1; IL-6, interleukin-6; RANKL, receptor activator of NFκB ligand; TGF- β, transforming growth factor β. [Keller E.T. and Brown J. 2004].

We have seen already that stressing the tissues can stop tumour proliferation in high-complexity tissues with a functional structure (TDPT). But a simple question emerges: **Can we relaxing the tissue stop tumour proliferation in tissues with a function limited to the cell activity (SDPT)?**

The work of [Bonnet S. et al 2007; Michelakis E.D. et al 2007] shows us that it is also possible. In this work, researchers reactive the mitochondrial normal function, avoiding glycolysis (aerobic glycolysis), reactivating the K voltage-dependent channels, hyperpolarizing the cell and conceivably carrying the cell out of the stress phenotype (Figure 53). This action stops tumour proliferation in cancers like non-small cell lung cancer[Bonnet S. et al 2007], colon [Bonnet S. et al 2007], mammary gland [Fantin V.R. et al 2006], endometrial gland [Wong J.Y. et al 2008] and hormone-sensitive prostate [Cao W. et al 2008].

The k voltage channels change the cell-membrane potential repolarising the cell getting it out of the stress phenotype. This action places the cell in a recovery post-stress phenotype, and kills most of the cells in the aforementioned tumours.

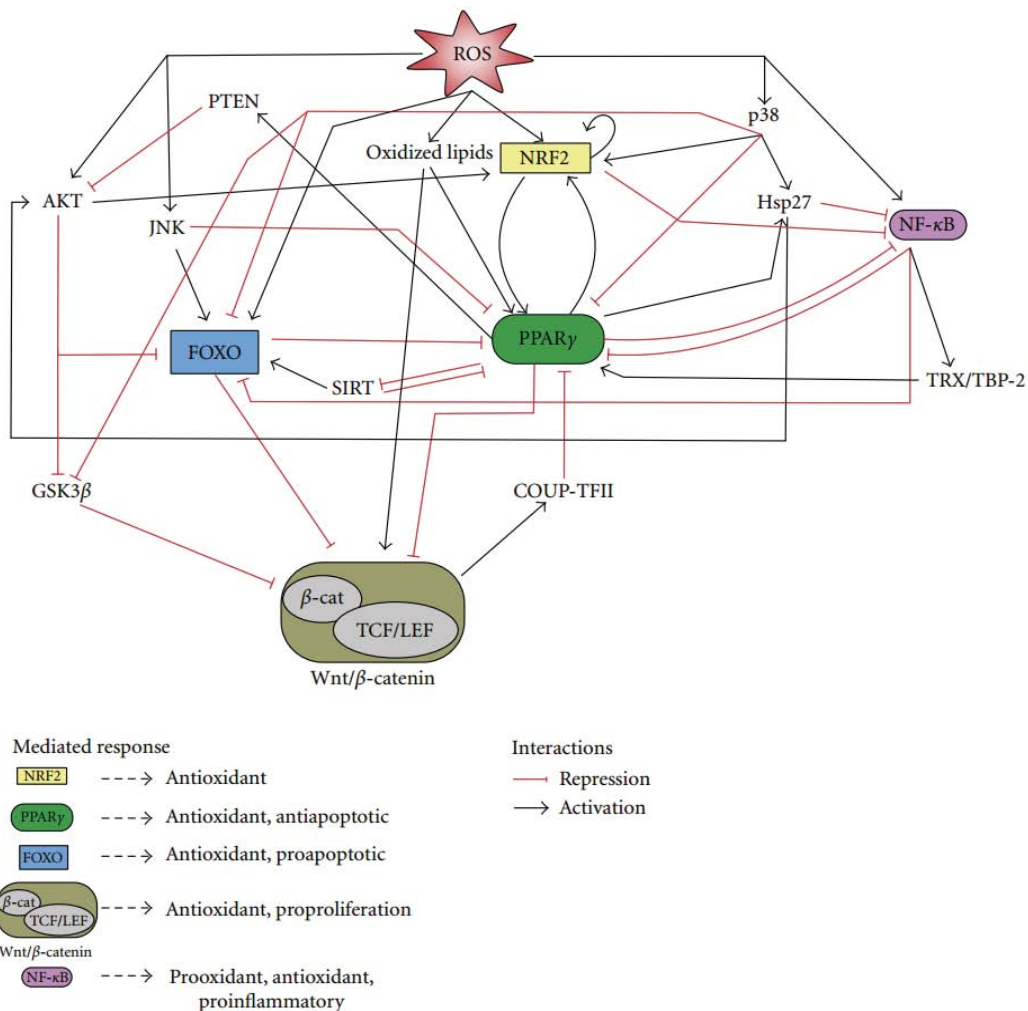




**Figure 53. A Reversible Metabolic-Electrical Remodelling in Cancer Contributes to Resistance to Apoptosis and Reveals Several Potential Therapeutic Targets.** In cancer, mitochondrial glucose oxidation is inhibited and energy production relies on the cytoplasmic glycolysis. This “inactivity” of the mitochondria likely induces a state of apoptosis resistance. Activation of PDH by DCA increases glucose oxidation by promoting the influx of acetyl-CoA into the mitochondria and the Krebs cycle, thus increasing NADH delivery to complex I of the electron transport chain, increasing the production of superoxide, which in the presence of MnSOD is dismutated to the more stable H<sub>2</sub>O<sub>2</sub>. Sustained increase in ROS generation can damage the redox-sensitive complex I, inhibiting H<sup>+</sup> efflux and allowing the efflux of cytochrome c. Both cytochrome c and H<sub>2</sub>O<sub>2</sub> open the redox-sensitive K<sup>+</sup> channel Kv1.5 in the plasma membrane and hyperpolarize the cell, inhibiting a voltage-dependent Ca<sup>2+</sup> entry. The decreased [Ca<sup>2+</sup>]<sub>i</sub> suppresses a tonic activation of NFAT, resulting in its removal from the nucleus, thus increasing Kv1.5 expression. The increased efflux of K<sup>+</sup> from the cell hyperpolarises the cell membrane [Bonnet S. et al 2007].

This unstressed phenotype is also induced by the activation of PPAR $\gamma$  [Panigrahy D. et al 2005] stopping tumours of tissues with a non-functional structure (SDPT) like colon [Sarraf P. et al 1998], mammary gland [Apostoli A.J. et al 2014], prostate [Grommes C. et al 2004], or endometrial cancers [Nickkho-Amiry M. et al 2012]. PPAR $\gamma$  is activated by the lipid oxidation to prevent it. Cellular oxidative stress results in the conversion of LDL to oxLDL, and PPAR $\gamma$  can modulate the effects of oxLDL signalling (Figure 54). As well as glucocorticoids, PPAR $\gamma$  has been reported to act both as a promoter and suppressor of neoplasia, but in this case of the opposite cancers. The PPAR $\gamma$  activation could revert the stress phenotype stopping the tumours in stressed tumour tissues (SDPT) [Panigrahy D. et al 2005], whereas PPAR $\gamma$  expression in ovarian carcinomas (TDPT) is increased when compared with benign ovarian tumours [Nickkho-Amiry M. et al 2012].





**Figure 54. Crosstalk of PPAR $\gamma$  with NRF2, Wnt/ $\beta$ -catenin, and FOXO signaling pathways in oxidative stress response.** In oxidative stress conditions the nuclear receptor PPAR $\gamma$  directly regulates a vast array of genes involved in the response to oxidative stress and exerts anti-inflammatory effects transrepressing NF- $\kappa$ B. PPAR $\gamma$  suppress phosphatidylinositol 3-kinase (PI3K)/Akt/Rac1 signaling axis via activation of PTEN resulting in decreased reactive oxygen species (ROS). ROS and other reactive species activate NRF2 and PPAR $\gamma$  that are linked by a positive feedback loop that sustains their expression. Through a negative feedback PPAR $\gamma$  inhibits Wnt/ $\beta$ -catenin and induces cell block. [Polvani S. et al 2012]. In tumours and other pathologies the equilibrium between PPAR $\gamma$  and Wnt/ $\beta$ -catenin is broken [Lecarpentier Y et al 2014].

An increasing of mitochondrial membrane potential is linked to a progressive stimulation of mitochondrial respiration and results in the cell stress of stressed tumour tissues (SDPT) [Fantin V. Et al 2006]. The stressed cells of stressed tumour tissues (SDPT) lose their normal mitochondrial function and change the way to obtain energy from mitochondrial respiration to the aerobic glycolysis, stressing more the cells stuck in a downward spiral [Fantin V. Et al 2006]. The difference between the ways to obtain energy of the two types of tumour proliferation can lie on the energy needs of each type of tumour proliferation. The stressed tumour tissues (SDPT) have an asynchronous demand of energy, dependent on the stress demand [Patra K.C. and Hay N. 2014], whereas non-stressed tumour tissues (TDPT) have a synchronous demand of energy, dependent on the tissue-remodelling timings. In this way, the priority for the stressed tumours (SDPT) seems to be to get energy rapidly, and for the non-stressed tumour tissues (TDPT) to obtain energy efficiently.

The different criterion to obtain energy of the two types of tumour proliferation found is also reflected in the expression data of study of the current work. PDP1 (pyruvate dehydrogenase phosphatase catalytic subunit 1) reverse the effects of pyruvate dehydrogenase kinases as previously seen with DCA (Figure 53). In our expression data, PDP1 appears over-expressed in the non-stressed tumour proliferation (blue sample clusters), maximising the mitochondria efficiency. PDP1 appears under-expressed in the chronic stress phenotype, pointing out a mitochondria deregulation.

This criterion to obtain energy seems to be also reflected in the number of mitochondria, which is larger in stressed tumour tissues (SDPT) than in non-stressed tumours (TDPT) [Langmesser S. and Albrecht U. 2006; Cuezva J.M. et al 2009]. After the quiescence, the number of mitochondria grows to satisfy the energy demands [Wagatsuma A. and Sakuma K. 2013]. PPAR $\gamma$  control mitochondrial biogenesis [Simmons G.E. et al 2015]. The action of PPAR $\gamma$  stopping the tumour proliferation of tissues like lung lies, at least in part, in increasing the number of mitochondria [Kim J. et al 2015], which recovers the mitochondrial respiration and finish the aerobic glycolysis, getting the cell out from the stress phenotype [Miglio G. et al 2009]. Again, as with DAC, after PPAR $\gamma$  activation, the cell go out from the stress phenotype by means of potassium channels like calcium-operated IK(ca) and BK(ca) potassium channels [LoVerme J1. et al 2006]. The different criterion followed by the two types of tumour proliferation to obtain energy determines the different membrane polarity of the two types of tumour proliferation.

Thus far, we know why the two types of tissues (SDPT and TDPT) have a different tumour proliferation, we also know that stressor agents stop tumour proliferation in tissues with a highly functional structure (TDPT), and relaxing agents stop tumour proliferation in tissues with a function limited to cell function (SDPT). We know that the two tumour proliferation ways follow a different criterion to obtain energy. Nevertheless, we **don't know how these two types of tumour proliferation ways are carried out.**

This could be linked to the aforementioned membrane potential and cell polarisation of stressed and unstressed cells. Whereas in the tumour proliferation of tissues with highly-functional structures (TDPT), the cell is hyperpolarised, a characteristic treat of non-stressed tissues (TDPT), in the tumour proliferation of tissues based on cell function (SDPT), the cell is depolarised, a characteristic treat of stressed tissues (SDPT) [Pokorný J. et al 2014].

The membrane potential of the cell membrane and the mitochondrial inner membrane are closely linked. Because both are sustained on the difference of potential at both sides of the membrane. Thus, a mitochondrial membrane depolarisation implies a cell membrane hyperpolarisation, and vice versa.

Again we can observe a difference between the two types of tumours found. The stressed tumours like non-small cell lung cancer or mammary-gland cancer have a hyperpolarised mitochondria and a depolarised cell membrane [Bonnet S. et al 2007] and the unstressed tumours have a depolarised mitochondria and a hyperpolarised cell membrane [Pokorný J. et al 2014]. As can be seen in the Table X. Remember that stressed tumours come from tissues which tissue function is mainly based on cell function and have proliferation areas and function areas (SDPT), and unstressed tumours come from tissues with a tissue function highly linked to the tissue structure (TDPT).

