Università degli Studi di Ferrara

DOTTORATO DI RICERCA IN
FARMACOLOGIA ED ONCOLOGIA MOLECOLARE

CICLO
XXI

COORDINATORE Prof. Pier Andrea Borea

Involvement of genes and non-coding RNAs in cancer: profiling using microarrays

Settore Scientifico Disciplinare BIO/14

Dottorando

Dott. ssa Rossi Simona

____

(firma)

Tutore

Prof. Volinia Stefano

____

(firma)

Anni 2006/2008

To **Stefano** *and* **George***, for their support, ideas and endless enthusiasm!*

Table of Contents

ABSTRACT

MicroRNAs (miRNAs) are small noncoding RNAs (ncRNAs, RNAs that do not code for proteins) that regulate the expression of target genes. MiRNAs can act as tumor suppressor genes or oncogenes in human cancers. Moreover, a large fraction of genomic ultraconserved regions (UCRs) encode a particular set of ncRNAs whose expression is altered in human cancers. Bioinformatics studies are emerging as important tools to identify associations between miRNAs/ncRNAs and CAGRs (Cancer Associated Genomic Regions). ncRNA profiling, the use of highly parallel devices like microarrays for expression, public resources like mapping, expression, functional databases, and prediction algorithms have allowed the identification of specific signatures associated with diagnosis, prognosis and response to treatment of human tumors.

SUMMARY

The massive production of biological data by means of highly parallel devices like microarrays for gene expression has paved the way to new possible approaches in molecular genetics. Among them the possibility of inferring biological answers by querying large sets of expression data. Based on this principle was implemented TOM, a web-based resource for the efficient extraction of candidate genes for hereditary diseases. The algorithm uses the information stored in public resources, including mapping, expression and functional databases to select one or more candidate genes. This approach allows the geneticist to bypass the costly and time consuming tracing of genetic markers through entire families and might improve the chance of identifying disease genes, particularly for rare diseases. Results were obtained on known benchmark and on hereditary predisposition to familial thyroid cancer.

After TOM project, the efforts were focused on miRNAs study. MiRNAs are small, ncRNAs that can contribute to cancer development and progression by acting as oncogenes or tumor suppressor genes. MiRNAs and UCRs are frequently located at fragile sites and genomic regions affected in various cancers, named cancer associated genomic regions (CAGRs). Bionformatics studies are important tools to identify associations and correlations between miRNAs/ncRNAs and CAGRs. An algorithm was implemented to calculate statistically significant Spearman correlations between UCRs and miRNAs and it supported biological experiments proving that certain UCRs whose expression may be regulated by miRNAs abnormally expressed in human chronic lymphocytic leukaemia (CLL).

Moreover, miRNAs microarray profiling was performed to study and support the hypothesis of a possible miRNA hypermethylaton profile characteristic of human metastasis. A pharmacological and genomic approach was used to reveal aberrant epigenetic silencing program by treating lymph node metastatic cancer cells with a DNA demethylating agent followed by hybridization to an expression microarray. Among the miRNAs that were reactivated upon drug treatment, miR-148a, miR-34b/c, and miR-9 were found to undergo specific hypermethylation –associated silencing in cancer cells compared with normal tissues. The findings indicate that DNA methylation-associated silencing of tumor suppressor miRNAs contributes to the development of human cancer metastasis. The miRNAs microarray profiling contributed to support those results.

LIST OF PUBLICATIONS

This thesis is based on the following original articles and reviews plus a regional project presented at BITS 2007.

I. GebbaLab project, 2006-2008, www.gebbalab.it

II. **Rossi S***, Masotti D, Nardini C, Bonora E, Romeo G, Macii E, Benini L, Volinia S. (2006) *TOM: a web-based integrated approach for identification of candidate disease genes.* Nucleic Acids Res.:W285-92.

III. Masotti D*, Nardini C, **Rossi S**, Bonora E, Romeo G, Volinia S, Benini L. (2007) *TOM: enhancement and extension of a tool suite for in silico approaches to multigenic hereditary disorders.* Bioinformatics. 24(3):428-9.

IV. Gamberoni G, Lamma E, Lodo G, Marchesini J, Mascellani N, **Rossi S**, Storari S, Tagliavini L, Volinia S*. (2007) *Fun&Co: identification of key functional differences in transcriptomes.* Bioinformatics. 23(20):2725-32.

V. Mascellani N*, Liu X, **Rossi S**, Marchesini J, Valentini D, Arcelli D, Taccioli C, Helmer Citterich M, Liu CG, Evangelisti R, Russo G, Santos JM, Croce CM, Volinia S*. (2007) *Compatible solutes from hyperthermophiles improve the quality of DNA microarrays.* BMC Biotechnol. 23;7:82.

VI. Calin GA*, Liu CG, Ferracin M, Hyslop T, Spizzo R, Sevignani C, Fabbri M, Cimmino A, Lee EJ, Wojcik SE, Shimizu M, Tili E, **Rossi S**, Taccioli C, Pichiorri F, Liu X, Zupo S, Herlea V, Gramantieri L, Lanza G, Alder H, Rassenti L, Volinia S, Schmittgen TD, Kipps TJ, Negrini M, Croce CM. (2007) *Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas.* Cancer Cell. 12(3):215-29.

VII. **Rossi S,** Macellani N, Marchesini J, Gamberoni G, Tagliavini L, Taccioli C, Volinia S*. (2007) *I microRNA: un nuovo attore sul palcoscenico cellulare*. Accademia Delle Scienze. Universita' di Ferrara.

VIII. **Rossi S**, Sevignani C, Nnadi SC, Siracusa LD, Calin GA*. (2008) *Cancer-associated genomic regions (CAGRs) and noncoding RNAs: bioinformatics and therapeutic implications.* Mamm Genome. 19(7-8):526-40. Review

IX. Lujambio A, Calin GA, Villanueva A, Ropero S, Sánchez-Céspedes M, Blanco D, Montuenga LM, **Rossi S**, Nicoloso MS, Faller WJ, Gallagher WM, Eccles SA, Croce CM, Esteller M*. (2008) *A microRNA DNA methylation signature for human cancer metastasis.* Proc Natl Acad Sci U S A. 105(36):13556-61.

X. **Rossi S,** Calin GA, Amoroso A, Mascellani N, Volinia S*.*OMiR: identification of associations between OMIM diseases and microRNAs***.** Submitted

XI. **Rossi S**, Kopetz S, Davuluri R, Hamilton S., Calin GA*. *MicroRNAs and ultraconserved genes: from the scientist bench to the bedside of colorectal cancer patients*. Review. Submitted

*Corresponding author

INTRODUCTION

Ceaseless advances in biotechnology, along with the growing experience cumulated by researchers in recent years, has allowed a continuous and faster blooming of the number of genomes being sequenced and, most importantly, annotated. The consequent necessities of storing, retrieving, sharing and, in particular, understanding this vast amount of data led to the creation of genome databases, an open source of genetic information for scientists worldwide. Whereas the genomic era opened the doors to the very existence of such large and comprehensive (omic) data repositories, the strongest urgency of the post-genomic era is now to interrelate various sources of biomedical information. Several parallel efforts are currently underway to achieve a better understanding of the human genome. These actions are turned to the extraction of high-throughput information from global approaches such as the International HapMap Project (The International HapMap Consortium, 2005) for identification of single nucleotide polymorphisms or the prosecution of the ENCODE [Encyclopedia Of DNA Elements (ENCODE Consortium, 2004)] for the identification of all functional elements in the genome sequence. Therefore the integration of various existing and upcoming efforts is going to be a key element for the full comprehension of the cellular machinery. We focused on two of the many fields that will strongly benefit from such an integration: the study of **hereditary diseases** and the study of **cancer**. Often more than one gene is involved in life threatening misfunctioning of cellular functions. To characterize such diseases the identification of all the responsible genes is eventually a crucial requirement. This process usually involves costly, time consuming and difficult tracing of large family lineages to follow the line of transmission of genes and thus to define the linkage areas where genes responsible for the disease could be located. Computational technologies can appropriately be employed to integrate available data and can, in principle, be used to save on the expensive process of candidate genes selection.

Furthermore, with the advent of high-throughput technologies for the global measurement of miRNAs, these post-transcriptional regulatory molecules are emerging as a new class of cancer biomarkers. Numerous studies have explored associations between miRNAs and cancer features (for a view see Jeffrey 2008).

Discovered in *Caenorhabditis elegans* in 1993 and formally named in 2001 (Ruvkun 2001), miRNAs have been identified in every plant and animal species examined. They are a class of noncoding RNAs, 18–25 nucleotides in length, that plays key roles in the regulation of fundamental cellular processes such as differentiation, proliferation, apoptosis and metabolic homeostasis. The functions and targets of most miRNAs await discovery. However, specific miRNAs show expression variation across different stages of organism development, in tissue-specific cell patterning and asymmetry during organogenesis, and in oncogenesis (Garzon et al., 2006). **Figure 1** shows the miRNAs biogenesis.

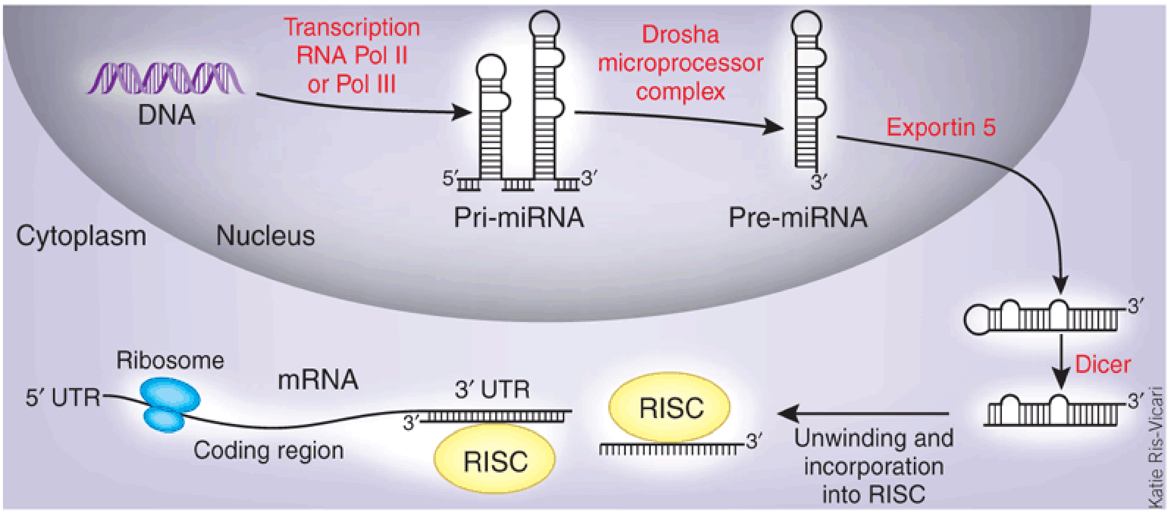

Figure 1. In the nucleus, a primary miRNA (pri-miRNA) is transcribed from DNA by the RNA polymerase II (Pol II) or Pol III enzyme. The long pri-miRNA transcript(~0.5–7kb) folds into a single or a cluster of multiple hairpin structure(s). The pri-miRNA is then cleaved by a microprocessor complex composed of the enzyme Drosha, a nuclear RNase III and the RNA binding protein cofactor Pasha (also known as DGCR8). The shorter stem-loop structure (~60–70 nucleotides), now termed precursor miRNA (pre-miRNA), is transported across the nuclear membrane by Exportin 5. In the cytoplasm, the pre-miRNA is further processed by a second RNase enzyme, Dicer. The hairpin loop is cropped off the double-stranded RNA, leaving a short miRNA duplex that is unwound by a helicase, cleaved into a mature miRNA (~18–25 nucleotides), and incorporated into an RNA-induced silencing complex (RISC), with an Argonaute protein as the catalytic component. The miRNA-RISC complex negatively regulates post-transcriptional gene expression by hybridizing to complimentary sequences in the 3' untranslated region (UTR) of a target mRNA and inhibiting protein translation or degrading the mRNA itself. (Figure from Jeffrey SS, Nat Biotechnol. 26:400-401, (2008))

Here I present several algorithms and web applications (TOM, Fun&Co and GebbaLab), implemented by me and my collaborators, to extract knowledge related to hereditary diseases and cancer, from public microarray messengerRNA and annotation data sets, and miRNAs microarray cancer studies. I report also several works which into my statistical analysis and bioinformatics procedures helped to quantify, confirm and visualize hypotheses and results, such as: improve the quality of DNA microarrays, ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas, miRNA DNA methylation signature for human cancer metastasis and miRNAs involved in colorectal cancer (CRC).

DNA microarrays

A DNA microarray is a multiplex technology used in molecular biology and in medicine. It consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides, called features, each containing picomoles of a specific DNA sequence. This can be a short section of a gene or other DNA element that are used as probes to hybridize a cDNA or cRNA sample (called target) under high-stringency conditions. Probe-target hybridization is usually detected and quantified by fluorescence-based detection of fluorophore-labeled targets to determine relative abundance of nucleic acid sequences in the target.

In standard microarrays, the probes are attached to a solid surface by a covalent bond to a chemical matrix (via epoxy-silane, amino-silane, lysine, polyacrylamide or others). The solid surface can be glass or a silicon chip, in which case they are commonly known as *gene chip*. Other microarray platforms, such as Illumina, use microscopic beads, instead of the large solid support. DNA arrays are different from other types of microarray only in that they either measure DNA or use DNA as part of its detection system.

DNA microarrays can be used to measure changes in expression levels, to detect single nucleotide polymorphisms (SNPs), in genotyping or in resquencing mutant genomes. Microarrays also differ in fabrication, workings, accuracy, efficiency, and cost. Additional factors for microarray experiments are the experimental design and the methods of analyzing the data. Arrays of DNA can be spatially arranged, as in the commonly known gene chip (also called genome chip, DNA chip or gene array), or can be specific DNA sequences labelled such that they can be independently identified in solution. The traditional solid-phase array is a collection of microscopic DNA spots attached to a solid surface, such as glass, plastic or silicon biochip. The affixed DNA segments are known as probes . Thousands of them can be placed in known locations on a single DNA microarray.

DNA microarrays can be used to detect DNA (as in comparative genomic hybridization), or detect RNA (most commonly as cDNA after reverse transcription) that may or may not be translated into proteins. The process of measuring gene expression via cDNA is called expression analysis or expression profiling.

Since an array can contain tens of thousands of probes, a microarray experiment can accomplish that many genetic tests in parallel. Therefore arrays have dramatically accelerated many types of investigation (**Table 1**).

Table 1 Microarray Applications

Technology or Application	**Synopsis**
Gene expression profiling	In an mRNA or gene expression profiling experiment the expression levels of thousands of genes are simultaneously monitored to study the effects of certain treatments, diseases, and developmental stages on gene expression. For example, microarray-based gene expression profiling can be used to identify genes whose expression is changed in response to pathogens or other organisms by comparing gene expression in infected to that in uninfected cells or tissues.
Comparative genomic hybridization	Assessing genome content in different cells or closely related organisms.
Chromatin immunoprecipitation on Chip	DNA sequences bound to a particular protein can be isolated by immunoprecipitating that protein (ChIP), these fragments can be then hybridized to a microarray (such as a tiling array) allowing the determination of protein binding site occupancy throughout the genome. Example protein to immunoprecipitate are histone modifications (H3K27me3, H3K4me2, H3K9me3, etc), Polycomb-group protein (PRC2:Suz12, PRC1:YY1) and trithorax-group protein (Ash1) to study the epigenetic landscape or RNA Polymerase II to study the transcription lanscape.
SNP detection	Identifying single nucleotide polymorphism among alleles within or between populations. Several applications of microarrays make use of SNP detection, including Genotyping, forensic analysis, measuring predisposition to disease, identifying drug-candidates, evaluating germline mutations in individuals or somatic mutations in cancers, assessing loss of heterozygosity, or genetic linkage analysis.
Alternative splicing detection	An exon junction *array* design uses probes specific to the expected or potential splice sites of predicted exons for a gene. It is of intermediate density, or coverage, to a typical gene expression array (with 1-3 probes per gene) and a genomic tiling array (with hundreds or thousands of probes per gene). It is used to assay the expression of alternative splice forms of a gene. Exon arrays have a different design, employing probes designed to detect each individual exon for known or predicted genes, and can be used for detecting different splicing isoforms.
Tiling array	Genome tiling arrays consist of overlapping probes designed to densely represent a genomic region of interest, sometimes as large as an entire human chromosome. The purpose is to empirically detect expression of transcripts or alternatively splice forms which may not have been previously known or predicted.

DNA microarray- two colors

DNA microarrays allow for rapid measurement and visualisation of differential expression between genes at the whole genome scale. The major steps involved in this process are:

i. Microarray production process
ii. Target preparation
iii. Hybridization
iv. Slide scanning
v. Data analysis
vi. Expression profile clustering

Microarray production process

DNA fragments amplified by PCR technique are spotted on a microscopic glass slide coated with polylysine prior to spotting process. The polylysine coating goal is to ensure DNA fixation through electrostatic interactions. Slide preparation is achieved by blocking the polylysine not fixed to DNA in order to avoid target binding. Prior to hybridisation, DNA is denatured to obtained a single strand DNA on the microarray, this will allow the probe to bind to the complementary strand from the target.

Target preparation

RNA are extracted from cultures/samples from which we want to compare expression level. Messengers RNA are then transformed in cDNA by reverse transcription. On this stage, DNA from the first culture with a green dye, whereas DNA from the second culture is labelled with a red dye.

Hybridisation

Green labelled cDNA and red labelled ones are mixed together (call the target) and put on the matrix of spotted single strand DNA (call the probe). The chip is then incubated one night at 60 degrees. At this temperature, a DNA strand that encounter the complementary strand and match together to create a double strand DNA. The fluorescent DNA will then hybridise on the spotted ones (**Figure 2**).

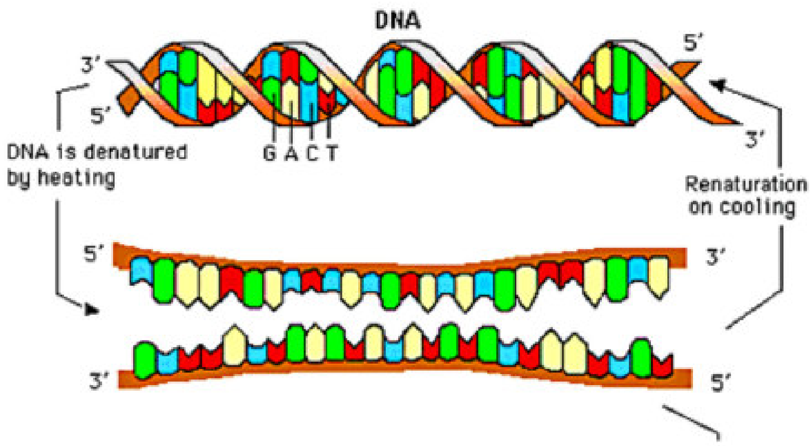

Figure 2 Hybridization. (Figure from http://transcriptome.ens.fr/sgdb/presentation/principle.php).

Slide scanning

A laser excites each spot and the fluorescent emission gather through a photo-multiplicator (PMT) coupled to a confocal microscope. We obtained two images where grey scales represent fluorescent intensities read. If we replace grey scales by green scales for the first image and red scales for the second one, we obtained by superimposing the two images one image composed of spots going from green ones (where only DNA from the first condition is fixed) to red (where only DNA from the second condition is fixed) passing through the yellow colour (where DNA from the two conditions are fixed on equal amount).

Data analysis

We have now two microarray images from which we have to calculate the number of DNA molecules in each experimental condition. To do so, we measure the signal amount in the green dye emission wavelength and the signal amount in the red dye emission wavelength. Then we normalise these amount according to various parameters (yeast amount in each culture condition, emission power of each dye, …). We suppose that the amount of fluorescent DNA fixed is proportional to the mRNA amount present in each cell at the beginning and we calculate the red/green fluorescence ratio. If this ratio is greater than 1 (red on the image), the gene expression is greater in the second experimental condition, if this ration is smaller than 1 (green on the image), the gene expression is greater in the first condition as shown in **Figure 3.**

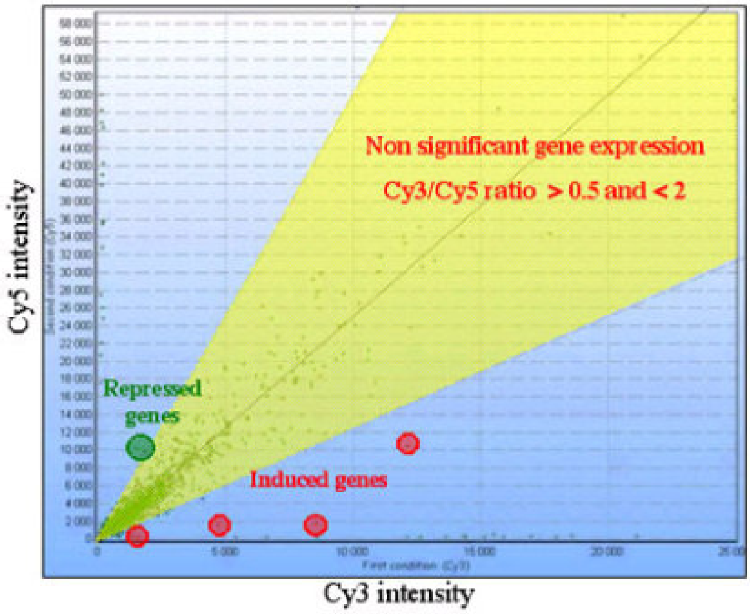

Figure 3 Two Color microarray, scatter plot of differentially expressed genes. (Figure from http://transcriptome.ens.fr/sgdb/presentation/principle.php).

Expression profile clustering

Then we can try to gather genes that share the same expression profile on several experiments. This clustering can be done gradually as for phylogenetic analysis, which consist in calculating similarity criteria between expression profiles and gather the most similar ones. We can also use more complex techniques as principal component analysis or neuronal networks.

DNA microarrays – one color - Affymetrix©

The GeneChip high-density oligonucleotide arrays are fabricated by using in-situ synthesis of short oligonucleotide sequences on a small glass chip using light directed synthesis. This technique allows for the precise construction of a highly ordered matrix of DNA oligomers on the chip.

Design overview

In the GeneChip system a known gene or potentially expressed sequence is represented on the chip by 11-20 unique oligomeric probes, each 25 bases in length. The group of probes corresponding to a given gene or small group of highly similar genes is known as the probe set and generally spans a region of about 600 bases, known as the target sequence. Many copies of each oligomer are synthesized in discrete features (or cells) on the GeneChip array. In addition, for each oligomer on the array there is a matched oligomer, synthesized in an adjacent cell that is identical with the exception of a mismatched base at the central position (i.e. base 13). These are designated Perfect Match (PM) and Mismatch (MM) probes, respectively. The MM probes serves as a control for non-specific hybridization.

GeneChip Array Fabrication

Probe arrays are manufactured by Affymetrix's proprietary, light-directed chemical synthesis process, which combines solid-phase chemical synthesis with photolithographic fabrication techniques employed in the semiconductor industry. Using a series of photolithographic masks to define chip exposure sites, followed by specific chemical synthesis steps, the process constructs high-density arrays of oligonucleotides, with each probe in a predefined position in the array. Multiple probe arrays are synthesized simultaneously on a large glass wafer. This parallel process enhances reproducibility and helps achieve economies of scale. The wafers are then diced, and individual probe arrays are packaged in injection-molded plastic cartridges, which protect them from the environment and serve as chambers for hybridization.

Data Overview

Affymetrix GeneChip experiments are managed using GCOS. GCOS interfaces with equipment to run a probe array experiment and is also used to generate preliminary analysis data from an experiment. The next section covers the basics of files generated by GCOS and also explains some of the most widely used variables generated by GCOS.

MAS File Types

The next section covers the basics of files generated by GCOS and also explains some of the most widely used variables generated by GCOS.

- **Experiment File *.EXP:** This file contains the parameters of the experiment such as Probe Array Type, Experiment Name, Equipment parameters, Sample Description, and others. This file is not used for analysis, but is required to open other GCOS files for the designated chip experiment.
- **Image Data File *.DAT:** This is an image file generated by the scanner from the Probe Array after processing on the Fluidics Station. This file can be viewed in GCOS to assess the quality of scanning event or exported as a *.TIFF image. It is used in GCOS to generate the *.CEL file.
- **Cell Intensity File *.CEL:** This binary file is the result of low level analysis performed from the *.DAT image file. It is exported from GCOS and is often used as the base file

for further analysis.