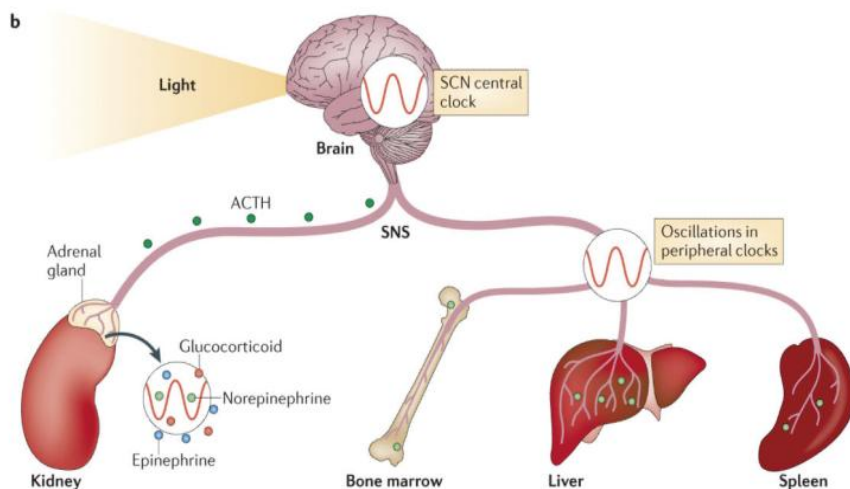
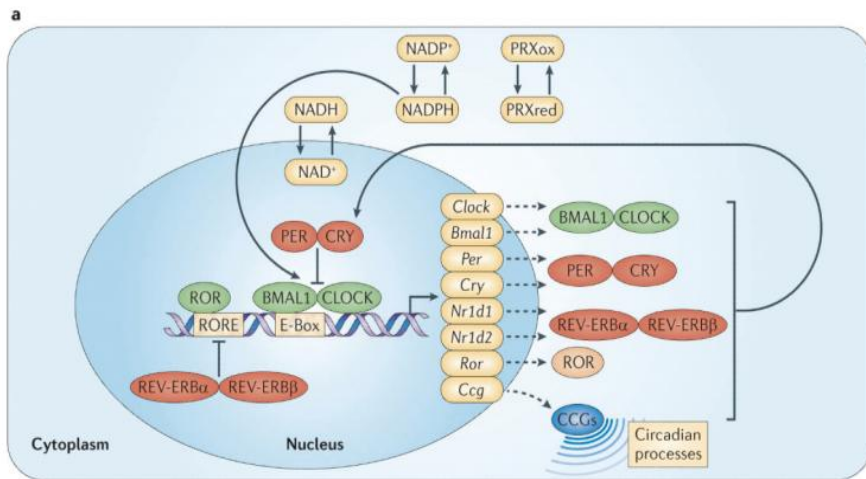
Low apparent potential	High apparent potential (hyperpolarized mitochondria)	Low or very low apparent potential
Normal epithelial cells	A great majority of: – adenocarcinoma – transitional cell carcinoma – squamous cell carcinoma – melanoma	– human oat cells – large cell carcinoma of lung – poorly differentiated carcinoma of the colon – leukemias – lymphomas – neuroblastomas – osteosarcomas
Low Rh123 uptake and retention	High Rh123 uptake and retention	Low or very low Rh123 uptake and retention

The data are taken from: Klingenberg and Rottenberg (1977), Modica-Napolitano and Aprille (1987), Chen (1988), O'Connor et al. (1988). Rh123 – Rhodamine 123

Table X. Mitochondrial inner membrane apparent potential of healthy and cancer cells. [Pokorný J et al 2014].

The transitions between the depolarisation and hyperpolarisation of the cells in healthy tissues are governed by circadian rhythms and activity, which alters these circadian rhythms prolonging the activity phase. The circadian rhythms guide the cyclic transitions between a stress phase and a post-stress tissue recovery phase. As a result of the changes in membrane potentials in tumour cells, these circadian rhythms are altered in tumour tissues [Lecarpentier Y. et al 2014]. In stressed tumour tissues (SDPT) the circadian stress phenotype is prolonged and in non-stressed tissues (TDPT), the circadian post-stress recovery phenotype is prolonged. In a similar way, a tumour in a peripheral organ alters the circadian rhythms of the organism [Savvidis C. and Koutsilieris M. 2006]. The stressed tumours prolong the activity phase of the organism, and the non-stressed tumours prolong the post-stress recovery phase of the organism. This affects, for example, the organism hormone production, retarding the cortisol production in early stress phenotype, or the melatonin production of the post-stress recovery phenotype [Savvidis C. and Koutsilieris M. 2006].

This fact also has medical implications, since hormones like melatonin or cortisol affect tumour development (with a different effect on stressed tumours (SDPT) and on non-stressed tumours (TDPT)). For instance, women with endometrial cancer (a stressed tumour) have lower melatonin levels [Viswanathan A. and Schernhammer E. 2008]. Releasing cortisol earlier in the day and melatonin at the beginning of the night, organism tries to synchronise the circadian clock of the different organs. In fact, when we administrate glucocorticoids to treat non-stressed tumour tissues (TDPT), we are acting in the same way than the suprachiasmatic nuclei (SNC) trying to adjust a disrupted circadian rhythm in a peripheral tissue (Figure 55.b). The hormone oxytocin launch the transition from the body's stress phase to the post-stress recovery phase [ Viero C. et al 2010; Lawson E.A. et al 2013]. Serotonin peaks follow oxytocin peak, and then serotonin is transformed into melatonin. Oxytocin also acts in peripheral tissues by means very different ways. It stops the tumour proliferation in stressed tumours (SDPT) like prostate [Whittington K. et al 2004]. And it allows the tissue remodelling in tissues that cannot combine proliferation and stress (TDPT) like bone. Oxytocin knockout mice have severe osteoporosis, and administration of oxytocin improves bone microarchitecture in these mice [Lawson E.A. et al 2013]. The oxytocin signal can be synchronous or asynchronous.



**Figure 55. (55.a) The molecular clock.** Transcription of the core clock genes *Bmal1* and *Clock* results in their heterodimerization in the cytoplasm and the translocation to the nucleus. These helix-loop-helix transcription factors bind to canonical E-Box sequences of clock-controlled genes (CCGs), driving circadian processes and their own expression. In addition, the expression of negative (PER and CRY) as well as positive (ROR) regulators of this cycle is induced. The PER/CRY complex represses the binding of BMAL1-CLOCK to target genes, whereas ROR induces *Bmal1* expression. After a period of time, the PER/CRY complex is degraded and BMAL1-CLOCK activates another transcription cycle. A second autoregulatory feedback loop is induced by transcription of REV-ERBs. The REV-ERB $\alpha$ -REV-ERB $\beta$  complex represses *Bmal1* transcription and competes with ROR for binding of ROREs. This pathway stabilizes the clock, and it can also directly drive circadian rhythms. The molecular clock does not only depend on transcriptional feedback but is also regulated by rhythms in the post-translational modifications of proteins, such as oxidation cycles of peroxiredoxins (PRXox(idized)/PRXred(uced)), as well as NADPH and NADH. NADPH and NADH can directly modulate the binding of the BMAL1/CLOCK complex to DNA. **(55.b) Entrainment and synchronisation.** Circadian rhythms are entrained by external cues, of which light is a major contributor. Light is processed via the retina, leading to synchronization of rhythms in hypothalamic suprachiasmatic nuclei (SCN), which comprise the master clock of the organism. From here, humoral and neural output systems modulate clocks in peripheral tissues via the hypothalamic-pituitary-adrenal (HPA) axis, setting a common circadian phase. Release of adrenocorticotrophic hormone (ACTH) from the pituitary gland cooperates with the sympathetic nervous system (SNS) to regulate rhythmic release of hormones (glucocorticoids, epinephrine and norepinephrine) from the adrenal glands. In addition, the SNS directly innervates tissues, and can modulate circadian rhythms locally via the cyclical release of norepinephrine from nerve varicosities [Scheiermann C. et al 2013].

The master regulator site of body circadian rhythms is the suprachiasmatic nucleus (SCN) inside the hypothalamus in which core clock genes are rhythmically expressed. In addition to this central clock, each organ has its own biological clock system, termed peripheral clock. The SCN and most peripheral tissues such as heart, blood vessels, skeletal muscles, kidneys, liver, and fat, govern numerous functions that are synchronized with the sleep-awake cycle. Peripheral clocks have their own regulatory mechanisms, which are specific to each peripheral organ by regulating the expression of clock-controlled genes. Both, central and peripheral clocks, are governed and modulated by the demands of activity through the changes in cell-membrane polarity [Colwell C.S. 2011]. These activity demands stress the cells and depolarise the cell membrane. The circadian switch of the central clock, between the stress phase and the post-stress recovery phase, can be induced by heat, as in our experiment on zebrafish. Zebrafishes can induce themselves a phenotypic change from stress phenotype to a post-stress recovery phenotype after a virus infection. They force this transition to the post-stress recovery phenotype swimming towards hotter waters, increasing their survival rates as a result [Boltana S. et al 2013]. Fever is a common response in mammals to induce the post-stress recovery phenotype (from a stress phenotype previously induced by a virus infection, among others). This recovery phenotype activates the immune system.

BMAL1 is the clock gene that initiates the gene activation downstream of the circadian activity phase. CRY1 retards the transition to the resting phase under demands of acute activity. PER2 tries to support the demands of acute activity and prepares the circadian resting phase to be proportional to the previous demand of activity. Bmal1 levels are high at the beginning of a subjective day and low at the beginning of a subjective night [Fu L. and Lee C.C. 2003]. Transcript levels of both, *Pers* and *Crys*, are anti-phase to *Bmal1* expression [Colwell C.S. 2011].

CRY depolarises the membrane through redox-based regulation of a K<sup>+</sup> channel conductance, stressing the cell in *Drosophila*. This CRY membrane depolarisation by potassium channel modulation depends on flavin-specific redox reactions [Fogle K.J. et al 2011]. BK channel expression is regulated by PER2. *Per2* mutants have shorter [Meredith A.L. et al 2006], hampering the transition to the recovery phase.

The affectation of the circadian rhythms by the two types of tumour proliferation found is also reflected in our expression data of study. In our expression data, CRY1 appears over-expressed in the stress phenotypes (red and yellow sample clusters) and in the stressed tumour proliferation phenotype (green clusters), expanding the circadian stress phase. CRY1 appears under-expressed in the non-stressed tumour proliferation, anticipating the circadian recovery phase in these tumours.

Circadian rhythms operate far-from- equilibrium and generate order spontaneously by exchanging energy with their external environment. PPAR $\gamma$  presents circadian properties which coordinate the interplay between metabolism and circadian rhythms. PPAR $\gamma$  is a peripheral regulator of rhythms like cardiovascular [Lecarpentier Y. et al 2014]. Deletion of PPAR $\gamma$  in mouse suppresses or diminishes the resting phase expanding the activity phase. Nocturnin binds to PPAR $\gamma$  and stimulates its transcriptional activity whereas its deletion suppresses PPAR $\gamma$  oscillations [Lecarpentier Y. et al 2014].

The relaxing action of PPAR $\gamma$  is not limited to the peripheral organs. Activation of central PPAR $\gamma$  reduces sympathetic excitation distressing the organism. PPAR $\gamma$  and sympathetic nerve activity (SNA) antagonistically regulate energy metabolism with the former promoting anabolism and the later favouring catabolism [Lecarpentier Y. et al 2014]. Global deletion of PPAR $\gamma$  in mice abolished or dampened circadian rhythms at both behavioural and cellular levels [Chen L. et al 2014].

Two major systems interfere with circadian genes: the canonical Wnt pathway, and the PPAR system. As shown in Figure 54 both are antagonists. PPAR $\gamma$  is linked to take the cell out of stress and Wnt pathway is linked to stressed-tumour proliferation. Activation of the Wnt/ $\beta$ -catenin signaling pathway decreases PPAR $\gamma$  activity in colon cancer cells and a loss-of-function mutations in PPAR $\gamma$  is associated with human colon cancer [Lecarpentier Y. et al 2014]. There is an opposite between activation of the Wnt pathway and PPAR $\gamma$ , and the gene finally activated will decide the future of the cell. Activation of the Wnt system with inactivation of PPAR $\gamma$  favours diabetes, hypertension, cancers in (SDPT) tissues that develop stressed tumours, and neurodegenerative diseases [Lecarpentier Y et al 2014]. All of these pathologies are linked to cell stress. The reverse is observed in osteoporosis, Alzheimer disease, bipolar disorder, schizophrenia, and myocardial ischemia [Lecarpentier Y et al 2014]. All of them are linked to a progressive loss in number of cells possibly due to inactivity.

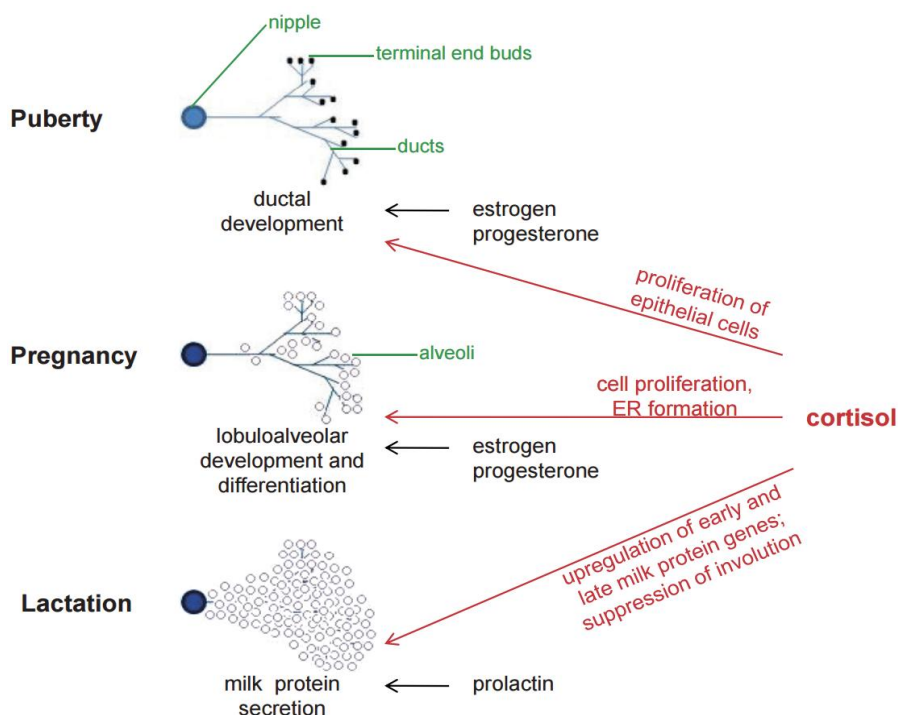
As aforementioned, colon cancer is linked to a circadian rhythm dysregulation. Concretely, extending the circadian active phase. Down-regulation of PER2 increases  $\beta$ -catenin protein levels and leads to cell proliferation in colon cancer cell lines and colonic polyp formation. PER2 gene activation suppresses tumorigenesis in colon by down-regulation of  $\beta$ -catenin. Increased  $\beta$ -catenin affects the circadian clock and enhances PER2 protein degradation in colon cancer [Lecarpentier Y et al 2014]. PPAR $\gamma$  can revert this dysregulation of the circadian rhythm, promote the change of circadian phase, and lead the cell from the stress phenotype to the post-stress recovery phenotype. PER proteins are considered to be inhibitors of cell cycle progression and  $\beta$ -catenin activation. BMAL1 on the other hand, activates cell proliferation,  $\beta$ -catenin pathway, and is strongly associated with colon cancer initiation and poor clinical outcome [Karantanos T. et al 2014]. PER2 has been reported to be a tumours suppressor gene whose expression inhibits the formation of a variety of tumours, including breast, prostate, lung, and lymphoma [Hill S.M. et al 2015].