- **Probe Array Results File *.CHP:** This binary file is a gene level summarization of the CEL file using the Affymetrix' MAS 5.0 or PLIER algorithms. It is exported from GCOS. It can also be used as the base file for further analysis; however one needs to know the settings of key parameters (alpha1, alpha2, tau, target signal etc.). There are many other algorithms that have been adopted by the community other than MAS 5.0 and PLIER, hence the reason why CEL files are often preferred over CHP files by Investigators for analysis.
- **Report File *.RPT:** The report file is generated from the chip file. This expression report summarizes information about expression analysis settings and probe set hybridization intensity data.
- **MAGE-ML *.XML:** This file contains information related the microarray experiment (one per experiment). This information can include biologically relevant information, array details, fluidics protocol details and the analysis settings. It also records the file hierarchy of an experiment.

GCOS uses a statistical algorithm to calculate signals and make significance calls for the data: MAS (MAS 5.0 computes local background in each of 16 squares, and then subtracts a weighted combination of these background estimates from each probe intensity).

MAS Analysis Metrics

- **Signal:** a measure of the abundance of transcript
- **Detection:** the call that indicates whether the transcript is detected (P present), undetected (A, absent), or at the limit of detection (M, marginal).
- **Detection p-value:** p-value that indicates the significance of the detection call.
- **Signal Log Ratio:** the change in expression level of a transcript between a baseline and an experiment array. This change is expressed as the log2 ratio.

Each probe set on a GeneChip array has a unique name known as the **Probe set ID**.

Affymetrix has upgraded their MicroArray Suite (MAS) software several times. MAS 4.0 was the standard until January 2002 and is still cited in published papers. MAS 4 calculates a weighted average of the probe-pair differences (PM – MM) for each probe pair representing a gene. MAS 5.0 improves in two important ways. First the intensities are transformed to a logarithmic scale before the average is taken; this equalizes the contribution of different probes. Secondly an estimate of background based on MM replaces MM itself in the difference PM-MM; this estimate is itself a weighted average of log probe pair differences: $\log(PM_j/MM_j)$.

Normalization

Biologists have long experience coping with systematic variation between experimental conditions (technical variation) that is unrelated to the biological differences. Normalization is the attempt to compensate for systematic technical differences between chips, to see more clearly the systematic biological differences between samples. For example, differences in treatment of two samples, especially in labelling and in hybridization, bias the relative measures on any two chips.

Normalization by Scaling and its Limitations

The simplest approach to normalizing Affymetrix data is to re-scale each chip in an experiment to equalize the average (or total) signal intensity across all chips. The reasoning behind this is that there should be with equal weights of RNA for all the samples; if the sizes of the RNA molecules are comparable, the number of RNA molecules should also be the roughly the same in each sample. Consequently, nearly the same number of labeled molecules from each sample should hybridise to the arrays and, if all other conditions were equal, the total hybridisation intensities summed over all elements in the arrays should be the same for each sample.

To do better, we examine in detail the relationships among replicate chips (chips hybridized to the same sample). Figure 4a shows a scatter plot of probes from one pair of chips; there is clearly a non-linear relation among probes. Figure 4b shows plots of probe distributions from a number of replicate chips on a log scale; these distributions have very different shapes; on a log scale, applying a scaling transform to a chip, shifts its distribution curve to the right or left, but doesn't change its shape.

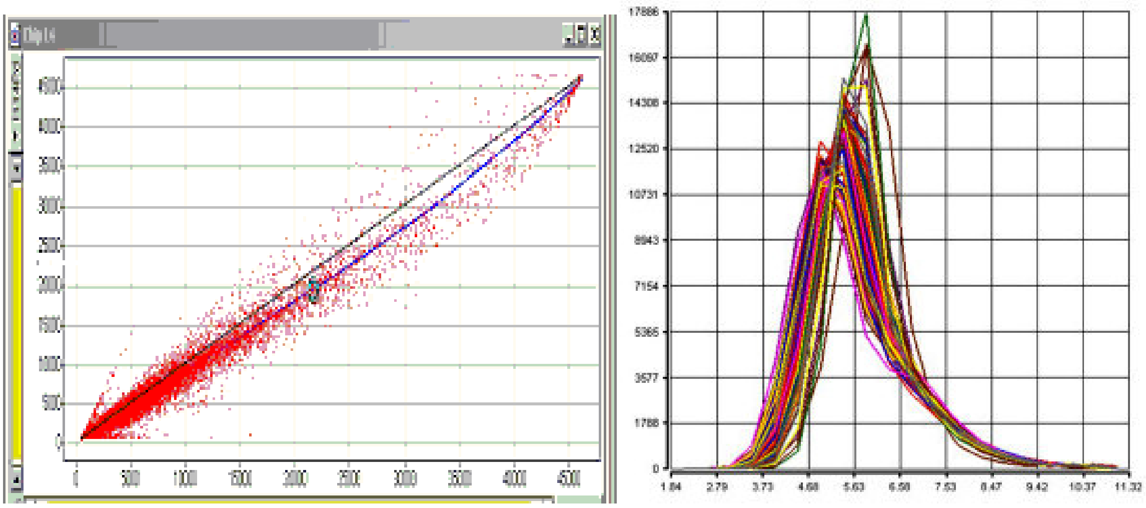

Figure 4. a) Plot of probe signals from two Affymetrix chips hybridized with identical mRNA samples. The black straight line represents equality, while the blue curve is a spline fit through the scatter plot. B) Density curves of global signal intensities. The plots show the overall signal density distribution of all probe sets represented on the HG-U133 Plus 2.0 microarray. Data from each microarray analysis is represented by a separate line. The plot is useful to visualize whether there are differences in the overall signal distributions of the experiments. (Figure from http://www.bea.ki.se/staff/reimers/Web.Pages/Affymetrix.Normalization.htm).

Quantile Normalization

Is a non-parametric procedure normalizing to a synthetic chip (Bolstad et al., 2003). It is a kind of normalization that works across arrays as well as within arrays. It turns out that quantile normalization works quite well at reducing variance between arrays, while compensating the intensity-dependent dye bias, as well as does lowess normalization.This method assumes that the distribution of gene abundances is nearly the same in all samples. The pooled distribution of probes on all chips are taken. Then to normalize each chip the algorithm compute for each value, the quantile of that value in the distribution of probe intensities; then it transform the original value to that quantile's value on the reference chip. In a formula, the transform is

$$x_{norm} = F_i^{-1}(F_{ref}(x)) ,$$

where F_i is the distribution function of chip i, and F_{ref} is the distribution function of the reference chip.

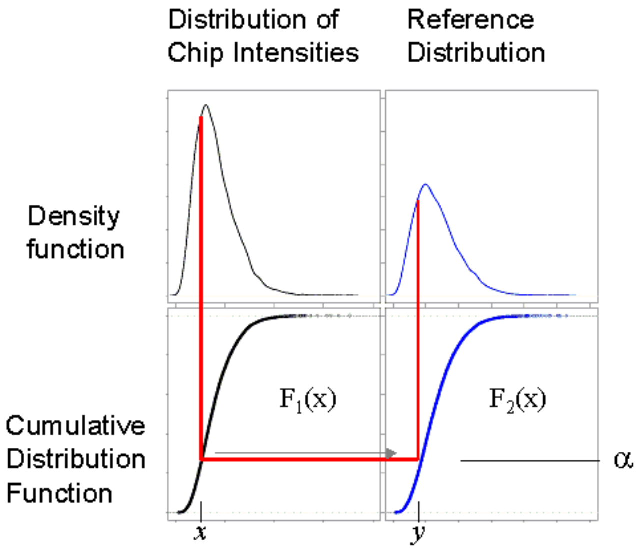

Figure 5. Schematic representation of quantile normalization: the value x, which is the α-th quantile of all probes on chip 1, is mapped to the value y, which is the α quantile of the reference distribution F_2. (Figure from http://www.bea.ki.se/staff/reimers/Web.Pages/Affymetrix.Normalization.htm).

Proportional Variance – RMA (Robust Multichip Average)

This is largely the work of Terry Speed's group at Berkeley, especially Ben Bolstad, and Rafael Irizarry (Irizarry et al., 2003). They work only with PM values, and ignore MM entirely. They take a log transform of equation () and find:

$$\log(PM_{ij}) = \log(a_i) + \log(f_j)$$

With errors proportional to intensity in the original scale, the errors on the log scale have constant variance. After background subtraction and normalization they fit:

$$\mathrm{nlog}(PM_{ij} - \mathrm{bg}) = a_i + b_j + \varepsilon_{ij}$$

where nlog is their terminology for 'normalize and then take logarithm'. They fit this model by iteratively re-weighted least squares, or by median polish. Code is available in the affy package on BioConductor, together with quantile normalization (http://www.bioconductor.org/packages/bioc/).

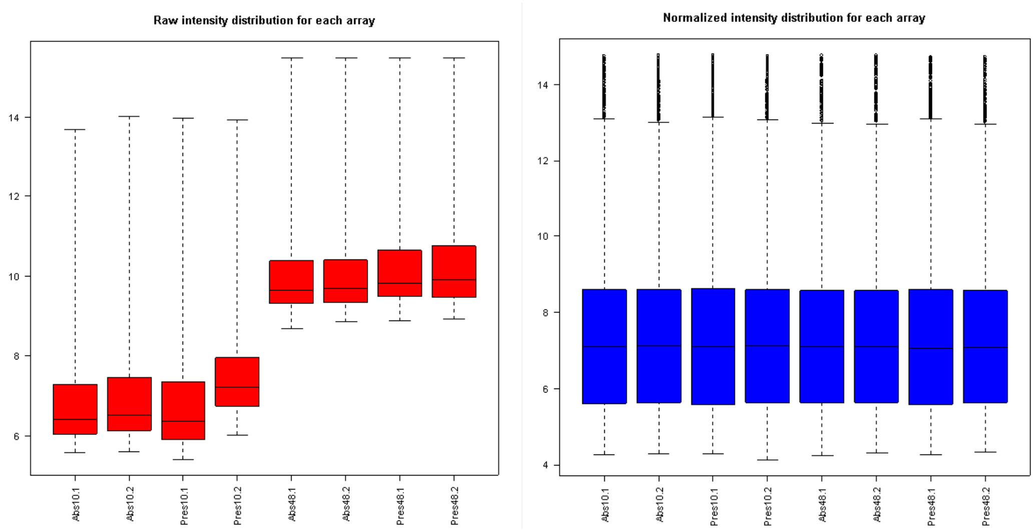

Figure 6. Raw Data Box Plot (left), the same intensity box plot using normalized data (right). The normalization remove effects seen in large proportions of the data (in this case a time effect is obvious) while still preserving effects seen in small proportions of the data. (Figure from http://bioinf.wehi.edu.au/affylmGUI/R/library/affylmGUI/doc/estrogen/estrogen.html).

GCRMA

Another proposed background correction method is GC-RMA (Wu et al., 2003). This method is based upon sequence information, such as GC content, for each probe and stochastic models for binding affinities. GC-RMA is a modified version of RMA that models intensity of probe level data as a function of GC-content. We expect to see higher intensity values for probes that are GC rich due to increased binding.

Exploratory Analysis

Pattern-Finding

Exploratory analysis aims to find patterns in the data that aren't predicted by the experimenter's current knowledge or pre-conceptions. Some typical goals are to identify groups of genes expression patterns across samples are closely related; or to find unknown subgroups among samples. A useful first step in all analyses is to identify outliers among samples – those that appear suspiciously far from others in their group. To address these questions, researchers have turned to methods such as cluster analysis, and principal components analysis.

Clustering

Suppose that we want to find groups of similar genes or similar samples, how do we go about it? Clustering depends on the idea that differences between gene expression profiles are like distances; however the user must make choices to compute a single measure of distance from many individual differences. Different procedures emphasize different types of similarities, and give different resulting clusters. Four choices we have to make are:

i. what scale to use: original scale, log scale, or another transform,
ii. whether to use all genes or to make a selection of genes,
iii. what metric (distance measure) to use to combine the scaled values of the selected genes,
iv. what clustering algorithm to use.

Scale

Differences measured on the linear scale will be strongly influenced by the one hundred or so highly expressed genes, and only moderately affected by the hundreds of moderate abundance genes; the thousands of low abundance genes will contribute little. Often the high-abundance genes are 'housekeeping' genes; these may or may not be diagnostic for the kinds of differences being sought. On the other hand, the log scale will amplify the noise among genes with low expression levels. If low-abundance genes are included then they should be down-weighted. The most useful measure of a single gene difference is the difference between two samples, relative to that gene's variability within experimental groups: this is like a t-score for difference between two individuals.

Gene Selection

It would be wise not to place much emphasis on genes whose values are uncertain. These are usually those with low signals in relation to noise, or which fail spot-level quality control. If the estimation software provides a measure of confidence in each gene estimate, this can be used to weight the contribution to distance of that gene overall. It's not wise to simply omit (that is, set to 0) distances which are not known accurately, but it is wise to down-weight relative distances if several are probably in error. A simple general rule is that genes whose signal falls within the background noise range are probably contributing just noise to your clustering (and any other global procedure); discard them.

Metrics

Usually, cluster programs give us a menu of distance measures: Euclidean, Manhattan distances, and some relational measures: correlation, and sometimes relative distance, and mutual information. The names describe how differences are combined:

Euclidean is straight-line distance: (root of sum of squares, as in geometry), Manhattan is sum of linear distances (like navigating in Manhattan). The correlation distance measure is actually 1-r, where r is the correlation coefficient. Probably a more useful version is 1 – |r|; negative correlation is as informative as positive correlation. We do get different results depending on the algorithm we use, as shown below (**Figure 7**) for a study with 10 samples: two normal samples and two groups of tumor samples.

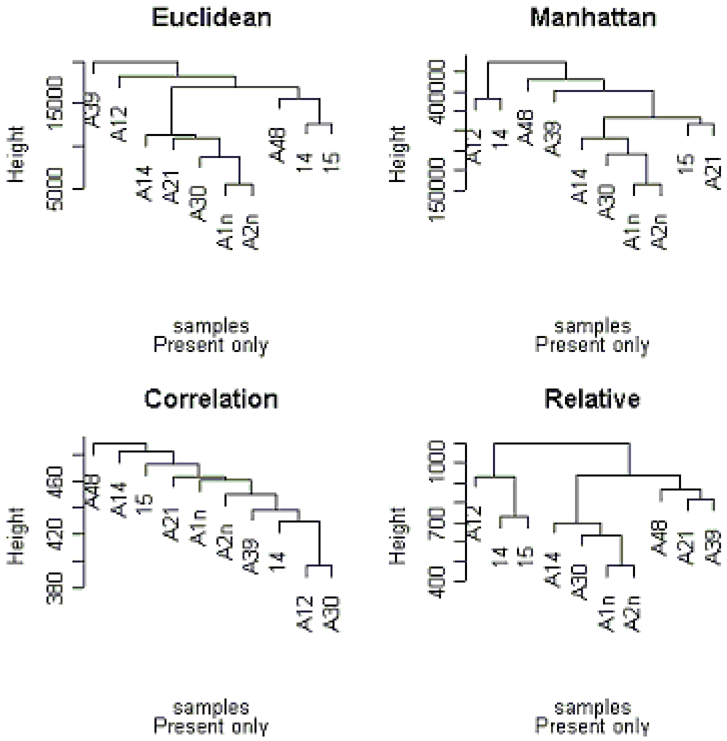

Figure 7 Clustering of the same data set using four different distance measures. (Figure from http://discover.nci.nih.gov/microarrayAnalysis/Exploratory.Analysis.jsp).

Principal Components and Multi-dimensional scaling

Several other good multivariate techniques can help with exploratory analysis. Many authors suggest principal components analysis (PCA) or singular value decomposition to find coherent patterns of genes, or 'metagenes', that discriminate groups. These techniques with a long history in the statistical arsenal rely on the idea that most variation in a data set can be explained by a smaller number of transformed variables; they each form linear combinations of the data, which represent most of the variation, and in principle these approaches, are well-suited for this purpose. These multivariate approaches are more useful for exploring relations among samples, and particularly for a diagnostic look at samples before formal statistical tests (see **Figure 8**).

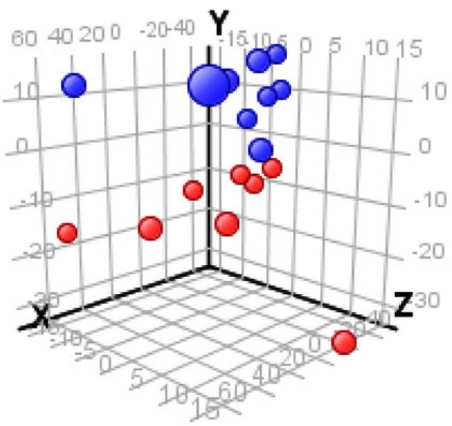

Figure 8 shows a GeneSpring© PCA plot of two groups in a comparative study; the control group is in red; treated samples are quite distinct from untreated and each other. (Figure obtained using 16 samples of treated and untreated samples).

Statistical Tests

The Purposes of Statistical Tests

Microarray studies often aim to identify genes that are differentially regulated across different classes of samples; examples are: finding the genes affected by a treatment, or finding marker genes that discriminate diseased from healthy subjects.

Microarray data is often used as a guide to further, more precise studies of gene expression by qt-PCR or other methods. Then the goal of the statistical analysis is heuristic: to provide the experimenter with an ordered list of good candidate genes to follow up. Sometimes the experimenter plans to publish microarray results as evidence for changes in gene abundance; in this case it is important to state the correct degree of evidence: the 'p-value'. Being many genes actually tested in parallel singles p-values are wrong in the context of testing thousands of genes. A better way to specify the confidence of microarray results is the 'false discovery rate'.

Transforms

Often the first step is transforming the values to log scale, and doing all subsequent steps on the log-transformed values. Although taking logarithms is common practice, and helpful in several ways, there are other options. The main justification for transforms in statistics is to better detect differences between groups whose within-group variances are very different. Most commonly the within-group variances are higher in those groups where the mean is also higher. A different kind of variation, the measurement error in expression level estimates, grows with the mean level. If the measurement error is proportional to the mean, then the log-transformed values will have consistent variance for all genes. For both reasons many researchers argue that gene expression measures should be analyzed on a logarithmic scale.

Comparison of Two Groups of Samples

The simplest and most common experimental set-up is to compare two groups: for example, Treatment vs. Control, or Mutant vs. Wild type.

The long-time standard test statistic for comparing two groups is the t-statistic:

$$t = (x_{i,1} - x_{i,2}) / s_i,$$

where $x_{i,1}$ is the mean value of gene i in group 1, $x_{i,2}$ is the mean in group 2, and s_i is the (non-pooled) within-groups standard error (SE) for gene i.

Another approach to detecting more of the differentially expressed genes is to use a more precise estimate of the variation between individuals, for each gene, in tests of that gene. If a good deal of prior data exist on the tissue and strain used in the wild-type (or control) group, measured on the same microarray platform – and this is sometimes the case now – then it is defensible to pool the estimates of wild-type variation from each of the prior studies, and use this as the denominator in the t-scores. The t-scores should then be compared to the t-distribution on a number of degrees of freedom, equal to that used in computing the pooled standard error. A variant of this approach may be used in a study where many groups are compared in parallel. The within-group variances for each gene may be pooled across the different groups to obtain a more accurate estimate of variation. This presumes that treatments applied to different groups affect mostly the mean expression levels, and not the variation among individuals. Of course one should test that the discrepancies in variance estimates are not too large for many of the genes that are selected as differentially expressed. This may be done by computing the ratios of variances between groups (F-ratios), and comparing to an F-distribution.

Permutation Tests

Permutation testing is an approach that is widely applicable and copes with distributions that are far from Normal; this approach is particularly useful for microarray studies because it can be easily adapted to estimate significance levels for many genes in parallel. Some recent software packages, notably SAM (Significance Analysis of Microarray, http://www-stat.stanford.edu/~tibs/SAM/), implement permutation testing in a menu-driven interface.

The meaning of a p-value from a permutation procedure differs from the meaning of a model-based p-value. The model-based p-value is the probability of the test statistic, assuming that the gene levels in both the treatment and control groups follow the model (eg. a Normal distribution). A permutation-based p-value tells how rare that test statistic is, among all the random partitions of the actual samples into pseudo-treatment and pseudo-control groups. The steps in a permutation-based computation of the significance level of a test statistic are as follows:

i. Choose a test statistic, eg. a t-score for a comparison of two groups,
ii. Compute the test statistic for the gene of interest,
iii. Permute the labels on samples at random, and re-compute the test statistic for the rearranged labels; repeat for a large number (perhaps 1,000) permutations, and finally,
iv. Compute the fraction of cases in which the test statistics from iii) exceed the real test statistic from ii).

Volcano Plot

However one chooses to compute the significance values (p-values) of the genes, it is interesting to compare the size of the fold change to the statistical significance level. The ‘volcano plot’ arrange genes along dimensions of biological and statistical significance. The first (horizontal) dimension is the fold change between the two groups (on a log scale, so that up and down regulation appear symmetric), and the second (vertical) axis represents the p-value for a t-test of differences between samples (most conveniently on a negative log scale – so smaller p-values appear higher up). The first axis indicates biological impact of the change; the second indicates the statistical evidence, or reliability of the change. In this way we can then make judgements about the most promising candidates for follow-up studies, by trading off both these criteria by eye (**Figure 9**).

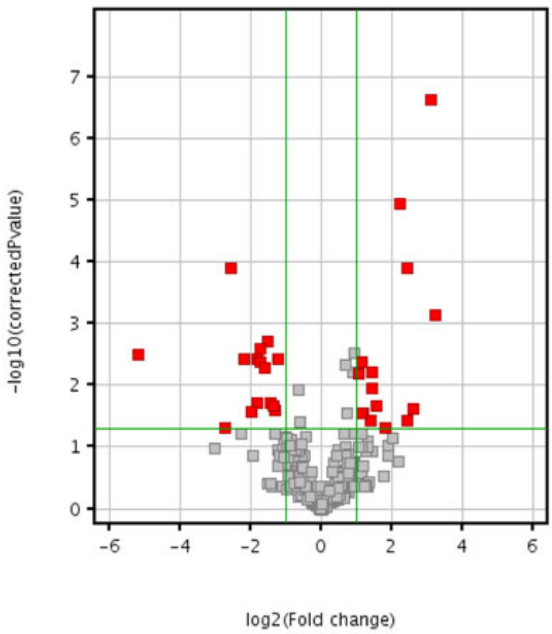

Figure 9. A volcano plot. Siaplaying entities satisfying p_value cutoff and Fold Change cut-off. (Figure obtained using 16 samples of treated and untreated samples).

Genome-Wide Comparisons, Corrected P-Values, and False Discovery Rates

P-Values and False Discovery Rates

Most scientific papers quote p-values, however few papers discuss their meaning. In order to understand what the problem is with quoting p-values for massively parallel comparisons, we need to be precise. Let's consider, for example, a t-test of differences between two samples. If there is no systematic (real, reproducible) difference between groups, nevertheless the t-score for differences between groups is never exactly 0. Common sense cannot decide whether a particular value provides strong evidence for a real difference. The natural question to ask is: how often a random sampling of a single group would produce a t value as far from 0 as the t we observed. When you declare an effect is significant at 5%, you say you are willing to let one false positive sneak in, roughly every twenty tests. We don't accept this for critical decisions; we won't long continue to cross the street, if we do so on a 95% confidence that there is a break in traffic. We may call this the false positive rate (FPR); the FPR of a procedure is the fraction of truly unchanged genes which appear as (false) positives.

If the aim of the microarray study is to select a few genes for more precise study, then the goal is an ordered list of genes, most of which are really different (true positives). Another way to say this is that the expected number of false positives is some reasonable fraction (for example less than .3) of the genes selected. This goal leads naturally to specifying the false discovery rate (FDR) for a list, rather than significance level (FPR). The FDR is the expected fraction of false positives in a list of genes selected following a particular statistical procedure.

Multiple Testing P-Values and False Positives

Suppose you compare two groups of samples drawn from the same larger group, using a chip with 10,000 genes on it. On average 500 genes will appear 'significantly different' at a 5% threshold. For these genes, the variation between samples will be large relative to the variation within groups due to random, but uneven allocation of the expression values to the treatment and control groups. Therefore the p-value appropriate to a single test situation is inappropriate to presenting evidence for a set of changed genes.

Statisticians have devised several procedures for adjusting p-values to correct for the multiple comparisons problem. The oldest is the Bonferroni correction; this is available as an option in many microarray software packages. The corrected p-value, p_i^* for gene i is set to: $p_i^* = Np_i$, if $Np_i < 1$, or 1, if $Np_i > 1$; where p_i is the p-value for a single test of gene i, and N is the number of genes being tested (which may be less than the number of genes on the array).

Calculating Permutation-Based Corrected P-values

To calculate corrected p-values, first calculate single-step p-values for all genes: $p_1, \ldots, p_N$. Then order the p-values: $p_{(1)}, \ldots, p_{(N)}$, from least to greatest. Next permute the sample labels at random, and compute the test statistics for all genes between the two (randomized) groups. For each position k, keep track of how often you get at least one p-value more significant than $p_{(k)}$, from gene k, or from any of the genes further down on the list: $k+1$, $k+2$, …, N. After all permutations, compute the fraction of permutations with at least one apparently more significant p-value less than $p_{(k)}$. This is the corrected p-value for gene k. Although this procedure is complicated, it is much more powerful than the other corrections: that is, the procedure gives a much smaller corrected p-value for each gene than the Bonferroni procedure, and therefore a bigger list of significant genes at any corrected significance level (specified risk of false positives). This is known as the Westfall–Young correction.

Several Groups – Analysis of Variance

Many current microarray studies compare more than two groups. Sometimes the question is to determine differences among three or more cell lines, or strains of experimental animal. Another common design compares the effect of a particular treatment (often a ligand for a receptor), on cell lines (or animals) with wild-type and mutant versions of the receptor. Usually the experimenter wants to know which genes are actively regulated during treatment in both cell lines, or wants some criterion for selecting those that are differentially regulated among groups. These questions belong in the tradition of analysis of variance (ANOVA). Generally, all of the procedures that were discussed above in the context of two-sample comparisons, carry over to analogues in ANOVA.

Microarray databases

There are many public databases for microarray data. The databases meant to be central repositories, among them, the most known:

i. ArrayExpress at EBI (http://www.ebi.ac.uk/arrayexpress/)

ii. GEO (Gene Expression Omnibus) at NCBI (http://www.ncbi.nlm.nih.gov/geo/)

They are not only public archives of microarray data but also enforcing standards to maintain data quality, providing powerful search methods to facilitate finding particular data, and providing analytical tools to facilitate comparison and/or visualization of large data. According to the publication about GEO (Barrett et al., 2005), "These data include single and multiple channel microarray-based experiments measuring the abundance of mRNA, genomic DNA and protein molecules. Data generated by innovative applications of microarray technology are also accepted, e.g. chromatin immunoprecipitation (ChIP-chips) for identifying protein-binding DNA regions and tiling arrays for genome annotation. Data from non-array-based highthroughput functional genomics and proteomics technologies are also archived, including serial analysis of gene expression (SAGE), and mass spectrometry peptide profiling." So, it is not only microarray data.

Let's look at GEO (**Figure 10**).

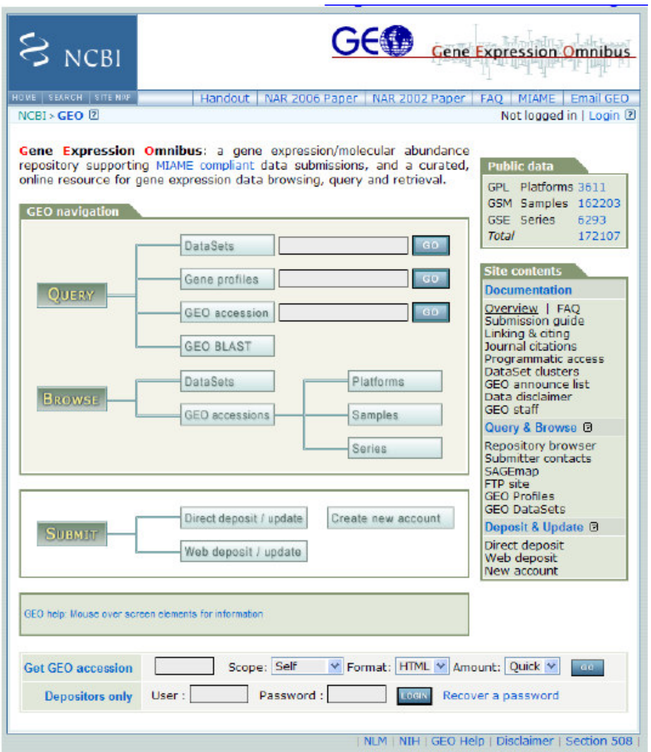

Figure 10 ncbi/GEO query interface. (Figure obtaine from http://www.ncbi.nlm.nih.gov/geo/).

We see we can make queries to get data of our interest or browse through them. The data sets are called GEO-DataSets. It is an experiment-centric view (organized according to each experiment). If, for example, we search for “Human[Organism] AND microRNAs” we obtain several records. Note that there are ‘GDS…’ records lines, ‘GSE…’ records, and a ‘GPL…’ record. Each GDS (dataset) record is the entire set of data from one experiment and comes with various tools to play. GSE (series) records explain the experiment, including RNA samples used. GPL (platform) records are descriptions of microarray platforms. We can also see ‘GSM…’ in GDS and GSE records. GSM (sample) records contain data from each sample used in the experiment.

List of several useful microarray data and not only:

- ArrayExpress—a public database of microarray experiments and gene expression profiles
- The Stanford Microarray Database
- Microarray retriever: a web-based tool for searching and large scale retrieval of public microarray data
- OligoArrayDb: pangenomic oligonucleotide microarray probe sets database
- CEBS—Chemical Effects in Biological Systems: a public data repository integrating study design and toxicity data with microarray and proteomics data
- Gene Aging Nexus: a web database and data mining platform for microarray data on aging
- ChipInfo: software for extracting gene annotation and gene ontology information for microarray analysis
- ITTACA: a new database for integrated tumor transcriptome array and clinical data analysis
- BarleyBase—an expression profiling database for plant genomics
- ArrayXPath: mapping and visualizing microarray gene-expression data with integrated biological pathway resources using Scalable Vector Graphics
- NASCArrays: a repository for microarray data generated by NASC’s transcriptomics service
- CanGEM: mining gene copy number changes in cancer
- CleanEx: a database of heterogeneous gene expression data based on a consistent gene nomenclature
- GEPAS: a web-based resource for microarray gene expression data analysis
- NetAffx: Affymetrix probesets and annotations

The term microarray database is usually used to describe a repository containing microarray gene expression data. The key features of a microarray database are to store the measurement data, manage a searchable index, and make the data available to other applications for analysis and interpretation (either directly, or via user downloads).

Microarray databases can fall into two distinct classes:

i. A peer reviewed, public repository that adheres to academic or industry standards and is designed to be used by many analysis applications and groups. A good example of this is the Gene Expression Omnibus (GEO) from NCBI or ArrayExpress from EBI.

ii. A specialized repository associated primarily with the brand of a particular entity (lab, company, university, consortium, group), an application suite, a topic, or an analysis method, whether it is commercial, non-profit, or academic. These databases may be characterized by:

   - A subscription or license may be needed to gain full access,
   - The content may come primarily from a specific group (e.g. SMD, or UPSC-BASE),
   - There may be limits on how who can use the data, and for what purpose,
   - Special permission may be required to submit new data, or there may be no obvious process at all,
   - Only certain applications may be equipped to use the data, often also associated with the same entity (for example, caArray at NCI is specialized for the caBIG),
   - Further processing or reformatting of the data may be required for standard applications or analysis,
   - They claim to address the 'urgent need' to have a standard, centralized repository for microarray data. (See YMD, last updated in 2003, for example),
   - There is a claim to an incremental improvement over one of the public repositories,
   - A meta-analysis *application*, which incorporates studies from one or more public databases (e.g. Gemma primarily uses GEO studies; NextBio uses various sources)

Some of the most known public, curated microarray *databases* are reported in **Table 2**.

Table 2 Microarray Databases

Database	Scope	Web site
Gene Expression Omnibus - NCBI	any curated MIAME compliant molecular abundance study	http://www.ncbi.nlm.nih.gov/geo/
Stanford Microarray database	stores raw and normalized data from microarray experiments, and provides data retrieval, analysis and visualization	http://genome-www5.stanford.edu/
Genevestigator database	Manually curated microarray data for expression meta-analysis	https://www.genevestigator.ethz.ch/gv/index.jsp
ArrayExpress at EBI	Any curated MIAME or MINSEQE compliant transcriptomics data	http://www.ebi.ac.uk/microarray-as/ae/
UPenn RAD database	MIAMI compliant public and private studies, associated with ArrayExpress	http://www.cbil.upenn.edu/RAD/php/index.php
UNC Microarray database	microarray data storage, retrieval, analysis, and visualization	https://genome.unc.edu/
MUSC database	repository for DNA microarray data generated by MUSC investigators	http://proteogenomics.musc.edu/ma/musc_madb.php?page=home&act=manage
caArray at NCI	Cancer data, prepared for analysis on caBIG	https://array.nci.nih.gov/caarray/home.action
UPSC-BASE	data generated by microarray analysis within Umeå Plant Science Centre (UPSC).	https://www.upscbase.db.umu.se/

MicroRNAs

MicroRNAs (miRNAs) are small noncoding RNAs (ncRNAs, RNAs that do not code for proteins) that regulate the expression of target genes at the posttranscriptional or posttranslational level. Many miRNAs have conserved sequences between distantly related organisms, suggesting that these molecules participate in essential developmental and physiologic processes. miRNAs can act as tumor suppressor genes or oncogenes in human cancers. Mutations, deletions, or amplifications have been found in human cancers and shown to alter expression levels of mature and/or precursor miRNA transcripts. Moreover, a large fraction of genomic ultraconserved regions (UCRs) encode a particular set of ncRNAs whose expression is altered in human cancers.

In genetics, miRNAs are single-stranded RNA molecules of about 21–23 nucleotides in length, which regulate gene expression. miRNAs are encoded by genes from whose DNA they are transcribed but miRNAs are not translated into protein (non-coding RNA); instead each primary transcript (a *pri-miRNA*) is processed into a short stem-loop structure called a *pre-miRNA* and finally into a functional miRNA.

Mature miRNA molecules are partially complementary to one or more messenger RNA (mRNA) molecules, and their main function is to down-regulate gene expression. They were first described in 1993 by Lee and colleagues in the Victor Ambros lab (Lee et al., 1993), yet the term *microRNA* was only introduced in 2001 in a set of three articles in Science (Ruvkun , 2001).

Microarrays for miRNA

The microRNAs expression study had in the past years large difficulties due to their small dimensions and the insufficient sensibility of the methods used, like the Northern blot, the cloning and the arrays on membrane revealed with a radioactive. The application of the technology of the microarrays to the analysis of the profile of expression of miRNA offered meaningful advantages like a greater sensibility and elevated comparative abilities. In the laboratory of Prof. Croce (OSUCCC, Ohio State University, Liu et al., 2008) has been developd a microarray (or chip) for the study of the alterations in the expression of all the miRNA known in the human cancer and it has been established as a reproductable detection method (Liu et Al, 2004). The levels of miRNA are obtained for quantification of the intensities of mark them with appropriate software as GenePix (**Figure 11**).

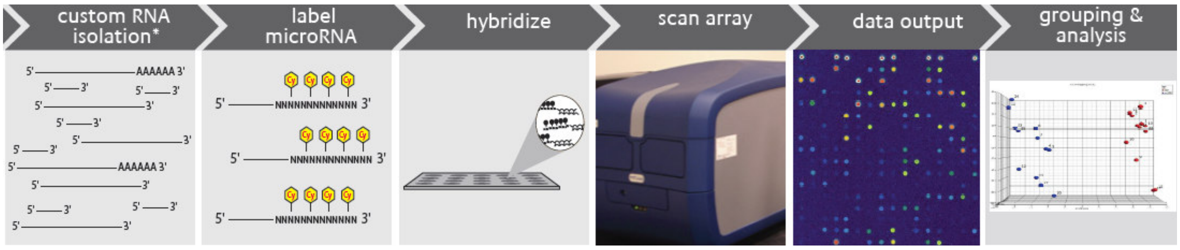

Figure 11. Phases of miRNA microarray experiment. (Figure from paper VII).

MiRNA and cancer

Several miRNAs has been found to have links with some types of cancer.

Recent studies demonstrated that in cancer the levels of some miRNA are altered (Volinia et al., 2006; Lu et al., 2005). Some miRNA, as miR-21 and miR-155 are overexpressed in solid tumors and in the leukaemias. In **figure 12** 6 solid tumors are represented (columns), and in every line the microRNA associated to the solid tumors. The figure is a graphical rapresentation of the obtained result by using approximately 500 biopsies. A red square represents an overexpression of the microRNAs in tumor; a green represents downregulation of the miRNA in tumor. For example, miR-21 is overwxpressed in all and the 6 considered tumors (breast, lung, colon, pancreas, prostate and stomach).

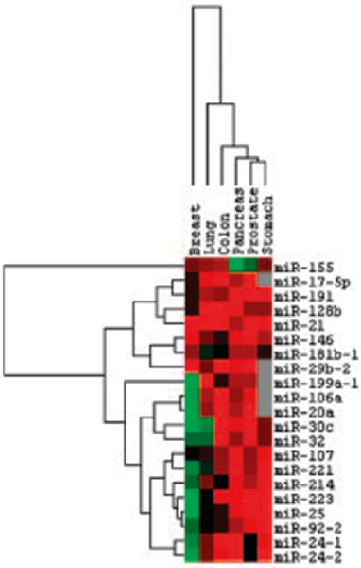

Figura 12. Fold changes (cancer vs. normal) of the miRNAs present in the signatures of at least 50% of the solid tumors. The tree displays the log2 transformation of the average fold changes (cancer over normal). The mean was computed over all samples from the same tissue or tumor histotype. Arrays were mean centered and normalized by using GENE CLUSTER 2.0. Average linkage clustering was performed by using uncentered correlation metric. (Figure from Volinia S et al., Proc Natl Acad Sci U S A., 103:2257-2261, (2006)).

The discovery of the association miRNA-cancer although revealed of a strong correlation, has not been enougth to establish a cause-effect connection. After, the responsibility of the microRNA in the insorgence of the cancer has been demonstrated in transgenic mouse with miR-155 (Costinean et al., 2006). In such mice in fact the insertion of additional copies of miR-155 provokes a high grade lymphoma with correspondent splenomegaly (**Figure 13**).

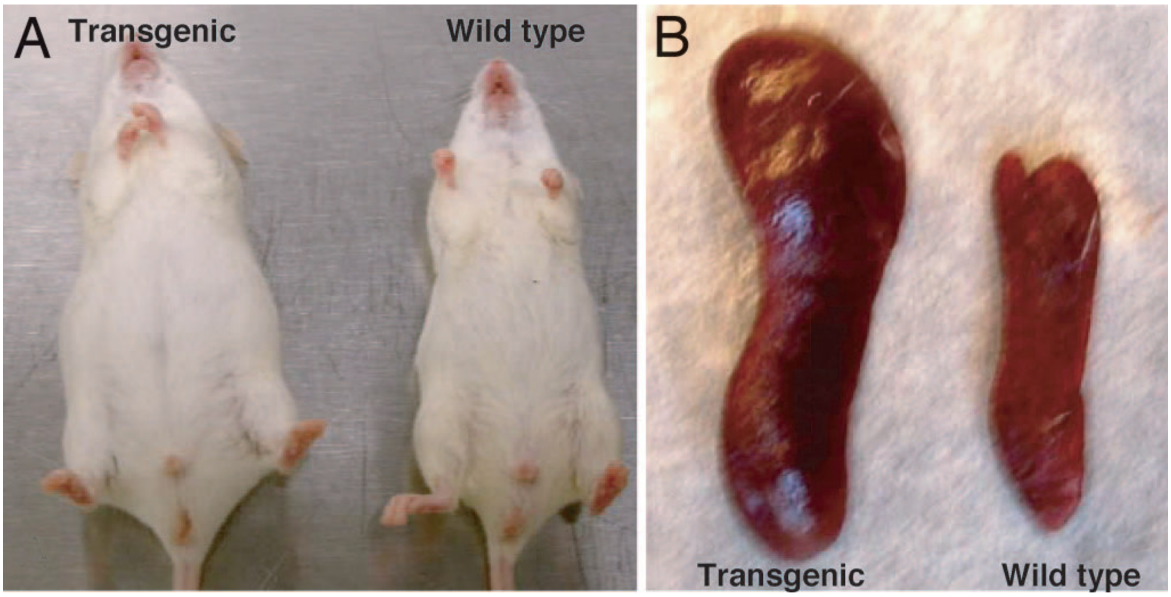

Figura 13. Transgenic mice, 6 months old, presented an enlarged abdomen and important splenomegaly. (*A*) Transgenic mice, 6 months old, had a considerably enlarged abdomen compared with wild-type mice, due to the clinically evident splenomegaly. (*B*) Spleens of the mice shown in *A*. The transgenic spleen is enlarged due to expansion of leukemic_lymphoma cells. (Figure from Costinean et al., Proc Natl Acad Sci U S A. 103:7024-7029, (2006)).

Moreover, for many genetic diseases, even if studied for a long time and for which the chromosomic region of linkage is well known, the gene- disease has not yet has discovered. For this group of genetic diseases has been hypotized a role of not conventional genes" , like miRNAs.

MiRNAs as cancer players – a balance between miRNA targets repression and miRNA expression regulation.

The classical models of tumorigenesis postulate alterations in protein coding oncogenes and tumor suppressor genes. MiRNAs are also contributors to oncogenesis, functioning as tumor suppressors, as is the case of *miR-15a* and *miR-16-1* (Cimmino et al., 2005) or *let-7* family (Johnson et al., 2007)) or as oncogenes, as is the case of *miR-155* (Volinia et al., 2006), *miR17-92* cluster (He L et al., 2005) or *miR-21* (Chan et al., 2005) (Volinia et al., 2006)).

Relatively minor variations in the levels of expression of a miRNAs or mutations that affect moderately the conformation of miRNA::mRNA pairing could have important consequences for the cell because of the large number of targets of each miRNA. “Traditional teaching” suggests that miRNA binds to target messenger RNA by imperfect complementarity, causing either mRNA degradation, or translation inhibition (Mathonnet et al. 2007). Recently, a deviation from the above point of view on miRNA function was found: the *miR-369-3* can up-regulate translation tumor necrosis factor alpha (TNFa) after binding the 3’ untranslated region of TNFa, suggesting an additional level of complexity on miRNA function (Vasudevan et al., 2007).

A growing list of publications proved that miRNAs play a critical role in cancer initiation and progression, and that miRNA alterations are ubiquitous in human cancers. Consequently, events activating or inactivating miRNAs were viewed to cooperate with protein coding genes (PCGs) abnormalities in human tumorigenesis (Calin and Croce, 2006). For example, recently it was shown by Nagel and colleagues that *miR-135a* and *miR-135b* directly target the 3' untranslated region of APC, suppress its expression, and induce downstream Wnt pathway activity (Nagel et al, 2008). Inactivation of the adenomatous polyposis coli (APC) gene is a major initiating event in colorectal tumorigenesis. Thus, these results uncover a miRNA-mediated mechanism for the control of APC expression and Wnt pathway activity, and suggest its contribution to colorectal cancer pathogenesis.

Much less was known about the upstream regulation of miRNA in cancer cells until recently, when a series of publications demonstrated that the TP53 tumor suppressor regulates the transcription of the *miR-34* family (for a review see He X et al., 2007), and that the *miR-34* family subsequently mediates induction of apoptosis, cell cycle arrest, and senescence. Using quantitative RT-PCR analysis, it was demonstrated that *miR-34a* was highly up-regulated in a human colon cancer cell line, HCT 116, treated with a DNA-damaging agent, adriamycin (Tazawa et al., 2007). Furthermore, it was shown that widespread miRNA repression by Myc contributes to tumorigenesis in general (Chang et al., 2008), and to repression of the *miR17-92* cluster in particular. MiRNAs from this cluster modulate tumor formation and function as oncogenes by influencing the translation of E2F1 mRNA (O’Donnell et al., 2005).

MicroRNAs and Colorectal Cancer

Cancer is a complex genetic disease caused by the accumulation of mutations, which lead to deregulation of gene expression and uncontrolled cell proliferation. Given the wide impact of miRNAs on gene expression, it is not surprising that a number of miRNAs have been implicated in cancer (Bueno et al., 2008). CRC accounts for 13% of all cancers and is the second most common cause of cancer death in the Western world (Aaltonen and Hamilton, 2000, Greenlee and et. 2001, Parkin et al, 2005). Early detection provides a significant survival advantage, and many efforts are focused on improving detection rates and screening utilization. Currently, surgery is the only curative approach for early stage adenocarcinomas, with chemotherapy providing a modest incremental survival benefit at the cost of additional toxicities (Rodriguez-Bigas et al., 2006)(Kopetz et al,, 2008). Therefore, the identification of improved diagnostic and prognostic markers as well as new therapeutic options for CRC patients is of great and immediate interest.

Cancer-associated genomic regions (CAGRs) and noncoding RNAs

MiRNAs and UCRs are frequently located at fragile sites and genomic regions affected in various cancers, named cancer-associated genomic regions (CAGRs). Bioinformatics studies are emerging as important tools to identify associations and/or correlations between miRNAs/ncRNAs and CAGRs. ncRNA profiling has allowed the identification of specific signatures associated with diagnosis, prognosis, and response to treatment of human tumors. Several abnormalities could contribute to the alteration of miRNA expression profiles in each kind of tumor and in each kind of tissue. Here we focused on the miRNAs and ncRNAs as genes affecting cancer risk, and we provided an updated catalog of miRNAs and UCRs located at fragile sites or at cancer susceptibility loci. These types of studies are the first step toward discoveries leading to novel approaches for cancer therapies.

Noncoding RNAs and bioinformatics

Recent biotechnology advances, along with a growing number of new biological-computational approaches, have allowed an expansion of the number of genomes being sequenced and annotated, as well as facilitated the development of databases to collect and analyze large amounts of genetic information. The consequent necessities of retrieving, sharing, and, in particular, understanding this vast amount of data led to the creation of genome databases, providing an open source of genetic information for scientists worldwide. Thus, there is now a strong urgency to integrate various sources of biomedical and clinical information.

One of the many fields that will strongly benefit from such integration is the study of noncoding RNAs (ncRNAs) (Barbarotto et al. *2008*; Calin and Croce *2006*; Esquela-Kerscher and Slack *2006*). The most studied ncRNAs are the miRNAs. Recently, the physiologic role of miRNAs during development, differentiation, cell cycle regulation, aging, and metabolism has begun to be elucidated (Ambros *2004*; Costa *2005*; Johnson et al. *2007*; Mendell *2005*). Consequently, miRNA deregulation has been found in many different human diseases, including cancer, diabetes, and immuno- or neurodegenerative disorders (Perwez Hussain and Harris *2007*; Sevignani et al. *2006*). Several lines of evidence implicate abnormalities in miRNA and ncRNA gene expression with cancer, such as (1) the location of ncRNAs at CAGRs, (2) the epigenetic regulation of miRNA expression, and (3) abnormalities in miRNA processing genes and proteins. A unique and specific miRNA signature is able to distinguish between different normal tissues, and more interestingly characterizes different types of cancers (Calin and Croce *2007*; Croce

2008; Dickins et al. *2005*; He H et al. *2007*; Scott et al. *2006*). The tumor specificity of this signature harbors diagnostic, prognostic, and therapeutic implications (Calin and Croce *2006*). Recently, a new class of ncRNAs, called the ultraconserved regions (UCRs), has been implicated in human tumorigenesis (Calin et al. *2007*). UCRs are a subset of conserved sequences that are strictly conserved among orthologous regions of the human, rat, and mouse genomes and, because of their noncoding nature, have been considered for a long time as the "dark matter" of the human genome (Bejerano et al. *2004*). The identification of regulatory functions of miRNAs on cancer-associated UCRs opens new investigational scenarios with theoretical clinical implications (Calin et al. *2007*). Another area highly influenced by the link between ncRNAs and bioinformatics is the study of cancer risk. The identification of genes responsible for cancer predisposition is a crucial requirement for better diagnosis, but this process usually involves costly, time–consuming, and difficult tracing of large family lineages to define the chromosomal region where genes responsible for the disease are located. Computational technologies can appropriately be employed to integrate available data and can, in principle, be used to save on the expensive process of linkage analysis. This approach has been used to study miRNAs by integrating existing information, analyzing data, and verifying biological results. For example, bioinformatics approaches allow the geneticist to determine that miRNAs and/or UCRs are located in fragile sites (Calin et al. *2004*), in regions involved in cancers (Makunin et al. *2007*), or in tumor susceptibility loci (Sevignani et al. *2007*).