Melatonin impedes the stress phenotype repressing the transcriptional activity of ROR $\alpha$ 1, a member of the NR/steroid hormone receptor superfamily, and a transcriptional inducer of the core clock gene BMAL1 (Figure 55.a). In breast cancer and human breast epithelial cells, melatonin administration significantly repressed ROR $\alpha$ 1 transactivation by inhibiting its induction of BMAL1 gene expression [Hill S.M. et al 2015]. Melatonin can restore the levels of Per and Cry, and lowering the levels of Bmal1, reducing the prostate cancer growth [Jung-Hynes B. et al 2010].

Actually the pathological disruption of the circadian rhythms is closely related with the circadian control of the healthy tissues. Sometimes, this circadian disruption is non pathological and entirely functional, but if the control fails can quickly become pathological. This close connection between the functional and pathological switches, between the stress phase and the recovery phase, can be clearly seen in the circadian alterations of mammary-gland cycles and endometrial cycles.

Mammary glands proliferate as required during pregnancy, and undergo apoptosis or reversible transdifferentiation during involution once lactation is complete. During pregnancy a non-stressed proliferation of undifferentiated mammary glands is performed [Briskin C. 2002; Owens T.W. et al 2014]. When the lactation phase starts, mammary glands become stressed, until lactation ends (Figure 56). As aforementioned, the tissue function of mammary glands is limited to their cell function (SDPT), which allows combining proliferation areas and functional areas (Figure 49). Thus, this tissue could develop stressed tumours, and the proliferation in the lactation phase becomes tumoural. Remember that the stressed proliferation of mammary glands can be induced by their specific stressor prolactin [Wennbo H. and Tömel J. 2000].

During pregnancy and lactation, the PPAR $\gamma$  mRNAs decreased [Gimble J.M. et al 1998]. After lactation is complete, the PPAR $\gamma$  rises up, and the mammary gland undergoes remodelling to a pre-pregnant-like resting structure wherein these milk-producing cells are cleared by apoptosis. PPAR $\gamma$  seems to be involved in this phenotype transition [Apostoli A.J. et al 2014]. Low levels of PPAR $\gamma$  at the end of lactation is related with a higher risk of breast tumorigenesis [Apostoli A.J. et al 2014] possibly because makes it difficult to leave the cell stress phase. PPAR $\gamma$  mRNA was not detected in several induced mammary tumours [Gimble J.M. et al 1998] and among cells expressing PPAR $\gamma$ , activation of PPAR $\gamma$ -dependent signalling suppresses mammary tumour growth [Apostoli A.J. et al 2014].



**Figure 56. Role of cortisol in mammary gland development.** The role of cortisol is shown for the different post-embryonic developmental stages of the mammary gland. Other hormones involved in the different developmental stages are also listed. Estrogen and progesterone promote ductal system proliferation during puberty. However, the DNA binding function of the glucocorticoid receptor also appears to be required. During pregnancy, cortisol contributes to lobuloalveolar development of the mammary gland, in conjunction with estrogen and progesterone. Prolactin and cortisol prepare the mammary cells for lactation and stimulate milk protein production following parturition. In addition, cortisol contributes stressing the cells to the maintenance of lactation by suppressing involution. ER, endoplasmic reticulum [Antonova L. et al 2011].

Suppression of mammary clock gene expression rhythms represents a physiological adaptation to suckling cues. Tissue-specific changes in molecular clocks occurred from late pregnant and early lactation mice. Attenuated rhythms appeared to be a physiological adaptation of mammary to lactation. Induction of differentiation of mammary epithelial cell line HC11 with dexamethasone, insulin, and prolactin resulted in similar changes, and prolactin induced phase shifts in HC11 Arntl expression rhythm [Casey T.M. et al 2014]. Changes in core molecular clock genes occurred across multiple tissues during the transition from pregnancy to lactation because these circadian alterations affect the entire organism. The metabolic output of the SCN during lactation may be increased and affect peripheral clock function. Changes in SCN PER2 levels coincided with an increase in plasma corticosterone rhythm amplitude from late pregnancy to lactation day 3, with basal corticosterone levels of lactating dams significantly greater than pregnant dams [Casey T.M. et al 2014]. This fact implies a stressed phenotype in mammary glands during lactation, a stressed phenotype that is reverted when lactation is finished by the expected PPAR $\gamma$  activation. Mammary clock underwent prominent changes during the transition from pregnancy to lactation, with the ratio of CLOCK to PER2 4-fold higher in lactating versus pregnant glands in the moment of transition. Next it is stabilised and PER2 expression during lactation becomes lower than in pregnancy. [Casey T.M. et al 2014].

Normal breast exhibits rhythmic properties linked to the hormonal environment of the gland in animals and humans. Breast cancer in humans is characterized by a modification of these circadian patterns [Coudert B. Et al 2002]. Induced mammary gland cancer in rodents lost PER expression. BMAL1 repression and restored PER expression previously lost have been linked to revert the previously induced mammary-gland cancer in rodents [Fang M. et al 2015]. This action is mediated via SIRT1 and also has been observed in prostate tumourigenesis, especially during early stages [Fang M. et al 2015]. Increased  $\beta$ -catenin and cell proliferation in murine breast cancer cells results from down regulation of PER1 and PER2 [Wood P.A. et al 2009]. Furthermore, constant exposure to light (Figure 55.b) showed significantly increased growth of induced mammary adenocarcinomas in female rats when they were compared to the light-dark groups [Kochan D.Z. and Kovalchuk O. 2015]. In summary, the mammary gland tumour (alveolar) seems to be linked to an elongation of the circadian stress phase.

The circadian alterations in endometrial-gland cycle seem to be similar than the ones seen in mammary-gland cycle (Figure 56). Their connection with tumourigenesis seems to be also similar. In humans, the differentiation of the endometrial glands starts to take place at the beginning of the luteal phase of the menstrual cycle [Maruyama T. and Yoshimura Y. 2008], previous to enter in the stressed phenotype from the undifferentiated proliferative phase. The circadian activity phase stresses the endometrium and is governed by progesterone signal (Figure 57). When the tissue backs to the post-stress circadian recovery phase, the mense happens. Endometrial carcinoma is believed to develop as a result of persistent unopposed progesterone stimulation on the endometrium resulting in increased risk of malignant transformation [Nickkho-Amiry M. et al 2012]. PPAR $\gamma$  activation reduces proliferation of endometrial cells and PPAR $\gamma$  may be reduced in endometrial cancers [Nickkho-Amiry M. et al 2012], reducing the capacity of the cells to leave the stress phenotype. The endometrial cancer could stop the stressed tumour proliferation leaving the stress phenotype [Wong J.Y. et al 2008]. PPAR $\gamma$  activation reduces proliferation of endometrial cells via regulation of PTEN (Figure 54). When the endometrium tumour ends, the endometrium



degrades and peels with the menstruation (a menstruation with big clots). Notice that estrogens and FSH are acting as growth factors, and LH and progesterone (and prolactin) as stressors (Figure 56 and 57).

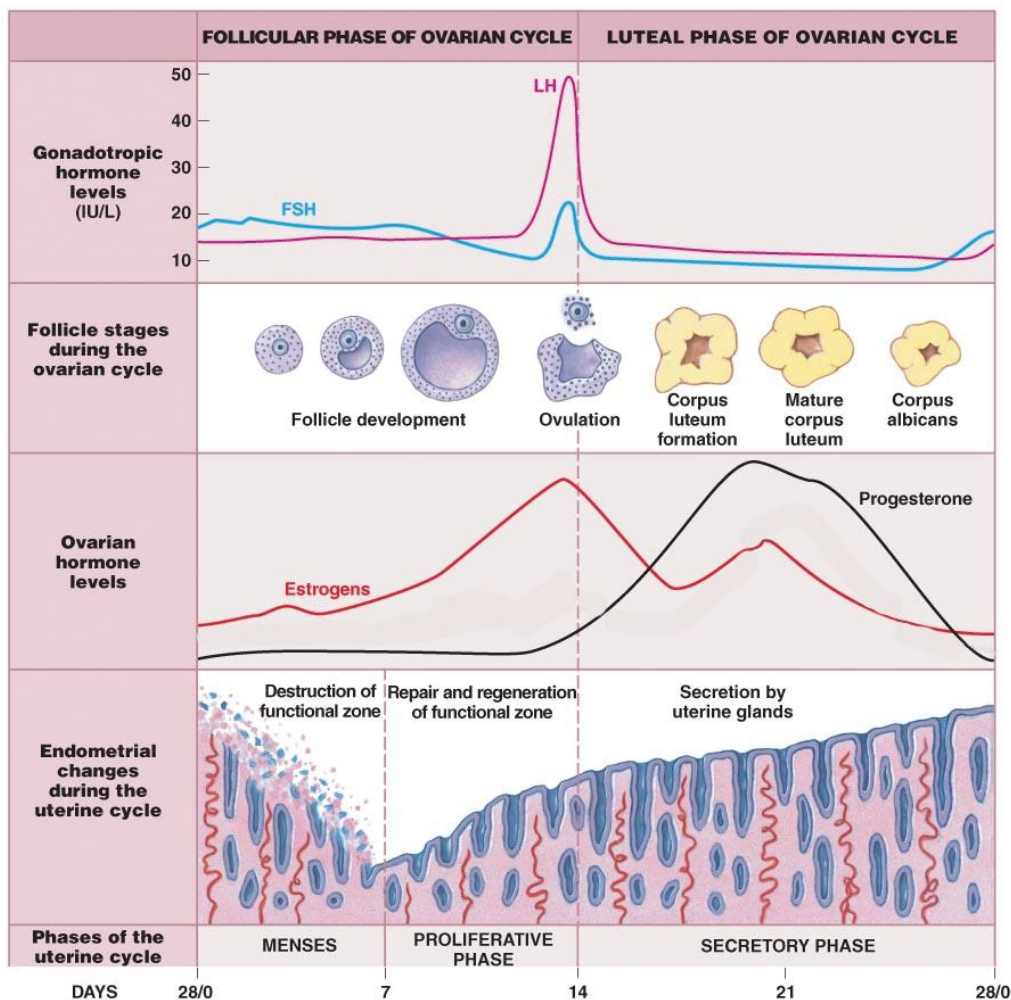


Figure 57. The human female menstrual cycle is divided into several phases, which vary in length among women and among cycles, and have an ovarian and endometrial cycles totally synchronized. The menstrual cycle begins the day that menstrual bleeding starts. A decrease in the levels of estrogen and progesterone makes the endometrium unstressed, and triggers the top layers of the thickened endometrium to break down and be shed, resulting in bleeding. Concomitant with this, levels of FSH increase very slightly, stimulating the development of several oocyte-containing follicles. FSH levels subsequently decrease and only one or two follicles continue to develop. The developing follicles release estrogen, and this initiates thickening of the endometrium in a non-estressed proliferative phase of the endometrium. The ovulatory phase begins when levels of LH and FSH increase dramatically; levels of estrogen also peak at this time, and levels of progesterone begin to increase. The high levels of LH stimulate ovulation. The final phase of the menstrual cycle is the luteal phase. During this phase, levels of LH and FSH decrease and the ruptured follicle forms the corpus luteum, which produces large amounts of progesterone. In the secretory phase of the endometrium, this is stressed by the higher levels of progesterone and is receptive to implantation of an embryo if fertilization has occurred. In the absence of fertilization, the corpus luteum degenerates and the loss of progesterone production stress out the endometrial glands and initiates a new menstrual cycle [Aitken R.J. et al 2008].

In the same way that the pass from the stress phenotype to the post-stress recovery phenotype is linked to a circadian switch and clock-genes expression variations, the pass from post-stress recovery phenotype to the stress phenotype is linked to a circadian switch and clock-genes expression variations. This fact affects the tumour proliferation of non-stressed tumour tissues (TDPT). When the non-stressed tumour stops, the cells are differentiated (remember that the cells of the non-stressed tumour proliferation are undifferentiated) and become functional again. The cell membrane is depolarised and the sodium channels are involved in this membrane depolarisation. The NEDD4L transition from its over-expression to its under-expression takes part in this sodium-channel increased activity [Araki N. et al 2009]. The depolarisation of the cell and the normal activity of the tissue reestablishes the normal circadian rhythms normalizing the BMAL1 expression. This BMAL1 expression is reestablished by the acute membrane depolarisations and hyperpolarisations of normal activity [Nitabach M.N. et al 2002]. As well as by means of direct gene action or other cycles like NADPH and NADH, or cAMP-cGMP cycles, characteristic of the activity phase (Figure 55). When the non-stressed proliferation stops, the mitochondria biogenesis is promoted being ready to be stressed [Wagatsuma A. and Sakuma K. 2013].