Human miRNA genes are frequently located at genomic loci involved in cancer

The first genome-wide link between miRNAs and cancer involved a search for correlations between the genomic positions of miRNAs and chromosomal regions that exhibited specific cancer-associated abnormalities; this study showed that about half of human miRNA genes are located at fragile sites (FRAs) and various types of cancer-associated genomic regions (CAGRs) (Calin et al. *2004*). This bioinformatics analysis was performed by establishing a new database that combined the positions of FRAs and CAGRs with the positions of known and predicted miRNAs across the human genome. When this study was performed (Calin et al. *2004*), no powerful bioinformatics tools for finding miRNA targets were available, so the genomic proximity approach served as a useful screen for such interactions. **Table A1** (see Appendix A) presents a complete list of miRNAs and UCRs, according to the most recent versions of databases, which are located at human FRAs. **Table A2** (see Appendix A) shows a list of databases, tools, resources, bioinformatics, and statistical analyses used in the studies presented in this review. **Table A3** (see Appendix A) presents a glossary of bioinformatics terms.

The effects of altering the expression of target genes that participate in apoptosis, cell cycle, invasion, or angiogenesis, cause the initiation, growth, and/or progression of tumors (**Figure 14**). Genetic alterations of miRNAs may therefore lead to changes in protein expression in cancer cells, which accelerate malignant transformation and/or enhance tumor growth (Calin et al. *2004*; Croce and Calin *2005*; Gregory and Shiekhattar *2005*; McManus *2003*).

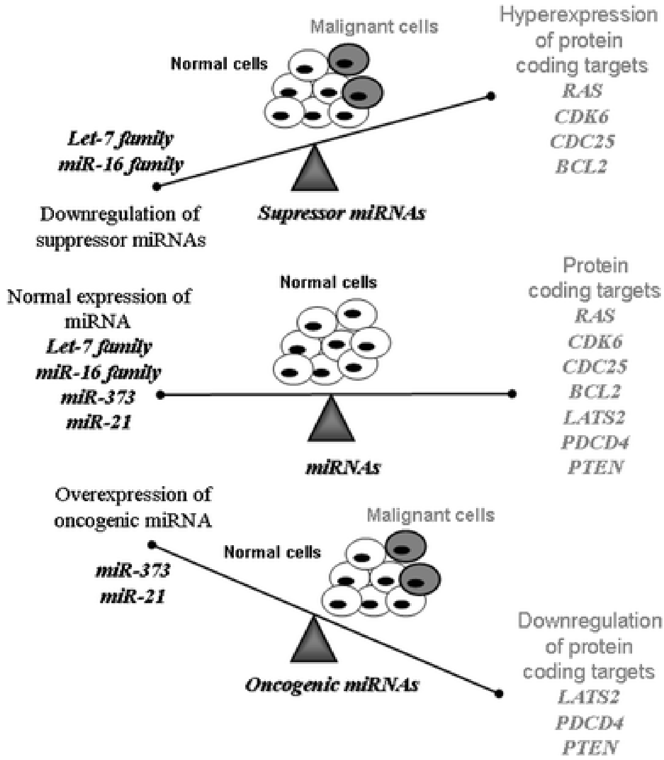

Figure. 14 An example of miRNAs as oncogenes or tumor suppressor genes and their protein-coding targets in tumorigenesis. The figure shows two examples of how miRNAs affect cancer development. The *let-7* and *miR-16* families are the best studied examples of suppressor miRNAs, while *miR-21* and *miR-373* are examples of oncogenic miRNAs. Targeting such miRNAs located in CAGR could be a potential future therapeutic option for cancer patients. The significant confirmed targets are included on the right side (Figure from paper VIII)

THESIS OBJECTIVES

The aim of the thesis was to answer the question whether coding genes and non-coding RNAs expression data can be used for the targeted extraction of a shortlist of candidate genes, thus saving resources for the following costly and time-consuming genetic analysis. Furthermore we wanted to investigate whether coding genes can be candidates for cancer and hereditary diseases developing new algorithms and tools using several statistical measures and public databases data and information. Several results were biologically confirmed or well known in literature. Then, we decided to apply our methods to study miRNAs in cancer; specifically to prove that expression data analysis may serve as markers of early and late stages in cancer progression. We expect our results to show evidences which support the use of miRNA data for cancer development analysis, not only by distinguishing classes of patients, but also by identifying markers of early and late cancer progression stages.

Therefore, the aims of the thesis were the followings:

i. Creation of a web-based tool to provide candidate disease genes
ii. Implementation of a tool to detect functional differences between tissues
iii. Development of an algorithm for the identification of associations between OMIM diseases and microRNAs
iv. Statistical validation and support to assess the quality of DNA microarrays, correlation between microRNAs and ultraconserved genes expression, microRNAs and ultraconserved genes involved in cancer metastasis.

In the following subsections the objectives of this projects are described.

MATERIALS AND METHODS

GebbaLab

Microarray data infrastructure using GebbaLab based on Alfresco technology

The efficient storage and analysis of microarray data is of considerable interest and there is much activity worldwide. In general most researchers adopt a “single workstation approach" for data management and analysing expression data. However this method is rapidly becoming inconvenient for many reasons:

- There is no provision for the systematic recording of experimental information
- Current PCs are not sufficiently powerful for analysing data
- Comparison with data from other researchers or public repositories is difficult
    - Careful consideration of these points has suggested the following criteria for the design of the microarray infrastructure.

- Users must be given the opportunity to use a wide range of common and user-friendly tools for data entry and for the different platforms available, e.g. Affymetrix, Agilent, Illumina etc.
- Data should be distributed
- Data must be recorded in a format which allows interoperability of all the data sources
- User-friendly portals or clients are required to access resources and powerful computational facilities to process datasets

To satisfy these criteria the infrastructure was structured into two distinct levels:

i. The data entry and storage level.
ii. The application level for running analysis applications.

The system consists of a “central” node and many “satellite” nodes, each of which with its own data store, potentially virtualized. The system has been designed in a modular way in order to work even in case of unavailability of the central node. In fact in our schema “central" merely indicates a central registry for distributed indexing and querying. Data is stored, analyzed and exchanged through a complex architecture build upon Alfresco, an advanced Open Source Enterprise Content Management that provides a common web interface and access to distributed data sources. Alfresco also includes user authentication and various levels of access privileges, thus allowing many degrees of data security and privacy. Our effort have been lead to build upon the Alfresco structure many software modules in order to manipulate MAGE-ML files, extract metadata from MAGE-MLs, to store and index metadata into the repository for querying microarray data according to different search criteria. All the modules are provided through a SOA (Service Oriented Architecture) layer among different web applications that allow users to choose, through a web portal, the appropriate software from those available that transparently invokes algorithms to fetch the data for analysis. High performance servers are available for CPU or memory intensive calculations.

TOM

The input data to TOM can be constituted by a (list of) gene(s) and one or two chromosomal areas of interest. In the One Locus option, while the chromosomal area represents the hypothesis to test, the input gene(s) are the queries of the search. For this reason almost invariably, but not exclusively, the seed of the search will belong to the repository of the Online Mendelian Inheritance in Man (OMIM) database. This repository stores a comprehensive collection of genes known to be related to human diseases. OMIM is updated daily and stores information (mainly genes) related to disorders inherited in a Mendelian manner, where traits are passed from parents to children. Any type of gene can be used with this method. Nevertheless, because most of the power in our procedure is given by expression data, we named the application TOM (Transcriptome of OMIM).

The three-step filtering algorithm

We describe here the detailed steps allowing the selection of the final candidate gene sets for an hereditary disease (see **Figure 15**).

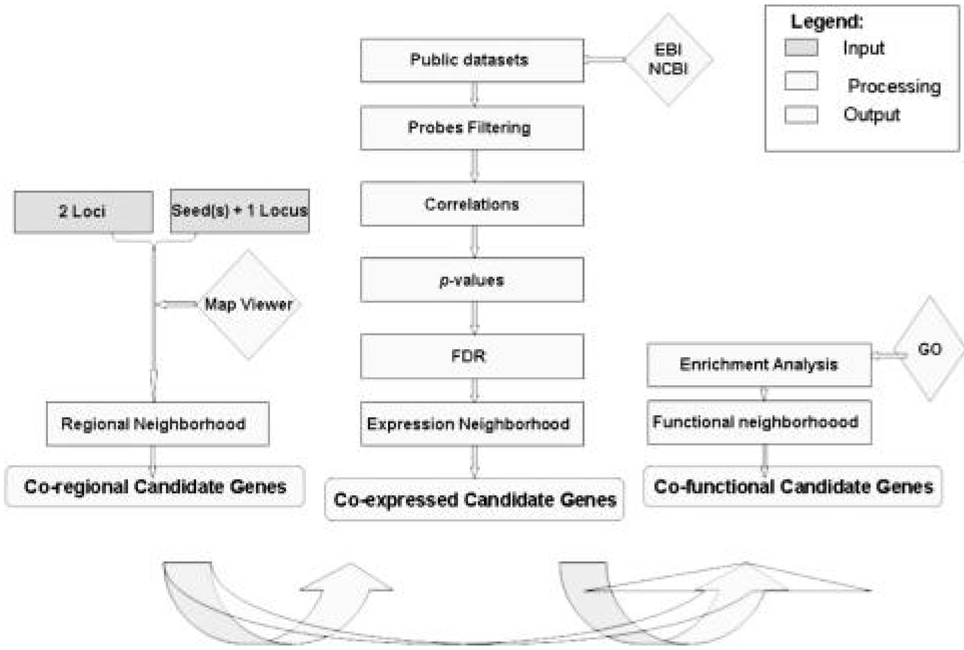

Figure 15. Global description of the process. The three steps of the algorithm, along with the databases and the intermediate and final results are shown in the figure. The output can be used at the end of the second step, in the form of co-expressed genes, or refined through the third step where the functional analysis (based on GO) is performed. The longest arrow depicts the alternate route to the functional analysis. (Figure from paper II).

The first step is designed to select the list of genes mapped on the chromosomal area/s of interest, using genome sequence information. Then, in the second step, TOM employs transcriptome data from public repositories. TOM retains here only the genes that have related expression variations in the datasets, either among them (Two Loci) or to the seeds (One Locus). Formally, this is achieved defining the expression neighborhood, i.e. the set of genes encoded in the genomic area of interest that are related among them or to the seeds, based on the similarity of their expression. We evaluate the *P*-values of the correlation tests and select the genes whose correlation value is significant at a given value of rejection. Evaluation of *P*-values is performed assuming that the correlation values are distributed using Student's *t* cumulative distribution, with a number of degrees of freedom corresponding to the number of samples in the microarray experiment.

Given the high number of correlation tests performed in TOM, *P*-values are corrected for multiple testing by using the false detection rate (FDR), as defined by Ref. (Benjamini. and Hochberg. 1995). FDR controls a different probability than that which is controlled with the better known *P*-value. In fact, *P*-values control the number of false positive over the number of truly null tests, while FDR controls the number of false positive over the number of significant tests. Several ways of estimating this number have been proposed, we adopted the solution devised by Tom Nichols (see http://froi.sourceforge.net/documents/technical/matlab/FDR.html), that rescales the *P*-value obtained on a single test multiplying it by a combination of indexes related to the total number of tests performed: $Kp_i / \left(i \sum_{i=1}^{K} i^{-1} \right)$, where p_i represents the *i*-th of the total *K* single *P*-values. Correction was performed on a seed by seed basis, this means that the genes in the seeds list or in the first chromosomal area are considered independent tests.

Finally, in the third optional step, we further filter candidates, based on their functional role(s). To this end we use GO [the more widespread controlled and hierarchically structured vocabulary for the description of genes and genes' products characteristics in any organism (The Gene Ontology Consortium, 2001)] to better interpret the genes selected in the expression neighborhood, and extract genes related to the same biological process(es). In particular, we use the hypergeometric distribution

$$p = 1 - \sum_{i=1}^{r-1} \frac{\binom{n_1}{i} \cdot \binom{N - n_1}{n \quad i}}{\binom{N}{n}},$$

to evaluate the probability that in a sample of size *n*, *r* items of a given type—a type characterizing n_1 items in a population of *N* items—can be selected without replacement (Rosner, 2000). This probability statistically validates the proportion of genes in the group of candidate genes (enriched for some known function), compared with what would be expected by chance alone. The third step can alternatively be performed directly after the first step, simply by skipping the transcriptional analysis. This statistically validated triple filtering allows the targeted extraction of a shortlist of candidate genes, thus saving resources for the following costly and time-consuming genetic analysis.

Implementation

The tool core is developed under R and the user interface is developed using Php. Users requests are initially stored in a database (MySql), where a batch scheduled task retrieves and processes them, while the user interface is waiting. For defining the position of the bands, we use the NCBI Map- Viewer (http://www.ncbi.nlm.nih.gov/mapview/). The BUILD.35.1 genome data are stored in the TOM database.

The stored microarrays can then be searched to retrieve the expression values for the correlation analysis by checking the Expression Correlation filtering check box and setting the FDR threshold field. Microarray expression values undergo a double filtering process on the basis of the calls (flexible filtering) and of a fold-change and absolute filter variation over samples (max/min < 3 and max _ min < 100, fixed filtering). This double filtering is meant to allow a stringent and constant selection based on the variation of the expression profile, while preserving the maximum amount of information, based on the general quality of the array. The quality of the array is related to the number of present calls available. For this reason, we filtered the array spots based on the assumption that a minimum amount of information can be extracted from the arrays. Experiments conducted on Affymetrix chips (Human Genome U133A, U95A, U133A and B and U1333plus2 chips with detection calls available) were downloaded from the repository of Gene Expression Ominbus (GEO, http://www.ncbi.nlm.nih.gov/geo/) and EBI repository ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) using a Perl script to capture the expression values. We performed these results loading 40% of the human microarray experiments with detection calls available. The functional analysis based on GO identifies distribution and can be performed by checking the Validate GO box. We applied the Biobase and GOstats bioconductor packages (Gentleman et al., 2004) to perform the functional analysis.

Fun&Co

This section describes all the steps used for identifying functional differences between different tissues by analyzing a pair of dataset groups, e.g. muscle and heart.

Dataset selection

In this first step, the users have to choose between the available GEO datasets the ones to use in the Fun&Co analysis. It is also possible to upload users' owned datasets.

Gene selection

Given the two dataset groups, the users have to specify a GO term (named GOmain). Fun&Co uses GOmain to filter the probes contained in the datasets, keeping only the probes annotated with such GO term or a more specific one (probe annotations are taken from the Affymetrix support library).

Correlation study

After the gene selection described in the previous section, Fun&Co studies any possible pair of probes in the dataset, valuing the correlation coefficient of each pair (for measuring the degree of relationship between two variables, i.e. two probes). Due to the nature of the Affymetrix microarrays, it is preferable to use a non-parametric correlation coefficient. For this reason, our tool uses Spearman's correlation.

False detection rate

Given the high number of correlation tests performed, *P*-values can be corrected for multiple testing by using the FDR (Benjamini and Hochberg, 1995). FDR controls a different probability than the one controlled with the better known *P*-value. In fact, *P*-values control the number of false positive out of the number of truly null tests, while FDR controls the number of false positive over the number of significant tests.

Correlation filtering

Users can decide when a probe pair correlation is significant or not by choosing a p-threshold (default is 0.05): if the *P*-value is less than the p-threshold, then the probes are considered correlated.

Group comparison based on correlation

The aim of the group comparison was to identify the GO terms that showed a significant difference between the two groups. This was performed in three steps. In the first step, for each dataset, Fun&Co studied the correlations as described. In the second step, for each dataset, the system counted the number of correlated pairs of probesets associated to the same GO term. In the last step, for each GO term, the system compared the number of correlated probe pairs detected in the two groups under study.

Comparison between two datasets

Given two gene expression datasets, named *A* and *B*, for each GO term under analysis, Fun&Co computed the number of correlated probe pairs found in each dataset [named *N*_couples(*A*)) and *N*_couples(*B*)]. At this point, the system discarded all the GO terms that had too few probe pairs in both datasets: a GO term was discarded if its pair count is below a user-defined threshold, named $Fitness_{th}$, in both datasets. This filtering was performed in order to avoid results that were not supported by a minimum number of found correlations. For each GO term not discarded by $Fitness_{th}$, Fun&Co identified if this term in the *A* dataset was over-correlated (under-correlated) with respect to the *B* dataset. In order to identify if a GO term was over-correlated or under-correlated, Fun&Co computed the *logratio* measure: $\log_2\left(\frac{N_{\text{couples}}(A)+1}{N_{\text{couples}}(B)+1}\right)$. Then it normalized this value, subtracting to it the mean, computed among all the GO terms. By introducing a threshold

(t), the GO terms over-correlated (under-correlated) in the D_a samples with respect to the D_b samples were the ones with $logratio - mean > t$($logratio - mean < -t$).

Dataset group comparison

Fun&Co may search for functional differences between two groups of datasets (A1, ..., A_n and B_1, ..., B_m) rather than simply considering two datasets. This feature requires an extension of the analysis approach described.

Fun&Co performed the extended analysis, combining the results of n x m comparisons (each dataset A_i of the first group compared with each dataset B_j from the second group). The results of these comparisons were collected in a table named 'continuous comparison', that had n x m columns (where n is the number of datasets in the first group and m is the number of datasets in the second group).
As described before, Fun&Co considered only the GO terms with at least $Fitness_{th}$ correlated probe pairs in at least one dataset. After computing the normalized logratio, it identified the over-correlated GO terms.
In order to provide synthetic results, Fun&Co built a 'consensus list' table, that includes all the GO terms over-correlated for at least 50% of comparisons. For each GO term, it provided also the group in which this term was over-correlated and the mean of its normalized logratio in all performed comparisons.

Implementation

The tool was structured as a web server which manages Fun&Co computations (also named job) requested by the users. The web interface was developed using JSP and the computation core was developed under JAVA. The user interface allows user registration, job submission and results retrieval. Users accounts and job requests werestored in a database (MySql).

Hypersolutes

Compatible solutes from hyperthermophiles improve the quality of DNA microarrays

Different series of transcriptome analysis using constant human RNAs and variable concentrations of hypersolutes were performed. Total RNA from HEK 293 cells was extracted by using NucleoSpin® RNA II Kit (Macherey- Nagel, Düren, Germany). Different batches of RNA were pooled together after quality assessment by spectrophotometric analysis supported by gel electrophoresis and Agilent Bioanalyzer™

(Palo Alto, CA, USA). The RNA Integrity Numbers (RINs) from the Bioanalyzer™ reports were all between 9.5 and 10.0. Compatible solutes were from BITOP (Witten, Germany). All operations were carried out according to the standard Affymetrix protocol, with the sole exception of adding compatible solutes to the hybridization buffer. The fragmented cRNA targets were hybridized onto Affymetrix GeneChip® Test3 Arrays (Santa Clara, CA, USA). The samples for hybridization were prepared by adding the hypersolute to the fragmented cRNAs in DEPC water. PolyA spike-ins were not used. Arrays were scanned by using the Affymetrix GeneChip® 3000 scanner. The CEL files were analyzed using the Affymetrix GeneChip® Operating Software, and standard array quality parameters such as raw Q, background, scaling factor and percent present calls [all defined in the Affymetrix GeneChip® Expression Analysis Technical Manual (Affymetrix Technical Documentation)], were measured. T-test was used to compare means for independent samples. In addition to the standard Affymetrix quality parameters listed above, we needed additional statistical measures to test chips quality and to evaluate and validate the results. Therefore, we used the Bioconductor package affyPLM (http://bioconductor.org/packages/1.9/bioc/vignettes/affyPLM/inst/doc/QualityAssess.pdf). This package performs quality Affymetrix array tests by a variety of procedures, such as pseudo images, standard error evaluation and relative log expression. Chip pseudo-images are very useful for detecting potential quality problems. For each hybridization we produced a pseudo-image, where areas of low quality were green and those of high quality were light grey. Another quality parameter we used was the normalized unscaled standard errors (NUSE). The estimated standard error obtained for each gene on each array from fitPLM was standardized across arrays so that the median standard error for that gene was 1. NUSE statistics (NUSE median and interquartile range IQR) were computed for each array. The relative log expression (RLE) was also studied. The RLE values were calculated for each probe-set by comparing the expression value on each array against the median expression value for that probe-set across all arrays. The RLE statistics (RLE median and IQR) were computed for each array. After computing NUSE and RLE statistics for each array, the results were resumed by M = (median+2*IQR); median represents a measure of central location of the data and IQR (inter-quantile range) is defined as the difference between the 75th percentile and the 25th percentile (i.e. the upper and the lower quantiles). M was used to identify confidence limits to evaluate RLEs and NUSEs. The mean of M measures was calculated for each group of replicates, and finally M means were normalized by the control mean. PM (perfect match) and MM (mismatch) were calculated by using PM and MM affy Bioconductor package functions (http://bioconductor.org/packages/1.9/bioc/vignettes/affy/inst/doc/affy.pdf). The PM/MM based quality comparisons were performed by calculating the percentage of PM larger than MM in each array.

MicroRNA DNA methylation

A microRNA DNA methylation signature for human cancer metastasis

SW620, 11B and IGR37 cell lines treated and associated controls were analyzed using microarray technology (triplicates for each sample were performed). The microarray dataset was normalized by quantile method (http://rss.acs.unt.edu/Rdoc/library/affy/html/normalize.quantiles.html), and a microRNA was excluded if less than 20% of expression data values had at least 1.5 fold change either directions higher or lower as the microRNA's median value.

Principal component analysis (PCA, Partek Genomic Suite®™) was performed to classify samples. PCA is a statistical method for exploring and making sense of datasets with a large number of measurements by reducing them to the few principal components (PCs that explain the main patterns) (Reich et al., 2008).

Unsupervised principal component analysis of normalized and filtered dataset made with cell treated lines and controls (**Figure 16**), was performed using all survived genes and showed clear separation between various classes (SW620, 11B, IGR37 cell lines treated and controls).

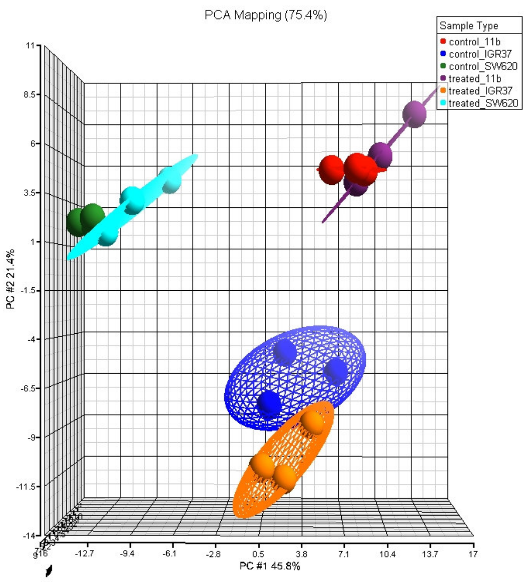

Figure 16, Unsupervised principal component analysis shows differential expression between 11B (red and purple), IGR37 (blue and orange) and SW620 (green and turquoise) cell lines, both treated and controls cell lines.

RESULTS

GebbaLab Project

Great progress has been made in recent years in integrating technologies and innovations in computer science with those of the life sciences. However, many activities in biological and especially clinical research still do not have access to the necessary computer technology. Hospitals, for example, often perform outstanding research but lack the bioinformatics tools which could fully exploit the activities carried out. The GeBBALab (www.gebbalab.it) project is addressing these problems by creating a "virtual laboratory" with contributions from both scientific and technological/industrial partners: IOR (Istituti Ortopedici Rizzoli), DAMA (Data Mining and Analysis) University of Ferrara, CINECA (Consorzio Interunivesitario per il calcolo automatico) and NSI (NIER Soluzioni Informatiche).

The project has identified two key areas:

i. Microarray data management and analysis

ii. Integration of patient and clinical data with genomics information

We concentrate on the GebbaLab (Genetics, Biotechnology and Applied Bioinformatics) infrastructure for microarray storage and analysis (GebbaMa), project wich into we directly contributed with microarray data analysis experience.

GebbaLab is a regional project, funded by Regione Emilia Romagna finished on June 2008.

GebbaMa

We implemented a user-oriented, powerful infrastructure for microarray data management and analysis. It allows the user to enter data and being distributed avoids the limitations of a centralised server. A prototype using Alfresco is already available (GebbaMa, http://gebbama.cineca.it/, **Figure 17**) and microarray researchers are invited to contact the authors if they wish to experiment with the system. It has been developed as a central node –CINECA- and two satellite nodes –DAMA, University of Ferrara and CINECA develop-.Future enhancements to Gebbalab will include analysis applications and, crucially, the possibility of integrating patient data.