The non-stressed tumour proliferation of tissues that cannot combine stress and proliferation (TDPT) can be stopped acting directly on the clock gene expression. BMAL1 is a clear target, since it controls the start of the circadian active phase and gene downstream. BMAL1 may be epigenetically silenced in ovarian cancer and induced over-expression of BMAL1 inhibited cell growth in ovarian cancer cells. This over-expression of BMAL1 restored the rhythmic activity of c-MYC in ovarian cancer cells [Yeh C.M. et al 2014]. If it is manipulated the expression of a non-voltage sensitive  $K^+$  channel assess the impact of chronically hyperpolarizing. This chronically hyperpolarizing blocked the rhythm in expression of the PER [Nitabach M.N. et al 2002] disrupting the circadian rhythm.

The sodium influx after the pass from the post-stress recovery phenotype to the stress phenotype depolarises the cell [Colwell C.S. 2011]. When the stress demands are being satisfied, the cell depolarisation increases, and the potassium/sodium exchange starts. The membrane polarity does not change with this action, but the cell is ready to open the potassium channels and hyperpolarize the cell, leaving the stressed phenotype if necessary. The inhibition of potassium channels and  $Na^+/K^+$ -ATPase results in cell depolarisation, opening of voltage-gated  $Ca^{2+}$  channels, and increase of intracellular  $Ca^{2+}$  concentration [Boukroun S. et al 2015]. The  $Na^+-K^+-2Cl^-$  cotransporter NKCC1 raises intracellular  $Cl^-$  levels and thus contributes to a more depolarised  $E_{Cl^-}$  [Colwell C.S. 2011]. In the mammalian SCN, the BKs are also circadian regulated, with peak activity at night. There is also a diurnal rhythm in  $Na^+/K^+$ -ATPase activity, which is higher during the day.  $Ca^{2+}$  levels peak occurs during the day. cAMP levels peak occurs also during the day and precede the peak in the activity rhythm. Reduced cAMP levels greatly reduce PER2 expression [Colwell C.S. 2011], possibly because a reduced post-stress recovery phase is necessary due to a previous reduced stress phase. Notice that the activity demands are who governs the circadian-clock entrainment and the post-stress recovery phase seems to be a response to the stress phase [Scheiermann C. et al 2013].

Deletion of BK channels has little systematic effect on gene expression. Furthermore, no broad transcriptional differences were revealed between light-dark versus constant darkness conditions [Meredith A.L. et al 2006]. This fact points out that the circadian phenotypic changes could be very

dramatic or unappreciable, possibly depending on the previous phenotype from which the tissue comes.

Bone is a tissue with a function highly dependent on tissue structure (TDPT) and develops non-stressed tumours. Circadian rhythms are prevalent in bone metabolism, and control the osteoclastic-osteoblastic switch. Output signals from the suprachiasmatic nucleus (SCN) are transmitted from the master circadian rhythm to peripheral osteoblasts through  $\beta$ -adrenergic and glucocorticoid signalling and mediate circadian timing to peripheral osteoclasts [Fujihara Y. et al 2014]. Osteoclast numbers and bone resorption are rapidly increased by glucocorticoids, extending the stress phase, together with osteoblast inactivation and decreased bone formation [Frenkel B. et al 2015]. That is, retarding the post-stress recovery phase. In bone tumours, the osteoblastic phase is largely increased, but preceded by an increased osteoclastic phase. Glucocorticoids can stop the osteoblastic phase stressing the tissue, even in tumour proliferation [Yamamoto T. et al 2002]. *Cry2* and *Per2* affect distinct pathways in the regulation of bone volume with *Cry2* influencing mostly the osteoclastic cellular component of bone and *Per2* acting on osteoblast parameters. *Per2*<sup>Brdm1</sup> animals display increased bone formation and *Cry2*<sup>-/-</sup> mice decreased bone resorption [Maronde E. et al 2010].

Here we have then, another open question; the response to the acute activity demands could be mediated by cGMP in the cAMP-cGMP cycle providing a proportional response to the acute activity demands. The response to the daily activity demands could be mediated by PER and the clock genes, that can provide a proportional response to the daily activity demands. The mechanisms that guide the circadian alterations of the endometrial and mammary gland cycles are well known. But, if there exist, how and who mediates when the post-stress recovery phase turns pathological like in non-stressed tumours? Going back to the case of Bone tumours, is the osteoblastic phase dependent on the osteoclastic phase in duration or intensity? Is the pathological post-stress recovery phase dependent on the previous stress phase? If the answer is yes, how is this response mediated?.

In bone, the proportional post-stress recovery phase to the previous stress phase is controlled by the microenvironment. The osteoblast activity is proportional to the previous osteoclast activity level in healthy tissues. This “memory” could be mediated by osteoclast-derived factors that can act on cells that would transmit further signals to osteoblasts, or by physical changes recognized by osteoblasts, among others [Sims N.A. and Martin T.J. 2015]. Cervix tumours are also non-stressed tumours and seem to be preceded in many cases by the infection with certain papillomavirus (HPV) [Foppoli C. et al 2015]. The majority of HPV infections are subclinical, and the development of a tumour may be subjected to the stress level achieved in the infection phase [Foppoli C. et al 2015]. In this way, the stress level would determine if a tumour does develop in the post-stress recovery phase.

The same peripheral organs have tissues of the two types described here (SDPT and TDPT), and sometimes it is no easy to recognise each one, event less in advanced cancers. As the tissues with the tissue function based on tissue structure (TDPT) are structurally more complex and imply an important evolutionary step with respect to the tissues with a tissue function mainly sustained on cell function (SDPT), a tissue classification based on embryonic origin can help us. A possible tissue classification would be as shown in the table XI.

ENDODERM	MESODERM		ECTODERM
	PARENCHYMAL	MESENCHYMAL	
Fharynx	Dermis (chorion)	Conjunctive tissue	Flat epithelium
Hypophysis	Mammary gland	Bones	Bronchia epithelium
Parotid	Pericardium	Lymphatic ganglia	Larynx epithelium
Sublingual gland	Pleura	Blood vessels	Coronary epithelium
Palate	Peritoneum	Lymphatic vessels	Uterus
Tonsils		Muscles	Lower part of the stomach and duodenal bulb
Thyroid		Ovaries	Bile duct epithelium
Parathyroid		Testicles	Intra-pancreatic ducts
Eustachian tubes		Endocardium	Rectum epithelium
Middle ear			Ureter epithelium
Lacrimal glands			Urethra
Oral submucosa			Skin (epidermis)
Alveoli			Hair
Lower third of the esophagus			Retina
Upper portion of the stomach			Cornea
Duodenum , except bulb			Crystalline
Liver			Breast ducts
Pancreas			Nervous system
Small intestine			Nasal and buccal mucosa
Large intestine			Upper esophagus
Endometrium			Tear ducts
Prostate			Buccal salivary ducts
Fallopian tubes			Vitreous bodies
Bladder submucosa			
Kidney collecting tubules			
<b>SDPT</b>		<b>TDPT</b>	

**Table XI. Tissues classified by embryonic origin.** The tissues in green have proliferative areas and functional areas, or are developing very simple tasks, thus, they can develop stressed proliferative tumours (SDPT). The function of the tissues in blue is linked to their structure or complex relation among their cells, so their proliferation cannot be stressed (TDPT).

In summary, the two types of tumour proliferation found, and the two tissue types that they are linked to, could be described as:

**SDPT** : Spatial Dependent Proliferation Tissues. The tissue has proliferation areas and function areas. This fact allows the tissue to separate stress and proliferation in space and not in time. Tissue function is mainly sustained on cell function. The cells are differentiated and stressed during the tumour proliferation. The tumour cells are depolarised. The tumour proliferation develops during the circadian stress. Relaxing agents can stop the tumour proliferation. The embryonic origin of the tissue is endoderm or parenchymal mesoderm.

**TDPT** : Time Dependent Proliferation Tissues. Tissue function is highly linked to the tissue structure. This fact forces the tissue to separate stress and proliferation in time. The cells are undifferentiated and cannot be stressed during tumour proliferation. The tumour cells are

hyperpolarised. The tumour proliferation develops during the circadian post-stress recovery. Stressor agents can stop the tumour proliferation. The embryonic origin of the tissue is ectoderm or mesenchymal mesoderm.

Understanding all these complex links between tumour prognosis and the original tissues physiology is the current target of our study. The problem is addressed from a genetic perspective, trying to understand how to control this. In the workflow of Figure 44 it can be observed statistical procedures that have been applied in the analysis, how the results of these procedures are crossed to facilitate new analyses, as well as different points in which a crossing with external databases is needed.

Even though there is an important amount of manual work analysing and contrasting the results, these would be impossible to obtain without the multiple analysis tools of the pipeline. It should be remembered that this analysis approach is addressed to important issues, very well documented but with an unsolved paradox. Like in the case of glucocorticoids use in cancer. The most important of methodologies and tools are the results that they are able to provide, as this is the ultimate goal of any methodology. Currently, agents like those cited in this work are administered as adjuvant to treat undesirable side-effects of radiotherapy, surgery or cancer itself. But there are still many remaining questions to answer.

All the supplementary material is available at: <http://platypus.uab.es/GCinC>

In the supplementary material there are more analyses trying to describe the two ways for tumour proliferation found, with more genes of our expression data of study being involved in it. We encourage visiting them.

0	6
---	---

<b>CONCLUSIONS</b>	
--------------------	--

Cl.	
-----	--

What we have achieved.
------------------------



## 1.

Gene networks based on non-linear expression relationships identify new hubs of genes with respect to consider only linear expression relationships. Furthermore, the researcher interests are present in all the operations of our tools: the genes linking the query ones in expression terms are supplied, the sample clusters are successively redefined in subclasses. The access to remote databases enables expanding the analysis beyond the expression relationships. Thus, the user can enrich his/her expression data and sample clusters, improve his/her future experimental design and check the hypotheses generated from the data. All the above, together with the analysis of the expression-dependence fluctuations form a new approach which helps to:

- Identify and analyse the biological role of unknown members of a genetic pathway.
- Gain insight into the relationships (known or hidden) among different pathways.
- Analyse how the genes could be affected when a query gene is in a particular level of expression.
- Helps gene/protein function-prediction.
- Identifies and clarifies the modulation of putative drug-target genes/proteins in relation with the molecular basis and action on a disease phenotype.
- Helps to model whole cell physiology (systems and integrative biology).

## 2.

This approach not only permits the analysis of the complex expression relationships between single genes, but also between sets of coexpressed genes. As aforementioned, we obtain networks of sets of coexpressed genes where each set performs a different cell function. In these networks, all the sets of coexpressed genes maintain a non-linear expression relationship with each and every one of the other sets of the network. That is, anytime one knows how to move from one process to another, passing by any other intermediate process. As a result, multiple networks are provided by a dynamic system that allows detecting sets of coexpressed genes related between them by complex activation dependences.

## 3.

In addition, clustering the samples based on non-linear expression relationships can help us to automatically discover the concurrent phenotypic changes hidden in expression data. Our procedure shows the intersections among transverse phenotypes involving concrete sets of genes, and the web tools allow us to compare these transverse-phenotype intersections with the sample clusters obtained by clustering methods. In this way, it is possible to study the similarities and differences among the sample clusters obtained by common clustering methods and explain the cellular processes behind each sample cluster obtained.

## 4.

All of this agrees with the principal objective of our research: "Reusing expression data for studying the cell behaviour and phenotypic changes from a holistic point of view". This holistic point of view can cover fields like: cellular stress, cell proliferation, tissue remodelling, mitochondrial activity, oxygen and ROS/NOS levels, cell differentiation and un-differentiation, membrane polarity,



intracellular and extracellular pH levels, circadian rhythms, the inference of sympathetic and parasympathetic nervous-system, metastasis, or interaction with virus and bacteria.

## 5.

Finally we show how we can solve paradoxes or apparently incongruent experimental results by the acute dissection of the interdependencies among the phenotypes described by sample clusters using our tools. Specifically, this capability has allowed the study of the contradictory effects of glucocorticoids in cancer progression. The correct identification and characterization of the two types of tumour proliferation may help to adjust the treatments correctly in cancer and other pathologies. It could also help in understanding and preventing future malignancy, recurrence and metastasis. As aforementioned, controlled randomized trials evaluating the impact of these stressor agents on growth and metastasis in stressed (SDPT) and non-stressed (TDPT) tumour tissues need to be performed including patient-survival evaluation. Remember that agents like those cited in the work are usually administered as adjuvant to treat undesirable side effects of cancer therapy. But beyond the glucocorticoid paradox, future researches in cancer, among others, should start to consider the two types of tumour proliferation found.

## 7. Publications

Delicado P and **Huerta M**. Principal curves of oriented points: Theoretical and computational improvements. *Computational Statistics*, 2003, 18:293-315.

Cedano J, **Huerta M**, Estrada I, Ballosera F, Conchillo O, Delicado P and Querol E. A web server for automatic analysis and extraction of relevant biological knowledge. *Comput Biol Med.*, 2007, 37:1672-1675.

Cedano J, **Huerta M** and Querol E. NCR-PCOPGene: An Exploratory Tool for Analysis of Sample-Classes Effect on Gene-Expression Relationships. *Adv Bioinformatics*, 2008, 789026.

**Huerta M**, Cedano J and Querol E. Analysis of non-linear relation between expression profiles by the Principal Curves of Oriented-Points approach. *Journal of Bioinformatics and Computational Biology*, 2008, 6(2):367-86.

**Huerta M**, Cedano J, Peña D, Rodriguez A and Querol E. PCOPGene-Net: holistic characterisation of cellular states from microarray data based on continuous and non-continuous analysis of gene-expression relationships. *BMC Bioinformatics*, 2009, 10:138.