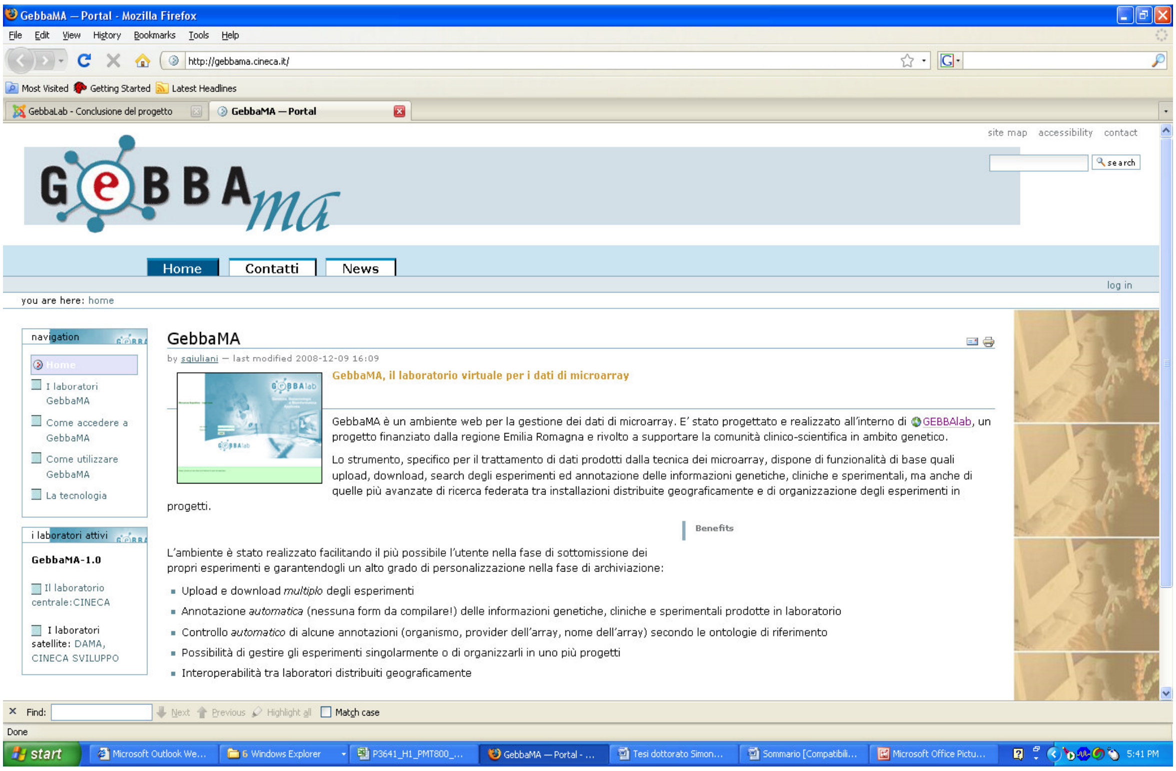

Figure 17. GebbaMa Web Interface. (Figure from www.gebbama.cineca.it).

TOM Project

We implemented TOM, a web-based resource for the efficient extraction of candidate genes for hereditary diseases, based on the principle that the massive production of biological data by highly parallel devices like microarrays paved the possibility of inferring biological answers by querying large amounts of expression data. The service requires the previous knowledge of at least another gene responsible for the disease and the linkage area, or else of two disease associated genetic intervals. The algorithm uses the information stored in public resources, including mapping, expression and functional databases. Given the queries, TOM will select and list one or more candidate genes. This approach allows the geneticist to bypass the costly and time consuming tracing of genetic markers through entire families and might improve the chance of identifying disease genes, particularly for rare diseases. We presented the tool and the results obtained on known benchmark and on hereditary predisposition to familial thyroid cancer. Our algorithm is available at **http://www-micrel.deis.unibo.it/~tom/.**

We present here some examples of the use of TOM for One Locus and Two Loci option. The following section presents results that show the ability of TOM to reproduce known genetic information (validation). We present three carefully documented benchmark tests whose results are summarized in the table of **Figure 18a**, and then broaden the validation with the analysis of five more examples. Global results are summarized in the rank distribution of **Figure 18b** that shows how the expected results rank in majority of the candidate genes list extracted with TOM. The Discovery section shows the results obtained on a Two Loci problem to gain further insight into a poorly characterized disease, namely familial thyroid cancer (discovery).

Validation

TOM was tested by searching for several genes known to interact with each other using the One Locus option of TOM. The aim of this approach was to ensure that the system correctly identifies gene–gene interplay. The examples used are reported in **Figure 18a**. Each gene was used as seed against the chromosomal region where the known interacting gene maps (ENSEMBL v.35), and vice versa. The examples considered for this first run of One Locus option were PKD1 and PKD2, T0MM70A and TIMM17A, ANNXA11 and PP1F. PKD1 and PKD2 are genes mutated in polycystic kidney disease. In a majority of cases, the gene involved is PKD1, which is located on chromosome 16 (16q13.3) and encodes polycystin-1, a large receptor-like integral membrane protein. In the remaining (10–15%) cases, the disease is caused by mutational changes in another gene (PKD2), which is locatedat chromosome 4 (4q21–23) and encodes polcystin-2, a transmembrane protein, which acts as a non-specific calciumpermeable channel. Both polycystins function together in a non-redundant fashion, through a common pathway, and produce cellular responses that regulate proliferation, migration, differentiation and kidney morphogenesis [for a review see Ref. (Al-Bhalal and Akhtar 2005)]. TOMM70A and TIM17A are part of the mitochondrial complexes, through the outer and inner membrane respectively, for the import inside the mitochondria of nuclear-encoded proteins (Rapaport, 2005). Annexin 11 (ANXA11) is member of the annexin family, Ca2+-binding, membrane-fusogenic proteins with diverse functions. PP1F and RPS19 are two known interacting proteins with Annexin. Annexin 11 during cell cycle progression translocates from the nucleus to the spindle poles in metaphase and to the spindle midzone in anaphase (Tomas et al., 2004).

We also tested the program for complex traits such as tumor predisposition and development. Since several genes are already known to be involved in predisposition to tumor, we tested TOM for the major gene for familial breast cancer, BRCA1. Loss of function of BRCA1 caused by inherited mutation and tissue-specific somatic mutation leads to breast and ovarian cancer. Nearly all BRCA1 germline mutations involve truncation or loss of the C-terminal BRCT transcriptional activation domain, suggesting that transcriptional regulation is a critical function of the wild-type gene. Several microarray analyses have been carried out to identify a peculiar gene expression profile characteristic of carriers of BRCA1 mutations, which would have an important impact also for diagnostic purpose [for an example see (van't Veer et . al., 2002)]. It has been shown that there is a link between the role of BRCA1 in transcriptional regulation and its role in tumor suppression. Previous microarray analyses comparing transcription profiles of epithelial cells with low endogenous levels of BRCA1 versus transcription profiles of cells with 2 to 4-fold higher induced levels of expression of BRCA1 identified several genes with at least a 2-fold increase in expression, such as JAK1, a tyrosine protein kinase with a key role in cytokine signal transduction pathway (Welcsh et al., 2002). We thus tested whether by using BRCA1 as seed we could identify JAK1, giving as chromosomal location chr1p13.3, the region where JAK1 maps. The interaction was correctly identified, and also the reciprocal, i.e. JAK1 as seed and the region 17q21.31, where BRCA1 maps. We also evaluated Tuberous Sclerosis with TSC1 andTSC2 involved genes, Fanconi Anemia with FANCA, FANCG and FANCL genes, Muscular Dystrophy with CAV3, CAPN3, TRIM32, SGCB, SGCG and DYSF genes, Myeloproliferative disorders with DTL and ZNF198, and finally the Neurotransmitter transport with NAT1 and NET1. We evaluated the correlations setting the threshold for FDR < 0.01. For these and previous results we ranked the candidate genes by correlation values (preserving the absolute values) using TOM automatic sorting of candidates genes based on correlation or corrected P-values. The ranking distribution is shown in **Figure 18b.**

(a)

Seed	Probe_ID	Chromosomal region	Known gene in region	Probe_ID	Array experiment	Correlation coefficient
PKD2	203688_at	16p13.3	PKD1	202328s_at	GSE1462	0.763182
ANXA11	206200_at	10q22.3	PP1F	201489_at	GSE974	0.889076
ANXA11	206200_at	19q13.2	RPS19	202649x_at	GSE974	0.646041
TOMM70A	201519_at	1q32.1	TIM17A	2151715_at	GSE974	0.945727
BRCA1	204531s_at	1p31.3	JAK1	201648_at	GSE1364	0.700752
BRCA1	211851s_at	1p31.3	JAK1	201648_at	GSE2149	0.663158
JAK1	201648_at	17q21.31	BRCA1	204531s_at	GSE1364	0.700752

(b)

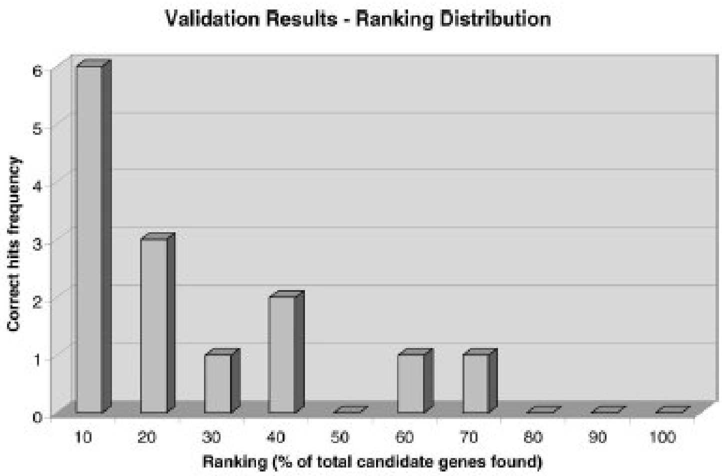

Figure 18. (a) The table summarizes the results of the first three examples. In the first four lines we record the results for One Locus problems for known interacting proteins. The last three lines show One Locus results for BRCA1-JAK1. (b) It shows a rank distribution of the genes known to be related to the eight examples discussed in Validation section, adding to the three described above five more benchmark examples, notably: Tuberous Sclerosis, Fanconi Anemia, Muscular Dystrophy, Myeloproliferative disorders and Neurotransmitter transport. The expected genes rank in majority within the first 20% of the list of candidate genes identified by TOM. (Figure from paper II).

Discovery—thyroid cancer

The TOM resource analyzes at the same time two different regions of interest and identifies the genes that are highly correlated and map to both regions. This approach proves very useful for genetic disorders in which a single gene has not yet been identified but genome scans provided regions of association on different chromosomes. We could hypothesize that genes with similar behavior might have a complementary effect on disease development. We tested our hypothesis on the familial form of nonmedullary thyroid carcinoma. Papillary thyroid carcinoma and follicular thyroid carcinoma are the most common forms of thyroid cancer accounting for between 80 and 90% of thyroid cancer patients. This disorder is associated with some of the highest familial risks among all cancer sites, with reported risks to first-degree relatives between 5- and 10-fold. Consequently, familial non-medullary thyroid cancer (fNMTC) has been recognized as a distinct clinicalentity, characterized by a higher degree of aggressiveness and mortality with respect to its sporadic counterpart (Alsanea and Clark, 2001). Transmission of susceptibility for fNMTC is compatible with an autosomal dominant mode of inheritance and incomplete penetrance. In collaboration with the International Consortium for the Genetics of fNMTC, two predisposing loci were previously mapped. The first one, TCO (Thyroid tumor with Cell Oxyphilia, MIM#603386), was mapped to the 19p13.2 region (Canzian et al., 1998) and confirmed in additional families. Oxiphilic thyroid tumors are a particular form of thyroid neoplasia, characterized by cells with mitochondrial proliferation and hyperplasia, (oxyphilic or Huerthle cells). The second locus, NMTC1 (non-medullary thyroid carcinoma1), was mapped to chr2q21 and was associated with the follicular variant of PTC (fvPTC-MIM# 606240) (McKay et. al., 2001). Evidence foran interaction between the two loci has been provided in a subset of fNMTC, and a two-locus mode of inheritance is consistent with stratification based on both the histological variants of oxyphilia and fvPTC (McKay et al., 2004). We thus performed a search using TOM to verify whether the genes mapping to the two areas of interest have any degree of correlation between them, and they might also be considered as potential candidate genes based on their functions. Among these genes, some look promising candidates for their biological function, such as UQCRFS1 on chromosome 19q, which is a mitochondrial ubiquitinol cytochrome c reductase iron–sulphur subunit and correlates with RAB3GAP on chromosome 2q, a Rab3 GTPase-protein involved in cell proliferation (correlation 0.626181; P-value < 0.001). This is very interesting since there is evidence of interaction between the two loci and the locus on chromosome 19 is associated with a mitochondrial phenotype. Thus, these genes could be considered plausible candidate genes based on position and function. Experimental studies will be needed to assess the presence of mutation/variants in affected individuals and prove an involvement in thyroid carcinoma predisposition. The advantage of using TOM here was to reduce the number of genes that can be selected for a first mutation screening, after having identified two regions of significant linkage.

Fun&Co Project

Microarray and other genome-wide technologies allow a global view of gene expression that can be used in several ways and whose potential has not been yet fully discovered. Functional insight into expression profiles is routinely obtained by using gene ontology terms associated to the cellular genes. In this article, we deal with functional data mining from expression profiles, proposing a novel approach that studies the correlations between genes and their relations to Gene Ontology (GO). We implemented this approach in a public web-based application named Fun&Co. By using Fun&Co, the user dissects in a pair-wise manner gene expression patterns and links correlated pairs to gene ontology terms. The proof of principle for our study was accomplished by dissecting molecular pathways in muscles. In particular, we identified specific cellular pathways by comparing the three different types of muscle in a pairwise fashion. In fact, we were interested in the specific molecular mechanisms regulating the cardiovascular system (cardiomyocytes and smooth muscle cells).

We applied Fun&Co to the molecular study of cardiovascular system and the identification of the specific molecular pathways in heart, skeletal and smooth muscles (using 317 microarrays) and to reveal functional differences between the three different kinds of muscle cells. Availability: Application is online at **http://tommy.unife.it**.

We used datasets produced on Affymetrix U133 platforms (U133a,b or U133plus) and available from the GEO public data repository.2 In particular, we used for heart 107 samples from GSE1145 and 70 samples from GSE2240 and for skeletal muscle 79 samples from GSE3307 and 35 from GSE4667. For the smooth muscle dataset, we merged the data from GSE1595, GSE2883 and GSE3356 in a single table, obtaining a 26 samples dataset. The goal of our model study was that of identifying specific cellular pathways by comparing the three types of muscle in a pairwise fashion. In particular, we were interested in the specific molecular mechanisms regulating the cardiovascular system (cardiomyocytes and smooth muscle cells).

Fun&Co: Approach

Fun&Co is an application for dissecting and comparing expression profiles at a functional level. We developed and applied it to the identification of molecular differences in the three skeletal, cardiac and smooth muscles. This study represents both a test model and a very important scientific and medical problem. Cardiovascular diseases represent one of the most common and serious health issues. Thus, a fine understanding of the molecular mechanisms underlying heart physiology is of great importance. We used here, the wealth of data generated by a number of laboratories on muscle transcriptomes to identify key functions and processes in human heart, and smooth cells when compared to each other and to the skeletal muscle. The rationale is that the specific functions and pathways will be of relevant medical importance. By linking the patterns of gene-pairs expression to the respective gene function (as provided by the Gene Ontology database), we can extract information to better understand genome-wide expression profiles and to help scientists in the subsequent design of focused experiments. As a proof-ofprinciples, we identified the GO terms which distinguish different tissues. In details, the functional correlations comparison aims to highlight changes in gene expression correlations, in order to identify relations involved in tissue differentiation. Merging these results with the GOannotations, we can immediately select functionally relevant biological entities associated to different tissues. We used the Spearman's correlation coefficient (Rosner, 1995) (described in 'Methods' section) to evaluate the correlation between the mRNA levels of all possible gene pairs. Finally, we linked these results with the GO terms and selected the significant functional differences. The approach, shown in **Figure 19**, consists of four steps: (i) gene selection; (ii) correlation computation; (iii) dataset comparison and (iv) result synthesis. These steps are described in details in the 'Methods' section.

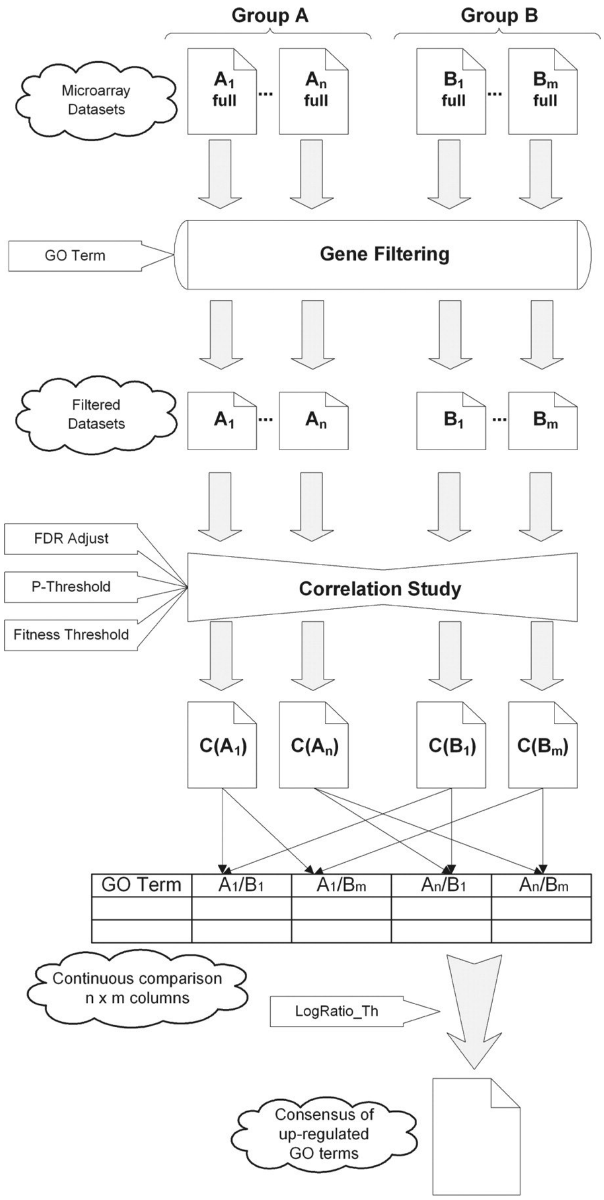

Figure 19. Process structure of Fun&Co, from the top, we can see the process steps: gene filtering, correlation computation, dataset comparison and result synthesis. (Figure from paper IV).

Our study was aimed to identify specific cellular pathways by comparing the three different muscles in a pairwise fashion. So we performed three pairwise comparisons between the datasets corresponding to the three tissues. Some GO terms related to general functions, processes and components were chosen, to avoid exploring all the possible and not relevant GO terms for the muscular tissues (see **Table 3**). In particular, we investigated the response to stimuli and extracellular environments.

Biological process	
Intracellular signaling cascade	GO:0007242
Cell surface receptor linked signal transduction	GO:0007166
Regulation of signal transduction	GO:0009966
Molecular function	
Kinase activity	GO:0016301
peptide receptor activity	GO:0001653
G-protein coupled receptor activity	GO:0004930
Cellular component	
Extracellular matrix	GO:0031012

Table 3. GO terms investigated (and the branching children) in skeletal and cardiac, and smooth muscles. (Table from paper IV).

For each of these GO terms, we performed all three possible pairwise comparisons (heart versus skeletal, heart versus smooth and skeletal versus smooth) applying Fun&Co to the datasets presented in Section 1. We used the default P-value (0.05) with the FDR adjustment and set $Fitness_{th}$ = 5 and $LogRatio_{th}$ = 1.. We obtained three consensus lists (one for each comparison). The number of significant terms in the consensus lists is shown in **Figure 20.**

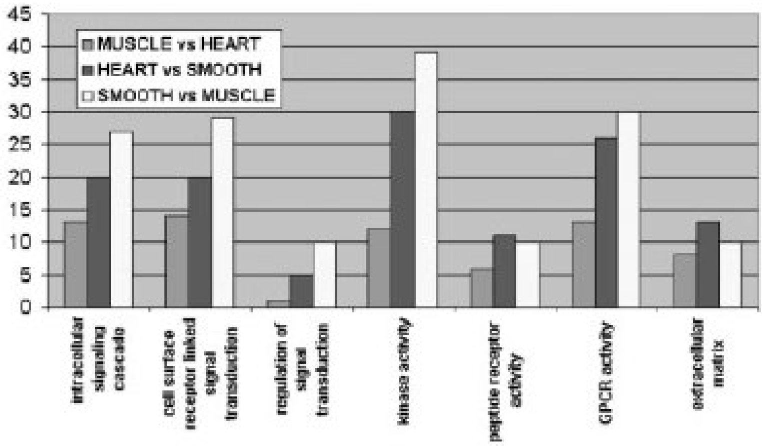

Figure 20. Number of terms found in the consensus lists for each pairwise (Figure from paper IV).

In order to assess the significance of the terms included in the consensus lists, we performed a bootstrap test. We randomly re-assigned the association table between the probeset ids and the GO terms, and generate 1000 bootstrapped annotation tables.

Hypersolutes

Additional hybridizations with the 3 hypersolutes were performed at 10 and 25 mM and repeated at 50 and 150 mM. This plan was set up to investigate the concentration effect and to determine the most effective working concentration of hypersolutes in the hybridization buffer. Each run was carried out in quadruplicate, in addition to a control test (without hypersolute) for each series. Quality assessment by normalized unscaled standard errors (NUSE), relative log expression (RLE) and pseudo images was performed with Bioconductor package affyPLM. All chips passed the quality control (QC) and were included in the following statistical analysis.

Further QC parameters were assessed by using the Affymetrix proprietary tools. Means of raw Q, background, scaling factor and percent present calls values with their standard deviations and p values (from t-test) are reported in **Table 4**

Table 4 Hypersolutes improve DNA microarray quality parameters. (Table from paper V).

Compatible solute	**HECT**				**DGP**				**MG**			
Concentration (mM)	***10***	***25***	***50***	***150***	***10***	***25***	***50***	***150***	***10***	***25***	***50***	***150***
Mean raw Q ± SD	2.1 ± 0.05	1.8 ± 0.1	2.6 ± 0.2	2.6 ± 0.2	2.0 ± 0.1	2.0 ± 0.1	2.6 ± 0.4	3.2 ± 0.4	1.9 ± 0.1	2.0 ± 0.1	2.4 ± 0.1	2.5 ± 0.2
% raw Q s *vs* control	+3.1	-8.3	+0.8	+4.6	-0.5	+1.0	+4.2	+26.5	-5.7	-0.4	-4.0	-0.8
p (t-Test s *vs* controls)	0.15	0.04*	0.45	0.19	0.43	0.38	0.31	0.01*	0.06	0.45	0.22	0.44
Mean bkg ± SD	69.9 ± 0.9	61.0 ± 3.4	84.5 ± 9.2	89.2 ± 8.4	67.2 ± 1.9	68.9 ± 3.8	91.0± 14.1	117. ± 18.7	62.2 ± 4.4	67.6 ± 1.2	81.3 ± 8.1	86.8 ± 6.0
% bkg s *vs* control	+3.0	-10.2	-4.1	+1.3	-1.0	+1.4	+3.3	+32.8	-8.4	-0.4	-7.7	-1.4
p (t-Test s *vs* controls)	0.12	0.01*	0.30	0.43	0.35	0.35	0.37	0.02*	0.04*	0.43	0.16	0.42
Mean SF ± SD	3.9 ± 0.3	4.2 ± 0.3	2.9 ± 0.3	3.2 ± 0.2	4.1 ± 0.1	4.2 ± 0.1	3.4 ± 0.6	2.8 ± 0.4	4.1 ± 0.2	4.0 ± 0.2	2.8 ± 0.1	3.2 ± 0.4
% SF s *vs* control	-8.0	-0.1	-9.6	-0.4	-1.7	-0.7	+7.9	-11.0	-1.5	-4.0	-11.4	-1.0
p (t-Test s *vs* controls)	0.04*	0.49	0.17	0.48	0.25	0.40	0.26	0.14	0.33	0.10	0.09	0.46
Mean %P ± SD	29.8 ± 0.4	30.5 ± 1.1	30.4 ± 0.8	29.6 ± 0.7	30.0 ± 0.8	30.1 ± 1.0	30.5 ± 1.6	30.4 ± 1.1	29.7 ± 1.1	29.7 ± 0.9	31.0 ± 1.0	30.8 ± 0.5
%P s *vs* control	+3.0	+5.3	+1.3	-1.1	+3.6	+3.9	+1.9	+1.3	+2.7	+2.5	+3.6	+2.9
p (t-Test s *vs* controls)	0.08	0.04*	0.30	0.33	0.08	0.08	0.29	0.32	0.16	0.17	0.11	0.12

Affymetrix quality control parameters; raw Q, background (bkg), scaling factor (SF) and percent present calls (%P): mean and standard deviation (SD) among replicates; % gain of solute(s) respect to the controls, with relative p values. HECT: hydroxyectoine; DGP: potassium diglycerol-phosphate; MG: potassium mannosylglycerate. * p-value < 0.05.