Boltana S, Rey S, Roher N, Vargas R, **Huerta M**, Huntingford FA, Goetz FW, Moore J, Garcia-Valtanen P, Estepa A and Mackenzie S. Behavioural fever is a synergic signal amplifying the innate immune response. *Proc Biol Sci.*, 2013, 280(1766):20131381.

**Huerta M**, Casanova O, Barchino R, Flores J, Querol E, Cedano J. Studying the complex expression dependences between sets of coexpressed genes. *Biomed Res Int.*, 2014, 940821.

**Huerta M**, Fernández-Márquez J, Cabello JL, Medrano A, Querol E and Cedano J. Analysis of gene expression for studying tumor progression: the case of glucocorticoid administration. *Gene*, 2014, 549(1):33-40.

**Huerta M**, Munyi M, Expósito D, Querol E and Cedano J. MGDB: crossing the marker genes of a user microarray with a database of public-microarrays marker genes. *Bioinformatics*, 2014, 30(12):1780-1.

**Huerta M**, Gispert B, Querol E, Cedano J. Deconstructing sample clusters in multiple concurrent phenotypic changes. *Plos One*, 2015 [submitted].

**Huerta M**, Hernandez C, Yesid J, Querol E, Cedano J. Crossing Clusters: Studying the relationships among the sample clusters obtained by clustering methods from expression data. *Bioinformatics*, 2015 [submitted].



## 8. Bibliography

Abdullah-Sayani A, Bueno-de-Mesquita JM and van de Vijver MJ. Technology Insight: tuning into the genetic orchestra using microarrays--limitations of DNA microarrays in clinical practice. *Nat Clin Pract Oncol.*, 2006 Sep;3(9):501-16.

Ahren K and Jacobsohn D. The Action of Cortisone on the Mammary Glands of Rats under Various States of Hormonal Imbalance. *Acta Physiologica Scandinavica*, 2008, 40(2-3):254-274.

Aitken RJ, Baker MA, Doncel GF, Matzuk MM, Mauck CK and Harper MJ. As the world grows: contraception in the 21st century. *J Clin Invest.*, 2008, 118(4):1330-43.

Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E, Buzadzija K, Cavero R, D'Abreo C, Donaldson I, Dorairajoo D, Dumontier MJ, Dumontier MR, Earles V, Farrall R, Feldman H, Garderman E, Gong Y, Gonzaga R, Grytsan V, Gryz E, Gu V, Haldorsen E, Halupa A, Haw R, Hrvojic A, Hurrell L, Isserlin R, Jack F, Juma F, Khan A, Kon T, Konopinsky S, Le V, Lee E, Ling S, Magidin M, Moniakis J, Montojo J, Moore S, Muskat B, Ng I, Paraiso JP, Parker B, Pintilie G, Pirone R, Salama JJ, Sgro S, Shan T, Shu Y, Siew J, Skinner D, Snyder K, Stasiuk R, Strumpf D, Tuekam B, Tao S, Wang Z, White M, Willis R, Wolting C, Wong S, Wrong A, Xin C, Yao R, Yates B, Zhang S, Zheng K, Pawson T, Ouellette BF, and Hogue CW. The Biomolecular Interaction Network Database and related tools. *Nucleic Acids Research*, 2005, 33:D418-424.

Alon U, Barkai N, Notterman D A, Gish K, Ybarra S, Mack D and Levine A J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, 1999, 96:6745-6750.

Andreopoulos B, An A, Wang X and Schroeder MA. Roadmap of clustering algorithms: finding a match for a biomedical application. *Brief Bioinform.*, 2009, 10:297-314.

Antonov AV, Tetko IV and Mewes HW. A systematic approach to infer biological relevance and biases of gene network structures. *Nucleic Acids Research*, 2006, 34(1):e6.

Antonova L, Aronson K and Mueller CR. Stress and breast cancer: from epidemiology to molecular biology. *Breast Cancer Res.*, 2011, Apr 21;13(2):208.

Apostoli AJ, Skelhorne-Gross GE, Rubino RE, Peterson NT, Di Lena MA, Schneider MM, SenGupta SK, and Nicol CJ. Loss of PPAR $\gamma$  expression in mammary secretory epithelial cells creates a pro-breast tumorigenic environment. *Int J Cancer.* 2014, 134(5):1055-66.

Araki N, Ishigami T, Ushio H, Minegishi S, Umemura M, Miyagi Y, Aoki I, Morinaga H, Tamura K, Toya Y, Uchino K and Umemura S. Identification of NPC2 protein as interaction molecule with C2 domain of human Nedd4L. *Biochem Biophys Res Commun*, 2009, 388(2):290-296.

Arkin A, Ross J and McAdams HH. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics*, 1998, 149:1633-48.

Ashburner M, Ball C A, Blake J A, Botstein D, Butler H, Cherry J M, Davis A P, Dolinski K, Dwight S S, Eppig J T, Harris M A, Hill D P, Issel-Tarver L, Kasarskis A, Lewis S, Matese J C, Richardson J E, Ringwald M, Rubin G M, and Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*. 2000, 25:25-9.

- Badawi AF, Eldeen MB, Liu Y, Ross EA and Badr MZ. Inhibition of rat mammary gland carcinogenesis by simultaneous targeting of cyclooxygenase-2 and peroxisome proliferator-activated receptor gamma. *Cancer Res.*, 2004, 64(3):1181-9.
- Barchuk AR, Cristino AS, Kucharski R, Costa LF, Simões ZL and Maleszka R. Molecular determinants of caste differentiation in the highly eusocial honeybee *Apis mellifera*. *BMC Dev Biol.*, 1998, 7: 70.
- Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W and Edgar R. NCBI GEO: mining millions of expression profiles, database and tools. *Nucleic Acids Research.*, 2005, 33:D562-566.
- Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muerter RN, Holko M, Ayanbule O, Yefanov A and Soboleva A. NCBI GEO: archive for functional genomics data sets--10 years on. *Nucleic Acids Research*, 2011, 39, D1005-1010.
- Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S and Soboleva A. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Research*, 2013, 41(Database issue):D991-995.
- Becksei A and Serrano L. Engineering stability in gene networks by autoregulation. *Nature*, 2000, 405:590-3.
- Bertucci F, Finetti P, Rougemont J, Charafe-Jauffret E, Cervera N, Tarpin C, Nguyen C, Xerri L, Houlgatte R, Jacquemier J, Viens P and Birnbaum D. Gene expression profiling identifies molecular subtypes of inflammatory breast cancer. *Cancer Res.*, 2005, 65(6):2170-8.
- Bielas J H, and Heddle J A. Elevated mutagenesis and decreased DNA repair at a transgene are associated with proliferation but not apoptosis in p53-deficient cells. *Proceedings of the National Academy of Sciences of the United States of America*, 2003, 100:12853-8.
- Billing AM, Fack F, Renaut J, Olinger CM, Schote AB, Turner JD and Muller CP. Proteomic analysis of the cortisol-mediated stress response in THP-1 monocytes using DIGE technology. *J Mass Spectrom.*, 2007, 42(11):1433-1444.
- Binder EB. The role of FKBP5, a co-chaperone of the glucocorticoid receptor in the pathogenesis and therapy of affective and anxiety disorders. *Psychoneuroendocrinology*, 2009, 34 Suppl 1:S186-195.
- Boltana S, Rey S, Roher N, Vargas R, Huerta M, Huntingford FA, Goetz FW, Moore J, Garcia-Valtanen P, Estepa A and Mackenzie S. Behavioural fever is a synergic signal amplifying the innate immune response. *Proc Biol Sci.*, 2013, 280(1766):20131381.
- Bonnet S, Archer SL, Allalunis-Turner J, Haromy A, Beaulieu C, Thompson R, Lee CT, Lopaschuk GD, Puttagunta L, Bonnet S, Harry G, Hashimoto K, Porter CJ, Andrade MA, Thebaud B and Michelakis ED. A mitochondria-K<sup>+</sup> channel axis is suppressed in cancer and its normalization promotes apoptosis and inhibits cancer growth. *Cancer Cell*, 2007, 11(1):37-51.
- Boulkroun S, Fernandes-Rosa FL and Zennaro MC. Molecular and Cellular Mechanisms of Aldosterone Producing Adenoma Development. *Front Endocrinol (Lausanne)*, 2015, 6:95.
- Brabletz S, Schmalhofer O and Brabletz T. Gastrointestinal stem cells in development and cancer. *J Pathol.*, 2009, 217: 307–317S.
- Braun R, Leibon G, Pauls S and Rockmore D. Partition decoupling for multi-gene analysis of gene expression profiling data. *BMC Bioinformatics*, 2011 Dec, 12:497.

- Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bähler J, Wood V, Dolinski K and Tyers M. The BioGRID Interaction Database. *Nucleic Acids Research*, 2008, 36:D637-640.
- Briskin C. Hormonal control of alveolar development and its implications for breast carcinogenesis. *J Mammary Gland Biol Neoplasia*. 2002 Jan;7(1):39-48.
- Brunel H, Gallardo-Chacón JJ, Buil A, Vallverdú M, Soria JM, Caminal P and Perera A. MISS: a non-linear methodology based on mutual information for genetic association studies in both population and sib-pairs analysis. *Bioinformatics*, 2010, 26: 1811-1818.
- Bullinger L, Döhner K, Bair E, Fröhling S, Schlenk RF, Tibshirani R, Döhner H and Pollack JR. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *N Engl J Med.*, 2004, 350(16):1605-1616.
- Burgarella S, Cattaneo D, Pinciroli F and Masseroli M. MicroGen: a MIAME compliant web system for microarray experiment information and workflow management. *BMC Bioinformatics* 2005, 6(Suppl 4):S6.
- C. Grommes, G.E. Landreth and M.T. Heneka. Antineoplastic effects of peroxisome proliferator-activated receptor gamma agonists. *Lancet Oncol.*, 2004, pp. 419–429.
- Cannistraci CV, Ravasi T, Montecchi FM, Ideker T and Alessio M. Nonlinear dimension reduction and clustering by Minimum Curvilinearity unfold neuropathic pain and tissue embryological classes. *Bioinformatics*, 2010, 26: i531-539.
- Cao W, Yacoub S, Shiverick KT, Namiki K, Sakai Y, Porvasnik S, Urbanek C and Rosser CJ. Dichloroacetate (DCA) sensitizes both wild-type and over expressing Bcl-2 prostate cancer cells *in vitro* to radiation. *Prostate*. 2008, 68:1223–1231.
- Casey TM, Crodian J, Erickson E, Kuropatwinski KK, Gleiberman AS and Antoch MP. Tissue-specific changes in molecular clocks during the transition from pregnancy to lactation in mice. *Biol Reprod.*, 2014, 90(6):127.
- Cedano J, Huerta M and Querol E. NCR-PCOPGene: An Exploratory Tool for Analysis of Sample-Classes Effect on Gene-Expression Relationships. *Adv Bioinformatics*, 2008, 789026.
- Cedano J, Huerta M, Estrada I, Ballosera F, Conchillo O, Delicado P and Querol E. A web server for automatic analysis and extraction of relevant biological knowledge. *Comput Biol Med.*, 2007, 37:1672-1675.
- Chang K, and Ghosh J. A unified model for probabilistic principal surfaces. *Transactions on Pattern Analysis and Machine Intelligence*, 2001, 23:22-41.
- Chao S and Lihui C. Feature Dimension Reduction For Microarray Data Analysis Using Locally Linear Embedding. *Proceedings Of The 3rd Asia-Pacific Bioinformatics Conference*, 2005, 1:211-218.
- Chen J, Huang X, Halicka D, Brodsky S, Avram A, Eskander J, Bloomgarden N A, Darzynkiewicz Z, and Goligorsky M S. Contribution of p16INK4a and p21CIP1 pathways to induction of premature senescence of human endothelial cells: permissive role of p53. *Am J Physiol Heart Circ Physiol*, 2006, 290:H1575-86.
- Chen L and Yang G. PPARs Integrate the Mammalian Clock and Energy Metabolism. *PPAR Res.*, 2014, 2014:653017.
- Chen R, Mallewar R and Thosar A, Venkatasubrahmanyam S and Butte AJ. GeneChaser: identifying all biological and clinical conditions in which genes of interest are differentially expressed. *BMC Bioinformatics*, 2008, 9:548.
- Chen YX, Li ZB, Diao F, Cao DM, Fu CC and Lu J. Up-regulation of RhoB by glucocorticoids and its effects on the cell proliferation and NF- $\kappa$ B transcriptional activity. *J Steroid Biochem Mol Biol.*, 2006, 101(4-5):179-187.

Clarke RM. The time-course of changes in mucosal architecture and epithelial cell production and cell shedding in the small intestine of the rat fed after fasting. *J Anat.*, 1975, 120(Pt 2):321-327.

Colwell CS. Linking neural activity and molecular oscillations in the SCN. *Nat Rev Neurosci.*, 2011, 12(10):553-69.

Costa V, Aprile M, Esposito R and Ciccodicola A. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. *Eur J Hum Genet.*, 2013, 21(2):134-42.

Crusio WE. My mouse has no phenotype. *Genes Brain Behav.*, 2002, 1(2):71.

Csardi G, Kutalik Z and Bergmann S. Modular analysis of gene expression data with R. *Bioinformatics*, 2010, 26: 1376-1377.

Cuezva JM, Ortega AD, Willers I, Sánchez-Cenizo L, Aldea M and Sánchez-Aragó M. The tumor suppressor function of mitochondria: translation into the clinics. *Biochim Biophys Acta*. 2009 Dec;1792(12):1145-58.