The graphs of the normalized unscaled standard errors (NUSE) and the relative log expression (RLE) are displayed as bar charts in **Figure 21** and **22**, respectively. In both cases, the values were normalized on the corresponding controls (the value 1 on the Y axis means 100% of the untreated control), as described in the Methods section. Poor quality chips have normalized NUSEs and RLEs higher than 1 (control value), while high quality chips have normalized NUSEs and RLEs lower than 1. NUSEs and RLEs for almost all 10 and 25 mM compatible solute concentrations were lower than 1, indicating improved arrays quality. Only in two cases, 10 mM DGP and HECT, the NUSEs were slightly higher than controls. On the other hand, the error indexes for the 50 and 150 mM solute concentrations were higher than the controls. The values reported in Figure 2 and 3 were referred to experiments run at the same site.

The hybridizations were performed on Affymetrix Gene- Chip Test3 arrays. These chips are commonly used for the assessment of target quality and contain probes representing a subset of genes from different organisms. The fragmented cRNA used in our assays hybridized to the human and the highly conserved probes. The results reported above were obtained by analyzing the whole array. In order to exclude solutes-induced cross-hybridization, we also measured PMs and MMs only for human probes. The mean values of PM > MM confirmed the higher quality of hybridizations with10 mM DGP, 25 mM HECT and 10 and 25 mM MG (**Figure 23**). Notice that i.e. 0.80 means 80% of PM larger than MM, according to the affy Bioconductor package (http://bioconductor.org/packages/1.9/bioc/vignettes/affy/inst/doc/affy.pdf).

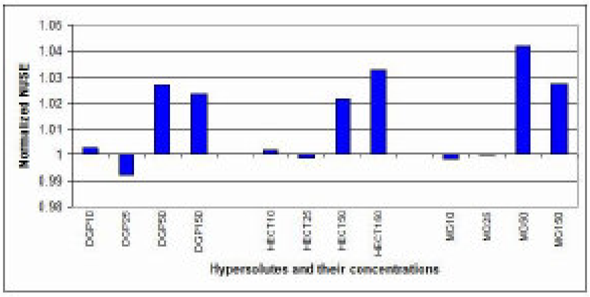

Figure 21. Normalized unscaled standard errors (NUSE). NUSE values were normalized on the controls (1 = 100% = untreated control). DGP25 (25 mM DGP), HECT25 (25 mM HECT) and MG10 (10 mM MG) arrays showed improved NUSE with respect to the control arrays. (Figure from paper V).

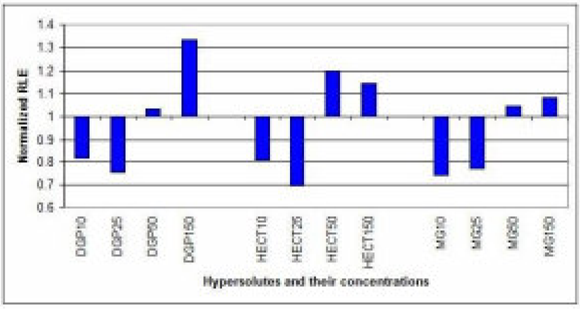

Figure 22. Relative log expression (RLE). RLE values were normalized on the controls (1 = 100% = untreated control). Notice that DGP10 (10 mM DGP), DGP25 (25 mM DGP), HECT10 (10 mM HECT), HECT25 (25 mM HECT), MG10 (10 mM HECT) and MG25 (25 mM MG) arrays displayed improved RLE with respect to the controls. (Figure from paper V).

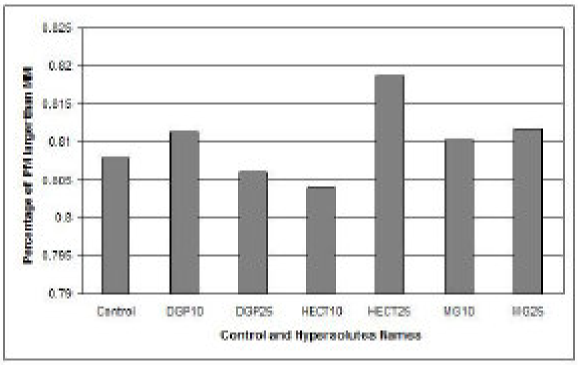

Figure 23. Percentage of PM > MM. Mean percentage of PM > MM for each hypersolute (as defined in the Bioconductor package). DGP10 (10 mM DGP), HECT25 (25 mM HECT), MG10 (10 mM MG) and MG25 (25 mM MG) showed higher percentage than control (i.e. 0.80 = 80% of PMs larger than MMs). (Figure from paper V).

OMiR

Loci for many genetic diseases have long been mapped on the human genome, but efforts by researchers to locate causative genes for a particular group of hereditary diseases in linkage areas have thus far been unsuccessful. We therefore developed and applied a novel approach called OMiR [Online Mendelian Inheritance in Man (OMIM) with microRNA (miRNA) associations] to test if miRNAs, which are not usually included in studies of candidate genes, were associated with these “orphan” Mendelian diseases. We used OMiR’s “location-comparison” approach to explore all OMIM identification number and miRNA pairs, chromosome by chromosome, and identify miRNAs that could be responsible for previously mapped human genetic diseases that have not yet been associated with a gene. We found that some loci for these genetic diseases were close to miRNAs more frequently than were some loci for genetic diseases with a known responsible gene, suggesting that miRNAs could be the genes responsible for those particular diseases. Furthermore, we found specific miRNAs associated with loci for cancer, diabetes and congenital diseases. Our results may improve the ability of geneticists to identify disease genes by including miRNAs, as candidates, particularly for rare diseases.

We used OMiR to obtain associations between 'orphan' OMIM IDs and miRNAs. **Table 5** shows the significant associations between OMIM IDs and the microRNAs clusters using 500 bootstrap cycles as validation method. Among the miRNAs identified in % OMIM loci, miR-210 was cloned from zebrafish and predicted by computational methods (Lim et al. 2003). It is among the most overexpressed miRNAs under hypoxia (Kulshreshtha et al., 2006) and is also overexpressed in solid cancers (Volinia et al., 2006). The sequence of human *miR-185* was predicted on the basis of homology to a verified mouse miRNA (Lagos-Quintana et al., 2003). The mature sequence of *miR-483* represents the form most commonly cloned in large-scale cloning studies (Landgraf et al., 2007) *mir-130b* is the predicted human homolog of mouse *miR-130b* cloned from mouse embryonic cells (Houbaviy et al., 2003, Weber 2005). The expression of *miR-210*, *miR-185* and *miR-130b* were verified in human BC-1 cells (Cai et al., 2005) and were also found to be related to cancer (Lui et al., 2007).

Table 5. Association between DELAY-class OMIM IDs and miRNAs. (Table from paper X).

OMIM ID and disease name	Gene map locus	miRNA Cluster name*
%125852 diabetes mellitus, insulin-dependent, 2	11p15.5	mir-210; mir-483
%194071 multiple tumor-associated chromosome region 1; MTACR1	11p15.5	mir-210; mir-483
%145410 hypertelorism with esophageal abnormality and hypospadias	22q11.2	mir-648; mir-185; mir-649;mir-130b; mir-650
%167870_1 panic disorder 1; PAND1	22q11.2	mir-648; mir-185; mir-649;mir-130b; mir-650

*Each cluster label consists of all the miRNA IDs in the cluster.

miRNA Target results

The miRgen interface provides access to unions and intersections of four target prediction programs and experimentally supported targets from TarBase (www.diana.pcbi.upenn.edu/tarbase.html). For each miRNA identified with an OMiR assay (*miR-210, miR-483, miR-648; miR-185, miR-649; miR-130b* and *miR-650*), we queried the database for predicted targets and associated GO terms. After that, we extracted from the lists only the GO terms related to the OMIM diseases by using key words: *serotonin, inositol, neuro, mental, anxia* and *synapse* for panic disorder; *diabetes, insulin* and *glucagons* for diabetes mellitus; *tumor, cancer, myeloid, lymphoid* and *carcinoma* for multiple tumor; *oncogene, cancer, abnormal, tumor, carcinoma, esophagus hypospadia* and *hypertelorism* for esophageal abnormality. We chose tumor-related keywords for esophageal abnormality and hypospadia, because several cases have shown a link between abnormality and tumor development (Habert et al., 2006), Maekawa et al., 2006), Mauduit et al., 2006). The same keywords were used to query the GO database and extract all the related GO terms. Finally, we computed the hypergeometric distribution to assess the probability of having the same list of terms by chance to obtain the lists of terms related to the miRNAs for Table 4 and all values of $p < 0.001$ (significant terms). The association with *miR-483* is the only one that was not significant ($p = 0.22$).

CRC and microRNAs

Colon cancer metastasis

The involvement of miRNAs in the development of metastases was initially reported by Li Ma, Julie Teruya-Feldstein and Robert Weinberg, who proved that *miR-10b* initiates breast cancer (BC) invasion and metastasis (Ma et al., 2007). Few months latter it was discovered that another miRNA, *miR-335,* suppresses metastasis and migration by targeting the transcription factor SOX4 and tenascin C, an extracellular matrix component with anti-adhesive properties (Tavazoie et al., 2008). At the same time, collaborative work between Huang's and Agami's groups reported that *miR-373* and *miR-520c* stimulated cancer cell migration and invasion and proposed as mechanism the suppression of CD44, which encodes a cell surface receptor for the extracellular matrix component hyaluronan (Huang et al., 2008). Taken together these landmark studies identify a fine balance of non-codingRNAs as stimulators and inhibitors of metastasis, and several targets that could potentially represent the molecular link between miRNA deregulation and a specific tumor behavior.

Recently, a miRNA hypermethylation profile characteristic of human metastasis, including CRCs suggesting that DNA methylation-associated silencing of tumor suppressor miRNAs contributes to the development of human cancer metastasis

(Lujambio et al, 2008). The reintroduction of *miR-148a* and *miR-34b* and *miR-34c* in cancer cells with epigenetic inactivation inhibited their motility, reduced tumor growth, and inhibited metastasis formation in xenograft models, with an associated down-regulation of the miRNA oncogenic target genes, such as C-MYC, E2F3, CDK6, and TGIF2. Most important, the involvement of these three miRNAs hypermethylation in metastasis formation was also suggested in human primary malignancies including colon, lung, breast, and head and neck carcinomas and melanomas, because it was significantly associated with the appearance of lymph node metastasis.

Although miRNAs represent the most widely studied of the non-coding RNAs (nc-RNAs), ther ncRNAs possibly involved in tumorigenesis are the ultraconservated genes (UCGs), a subset of genomic sequences that are located in both intra- and intergenic regions and are absolutely conserved (100%) between orthologous regions of the human, rat, and mouse genomes (Bejerano et al., 2004). Because of the high degree of conservation, the UCGs may have fundamental functional importance for the ontogeny and phylogeny of mammals and other vertebrates. Just as miRNAs may regulate mRNA levels, miRNAs may also CLL, as well as in CRC. Further expanding the involvement of UCG in human cancers, we were able to prove an oncogenic function for *uc.73(P)* in colon cancer, as diminution of its over-expression induced apoptosis and had antiproliferative effects specifically in colon cancer cells abnormally expressing uc.73(P), while no effects was found in cells with normal levels of this gene (Calin et al., 2007).

However, the role of UCGs in CRC metastases has not yet been defined. To further evaluate this, we analyzed miRNA and UCG expression in 6 cell lines: 3 non metastatic (COLO201, COLO205 and SW620) and 3 metastatic (COLO320, SW480 and HT29) for genome wide approach by our array technology. Data were imported into Biometric Research Branch (BRB, http://linus.nci.nih.gov/~brb/download.html) and SAM analysis was performed. As expected, several miRNAs and UCGs were found to be differentially expressed between metastatic and non-metastatic colon cancer cells at a statistically significant level. These ncRNAs were used to perform two supervised clusters, one related to human miRNAs and one to UCGs (**Figure 24**).

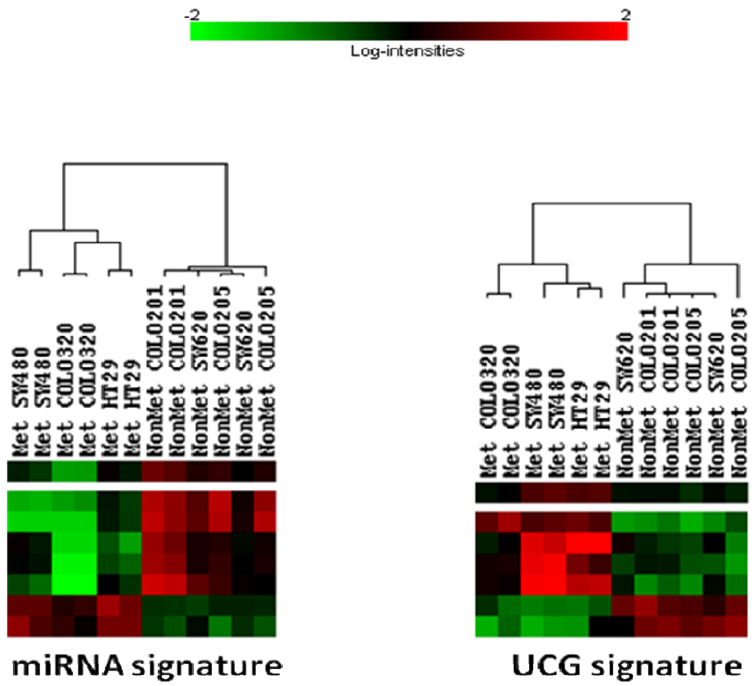

Figure 24. MicroRNAs and UCGs differentially expressed in metastatic versus non-metastatic colon cancer cell lines. Expression of the differentially regulated miRNAs and UCGs across colon cancer metastatic and non-metastatic cell lines. SAM analysis was performed to identify differentially expressed human microRNAs and UCGs using 0.1 as target proportion of false discoveries, 100 permutations and 90th percentile. We found 2 miRNAs upregulated in metastatic cell lines, 5 miRNAs downregulated in metastatic samples (Figure 1A), 2 UCGs downregulated in metastatic samples and 4 UCGs upregulated in metastatic cell lines (Figure 1B). MiRNAs and arrays were mean-centered using CLUSTER 3.0. Single linkage cluster was performed by using Spearman Correlation measure. (Figure from paper XI).

Finally, a principal component analysis was performed to highlight the power to differentiate metastatic from non-metastatic cell lines of the differentially expressed ncRNAs (**Figure 25**). A principal component analysis is a statistical method for exploring and making sense of datasets with a large number of measurements by reducing them to the few principal components that explain the main patterns.

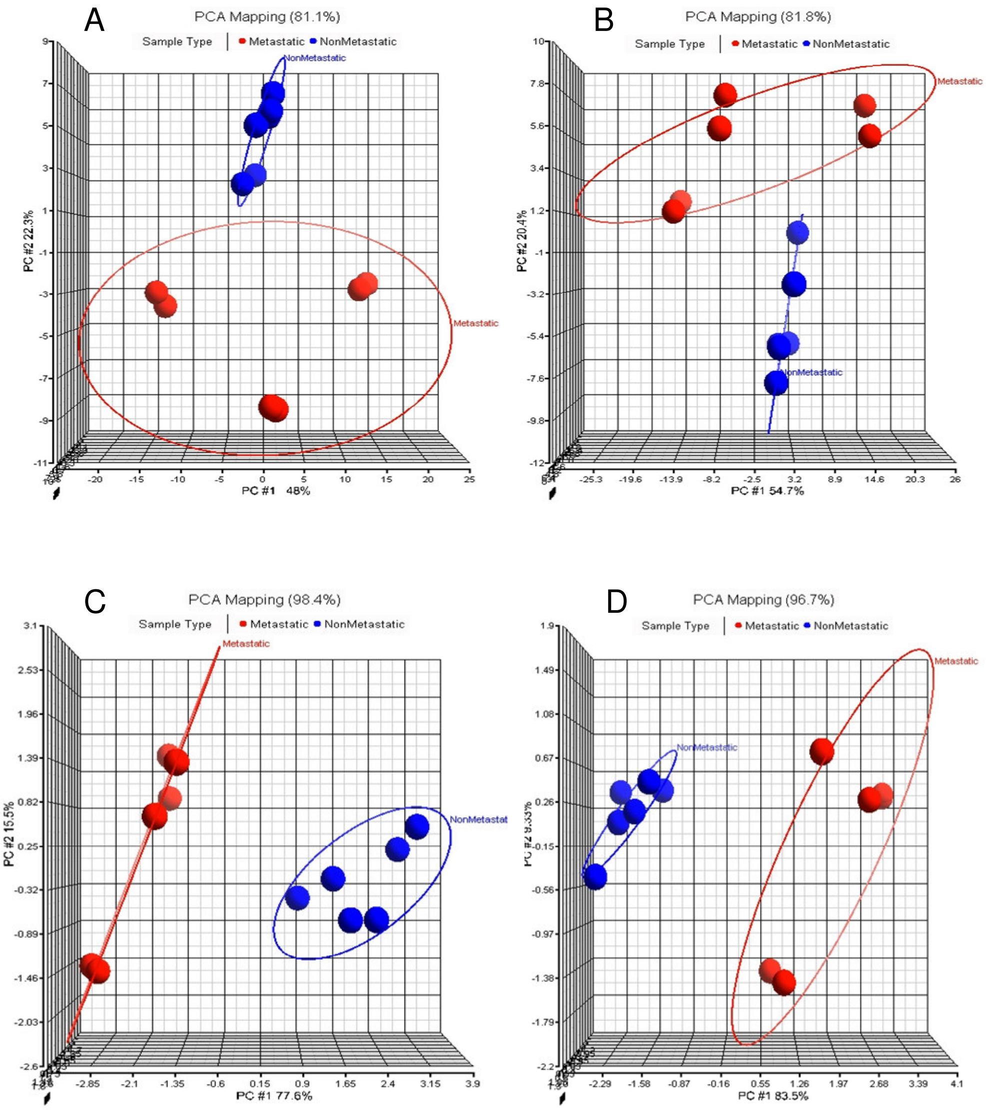

Figure 25. Principal Component Analysis (PCA). Unsupervised Sample Classification with PCA of normalized dataset was performed: Figure 4A reports as miRNAs and Figure 4B as UCGs separate metastatic from non-metastatic cell lines, the PCA mapping was 81.1% for miRNAs and 81.8% for UCGs and it is possible to see that the two classes of samples are already distinguished. Supervised classification using PCA analysis with significantly differentially expressed miRNAs (Figure 4B) and UCGs (Figure 4 C) were able to improve the similarity among samples that belong to the same group (see the axes scale) and also confer a better distinction between classes. With differentially expressed miRNAs the PCA mapping was 98.4% and for UCGs was 96.7%. (Figure from paper XI).

Ultraconserved Regions and MicroRNAs

Noncoding RNA (ncRNA) transcripts are thought to be involved in human tumorigenesis. We report that a large fraction of genomic ultraconserved regions (UCRs) encode a particular set of ncRNAs whose expression is altered in human cancers. Genome-wide profiling revealed that UCRs have distinct signatures in human leukemias and carcinomas. UCRs are frequently located at fragile sites and genomic regions involved in cancers. Certain UCRs whose expression may be regulated by microRNAs abnormally expressed in human chronic lymphocytic leukaemia were identified, and was proved that the inhibition of an overexpressed UCR induces apoptosis in colon cancer cells (Calin et al., 2007). Calin and colleagues findings argue that ncRNAs and interaction between noncoding genes are involved in tumorigenesis to a greater extent than previously thought. **Our statistical analysis supported their findings.**

Statistical Analyses for Correlations between Microarray Expression of UCRs and miRNAs

A detailed description is provided in the Discussion session. Briefly, The input data was constituted by a list of T-UCRs and by a list of miRNAs (the “seeds”) and the corresponding matrix of expression values. We calculated r, the Spearman rank coefficient of correlation for each pair of (miR, UC) genes; namely, we evaluate the p values of the correlation tests and select the genes whose correlation value is significant at a given value of rejection. Given the high number of correlation tests performed, p values were corrected for multiple testing by using the false detection rate (FDR), as in n this way, p values control the number of false positive over the number of truly null tests, while FDR controls the number of false positive over the number of significant tests.

DISCUSSION

TOM has been then improved and extended

We describe here three main improvements to TOM original algorithm: (i) enhancement of the statistical scores that define the associations; (ii) addition of murine expression data for more comprehensive analyses; (iii) introduction of advanced flexible enrichment analysis.

Statistical Scores Enhancement

Besides the p-value for assessing the significance of the correlation between two of the n genes profiles in the experiment (enhanced correction for multiple hypotheses for all the n–n couples), in this enhanced version another statistic is offered that gives a measure of the robustness of the result, here called R. The rationale of this score is the same of the enrichment (Zhang et al., 2005), and namely relies on the assumption that the more often a gene is found to be related to another gene, the more the association between the genes can be assumed to be robust. However, the frequency of positive results needs to be normalized on the total number of tests that genes appears to be statistically related in a small number of experiments, but corresponding to the total number of experiment tested on the same two genes, the result is more relevant than if it represents only a fraction of the tests on the same two genes performed. Namely, it is defined as: $R = |\{\rho_{x,\ y}\}|p_{x,\ y} < \theta|/|\{\rho_{x,\ y}\}|$, where x and y are the expression profiles of two genes, $\rho$is the correlation score, p the p-value corresponding to ρ, θ the user-defined statistical significance threshold.This statistic expands the broadness of the analysis, taking into account the robustness of the relationship found, based on its redundancy, across different sets of experiments.

Murine-Human data

Besides adding more human data, the database has been enriched in terms of murine expression data. As for the human data, information is obtained by pre-computed correlation among gene expression profiles, for series obtained from the GEO database. Once the query is performed on one or tow loci, TOM extracts two lists of candidate genes. These two lists represent either the correlating genes on the two loci or the list of seeds plus the correlating genes on the single locus. To identify more stringent correlations TOM proceeds to another query, based on murine data. The advantage of inserting also the mouse expression data allows to identify new correlations that might be not visible by comparing the human data alone. Extensive work has been done in mice in order to study human disorders and today technology permits to target virtually any mouse candidate gene that has a human homologue (Capecchi, 2005). Using the correlations in mice, TOM retrieves the human homologues (from Homologene at NCBI) and searches them in the list identified by comparison in human array data.

Extended Enrichment Analysis

Finally, we integrated and improved a second tool, FIT (Nardini et al., 2006) to help geneticists understand the role of the most significant genes. FIT measures the similarity between any list of candidate genes extracted with TOM (test list) and any number of lists (reference lists), extracted in the same way, or representing a signature, a pathway, obtained from literature, custom defined or annotated in KEGG (Kanehisa and Goto, 2000) or GenMAPP (Salomonis et al., 2002). This measure of similarity consists of a sequence of three statistical tests (enrichment, specificity of the enrichment and fit), for the quantification and the ranking of the relationship between any two sets of genes. Statistical significance of enrichment (p_{η}) is evaluated by means of the hypergeometric distribution (Sokal and Rohlf, 2003). It assesses if the number of relevant items in a set is greater than the one that would be obtained by chance. The specificity of the enrichment (p_{τ}) assesses if the enrichment is specific to the given category. Namely, specificity informs the researcher if the meaning, besides being statistically significant, is specific to a given set of genes, or if it is shared or distributed with others. In particular it can tell not only if the number of items falling in a given category is greater than what could be expected by chance, but also if it is unique to a given set of genes. To do so, the candidate gene list is represented as a distribution of all its genes across the bins defined by the categories (references) we want to compare to (sub-ontologies, pathways, other custom sets). The same is also done for any reference list, that is generally represented as an ‘impulsive’ distribution (almost all the genes fall in the same bin). The specificity is then defined as a significant value of correlation among the distributions profiles. This score also helps to disambiguate particular cases with identical enrichment, but different distributions of the genes (Nardini et al., 2006). The significance of the final fit score is obtained from the Fisher inverse χ^2 method (Hedges and Olkin, 1985) and is defined as $p_{\phi} = -2(\log(p_{\eta}) + \log(p_{\tau}))$. Globally, this analysis allows to define statistical scores to rank and thus help disambiguate the enrichment for the list of candidates genes for meaningful sets of known or annotates genes. To delimitate better the genomic area, it is then central to compare these same genes to others that might contribute to the same cellular pathway or are part of the same expression set. FIT allows this quantitative automated comparison and can list for example the *pϕ*-values for the enriched comparison analysis with all KEGG Pathways. The user can then for example choose to give priority in the candidate gene list to the ones that are related to the most enriched function and thus make the analysis more efficient. To make this comparison feasible, automated approaches are crucial to allow for the high-throughput quantification of these comparison. Given these necessities, we expect this approach to provide an integrative and efficient tool for enhanced effective hypothesis-driven research.

Fun&Co

We investigated the signaling properties of heart, skeletal and smooth muscles. Considering 'cell surface receptor linked signal transduction' (GO:0007166) as GOmain, we noticed that transmembrane receptor protein tyrosine phosphatase and dopamine receptors appeared associated to skeletal muscle, while IGF1R is associated to heart. Transgenic mice over-expressing IGF1R (Insuline growth factor like 1 receptor) in the heart displayed cardiac hypertrophy which was due to an increase in myocyte size, and there was no evidence of histopathology. This study suggests that targeting the cardiac IGF1R-PI3K(p110alpha) pathway could be a potential therapeutic strategy for the treatment of heart failure (Canicio and Kaliman, 2001). For insulin receptor signaling pathway, we obtained a P-value of 0.303 in the comparison between heart and skeletal muscle and a P-value of 0.207 in the comparison between heart and smooth muscle. Performing a similar experiment with GlobalTest, the comparative P-values were respectively, 0.508 and 0.226. In skeletal muscle, we noticed the presence of transmembrane receptor protein tyrosine phosphatase signaling pathway seems to be obvious: protein-tyrosine phosphatases (PTPases) have an important role in the regulation of insulin signal transduction, and the skeletal muscle is the major site of tissue insulin resistance in obesity and diabetes. The PTPase activity in skeletal muscle from non-diabetic obese subjects was increased significantly by 40–70% compared to the level in controls (Ahmad et al., 1997). We obtained a P-value of 0.149 in the comparison between heart and skeletal muscle and a P-value of 0.379 in the comparison between heart and smooth muscle. Performing a similar experiment with GlobalTest, the comparative P-values were respectively, 0.297 and 0.918. In 'intracellular signaling cascade' (GO:007242), the activation of NF-_B-inducing kinase pathway appears to be skeletal muscle associated. We obtained a P-value of 0.279 in the comparison between cardiac and skeletal muscle. Performing a similar experiment with GlobalTest, the comparative p was 0.748. The third and last GO category analyzed, molecular functions, yielded many additional interesting results. The first general GO term analyzed in this category was ‘kinase activity’ (GO:0016301). The ephrin receptor activity was the only one specific to cardiomyocytes: this protein is involved in cell–cell communication during development and in particular EphA3 plays a critical role in heart development (Stephen et al., 2007). In striated muscle, which includes both heart and skeletal muscle, the IGF receptor activity is listed: this receptor regulates the cell growth and development in muscles and other tissues. In heart, IGFs are locally produced and modulates cardiomyocyte growth and maturation. Biochemical alteration (expression variation of IGFs) may be associated to fetal/neonatal growth abnormalities of rats (Engelmann et al., 1989). We obtained a P-value of 0.561 in the comparison between skeletal and smooth muscle and a P-value of 0.524 in the comparison between heart and smooth muscle. Performing a similar experiment with GlobalTest, the comparative P-values were respectively, 0.699 and 0.754. In smooth muscle on the other hand, MAP kinase activity was apparent in its multiple steps. MAP kinase is part of a signal transduction pathway that promotes cell divisions in response to extracellular stimuli. MAP kinase pathway, activated by angiotensin II, is involved in hypertensive vascular remodeling, associated with cell growth and increased deposition of extracellular matrix, in particular collagen (Touyz et al., 2001). Furthermore, the G protein coupled receptors kinase activity, also short-listed in smooth muscle cells, mediates, via the MAPK pathways, the mitogenic effects of oxidized lowdensity lipoprotein on vascular smooth muscle cells (Yang et al., 2001). This mechanism is proven to be involved in pathogenesis of atherosclerosis. Other kinase activities could be particularly related to the different metabolic pathways of these three different muscle tissues. In skeletal muscles, the phosphorylase kinase activity is associated with the glycogen metabolism. Glycogen representing an easily available source of glucose. The liver and the skeletal muscles are in fact the main tissues that stock the glycogen, and the glycogen phosphorylase kinase is the key regulatory enzyme in this process. In striated, a range of kinase activities related to glucose metabolism (glycolisis and gluconeogenesis) were apparent, e.g. hexokinase and other enzymes, like pantothenate kinase, involved in synthetic pathways of acetyl-CoA, the point of connection of the main metabolic oxidative pathways (amino acids, fatty acids and carbohydrates). In stimulated smooth cells, the concentration of diacylglycerol (DAG) rises rapidly, and DAG functions as a second messenger by activating protein kinase C, which in turn regulates many cellular responses, including growth and differentiation. The attenuation of the DAG signal and phospholipid synthesis, by the conversion of DAG to phosphatidic acid (PA), is regulated by DAG kinases (DGKs). PA can also serve as a lipid messenger and the net effect of conversion of DAG to PA might vary from cell to cell and condition to condition. Diacylglycerol kinase (DGK) phosphorylates the lipid second messenger DAG to phosphatidic acid. DGK-theta is present both in smooth muscle and in endothelial cells of the small blood vessels. DGK-theta activity can be increased by noradrenaline (NA) and this pathway is thought to have a physiological role in vascular smooth-muscle responses (Walker et al., 2001). Guanylate kinase catalyzes the phosphorylation of either GMP to GDP or dGMP to dGDP and is an important enzyme in nucleotide metabolic pathways. Co-expression of guanylate kinase with thymidine kinase enhances pro-drug cell killing in vitro and suppresses vascular smooth muscle cell proliferation in vivo (Akyurek et al., 2001). In striated, the inositol trisphosphate 3-kinase converts Ins- 1,4,5-P3 to Ins-1,3,4,5-P4, that modulates the entry of Ca2þ from an extracellular source. The 3-kinase activity is significantly activated by the Ca2þ/calmodulin complex. In some experiments, the IP3 kinase activity was increased in SHRSP (stroke-prone spontaneously hypertensive rats) and its activity was markedly affected by divalent cations. These data suggest that the accumulations of IP3 and IP4 after hormonal stimulation play a physiologic role, possibly by alteration of Ca2þ levels in cardiac tissue (Kawaguchi et al., 1990). Another GO term that provided interesting results was ‘G-Protein coupled receptor activity‘ (GO:0004930). In heart, the thrombin activity induces IP3 formation associated to increase in cytosolic calcium, enhanced automaticity and prolong repolarization: this can be related to the electrical abnormalities observed in ischemia and infarction (Steinberg et al., 1991). In skeletal muscle the prostaglandin activity is important: arachidonic acid metabolites, such as prostaglandins (PG), are regulatory of vascular tone and can be released from the contracting muscles under the influence of dynamic exercise (Karamouzis et al., 2001). This can in turn augment the blood flow and allow the incoming of nutrients and oxygen, to support an increasing metabolic muscle request. Furthermore, prostaglandin F2 is involved in the multi-step process leading to the formation of large multinucleated muscle cells. Therefore, the use of prostaglandins might be therapeutic for treatment of muscle loss due to aging, injury and disease. And conversely caution should be taken in using drugs that inhibit PG production (like e.g. non-steroidal anti inflammatory drugs) which may be deleterious for muscle growth (Horsley and Pavlath, 2003). Also in smooth prostaglandins are active, but in this case prostaglandin E receptor activity induces relaxation, e.g. in trachea Platelet activating factor (PAF) receptor activity is another term specific to skeletal muscle: PAF cumulative effects in skeletal muscle reduce protein synthesis during endo-toxic and septic shock (Karlstad et al., 2000). In addition, PAF seems to be involved in skeletal muscle ischemia-reperfusion injury (IRI): infusion of PAF antagonists into the muscle prior to reperfusion can indeed reduce muscle necrosis (Silver et al., 1996). On the basis of opioid-stimulated contraction of dispersed gastric smooth muscle cells, it has been suggested that these cells possess opioid receptors of three subtypes: kappa, mu and delta. In smooth cells, the disorder of Ca2þ regulation induced by hemorrhagic shock was mediated by opioid receptor and alphaadrenoceptor, which may be partly responsible for the vascular hyporesponse, and opioid receptor antagonists improved the response of resistance arteries to vascular stimulants in decompensatory stage of hemorrhagic shock (Kai et al., 2004). Vaso-active intestinal polypeptide receptor is involved in smooth muscle relaxation and, in particular, in bladder, stomach and the esophageal sphincter. Tachykinin receptor activity in striated is due to its role: takykinin has cardio-acceleratory effect (Sliwowska et al., 2001) and probably some coordinated effect on skeletal muscle.

Ultraconserved Regions and MicroRNAs

The input data were constituted by a list of T-UCRs and by a list of miRNAs (the "seeds") and the corresponding matrix of expression values. We calculated r, the Spearman rank coefficient of correlation, a non-parametric measure of data trend correlation based on rankings, for each pair of (miR, UCR) genes; namely, we evaluate the P-values of the correlation tests and selected the genes whose correlation value is significant at a given value of rejection. Evaluation of P-values was performed assuming that the correlation values are distributed using Student's t cumulative distribution, with a number of degrees of freedom corresponding to the number of samples in the microarray experiment. The P-values measure the 'goodness' of the single correlations (among couples of genes), therefore, to understand if the real correlation derives by chance or represents a biologically important information, we choose the method of permutations, changing the order of the samples for each row (miR or UCR) and calculating the correlations between pair of genes (miR, UCR) with different changed samples orders. We repeated the samples permutation and computed correlations 100 times, in this way, every real correlation has 100 random correlations to compare with. Using all (100 * n° MIR * n° UCR) random correlations and real correlations, we recalculated P-values based on random correlations ranking and position of the real correlations. Given the high number of correlation tests performed, P-values were corrected for multiple testing by using the false detection rate (FDR), as defined by (Benjamini and Hochberg, 1995). In this way, P-values control the number of false positive over the number of truly null tests, while FDR controls the number of false positive over the number of significant tests. Several ways of estimating this number have been proposed, and we adopted the solution devised by Tom Nichols (see http://froi.sourceforge.net/documents/technical/matlab/FDR.html), that rescales the P-value obtained on a single test multiplying it by a combination of indexes related to the total number of tests performed. Correction was performed on a seed by seed basis, meaning that the genes in the seeds list were considered independent tests.

This statistically validated tripe filtering allows the targeted extraction of a shortlist of candidate genes, thus saving resources for the following costly and time-consuming genetic analysis. To build a scatter plot between *miR-24-1* and *uc.160* expression values and between miR-155 and uc.346+(A) expression values, respectively, we plotted a regression line by using MatLab function ROBUSTFIT to explain hypotheses of negative correlation between these two genes (see **Figure 26**).

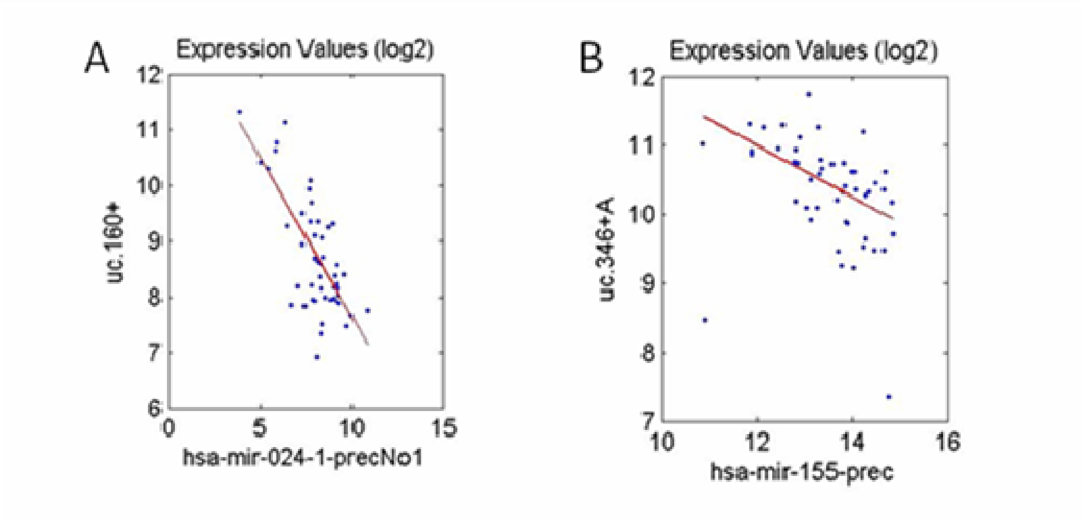

Figure 26 Two scatter plots between expression values of mir-24-1 and uc.160 and of miR-155 and uc.346A are presented. The regression line shows the negative correlation between these two genes. The name of the corresponding array probes are presented on the Y and X axes. Both probes recognize the mature form of the miRNA gene. (Original figure used for paper VI).

CONCLUSIONS

Bioinformatics is an important field that discover and support research.
Genomic research will continue to benefit from the productive interaction of biologists, geneticists, statisticians, and bioinformaticians. Several tools, web-based applications, algorithms and data analysis has been described in this thesis, for each one of them we report a specific conclusion.

GebbaMa

GebbaMA, the virtual laboratory for microarray data, is a web environment web for data management.

The tool, specific for the treatment of microarray data has several functionality like:

- upload
- download
- search of the experiments and annotation of the genetic information, clinics and experimental
- federate search among installations geographically distributed

The environment has been realized in a user friendly way guaranteeing a high degree of personalization in phase of storage:

- Upload and multiple download of the experiments
- automatic annotation of the genetic information, clinics and experimental, produced in the laboratory
- automatic control of some annotations (organism, provider of the array, name of the array) according to the reference ontologies
- Possibility to singly manage the experiments or to re-organize them in projects
- Interoperability among laboratories geographically distributed

TOM

We devised and implemented TOM, an algorithm for the identification of candidate genes responsible for genetic diseases. We took advantage of the microarray datasets available online to exploit novel computational biology approaches to molecular genetics. TOM allows a user to seamlessly associate functional and mapping data and to efficiently employ them in a quest for novel candidate genes in hereditary diseases. Additional selection principles can be implemented to extend TOM, such as declaring a putative pathway for the candidate gene, in the case of poorly characterized diseases. Moreover, constant updating TOM with new expression datasets will increase the robustness of the assay. Our work represents a novel computational tool for gene hunters and could help to integrate and improve the comprehension of the genetic roots and the molecular mechanisms of complex life threatening diseases.

Fun&Co

Fun&Co is a novel and very efficient way of mining functional differences from a number of datasets in a pairwise manner. The application extracts the most significant differences from the molecular expression data, as shown in this article on skeletal, heart and smooth muscles. The results are highly informative and synthetic. Important, it is apparent that as many as a dozen critical points were correctly detected by Fun&Co in common with the heart signaling network of Heineke and colleague (Heineke and Molkentin, 2006). This finding supports the potential usefulness of this application in the high level analysis of transcriptome.

Hypersolutes

Low millimolar concentrations of hydroxyectoine, potassium diglycerol phosphate and potassium mannosylglycerate reduced DNA microarray background and improved hybridization efficiency. The results were highly significant when analyzed by comparing different quality control measures: raw Q, background (bkg), scaling factor (SF), percent present calls (%P), chips pseudo-images, normalized unscaled standard errors (NUSE) and relative log expression (RLE). Twenty five mM DGP, 10 mM HECT and 10 mM MG were shown to be the optimal solutes and concentrations. The experiments were carried out and confirmed in two different Affymetrix facilities. The application of this finding to hybridization protocols could result in a significant improvement of microarray experiments, not limited to expression profiling.

OmiR

We devised and implemented OMiR, an algorithm to compute and study the associations between miRNAs and genetic diseases. We took advantage of the OMIM database to retrieve disease information. Additional association options can be implemented to extend the reach of OMiR, such as the possibility of using markers known by the user that have not yet been published. Moreover, updating OMiR with new OMIM IDs and new miRNAs will increase the robustness of the assay. Our work represents a novel way to extract information on the relationships between miRNAs and diseases with OMIM IDs not associated with a specific disease gene.

CAGRs and microRNAs

The genome-wide correlation studies described are the first steps toward defining a link between the role of noncoding RNAs (ncRNAs such as miRNAs and RNAs from UCRs) and the development of cancer. Bioinformatics and statistical tools were highly utilized to reveal associations between the ncRNAs and CAGRs and/or FRAs. Genome-wide expression profiling of miRNAs by various techniques has already begun to identify new diagnostic and prognostic tools for cancer patients. Several ncRNAs were shown to affect tumor development and progression. The implementation of new algorithms and tools specifically focused on miRNAs and UCGs will lead to important findings regarding ncRNAs as tumor biomarkers and therapeutic targets. Targeting of ncRNAs may thus provide an important therapeutic strategy for treatment of human cancer. From a therapeutic standpoint, restoration of the expression of a downregulated (or functionally deficient) ncRNA or, alternatively, inhibition of an overexpressed ncRNA could reverse the tumor phenotype. Knowing the location of such ncRNAs in CAGR could help in selecting the best candidates for starting the ncRNA-based gene therapy trials.

.

REFERENCES

Aaltonen LA and Hamilton SR. (2000) World Health Organization Classification of Tumors. Pathology and Genetics of Tumours of the Digestive System. Lyon, France: International Agency for Research on Cancer Press.

Ahmad F, Azevedo JL, Cortright R, Dohm GL, Goldstein BJ (1997) Alterations in skeletal muscle protein-tyrosine phosphatase activity and expression in insulin-resistant human obesity and diabetes. J. Clin. Invest., 100, 449–458.

Akyürek LM, Nallamshetty S, Aoki K, San H, Yang ZY, Nabel GJ, Nabel EG (2001) Coexpression of guanylate kinase with thymidine kinase enhances prodrug cell killing in vitro and suppresses vascular smooth muscle cell proliferation in vivo. Mol. Ther., 3, 779–786.

Al-Bhalal,L. and Akhtar,M. (2005) Molecular basis of autosomal dominant polycystic kidney disease. Adv. Anat. Pathol., 12, 126–133.

Alsanea,O. and Clark,O.H. (2001) Familial thyroid cancer. Curr. Opin. Oncol., 13, 44–51.

Ambros V. (2004) The functions of animal microRNAs. Nature. 431, 350-355.

Barbarotto E, Schmittgen TD, Calin GA (2008) MicroRNAs and cancer: profile, profile, profile. Int J Cancer 122, 969–977

Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R (2005) NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Res. 1;33(Database issue):D562-6.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. (2004) Ultraconserved elements in the human genome. Science. 304, 1321-1325.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B 57, 289–300.

Bolstad, B. M., Irizarry R. A., Astrand, M, and Speed, T. P. (2003) *A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance.* Bioinformatics 19, 185-193.

Bueno MJ, de Castro IP, Malumbres M. (2008) Control of cell proliferation pathways by microRNAs. Cell Cycle. 7, 3143-3148.

Cai X, Lu S, Zhang Z, Gonzalez CM, Damania B, Cullen BR. (2005) Kaposi's sarcoma-associated herpesvirus expresses an array of viral microRNAs in latently infected cells. Proc. Natl. Acad. Sci. USA. 102, 5570-5575.

Calin, G.A., and Croce, C.M. (2006). MicroRNA signatures in human cancers. Nat. Rev. Cancer 6, 857–866.

Calin GA, Croce CM. (2007) Investigation of microRNA alterations in leukemias and lymphomas. Methods Enzymol. 427:193-213.

Calin GA, Liu CG, Ferracin M, Hyslop T, Spizzo R, Sevignani C, Fabbri M, Cimmino A, Lee EJ, Wojcik SE, Shimizu M, Tili E, Rossi S, Taccioli C, Pichiorri F, Liu X, Zupo S, Herlea V, Gramantieri L, Lanza G, Alder H, Rassenti L, Volinia S, Schmittgen TD, Kipps TJ, Negrini M, Croce CM. (2007) Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. Cancer Cell. 12, 215-229.

Calin GA, Sevignani C, Dumitru CD, Hyslop T, Noch E, Yendamuri S, Shimizu M, Rattan S, Bullrich F, Negrini M, Croce CM. (2004) Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. Proc Natl Acad Sci U S A. 101, 2999-3004.

Canicio,J. and Kaliman,P. (2001) Nuclear factor kappa B-inducing kinase and Ikappa B kinase-alpha signal skeletal muscle cell differentiation. J. Biol. Chem., 276, 20228–20233.

Canzian,F., Amati,P., Harach,H.R., Kraimps,J.L., Lesueur,F., Barbier,J., Levillain,P., Romeo,G. and Bonneau,D. (1998) A gene predisposing to familial thyroid tumors with cell oxyphilia maps to chromosome 19p13.2. Am. J. Hum. Genet., 63, 1743–1748.

Capecchi,M.R. (2005) Gene targeting in mice: functional analysis of the mammalian genome for the twenty-first century. Nat. Rev. Genet., 6, 507–512.

Chan JA, Krichevsky AM, Kosik KS (2005) MicroRNA-21 is an antiapoptotic factor in human glioblastoma cells. Cancer Res 65, 6029–6033

Chang TC, Yu D, Lee YS, Wentzel EA, Arking DE, West KM, Dang CV, Thomas-Tikhonenko A, Mendell JT. (2008) Widespread microRNA repression by Myc contributes to tumorigenesis. Nat Genet. 40, 43-50.

Cimmino A, Calin GA, Fabbri M, Iorio MV, Ferracin M et al (2005) miR-15 and miR-16 induce apoptosis by targeting BCL2. Proc Natl Acad Sci USA 102, 13944–13949

Costa FF (2005) Non-coding RNAs: new players in eukaryotic biology. Gene 357, 83–94

Costinean S, Zanesi N, Pekarsky Y, Tili E, Volinia S, Heerema N, Croce CM. (2006) “Pre-B cell proliferation and lymphoblastic leukemia/high-grade lymphoma in E(mu)-miR155 transgenic mice”. Proc Natl Acad Sci U S A, 103, 7024-7029.

Croce, C.M., and Calin, G.A. (2005). miRNAs, Cancer, and Stem Cell Division. Cell 122, 6–7.

Dickins RA, Hemann MT, Zilfou JT, Simpson DR, Ibarra I et al (2005) Probing tumor phenotypes using stable and regulated synthetic microRNA precursors. Nat Genet 37, 1289–1295

ENCODE Consortium. (2004) The ENCODE (ENCylopedia Of DNA Elements) Project. Science, 5696, 636–640.

Engelmann GL, Boehm KD, Haskell JF, Khairallah PA, Ilan J. (1989) Insulin-like growth factors and neonatal cardiomyocyte development: ventricular gene expression and membrane receptor variations in normotensive and hypertensive rats. Mol. Cell Endocrinol., 63, 1–14.

Esquela-Kerscher A, Slack FJ (2006) Oncomirs—microRNAs with a role in cancer. Nat Rev Cancer 6, 259–269

Garzon R, Fabbri M, Cimmino A, Calin GA, Croce CM. (2006) MicroRNA expression and function in cancer. Trends Mol. Med. 12, 580–587.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol., 5, R80.

Greenlee RT, Hill-Harmon MB, Murray T, Thun M. (2001) Cancer statistics, 2001. CA Cancer J Clin 51, 15–36.

Gregory RI, Shiekhattar R (2005) MicroRNA biogenesis and cancer. Cancer Res 65, 3509–3512

Habert R, Delbes G, Duquenne C, Livera G, Levacher C (2006) Effects of estrogens on the development of the testis during fetal and neonatal life. Gynecol Obstet Fertil. 34, 970-977.

He H, Jazdzewski K, Li W, Liyanarachchi S, Nagy R, Volinia S, Calin GA, Liu CG, Franssila K, Suster S, Kloos RT, Croce CM, de la Chapelle A. (2005) The role of microRNA genes in papillary thyroid carcinoma. Proc Natl Acad Sci U S A. 102, 19075-19080.