Danza K, De Summa S, Pilato B, Carella M, Palumbo O, Popescu O, Paradiso A, Pinto R and Tommasi S. Combined microRNA and ER expression: a new classifier for familial and sporadic breast cancer patients. *Journal of Translational Medicine*, 2014, 12:319.

Davidson E H, Rast J P, Oliveri P, Ransick A, Calestani C, Yuh C H, Minokawa T, Amore G, Hinman V, Arenas-Mena C, Otim O, Brown C T, Livi C B, Lee P Y, Revilla R, Rust A G, Pan Z, Schilstra M J, Clarke P J, Arnone M I, Rowen L, Cameron R A, McClay D R, Hood L and Bolouri H. A genomic regulatory network for development. *Science*, 2002, 295:1669-78.

Delicado P and Huerta M. Principal curves of oriented points: Theoretical and computational improvements. *Computational Statistics*, 2003, 18:293-315.

Delicado P. Another look at principal curves and surfaces. *Journal of Multivariate Analysis*, 2001, 77:84-116.

D'Haeseleer P, Liang S, and Somogyi R. Genetic network inference: from coexpression clustering to reverse engineering. *Bioinformatics*, 2000, 16:707-26.

Dong D and McAvoy T J. Nonlinear principal component analysis - Based on principal curves and neural networks. *Computers and Chemical Engineering*, 1996, 20:65-78.

Eisen MB, Spellman PT, Brown PO and Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 1998, 95(25):14863-8.

Fang M, Guo WR1, Park Y6, Kang HG and Zarbl H. Enhancement of NAD<sup>+</sup>-dependent SIRT1 deacetylase activity by methylselenocysteine resets the circadian clock in carcinogen-treated mammary epithelial cells. *Oncotarget*, 2015 Oct 26.

Fantin VR, St-Pierre J and Leder P. Attenuation of LDH-A expression uncovers a link between glycolysis, mitochondrial physiology, and tumor maintenance. *Cancer Cell*, 2006, 9(6):425-34.

Feng C, Araki M, Kunimoto R, Tamon A, Makiguchi H, Niiijima S, Tsujimoto G and Okuno Y. GEM-TREND: a web tool for gene expression data mining toward relevant network discovery. *BMC Genomics*, 2009, 10:411.

Fogle KJ, Parson KG, Dahm NA and Holmes TC. CRYPTOCHROME is a blue-light sensor that regulates neuronal firing rate. *Science*, 2011, 331(6023):1409-13.

Foppoli C, De Marco F, Cini C and Perluigi M. Redox control of viral carcinogenesis: The human papillomavirus paradigm. *Biochim Biophys Acta*. 2015 Aug;1850(8):1622-32.

- Frenkel B, White W and Tuckermann J. Glucocorticoid-Induced Osteoporosis. *Adv Exp Med Biol.* 2015;872:179-215.
- Frickey T and Lupas A. CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, 2004, 20:3702-3704.
- Frickey T and Weiller G. Analyzing microarray data using CLANS. *Bioinformatics*, 2007, 23:1170-1171.
- Fu L, Lee CC. The circadian clock: pacemaker and tumour suppressor. *Nat Rev Cancer*, 2003, 3(5):350-61.
- Fujihara Y, Kondo H, Noguchi T and Togari A. Glucocorticoids mediate circadian timing in peripheral osteoclasts resulting in the circadian expression rhythm of osteoclast-related genes. *Bone*. 2014 Apr;61:1-9.
- Furukawa T, Sunamura M, Motoi F, Matsuno S and Horii A. Potential tumor suppressive pathway involving DUSP6/MKP-3 in pancreatic cancer. *Am J Pathol.*, 2003, 162(6):1807-1815.
- Garcia O, Saveanu C, Cline M, Fromont-Racine M, Jacquier A, Schwikowski B and Aittokallio T. Golorize: a Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring. *Bioinformatics*, 2007, 23, 394-396.
- Gargett CE, Chan RW and Schwab KE. Hormone and growth factor signaling in endometrial renewal: role of stem/progenitor cells. *Mol Cell Endocrinol*, 2008, 288(1-2):22-29.
- Getz, G. and Domany, E. Coupled two-way clustering server. *Bioinformatics*, 2003, 19, 1153-1154.
- Giancarlo R and Utro F. Speeding up the Consensus Clustering methodology for microarray data analysis. *Algorithms Mol Biol.*, 2011, 6(1):1.
- Gimble JM, Pighetti GM, Lerner MR, Wu X, Lightfoot SA, Brackett DJ, Darcy K and Hollingsworth AB. Expression of peroxisome proliferator activated receptor mRNA in normal and tumorigenic rodent mammary glands. *Biochem Biophys Res Commun.*, 1998, 253(3):813-7.
- Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD and Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 1999, 286(5439):531-7.
- Goryanin I, Hodgman T C, and Selkov E. Mathematical simulation and analysis of cellular metabolism and regulation. *Bioinformatics*, 1999, 15:749-58.
- Gower JC and Ross GJS. Minimum Spanning Trees and Single Linkage Cluster Analysis. *The Royal Statistical Society Series C-Applied Statistics*, 1969, 18:54.
- Gozgit JM, Pentecost BT, Marconi SA, Otis CN, Wu C and Arcaro KF. Use of an aggressive MCF-7 cell line variant, TMX2-28, to study cell invasion in breast cancer. *Mol Cancer Res.*, 2006, 4(12):905-913.
- Guet C C, Elowitz M B, Hsing W, and Leibler S. Combinatorial synthesis of genetic networks. *Science*, 2002, 296:1466-70.
- Guo X, Zhang Y, Hu W, Tan H and Wang X. Inferring nonlinear gene regulatory networks from gene expression data based on distance correlation. *PLoS One*, 2014, 9(2):e87446.
- Hall PA, Todd CB, Hyland PL, McDade SS, Grabsch H, Dattani M, Hillan KJ and Russell SE. The septin-binding protein anillin is overexpressed in diverse human tumors. *Clinical cancer research*, 2005, 11:6780-6786.



- Hamosh A, Scott AF, Amberger JS, Bocchini CA and McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, 2005, 33:D514-517.
- Hartigan J A. Consistency of Single Linkage for High-Density Clusters. *Journal of the American Statistical Association*, 1981, 76:388-394.
- Hastie T and Stuetzle W. Principal Curves. *Journal of the American Statistical Association*, 1989, 84:502-516.
- Heidtke K R and Schulze-Kremer S. Design and implementation of a qualitative simulation model of lambda phage infection. *Bioinformatics*, 1998, 14:81-91.
- Heine VM, Priller M, Ling J, Rowitch DH and Schuller U. Dexamethasone destabilizes Nmyc to inhibit the growth of hedgehog-associated medulloblastoma. *Cancer Res.*, 2010, 70(13):5220-5225.
- Herr I and Pfitzenmaier J. Glucocorticoid use in prostate cancer and other solid tumours: implications for effectiveness of cytotoxic treatment and metastases. *Lancet Oncol.*, 2006, 7(5):425-430.
- Herr I, Ucur E, Herzer K, Okouoyo S, Ridder R, Krammer PH, von Knebel Doeberitz M and Debatin KM. Glucocorticoid cotreatment induces apoptosis resistance toward cancer therapy in carcinomas. *Cancer Res.*, 2003, 63(12):3112-3120.
- Hill SM, Belancio VP, Dauchy RT, Xiang S, Brimer S, Mao L, Hauch A, Lundberg PW, Summers W, Yuan L, Frasch T and Blask DE. Melatonin: an inhibitor of breast cancer. *Endocr Relat Cancer*, 2015, 22(3):R183-204.
- Hinchcliffe E H, Li C, Thompson E A, Maller J L and Sluder G. Requirement of Cdk2-cyclin E activity for repeated centrosome reproduction in *Xenopus* egg extracts. *Science*, 1999, 283:851-4.
- Huerta M, Casanova O, Barchino R, Flores J, Querol E and Cedano J. Studying the complex expression dependences between sets of coexpressed genes. *Biomed Res Int.*, 2014, 940821.
- Huerta M, Cedano J and Querol E. Analysis of non-linear relation between expression profiles by the Principal Curves of Oriented-Points approach. *Journal of Bioinformatics and Computational Biology*, 2008, 6(2):367-86.
- Huerta M, Cedano J, Pe-a D, Rodriguez A and Querol E. PCOPGene-Net: holistic characterisation of cellular states from microarray data based on continuous and non-continuous analysis of gene-expression relationships. *BMC Bioinformatics*, 2009, 10:138.
- Huerta M, Fernández-Márquez J, Cabello JL, Medrano A, Querol E and Cedano J. Analysis of gene expression for studying tumor progression: the case of glucocorticoid administration. *Gene*, 2014, 549(1):33-40.
- Huerta M, Gispert B, Querol E and Cedano J. Deconstructing sample clusters in multiple concurrent phenotypic changes. *Plos One* [submitted]
- Huerta M, Hernandez C, Yesid J, Querol E and Cedano J. Crossing Clusters: Studying the relationships among the sample clusters obtained by clustering methods from expression data. *Bioinformatics* [submitted]
- Huerta M, Munyi M, Expósito D, Querol E and Cedano J. MGDB: crossing the marker genes of a user microarray with a database of public-microarrays marker genes. *Bioinformatics*, 2014, 30(12):1780-1.
- Hwang T, Atluri G, Xie M, Dey S, Hong C, Kumar V and Kuang R. Co-clustering phenome-genome for phenotype classification and disease gene discovery. *Nucleic Acids Res.*, 2012, 40(19):e146.

Jasonni VM, La Marca A and Santini D. Progestin effects on epidermal growth factor receptor (EGFR) endometrial expression in normal and hyperplastic endometrium. *Int J Gynaecol Obstet.*, 2005, 89(3):297-298.

Jung-Hynes B, Huang W, Reiter RJ and Ahmad N. Melatonin resynchronizes dysregulated circadian rhythm circuitry in human prostate cancer cells. *J Pineal Res.*, 2010, 49: 60-8.

Karantanos T, Theodoropoulos G, Pektasides D and Gazouli M. Clock genes: their role in colorectal cancer. *World J Gastroenterol.*, 2014, 20(8):1986-92.

Kastner J, Solomon J and Fraser S. Modelling a hox gene network in silico using a stochastic simulation algorithm. *Development Biology*, 2002, 246:122-31.

Kaup B, Schindler I, Knupfer H, Schlenzka A, Preiss R and Knupfer MM. Time-dependent inhibition of glioblastoma cell proliferation by dexamethasone. *J Neurooncol.*, 2001, 51(2):105-110.

Kegl B, Krzyzak A, Linder T and Zeger K. Learning and design of principal curves. *Transactions on Pattern Analysis and Machine Intelligence*, 2000, 22:281-297.

Keller ET and Brown J. Prostate cancer bone metastases promote both osteolytic and osteoblastic activity. *J Cell Biochem*, 2004, 91 718-29.

Kim J, Sato M, Choi JW, Kim HW, Yeh BI, Larsen JE, Minna JD, Cha JH, Jeong Y. Nuclear Receptor Expression and Function in Human Lung Cancer Pathogenesis. *PLoS One*, 2015, 10(8):e0134842.

Knoblach B, Keller B O, Groenendyk J, Aldred S, Zheng J, Lemire B D, Li L and Michalak M. ERp19 and ERp46, New Members of the Thioredoxin Family of Endoplasmic Reticulum Proteins. *Molecular and Cellular Proteomics*, 2003, 2:1104-19.

Kochan DZ and Kovalchuk O. Circadian disruption and breast cancer: An epigenetic link? *Oncotarget*, 2015, 6(19):16866-82.

Koestler DC, Marsit CJ, Christensen BC, Karagas MR, Bueno R, Sugarbaker DJ, Kelsey KT and Houseman EA. Semi-supervised recursively partitioned mixture models for identifying cancer subtypes. *Bioinformatics*, 2010, 26(20):2578-2585.

Kohonen T and Somervuo P. How to make large self-organizing maps for nonvectorial data. *Neural Netw.*, 2002, 15(8-9):945-952.

Kohonen T. Comparison of SOM point densities based on different criteria. *Neural Comput.*, 1999, 11:2081-2095.

Kolbe D, DeLoia J, Porter-Gill P, Strange M, Petrykowska H, Guirguis A, Krivak T, Brody L and Elnitski L. Differential Analysis of Ovarian and Endometrial Cancers Identifies a Methylator Phenotype. *PLoS One*, 2012, 7(3): e32941.

Kucharski R, Maleszka J, Foret S and Maleszka R. Nutritional control of reproductive status in honeybees via DNA methylation. *Science*, 2008, 319: 1827-1830.

Kumar V, Abbas AK, Fausto N and Aster J. Robbins and Cotran Pathologic Basis of Disease, Professional Edition: Expert Consult. Saunders Elsevier, 2009, 9th(3):74-1009.

Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES and Golub TR. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 2006, 293:1929-35.

- Lamb J. The Connectivity Map: a new tool for biomedical research. *Nature Reviews Cancer*, 2007, 7: 54-60.
- Langmesser S and Albrecht U. Life time-circadian clocks, mitochondria and metabolism. *Chronobiol Int.*, 2006, 23(1-2):151-7.
- Larsson O and Sandberg R. Lack of correct data format and comparability limits future integrative microarray research. *Nature Biotechnology*, 2006, 24(11):1322D1323.
- Lawrence JW, Mason ST, Schomer K and Klein MB. Epidemiology and impact of scarring after burn injury: a systematic review of the literature. *J Burn Care Res.*, 2012, 33(1):136-146.
- Lawson EA, Ackerman KE, Estella NM, Guereca G, Pierce L, Sluss PM, Boussein ML, Klibanski A and Misra M. Nocturnal oxytocin secretion is lower in amenorrheic athletes than nonathletes and associated with bone microarchitecture and finite element analysis parameters. *Eur J Endocrinol.* 2013 Feb 20;168(3):457-64.
- Lecarpentier Y, Claes V, Duthoit G and Hébert JL. Circadian rhythms, Wnt/  $\beta$ -catenin pathway and PPAR alpha/gamma profiles in diseases with primary or secondary cardiac dysfunction. *Front Physiol.*, 2014, 5:429.
- Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y and Pietenpol JA. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest*, 2011.
- Liang Y and Kelemen A. Associating phenotypes with molecular events: recent statistical advances and challenges underpinning microarray experiments. *Funct Integr Genomics*, 2006, 6: 1-13.
- Liu S, Jiang L, Li H, Shi H, Luo H, Zhang Y, Yu C and Jin Y. Mesenchymal Stem Cells Prevent Hypertrophic Scar Formation via Inflammatory Regulation when Undergoing Apoptosis. *J Invest Dermatol.*, 2014 Oct;134(10):2648-57.
- Liu Y, Gu Q, Hou JP, Han J and Ma J. A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression. *BMC Bioinformatics*, 2014, 15:37.
- Lorrain S, Lin B, Auriac MC, Kroj T, Saindrenan P, Nicole M, Balague C and Roby D. Vascular associated death1, a novel GRAM domain-containing protein, is a regulator of cell death and defense responses in vascular tissues. *Plant Cell*, 2004, 16(8):2217-2232.
- LoVerme J, Russo R, La Rana G, Fu J, Farthing J, Mattace-Raso G, Meli R, Hohmann A, Calignano A and Piomelli D. Rapid broad-spectrum analgesia through activation of peroxisome proliferator-activated receptor-alpha. *J Pharmacol Exp Ther.*, 2006, 319(3):1051-61.
- Lu YS, Lien HC, Yeh PY, Kuo SH, Chang WC, Kuo ML and Cheng AL. Glucocorticoid receptor expression in advanced non-small cell lung cancer: clinicopathological correlation and in vitro effect of glucocorticoid on cell growth and chemosensitivity. *Lung Cancer*, 2006, 53(3):303-310.
- Maere S, Heymans K and Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 2005, 21, 3448D3449.
- Mahner M and Kary M. What exactly are genomes, genotypes and phenotypes? And what about phenomes? *J Theor Biol*, 1997, 186: 55-63.
- Mao L and Resat H. Probabilistic representation of gene regulatory networks. *Bioinformatics*, 2004, 20:2258-69.
- Maronde E, Schilling AF, Seitz S, Schinke T, Schmutz I, van der Horst G, Amling M and Albrecht U. The clock genes Period 2 and Cryptochrome 2 differentially balance bone formation. *PLoS One.* 2010 Jul 12;5(7):e11527.

- Maruyama T and Yoshimura Y. Molecular and cellular mechanisms for differentiation and regeneration of the uterine endometrium. *Endocr J.*, 2008, 55(5):795-810.
- Mattern J, Buchler MW and Herr I. Cell cycle arrest by glucocorticoids may protect normal tissue and solid tumors from cancer therapy. *Cancer Biol Ther.*, 2007, 6(9):1345-1354.
- Mazzitelli L, Hancock RD, Haupt S, Walker PG, Pont SD, McNicol J, Cardle L, Morris J, Viola R, Brennan R, Hedley PE and Taylor MA. Co-ordinated gene expression during phases of dormancy release in raspberry (*Rubus idaeus* L.) buds. *J Exp Bot.*, 2007, 58: 1035-1045.
- McAdams H, and Arkin A. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 1997, 94:814-9.
- McCarty SM, Syed F and Bayat A. Influence of the human leukocyte antigen complex on the development of cutaneous fibrosis: an immunogenetic perspective. *Acta Derm Venereol*, 2010, 90(6):563-574.
- Meijer E and Sonneveld P. Hematology: Lenalidomide plus dexamethasone is effective in multiple myeloma. *Nat Rev Clin Oncol.*, 2009, 6(5):247-248.
- Meredith AL, Wiler SW, Miller BH, Takahashi JS, Fodor AA, Ruby NF and Aldrich RW. BK calcium-activated potassium channels regulate circadian behavioral rhythms and pacemaker output. *Nat Neurosci.*, 2006, 9(8):1041-9.
- Mestl T, Plahte E and Omholt S W. A mathematical framework for describing and analysing gene regulatory networks. *Journal of Theoretical Biology*, 1995, 176:291-300.
- Michaud J, Scott HS and Escher R. AML1 interconnected pathways of leukemogenesis. *Cancer Invest.*, 2003, 21, 105-136.
- Michelakis ED, Webster L and Mackey JR. Dichloroacetate (DCA) as a potential metabolic-targeting therapy for cancer. *Br J Cancer*, 2008, 99(7):989-94.
- Miglio G, Rosa AC, Rattazzi L, Collino M, Lombardi G and Fantozzi R. PPAR $\gamma$  stimulation promotes mitochondrial biogenesis and prevents glucose deprivation-induced neuronal cell loss. *Neurochem Int.*, 2009, 55(7):496-504.
- Miller DJ, Wang Y and Kesidis G. Emergent unsupervised clustering paradigms with potential application to bioinformatics. *Front Biosci*, 2008, 13:677-90.
- Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, Menon S, Hanumanthu G, Gupta M, Upendran S, Gupta S, Mahesh M, Jacob B, Mathew P, Chatterjee P, Arun KS, Sharma S, Chandrika KN, Deshpande N, Palvankar K, Raghavnath R, Krishnakanth R, Karathia H, Rekha B, Nayak R, Vishnupriya G, Kumar HG, Nagini M, Kumar GS, Jose R, Deepthi P, Mohan SS, Gandhi TK, Harsha HC, Deshpande KS, Sarker M, Prasad TS and Pandey A. Human protein reference database. *Nucleic Acids Res.*, 2006, 34:D411-414.
- Mohammadi A, Saraee MH and Salehi M. Identification of disease-causing genes using microarray data mining and Gene Ontology. *BMC Med Genomics*, 2011, 4:12.
- Morgan G, Ward R and Barton M. The Contribution of Cytotoxic Chemotherapy to 5-year Survival in Adult Malignancies. *Clinical Oncology*, 2004, 16: 549-560.
- Motschall E and Falck-Ytter Y. Searching the MEDLINE literature database through PubMed: a short guide. *Onkologie*, 2005, 28:517-522.

Mukaide H, Adachi Y, Taketani S, Iwasaki M, Koike-Kiriyama N, Shigematsu A, Shi M, Yanai S, Yoshioka K, Kamiyama Y and Ikehara S. FKBP51 expressed by both normal epithelial cells and adenocarcinoma of colon suppresses proliferation of colorectal adenocarcinoma. *Cancer Invest.*, 2008, 26(4):385-390.

Mulier F and Cherkassky V. Self-Organization as an Iterative Kernel Smoothing Process. *Neural Computation*, 1995, 7:1165-1177.

Müller FJ, Laurent LC, Kostka D, Ulitsky I, Williams R, Lu C, Park I, Rao MS, Shamir R, Schwartz PH, Schmidt NO and Loring JF. Regulatory networks define phenotypic classes of human stem cell lines. *Nature*, 2008 September 18; 455(7211): 401D405.

Nam D, Kim SB, Kim SK, Yang S, Kim SY and Chu IS. ADGO: analysis of differentially expressed gene sets using composite GO annotation. *Bioinformatics*, 2006, 22:2249-2253.

Nature. Microarray standards at last. *Nature*, 2002, 419, 323.

Newmark HL, Lipkin M and Maheshwari N. Colonic hyperproliferation induced in rats and mice by nutritional-stress diets containing four components of a human Western-style diet (series 2). *Am J Clin Nutr.*, 1991, 54(1 Suppl):209S-214S.

Nickkho-Amiry M, McVey R and Holland C. Peroxisome proliferator-activated receptors modulate proliferation and angiogenesis in human endometrial carcinoma. *Mol Cancer Res.*, 2012, 10(3):441-53.

Nitabach MN, Blau J and Holmes TC. Electrical silencing of *Drosophila* pacemaker neurons stops the free-running circadian clock. *Cell.*, 2002, 109(4):485-95.

Norton SM, Huyn P, Hastings CA and Heller JC. Data mining of spectroscopic data for biomarker discovery. *Curr Opin Drug Discov Devel.*, 2001, 4(3):325-31.

Okudela K, Yazawa T, Woo T, Sakaeda M, Ishii J, Mitsui H, Shimoyamada H, Sato H, Tajiri M, Ogawa N, Masuda M, Takahashi T, Sugimura H and Kitamura H. Down-regulation of DUSP6 expression in lung cancer: its mechanism and potential role in carcinogenesis. *Am J Pathol.*, 2009, 175(2):867-881.

O'Madadhain J, Fisher D, Smyth P, White S and Boey YB. Analysis and visualization of network data using JUNG. *Journal of Statistical Software*, 2005, VV:1-35.

Owens TW, Rogers RL, Best SA, Ledger A, Mooney AM, Ferguson A, Shore P, Swarbrick A, Ormandy CJ, Simpson PT, Carroll JS, Visvader JE and Naylor MJ. Runx2 is a novel regulator of mammary epithelial cell fate in development and breast cancer. *Cancer Res.* 2014 Sep 15;74(18):5277-86.

Pan JG and Mak TW. Metabolic targeting as an anticancer strategy: dawn of a new era? *Sci STKE.* 2007, 2007(381):pe14.

Panigrahy D, Huang S, Kieran MW and Kaipainen A. PPAR $\gamma$  as a therapeutic target for tumor angiogenesis and metastasis. *Cancer Biol Ther.*, 2005, 4(7):687-93.

Parkes M, Cortes A, van Heel DA and Brown MA. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat Rev Genet.*, 2013, 14(9):661-673.

Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, Mani R, Rayner T, Sharma A, William E, Sarkans U and Brazma A. ArrayExpress--a public database of microarray experiments and gene expression profiles, *Nucleic Acids Res.*, 2007, 35, D747-750.

Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, Coulson R, Farne A, Lara GG, Holloway E, Kapushesky M, Lilja P, Mukherjee G, Oezcimen A, Rayner T, Rocca-Serra P, Sharma A, Sansone S and Brazma A. ArrayExpress: a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, 2005, 33:D553-555.

Patra KC and Hay N. The pentose phosphate pathway and cancer. *Trends Biochem Sci.*, 2014, 39(8):347-54.

Persaud A, Alberts P, Amsen EM, Xiong X, Wasmuth J, Saadon Z, Fladd C, Parkinson J and Rotin D. Comparison of substrate specificity of the ubiquitin ligases Nedd4 and Nedd4-2 using proteome arrays. *Mol Syst Biol.*, 2009, 5:333.

Pheesse TJ and Clarke AR. Normal stem cells in cancer prone epithelial tissues. *Br J Cancer*, 2009, 100(2):221-227.

Phillips RS, Gopaul S, Gibson F, Houghton E, Craig JV, Light K and Pizer B. Antiemetic medication for prevention and treatment of chemotherapy induced nausea and vomiting in childhood. *Cochrane Database Syst Rev.*, 2010, (9):CD007786.

Pokorný J, Pokorný J, Kobilková J, Jandová A, Vrba J and Vrba J. Targeting mitochondria for cancer treatment - two types of mitochondrial dysfunction. *Prague Med Rep.* 2014, 115(3-4):104-19.

Polvani S, Tarocchi M and Galli A. PPAR $\gamma$  and Oxidative Stress: Con( $\beta$ ) Catenating NRF2 and FOXO. *PPAR Res.*, 2012, 2012:641087.

Poole AJ, Li Y, Kim Y, Lin SC, Lee WH and Lee EY. Prevention of Brca1-mediated mammary tumorigenesis in mice by a progesterone antagonist. *Science*, 2006, 314(5804):1467-1470.

Powe D. Beta-blockers help reduce metastasis and improve survival in patients. 7th European Breast Cancer Conference, 2010, EBCC7.

Qiu Q, Lu P, Xiang Y, Shyr Y, Chen X, Lehmann BD, Viox DJ, George AL Jr and Yi Y. A data similarity-based strategy for meta-analysis of transcriptional profiles in cancer. *PLoS One*, 2013, 8(1):e54979.

Reiche EM, Morimoto HK and Nunes SM. Stress and depression-induced immune dysfunction: implications for the development and progression of cancer. *Int Rev Psychiatry*, 2005, 17(6):515-527.

Ripley B. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

Roelofs J and Van Haastert P.J. Characterization of two unusual guanylyl cyclases from dictyostelium. *J Biol Chem*, 2002, 277, 9167-9174.

Rosenfeld N, Elowitz M B and Alon U. Negative autoregulation speeds the response times of transcription networks. *Journal of Molecular Biology*, 2002, 323:785-793.

Roweis ST and Saul LK. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000, 290, 2323-2326.

Sandilya S and Kulkarni S R. Principal curves with bounded turn. *Transactions on Information Theory*, 2002, 48:2789-2793.

Sarraf P, Mueller E, Jones D, King FJ, DeAngelo DJ, Partridge JB, Holden SA, Chen LB, Singer S, Fletcher C and Spiegelman BM. Differentiation and reversal of malignant changes in colon cancer through PPAR $\gamma$ . *Nat Med.*, 1998, 4(9):1046-52.

Savvidis C and Koutsilieris M. Circadian rhythm disruption in cancer biology. *Mol Med.*, 2012, 18:1249-60.

- Scheiermann C, Kunisaki Y and Frenette PS. Circadian control of the immune system. *Nat Rev Immunol.*, 2013, 13(3):190-8.
- Scherf U, Ross DT, Waltham M, Smith LH, Lee J K, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO and Weinstein JN. A gene expression database for the molecular pharmacology of cancer. *Nature Genetics*, 2000, 24:236-44.
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin li , Schwikowski B and Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 2003, 13, 2498D2504.
- Sherlock P and Hartmann WH. Adrenal steroids and the pattern of metastases of breast cancer. *Jama*, 1962, 181:313-317.
- Simmons GE Jr, Pruitt WM, Pruitt K. Diverse roles of SIRT1 in cancer biology and lipid metabolism. *Int J Mol Sci.*, 2015, 16(1):950-65.
- Sims NA and Martin TJ. Coupling Signals between the Osteoclast and Osteoblast: How are Messages Transmitted between These Temporary Visitors to the Bone Surface?. *Front Endocrinol (Lausanne)*. 2015 Mar 24;6:41.
- Singh P, Velasco M, Given R, Wargovich M, Varro A and Wang TC: Mice overexpressing progastrin are predisposed for developing aberrant colonic crypt foci in response to AOM. *Am J Physiol Gastrointest Liver Physiol.*, 2000, 278(3):G390-399.
- Sloan EK, Priceman SJ, Cox BF, Yu S, Pimentel MA, Tangkanangnukul V, Arevalo JM, Morizono K, Karanikolas BD, Wu L, Sood AK and Cole SW. The sympathetic nervous system induces a metastatic switch in primary breast cancer. *Cancer Res.*, 2010, 70(18):7042-7052.
- Snyder PM, Olson DR, Kabra R, Zhou R and Steines JC. CAMP and serum and glucocorticoid-inducible kinase (SGK) regulate the epithelial Na(+) channel through convergent phosphorylation of Nedd4-2. *J Biol Chem.*, 2004, 279(44):45753-45758.
- Spruck CH, Won K A and Reed SI. Deregulated cyclin E induces chromosome instability. *Nature*, 1999, 401:297-300.
- Street NR, Jansson S, Hvidsten TR. A systems biology model of the regulatory network in *Populus* leaves reveals interacting regulators and conserved regulation. *BMC Plant Biol.*, 2011, 11:13.
- Stuart JM, Segal E, Koller D and Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 2003, 302(5643):249-255.
- Sui M, Chen F, Chen Z and Fan W. Glucocorticoids interfere with therapeutic efficacy of paclitaxel against human breast and ovarian xenograft tumors. *Int J Cancer*, 2006, 119(3):712-717.
- Sullivan DC, Huminiecki L, Moore J W, Boyle JJ, Poulosom R, Creamer D, Barker J and Bicknell R. EndoPDI, a novel protein-disulfide isomerase-like protein that is preferentially expressed in endothelial cells acts as a stress survival factor. *Journal of Biological Chemistry*, 2003, 278:47079-47088.
- Swain P S, Elowitz M B and Siggia E D. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 2002, 99:12795-12800.
- Swamidass SJ. Mining small-molecule screens to repurpose drugs. *Brief Bioinform*, 2011, 12(4):327-35.
- Taminau J, Lazar C, Meganck S and NowŽ A. Comparison of merging and meta-analysis as alternative approaches for integrative gene expression analysis. *ISRN Bioinform*, 2014, 2014:345106.

- Tan J, Peng X, Luo G, Ma B, Cao C, He W, Yuan S, Li S, Wilkins JA and Wu J: Investigating the role of P311 in the hypertrophic scar. *PLoS One*, 2010, 5(4):e9995.
- Tanay A, Sharan R and Shamir R. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 2002, 18(Suppl 1):S136-144.
- Tanic M, Andres E, Rodriguez-Pinilla SM, Marquez-Rodas I, Cebollero-Presmanes M, Fernandez V, Osorio A, Benitez J and Martinez-Delgado B. MicroRNA-based molecular classification of non-BRCA1/2 hereditary breast tumours. *Br J Cancer*, 2013, 109(10):2724-2734.
- Tárraga J, Medina I, Carbonell J, Huerta-Cepas J, Minguéz P, Alloza E, Al-Shahrour F, Vegas-Azcárate S, Goetz S, Escobar P, Garcia-Garcia F, Conesa A, Montaner D and Dopazo J. GEPAS, a web-based tool for microarray data analysis and interpretation. *Nucleic Acids Res.*, 2008, 36:W308-314.
- Thalamuthu A, Mukhopadhyay I, Zheng X and Tseng GC. Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, 2006, 22(19):2405-2412.
- Theocharis S, Kouraklis G, Margeli A, Agapitos E, Ninos S, Karatzas G and Koutselinis A. Glucocorticoid receptor (GR) immunohistochemical expression is correlated with cell cycle-related molecules in human colon cancer. *Dig Dis Sci.*, 2003, 48(9):1745-1750.
- Thieffry D and Thomas R. Qualitative analysis of gene networks. *Pacific Symposium on Biocomputing*, 1998, 77-88.
- Thimm O, Blasing O, Gibon Y, Nagel A, Meyer S, Kruger P, Selbig J, Muller LA, Rhee SY and Stitt M. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.*, 2004, 37:914-939.
- Thorpe C, Hooper, KL, Raje S, Glynn NM, Burnside J, Turi GK and Coppock DL. Sulfhydryl oxidases: emerging catalysts of protein disulfide bond formation in eukaryotes. *Arch Biochem Biophys*, 2002, 405, 1-12.
- Tien ES, Davis JW and Vanden Heuvel JP. Identification of the CREB-binding protein/p300-interacting protein CITED2 as a peroxisome proliferator-activated receptor alpha coregulator. *J Biol Chem.*, 2004, 279(23):24053-24063.
- Tissier F, Louvel A, Grabar S, Hagnere AM, Bertherat J, Vacher-Lavenu MC, Dousset B, Chapuis Y, Bertagna X and Gicquel C. Cyclin E correlates with malignancy and adverse prognosis in adrenocortical tumors. *European Journal of Endocrinology*, 2004, 150:809-17.
- Tu Y, Stolovitzky G and Klein U. Quantitative noise analysis for gene expression microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 2002, 99:14031-6.
- Tutton PJ and Barkla DH. Effects of glucocorticoid hormones on cell proliferation in dimethylhydrazine-induced tumours in rat colon. *Virchows Arch B Cell Pathol Incl Mol Pathol.*, 1981, 38(2):247-251.
- Usary J, Llaca V, Karaca G, Presswala S, Karaca M, He X, Langerød A, Kåresen R, Oh DS, Dressler LG, Lønning PE, Strausberg RL, Chanock S, Børresen-Dale AL and Perou CM. Mutation of GATA3 in human breast tumors. *Oncogene*, 2004, 23, 7669-7678.
- Valent P, Cerny-Reiterer S, Herrmann H, Mirkina I, George TI, Sotlar K, Sperr WR and Horny HP. Phenotypic heterogeneity, novel diagnostic markers, and target expression profiles in normal and neoplastic human mast cells. *Best Pract Res Clin Haematol.*, 2010, 23(3):369-78.
- Valentini G. Clusterv: A tool for assessing the reliability of clusters discovered in DNA microarray data. *Bioinformatics*, 2006, 22:369-70.



- Van Dam S, Cordeiro R, Craig T, van Dam J, Wood SH and de Magalhaes JP. GeneFriends: an online coexpression analysis tool to identify novel gene targets for aging and complex diseases. *BMC Genomics*, 2012, 13:535.
- Van Laere S, Van der Auwera I, Van den Eynden G, Van Hummelen P, Van Dam P, Van Marck E, Vermeulen PB and Dirix L. Distinct molecular phenotype of inflammatory breast cancer compared to non-inflammatory breast cancer using Affymetrix-based genome-wide gene-expression analysis. *Br J Cancer*, 2007, 97(8): 1165D1174.
- Van Loey NE and Van Son MJ. Psychopathology and psychological problems in patients with burn scars: epidemiology and management. *Am J Clin Dermatol.*, 2003, 4(4):245-272.
- Vazquez M, Nogales-Cadenas R, Arroyo J, Botías P, García R, Carazo JM, Tirado F, Pascual-Montano A and Carmona-Saez P. MARQ: an online tool to mine GEO for experiments with similar or opposite gene expression signatures. *Nucleic Acids Res.*, 2010, 38(Web Server issue):W228-32.
- Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, Alexe G, Lawrence M, O'Kelly M, Tamayo P, Weir BA, Gabriel S, Winckler W, Gupta S, Jakkula L, Feiler HS, Hodgson JG, James CD, Sarkaria JN, Brennan C, Kahn A, Spellman PT, Wilson RK, Speed TP, Gray JW, Meyerson M, Getz G, Perou CM and Hayes DN. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell.*, 2010, 17(1):98D110.
- Viero C, Shibuya I, Kitamura N, Verkhatsky A, Fujihara H, Katoh A, Ueta Y, Zingg HH, Chvatal A, Sykova E and Dayanithi G. REVIEW: Oxytocin: Crossing the bridge between basic science and pharmacotherapy. *CNS Neurosci Ther.* 2010 Oct;16(5):e138-56.
- Viswanathan AN and Schernhammer ES. Circulating melatonin and the risk of breast and endometrial cancer in women. *Cancer Lett.*, 2009, 281(1):1-7.
- Vohradsky J. Neural network model of gene expression. *Faseb Journal*, 2001, 15:846-54.
- Vu T and Vohradsky J. Genexp--a genetic network simulation environment. *Bioinformatics*, 2002, 18:1400-1.
- Wagatsuma A and Sakuma K.. Mitochondria as a potential regulator of myogenesis. *ScientificWorldJournal*. 2013;2013:593267.
- Warnat P, Eils R and Brors B. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, 2005, 6:265.
- Watanabe K, Tachibana O, Sata K, Yonekawa Y, Kleihues P and Ohgaki H. Overexpression of the EGF receptor and p53 mutations are mutually exclusive in the evolution of primary and secondary glioblastomas. *Brain Pathol.*, 1996, 6(3):217-223; discussion 223-214.
- Weaver DC, Workman CT and Stormo GD. Modeling regulatory networks with weight matrices. *Pacific Symposium on Biocomputing*, 1999, 112-23.
- Wennbo H and Törnell J. The role of prolactin and growth hormone in breast cancer. *Oncogene.*, 2000, 19(8):1072-6.
- Wessels LF, van Someren EP and Reinders MJ. A comparison of genetic network models. *Pacific Symposium on Biocomputing*, 2001, 508-19.
- Wheeler D. Using GenBank. *Methods Mol Biol.*, 2007, 406:23-60.

- Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmsberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Ostell J, Pruitt KD, Schuler GD, Shumway M, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L and Yaschenko E. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 2008, 36:D13-21.
- Whittington K, Assinder S, Gould M and Nicholson H. Oxytocin, oxytocin-associated neurophysin and the oxytocin receptor in the human prostate. *Cell Tissue Res.* 2004 Nov;318(2):375-82.
- Wixon J and Kell D. The Kyoto encyclopedia of genes and genomes- KEGG, 2000, 17:48-55.
- Wong JY, Huggins GS, Debidda M, Munshi NC and De Vivo I. Dichloroacetate induces apoptosis in endometrial cancer cells. *Gynecol Oncol.*, 2008, 109:394-402.
- Wong P, Gladney S and Keasling JD. Mathematical model of the lac operon: inducer exclusion, catabolite repression, and diauxic growth on glucose and lactose. *Biotechnology Progress*, 1997, 13:132-43.
- Wood P.A, Yang X and Hrushesky W.J.M. Clock genes and cancer. *Integrative Cancer Therapies* Vol. 8(4), 2009.
- Wu B. Cancer outlier differential gene expression detection. *Biostatistics*, 2007, 8: 566-575.
- Wu CJ and Kasif S. GEMS: a web server for biclustering analysis of expression data. *Nucleic Acids Res.*, 2005, 33:W596-599.
- Yamamoto T, Nishiguchi M, Inoue N, Goto HG, Kudawara I, Ueda T, Yoshikawa H, Tanigaki Y and Nishizawa Y. Inhibition of murine osteosarcoma cell proliferation by glucocorticoid. *Anticancer Res.*, 2002, 22(6C):4151-4156.
- Yeh CM, Shay J, Zeng TC, Chou JL, Huang TH, Lai HC and Chan MW. Epigenetic silencing of ARNTL, a circadian gene and potential tumor suppressor in ovarian cancer. *Int J Oncol.*, 2014, 45(5):2101-7.
- Yin L, Huang CH and Ni J. Clustering of gene expression data: performance and similarity analysis. *BMC Bioinformatics*, 2006, 7 Suppl 4: S19.
- Yu Y, Tu K, Zheng S, Li Y, Ding G, Ping J, Hao P and Li Y. GEOGLE: context mining tool for the correlation between gene expression and the phenotypic distinction. *BMC Bioinformatics*, 2009, 10:264.
- Zhang C, Kolb A, Büchler P, Cato AC, Mattern J, Rittgen W, Edler L, Debatin KM, Büchler MW, Friess H and Herr I. Corticosteroid co-treatment induces resistance to chemotherapy in surgical resections, xenografts and established cell lines of pancreatic cancer. *BMC Cancer*, 2006, 6:61.
- Zhao H, Ma Z, Tibshirani R, Higgins J, Ljungberg B and Brooks J. Alteration of Gene Expression Signatures of Cortical Differentiation and Wound Response in Lethal Clear Cell Renal Cell Carcinomas. *PLoS One*, 2009, 4(6): e6039.