He L, Thomson JM, Hemann MT, Hernando-Monge E, Mu D, Goodson S, Powers S, Cordon-Cardo C, Lowe SW, Hannon GJ, Hammond SM (2005) A microRNA polycistron as a potential human oncogene. *Nature.* 435, 828-833.

He X, He L, Hannon GJ. The guardian's little helper: microRNAs in the p53 tumor suppressor network. Cancer Res. 2007 Dec 1, 11099-11101. Review.

Hedges,L. and Olkin,I. (1985) Statistical Methods for Meta-Analysis. Academic Press, New York.

Heineke J, Molkentin JD. (2006) Regulation of cardiac hypertrophy by intracellular signalling pathways. Nat Rev Mol Cell Biol. 7:589-600.

Horsley,V. and Pavlath,G.K. (2003) Prostaglandin F2(alpha) stimulates growth of skeletal muscle cells via an NFATC2-dependent pathway. J. Cell Biol., 161, 111–118.

Houbaviy HB, Murray MF, Sharp PA. (2003) Embryonic stem cell-specific MicroRNAs Dev. Cell. 5, 351-358.

Huang Q, Gumireddy K, Schrier M, le Sage C, Nagel R, Nair S, Egan DA, Li A, Huang G, Klein-Szanto AJ, Gimotty PA, Katsaros D, Coukos G, Zhang L, Puré E, Agami R. (2008) The microRNAs miR-373 and miR-520c promote tumour invasion and metastasis. *Nat Cell Biol* 10, 202–210.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 4, 249-264.

Jeffrey SS , (2008) Cancer biomarker profiling with microRNAs. Nat Biotechnol. 26, 400-401.

Johnson CD, Esquela-Kerscher A, Stefani G, Byrom M, Kelnar K, Ovcharenko D, Wilson M, Wang X, Shelton J, Shingara J, Chin L, Brown D, Slack FJ. (2007) The let-7 microRNA represses cell proliferation pathways in human cells. Cancer Res 67, 7713–7722

Kai L, Wang ZF, Shi YL, Liu LM, Hu DY. (2004) Opioid receptor antagonists increase [Ca2+]i in rat arterial smooth muscle cells in hemorrhagic shock. Acta Pharmacol. Sin., 25, 395–400.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res., 28, 27–30.

Karamouzis M, Langberg H, Skovgaard D, Bülow J, Kjaer M, Saltin B. (2001) In situ microdialysis of intramuscular prostaglandin and thromboxane in contracting skeletal muscle in humans. Acta Physiol. Scand., 171, 71–76.

Karlstad MD, Buripakdi D, Carroll RC. (2000) Platelet-activating factor (PAF)-induced decreases in whole-body and skeletal muscle protein synthesis. Shock, 14, 490–498.

Kawaguchi H, Iizuka K, Takahashi H, Yasuda H (1990) Inositol trisphosphate kinase activity in hypertrophied rat heart. Biochem. Med. Metab. Biol., 44, 42–50.

Kopetz S, Freitas D, Calabrich AF, Hoff PM. Adjuvant chemotherapy for stage II colon cancer. Oncology (Williston Park) 2008;22:260-70; discussion 70, 73, 75.

Kulshreshtha R, Ferracin M, Wojcik SE, Garzon R, Alder H, Agosto-Perez FJ, Davuluri R, Liu CG, Croce CM, Negrini M, Calin GA, Ivan M. (2006) A microRNA signature of hypoxia. Mol Cell Biol. 27, 1859-1867.

Lagos-Quintana M, Rauhut R, Meyer J, Borkhardt A, Tuschl T (2003) New microRNAs from mouse and human. RNA. 9, 175-179.

Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, Pfeffer S, Rice A, Kamphorst AO, Landthaler M, Lin C, Socci ND, Hermida L, Fulci V, Chiaretti S, Foà R, Schliwka J, Fuchs U, Novosel A, Müller RU, Schermer B, Bissels U, Inman J, Phan Q, Chien M, Weir DB, Choksi R, De Vita G, Frezzetti D, Trompeter HI, Hornung V, Teng G, Hartmann G, Palkovits M, Di Lauro R, Wernet P, Macino G, Rogler CE, Nagle JW, Ju J, Papavasiliou FN, Benzing T, Lichter P, Tam W, Brownstein MJ, Bosio A, Borkhardt A, Russo JJ, Sander C, Zavolan M, Tuschl T. (2007) A mammalian microRNA expression atlas based on small RNA library sequencing Cell. 129, 1401-1414.

Lee RC, Feinbaum RL, Ambros V (December 1993). "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14". *Cell* 75, 843–854.

Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP. (2003) Vertebrate microRNA genes. Science. 299, 1540.

Liu CG, Calin GA, Meloon B, Gamliel N, Sevignani C, Ferracin M, Dumitru CM, Shimizu M, Zupo S, Dono M, Alder H, Bullrich F, Negrini M, Croce CM. (2004) "An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues". Proc Natl Acad Sci U S A, 2004, 101, 9740–9744.

Liu CG, Calin GA, Volinia S, Croce CM. Nat Protoc. (2008) MicroRNA expression profiling using microarrays.;3, 563-578.

Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, Golub TR. (2005) MicroRNA expression profiles classify human cancers. *Nature.* 435, 834–838.

Lui WO, Pourmand N, Patterson BK, Fire A. (2007) Patterns of known and novel small RNAs in human cervical cancer Cancer Res. 67, 6031-6043.

Lujambio A, Esteller M (2007) CpG island hypermethylation of tumor suppressor microRNAs in human cancer. *Cell Cycle* 6, 1455–1459.

Ma L, Teruya-Feldstein J, Weinberg RA. (2007) Tumour invasion and metastasis initiated by microRNA-10b in breast cancer. Nature. 449, 682-688.

Makunin IV, Pheasant M, Simons C, Mattick JS (2007) Orthologous microRNA genes are located in cancer-associated genomic regions in human and mouse. PLoS ONE 2, e1133

Maekawa H, Tanaka N, Hashimoto N, Yamada H, Mitsui H, Ikeda H, Maruyama T, Mori M, Nagawa H, Kimura S. (2006) Esophageal smooth muscle tumor in a 25-year-old woman with congenital malformations. J Gastroenterol. 36, 700-703.

Mathonnet G, Fabian MR, Svitkin YV, Parsyan A, Huck L, Murata T, Biffo S, Merrick WC, Darzynkiewicz E, Pillai RS, Filipowicz W, Duchaine TF, Sonenberg N. (2007) MicroRNA inhibition of translation initiation in vitro by targeting the cap-binding complex eIF4F. Science. 317, 1764-1767.

Mauduit C, Florin A, Amara S, Bozec A, Siddeek B, Cunha S, Meunier L, Selva J, Albert M, Vialard F, Bailly M, Benahmed M. (2006) Long-term effects of environmental endocrine disruptors on male fertility. Gynecol Obstet Fertil. 34, 978-984.

McKay,J.D., Lesueur,F., Jonard,L., Pastore,A., Williamson,J., Hoffman,L., Burgess,J., Duffield,A., Papotti,M., Stark,M. et al. (2001) Localization of a susceptibility gene for familial nonmedullary thyroid carcinoma to chromosome 2q21. Am. J. Hum. Genet., 69, 440–446.

McKay,J.D., Thompson,D., Lesueur,F., Stankov,K., Pastore,A., Watfah,C., Strolz,S., Riccabona,G., Moncayo,R., Romeo,G. and Goldgar,D.E. (2004) Evidence for interaction between the TCO and NMTC1 loci in familial non-medullary thyroid cancer. J. Med. Genet., 41, 407–412.

McManus MT (2003) MicroRNAs and cancer. Semin Cancer Biol 13, 253–258

Mendell JT (2005) MicroRNAs: critical regulators of development, cellular physiology and malignancy. Cell Cycle 4, 1179–1184

Nagel R, le Sage C, Diosdado B, van der Wall M, Oude Vrielink JA, Bolijn A, Meijer GA, Agami R. (2008) Regulation of the adenomatous polyposis coli gene by the miR-135 family in colorectal cancer. Cancer Res. 68, 5795-5802.

Nardini,C., Masotti D., Yoon S., Macii E., Kuo MD., De Micheli G., Benini L. (2006) Mining gene sets for measuring similarities. In Proceedings of IEEE Symposium on Computers and Communications (ISCC). 227–232.

O'Donnell KA, Wentzel EA, Zeller KI, Dang CV, Mendell JT. (2005) c-Mycregulated microRNAs modulate E2F1 expression. *Nature* 435, 839-843.

Parkin DM, Bray F, Ferlay J, Pisani P. (2005) Global cancer statistics, CA Cancer J Clin 55, 74–108

Perwez Hussain S, Harris CC (2007) Inflammation and cancer: an ancient link with novel potentials. Int J Cancer 121, 2373–2380

Rapaport,D. (2005) How does the TOM complex mediate insertion of precursor proteins into the mitochondrial outer membrane? J. Cell Biol., 171, 419–423.

Reich D, Price AL, Patterson N. (2008) Principal component analysis of genetic data. Nat Genet. 40, 491-492.

Rodriguez-Bigas MA, Hoff P, Crane CH. (2006) Carcinoma of the colon and rectum. In: Kufe DW, Bast RC, Hait WN, et al, eds. Holland-Frei Cancer Medicine 7. 7th ed. Hamilton, Ontario: BC Decker Inc; 1369-1391.

Rosner,B. (2000) Fundamentals of Biostatistics. Duxbury, Pacific Grove, CA.

Ruvkun G (2001). "Molecular biology. Glimpses of a tiny RNA world". *Science* 294, 797–799.

Salomonis N, Hanspers K, Zambon AC, Vranizan K, Lawlor SC, Dahlquist KD, Doniger SW, Stuart J, Conklin BR, Pico AR. (2007) GenMAPP 2: new features and resources for pathway analysis. BMC Bioinformatics, 8, 217–229.

Scott GK, Mattie MD, Berger CE, Benz SC, Benz CC (2006) Rapid alteration of microRNA levels by histone deacetylase inhibition. Cancer Res 66, 1277–1281

Sevignani C, Calin GA, Siracusa LD, Croce CM (2006) Mammalian microRNAs: a small world for fine-tuning gene expression. Mamm Genome 17, 189–202

Sevignani C, Calin GA, Nnadi SC, Shimizu M, Davuluri RV, Hyslop T, Demant P, Croce CM, Siracusa LD. (2007) MicroRNA genes are frequently located near mouse cancer susceptibility loci. Proc Natl Acad Sci U S A. 104:8017-8022.

Silver D, Dhar A, Slocum M, Adams JG Jr, Shukla S. (1996) Role of platelet-activating factor in skeletal muscle ischemia-reperfusion injury. Adv. Exp. Med. Biol., 416, 217–221.

Sliwowska,J. et al. (2001) Cardioacceleratory action of tachykinin-related neuropeptides and proctolin in two coleopteran insect species. Peptides., 22, 209–217.

Sokal,R.R. and Rohlf,F.J. (2003) Biometry. Freeman, New York.

Steinberg SF, Robinson RB, Lieberman HB, Stern DM, Rosen MR. (1991) Thrombin modulates phosphoinositide metabolism, cytosolic calcium, and impulse initiation in the heart. Circ. Res., 68, 1216–1229 .

Stephen LJ, Fawkes AL, Verhoeve A, Lemke G, Brown A. (2007) A critical role for the EphA3 receptor tyrosine kinase in heart development. Dev. Biol., 302, 66–79.

Tavazoie SF, Alarcón C, Oskarsson T, Padua D, Wang Q, Bos PD, Gerald WL, Massagué J. (2008) Endogenous human microRNAs that suppress breast cancer metastasis. Nature. 451, 147-152.

Tazawa H, Tsuchiya N, Izumiya M, Nakagama H. (2007) Tumor-suppressive miR-34a induces senescence-like growth arrest through modulation of the E2F pathway in human colon cancer cells. Proc Natl Acad Sci U S A. 104, 15472-15477.

The Gene Ontology Consortium (2001) Creating the Gene Ontology resource: design and implementation. Genome Res., 11, 1425–1433.

The International HapMap Consortium. (2005) A haplotype map of the human genome. Nature, 437, 1299–1320.

Tomas,A., Futter,C. and Moss,S.E. (2004) Annexin 11 is required for midbody formation and completion of the terminal phase of cytokinesis. J. Cell Biol., 165, 813–822.

Touyz RM, He G, El Mabrouk M, Schiffrin EL. (2001) p38 Map kinase regulates vascular smooth muscle cell collagen synthesis by angiotensin II in SHR but not in WKY. Hypertension, 37, 574–580.

van't Veer,L.J., Dai,H., van de Vijver,M.J., He,Y.D., Hart,A.A., Mao,M., Peterse,H.L., van der Kooy,K., Marton,M.J., Witteveen,A.T. et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415, 530–536.

Vasudevan S, Tong Y, Steitz JA. (2007) Switching from repression to activation: microRNAs can up-regulate translation. Science 318, 1931-1934.

Volinia S, Calin GA, Liu CG, Ambs S, Cimmino A, Petrocca F, Visone R, Iorio M, Roldo C, Ferracin M, Prueitt RL, Yanaihara N, Lanza G, Scarpa A, Vecchione A, Negrini M, Harris CC, Croce CM. (2006) "A microRNA expression signature of human solid tumors defines cancer gene targets". Proc Natl Acad Sci U S A, 103, 2257-2261.

Walker AJ, Draeger A, Houssa B, van Blitterswijk WJ, Ohanian V, Ohanian J. (2001) Diacylglycerol kinase theta is translocated and phosphoinositide 3-kinase-dependently activated by noradrenaline but not angiotensin II in intact small arteries. Biochem. J., 353, 129–137.

Weber, M.J. (2005) New human and mouse microRNA genes found by homology search. FEBS J. 272, 59-73.

Welcsh,P.L., Lee,M.K., Gonzalez-Hernandez,R.M., Black,D.J., Mahadevappa,M., Swisher,E.M., Warrington,J.A. and King,M.C. (2002) BRCA1 transcriptionally regulates genes involved in breast tumorigenesis. Proc. Natl Acad. Sci. USA, 99, 7560–7565.

Wu Z., Irizarrt RA., Gentleman R., Martinez Murillo F., Spencer F. (2003) A Model Based Background Adjustment for Oligonucleotide Expression Arrays. Cobra http://www.bepress.com/jhubiostat/paper1/

Yang CM, Chien CS, Hsiao LD, Pan SL, Wang CC, Chiu CT, Lin CC. (2001) Mitogenic effect of oxidized low-density lipoprotein on vascular smooth muscle cells mediated by activation of Ras/Raf/MEK/MAPK pathway. Br. J. Pharmacol., 132, 1531–1541.

Zhang B, Kirov S, Snoddy J (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res., 33 (Web Server issue), W741–W748, doi:10.1093/nar/gki475.

APPENDIX A

Table A1 Table 1 List of human fragile regions (FRAs) containing miRNAs and UCRs. (Table from paper VIII).

FRA	Chr band	Clone/BAC/Gene/Marker	Start (bp)	End (bp)	Associated miRNAs	Associated UCRs
FRA1A	1p36	FGR	26322852	29322852	hsa-miR-801	
FRA1C	1p31	W72033	66784394	69784394		
FRA1F	1q21	AA007419	159812139	162812139	hsa-miR-556	uc.38;uc.39;uc.40
FRA1H	1q42.1	AC096642	216068384	219068384	hsa-miR-194-1;hsa-miR-215	
FRA2G	2q31	AC009475–AC093899	167706756	171680166		
FRA2I	2q33	FZD7	206835577	209842388		
FRA3B	3p14.2	D3S1287	62682121	65682121		
FRA4B	4q12	H23235	53358992	56358992		
FRA4C	4q31.1	ZNF330	140868420	143868420		
FRA5C	5q31.1	AC010238.6	133096692	136096692	hsa-miR-886	uc.173
FRA5E	5p14	AA701860	51317138	54317138	hsa-miR-581	
FRA6E	6q26	IGF2R; SLC22A3; PLG	159543260	162594328		
FRA6F	6q21	D6S1698–D6S1066	109922552	112922710		
FRA7E	7q21.11		79750000	82750000		
FRA7F	7q22	MUC3; TRIP6; DRA	105722132	108722244	hsa-miR-106b;hsa-miR-25;hsa-miR-93	

FRA7G	7q31.2	D7S486–D7S522	114182101	117360113		uc.228;uc.229;uc.230;uc.231
FRA7H	7q32.3	D7S786–D7S649	128564331	132008317	hsa-miR-182;hsa-miR-183;hsa-miR-29a;hsa-miR-29b-1;hsa-miR-335;hsa-miR-96	
FRA7I	7q35	AC004981, AC004911, AC006315	143130358	146130358		
FRA8B	8q22.1	AC004459	113928640	116928640		
FRA8E	8q24.1	EXT1	117534515	120534515		uc.246
FRA9D	9q22.1	NTRK2; GAS1	87250511	90250511	hsa-miR-7-1	
FRA9E	9q32–33.1	PAPPA; D9S1866–D9S177	116455904	119704422	hsa-miR-32;hsa-miR-455	
FRA10B	10q25.2	D10S597–D10S88	109720779	114308329		uc.310
FRA10C	10q21	N72215	71750934	74750934		
FRA10D	10q22.1	AC010163	78241247	81241247		
FRA11A	11q13.3	PC (distal=telom); D11S913–ACTN3	64192922	67587373	hsa-miR-192;hsa-miR-194-2;hsa-miR-612	uc.330
FRA11B	11q23.3	CBL2	117133133	120133133		
FRA12A	12q13.1	PCBP2(distal toFS)	50645950	53645950	hsa-miR-148b;hsa-miR-196a-2;hsa-miR-615	uc.338;uc.339;uc.340;uc.341;uc.342;uc.343;uc.344;uc.345
FRA13C	13q21.2	AL162376	71283245	74283245		
FRA15A	15q22	W58092	59645356	62645356	hsa-miR-190;hsa-miR-422a	
FRA16D	16q23.2	D16S518–D16S3029;	75191052	79303194		

		WWOX				
FRA16E	16p12.1	D16S299	26727578	29727578		
FRA17B	17q23.1	AC004686	53827242	56827242	hsa-miR-21;hsa-miR-301a;hsa-miR-454	uc.418;uc.419
FRA18A	18q12.2	D18S978	35092257	38092257	hsa-miR-924	
FRA22A	22q13	CSF2RB (proximal=centrom)	34156466	37156466	hsa-miR-658;hsa-miR-659;	uc.458
FRAXA	Xq27.31	DXS548	145111238	148111238	hsa-miR-506;hsa-miR-507;hsa-miR-508;hsa-miR-509-1;hsa-miR-509-2;hsa-miR-509-3;hsa-miR-510;hsa-miR-513-1;hsa-miR-513-2;hsa-miR-514-1;hsa-miR-514-2;hsa-miR-514-3	
FRAXB	Xp22.3	DXS1130–DXS237	5768190	9058235	hsa-miR-651	
FRAXE	Xq28	FMR2	145889831	149139865	hsa-miR-506;hsa-miR-507;hsa-miR-508;hsa-miR-509-1;hsa-miR-509-2;hsa-miR-509-3;hsa-miR-510;hsa-miR-513-1;hsa-miR-513-2;hsa-miR-514-1;hsa-miR-514-2;hsa-miR-514-3	
FRAXF	Xq28	FAM11A	147003696	150003696		

[a]Clone/BAC/Gene/Markers chromosome positions were determined by using http://www.Ensemble.org (release March 2008) and NCBI website (http://www.ncbi.nlm.nih.gov/). We considered by analogy with the length of a FRA that a distance of <2 Mb can define "close" vicinity (Calin et al. *2004*); in this way, association analysis considered 2 Mb as "close" vicinity. A Perl algorithm was implemented to find the association between FRAs and miRNAs as well as between FRAs and UCRs

Table A2 List of databases, bioinformatic/statistical analysis, and programs used to study miRNA functions, characteristics, associations, and correlations. (Table from paper VIII).

Reference	Database	Bioinformatic/statistical analysis	Program used
Calin et al. (2004)	MiRNA registry (http://www.microrna.sanger.ac.uk/registry/)	Random-effect Poisson regression models	BLAST
	PubMed (http://www.pubmed.com)	IRR (incidence rate ratio)	Perl
	GenBank (http://www.ncbi.nlm.nih.gov/Genbank/)		Bioperl Modules
	RNA folding program http://www.bioinfo.rpi.edu/applications/mfold/old/ma)		STATA 7.0
	Homo Sapiens Genome (http://www.ncbi.nlm.nih.gov)		
	FRA database (Calin et al. 2004)		
Zhang et al. (2006)	MiRNA Registry (http://www.microrna.sanger.ac.uk/registry/)	Target prediction	DIANA-MICRO T
			TARGET SCAN
			MIRANDA
			PICTAR
Calin et al. (2007)	UCR database (Calin et al. 2007)	ANOVA	GeneSping GX 7.3
			MatLab 6.5
	FRA database (Calin et al. 2004)	SAM (Significance Analysis of Microarrays)	
	CAGR database (Calin et al. 2004)	PAM (Prediction Analysis of	

		Microarrays)	
		Spearman rank correlation	
		FDR (False Discovery Rate)	
		Random-effect Poisson model	
		Binomial regression model	
		Fisher exact test	
Gaur et al. (2007)	Stanford MicroarrayDatabase (http://www.genome-www5.stanford.edu/)	Leave-one-out sensitivity analysis	TM4MeV v4 (Microarray Software Suite - http://www.tm4.org/mev.html)
	Sanger MirBase (http://www.microrna.sanger.ac.uk/)	Student's t-test	R (free software environment for statistical computing and graphics - http://www.r-project.org/)
		Spearman rank correlation	
		Multiscale bootstrapping analysis	
Makunin et al. (2007)	Retroviral Tagged Cancer Gene Database (http://www.rtcgd.ncifcrf.gov)	Bootstrap analysis	UCSC utilities
	miRNA Registry (http://www.microrna.sanger.ac.uk/registry/)		phastCons conservation scores
	Sanger miRBase (http://www.microrna.sanger.ac.uk/)		Microsoft Excel
	UCSC Genome Browser (http://www.genome.ucsc.edu)		
Sevignani et al. (2007)	MUSMIRSUS database (http://www.kimmelcancercenter.org/siracusa/musmirsus.htm)	Random effect Poisson regression model	BLAT

	Ensembl (http://www.ensembl.org)	IRR (Incidence Rate Ratio)	FirstEF (first-exon and promoter prediction program for human DNA - http://www.rulai.cshl.org/tools/FirstEF/)
	Mouse Genome Informatics Database (http://www.informatics.jax.org)		STATA 7.0
	PubMed (http://www.pubmed.com)		
	UCSC Genome Browser (http://www.genome.ucsc.edu)		

Table A3 Glossary of bioinformatics terms (Table from paper VIII).

Term	Definition
Agglomerative hierarchical clustering	The classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait, often proximity according to some defined distance measure.
axtAndBed, axtCalcMatrix	UCSC utilities that allow calculating genome-genome base-pair identity scores.
Bootstrap analysis	A general-purpose approach to statistical inference, falling within a broader class of resampling methods.
Circular binary segmentation algorithm	A modification of binary segmentation algorithm, used to translate noisy intensity measurements into regions of equal copy number.
FDR (False Discovery Rate)	A statistical method used in multiple-hypothesis testing to correct for multiple comparisons. In a list of rejected hypotheses, FDR controls the expected proportion of incorrectly rejected null hypotheses.
hgWiggle, -doStats	UCSC utility and flag that allow to fetch wiggle data from database or file.
Leave-one-out sensitivity analyses	Statistical diagnostics performed to investigate the validity and robustness of the meta-analysis applying an approach to subsets of the K studies, the leave-one-out method. The steps for the leave-one-out method are as follows:
	Remove the first of the K studies and conduct the meta-analysis on the remaining K – 1 studies
	Remove the second of the K studies and conduct the meta-analysis on the remaining K – 1 studies
	Continue this process until there are K distinct meta-analyses (each with K – 1 studies)
Permutation-generated reference distribution	A type of statistical significance test in which a reference distribution is obtained by calculating all possible values of the test statistic under rearrangements of the labels on the observed data points
Random-effect Poisson regression	A form of regression analysis used to model count data and contingency tables. Poisson regression assumes the response variable *Y* has a Poisson distribution and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters.

RefSeq	The Reference Sequence (RefSeq) database is an open access, annotated collection of publicly available nucleotide sequences (DNA, RNA) and their protein translations. This database is built by National Center for Biotechnology Information (NCBI).
Spearman rank correlation	A nonparametric measure of correlation; it assesses how well an arbitrary monotonic function could describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables.