



Gene Expression: From Microarrays to Functional Genomics

DARIO GRECO

**Institute of Biotechnology
and
Department of Biological and Environmental Sciences
Faculty of Biosciences
and
Viikki Graduate School in Biosciences
University of Helsinki**

Dissertationes bioscientiarum molecularium Universitatis Helsingiensis in Viikki

14/2009

GENE EXPRESSION: FROM MICROARRAYS TO FUNCTIONAL GENOMICS.

Dario Greco

Institute of Biotechnology
And
Department of Biological and Environmental Sciences
Faculty of Biosciences
And
Viikki Graduate School in Biosciences
University of Helsinki

Academic Dissertation in Genetics

To be presented for public examination with the permission of the Faculty of
Biosciences of the University of Helsinki in the Auditorium 1041 of the Biocenter 2,
Viikinkaari 5, Helsinki, on 28.05.2009 at 13:30.

Supervisor: Docent Petri Auvinen
Institute of Biotechnology
University of Helsinki
Helsinki, Finland

Reviewers: Professor Jukka Corander
Department of Mathematics
Åbo Akademi University
Turku, Finland

Docent Iris Hovatta
Research program of Molecular Neurology,
Faculty of Medicine
University of Helsinki
Helsinki, Finland

Opponent: Docent Outi Monni
Institute of Biomedicine
University of Helsinki
Helsinki, Finland

Custos: Professor Tapio Palva
Department of Biological and Environmental Sciences,
Faculty of Biosciences
University of Helsinki
Helsinki, Finland

ISSN 1795-7079

ISBN 978-952-10-5446-4 (paperback)

ISBN 978-952-10-5447-1 (PDF)

Email: dario.greco@helsinki.fi

On the cover: Leena Kleemola, *Untitled* (2005),
110x110 cm, acrylics on canvas.

Layout: Tinde Päivärinta

Printed: Helsinki University Press, Helsinki 2009

I don't believe there would be any science at all without intuition.
Rita Levi Montalcini

TABLE OF CONTENTS

LIST OF ORIGINAL ARTICLES

ABSTRACT

ABBREVIATIONS

1. INTRODUCTION	1
1.1 Functional Genomics	1
1.2 Methods to analyze gene expression	1
1.3 Regulation of gene expression	2
1.4 Gene expression in complex organisms	2
1.5 DNA microarrays	3
1.6 Experimental design	4
1.7 Microarray platforms	5
1.7.1 Agilent microarray technology.....	5
1.7.2 Affymetrix GeneChip technology	6
1.7.2.1 The mismatch probes.....	7
1.7.2.2 The annotation of the probes	7
1.7.2.3 Preprocessing of Affymetrix GeneChips.....	7
1.7.2.4 Complex tissues and probe pre-filtering.....	9
1.8 Microarray analysis of differential gene expression	10
1.8.1 Microarray functional analysis.....	10
1.8.2 Gene regulatory networks	11
1.9 Microarray meta-analysis	12
2. AIMS OF THE STUDY	12
3. METHODS	13
3.1 Microarray data collection from public repositories (III)	13
3.2 Microarray quality control (I, II, III, IV)	13
3.3 Affymetrix probes re-annotation (III, IV)	13
3.4 Affymetrix GeneChips preprocessing (I, III, IV)	14
3.5 Affymetrix GeneChips pre-filtering (I)	14
3.6 Agilent microarray preprocessing (II)	14
3.7 Differential gene expression analysis (I, II, IV)	14
3.8 Tissue-selective gene selection (III)	14
3.9 Microarray results functional analysis (I, III, IV)	15
3.10 Microarray functional global-testing (II)	15
3.11 Literature-based gene network analysis (III, IV)	15
3.12 Promoter computational analysis (III, IV)	15

4. RESULTS	16
4.1 Pre-filtering improves the reliability of Affymetrix GeneChip experiments in complex tissues as tested by qPCR (I)	16
4.2 Integrating global testing and gene-wise analysis in gene expression data (II)	16
4.3 Building a catalog of tissue-selective genes (III)	17
4.4 Gene expression as screening for characterizing embryonic mesencephalon and neuronal primary cultures (IV)	19
5. DISCUSSION	22
6. CONCLUSIONS	27
7. ACKNOWLEDGEMENTS	28
8. REFERENCES	30

LIST OF ORIGINAL ARTICLES

The thesis is based on the following articles, which are referred to in the text by their Roman numerals.

- I. **Greco D**, Leo D, di Porzio U, Perrone Capano C, Auvinen P. 2008. Pre-filtering improves reliability of Affymetrix GeneChips results when used to analyze gene expression in complex tissues. *Mol Cell Probes*. 22(2):115-21.
- II. Alvesalo J, **Greco D**, Leinonen M, Raitila T, Vuorela P, Auvinen P. 2008. Microarray analysis of a *Chlamydia pneumoniae*-infected human epithelial cell line by use of gene ontology hierarchy. *J Infect Dis*. 197(1):156-62.
- III. **Greco D**, Somervuo P, Di Lieto A, Raitila T, Nitsch L, Castrén E, Auvinen P. 2008. Physiology, pathology and relatedness of human tissues from gene expression meta-analysis. *PLoS ONE*. 3(4):e1880.
- IV. **Greco D**, Volpicelli F, Di Lieto A, Leo D, Perrone Capano C, Auvinen P, di Porzio U. Comparison of gene expression profile in embryonic mesencephalon and neuronal primary cultures. *Manuscript Submitted*.

These articles are reproduced with the permission of their copyright holders.

AUTHOR'S CONTRIBUTION TO EACH PUBLICATION

- I. DG has designed the microarray experiment and the PCR assays, performed all the computational analyses, and written the manuscript.
- II. DG has designed the microarray experiment, carried out all the computational analyses; he has also participated in the design of the other experiments and to writing the manuscript.
- III. DG has designed the study, collected the data, carried out all the analyses, and written the manuscript.
- IV. DG has designed the microarray experiment, carried out all the computational analyses, and contributed the design of the other experiments as well as writing the manuscript.

ABSTRACT

The time of the large sequencing projects has enabled unprecedented possibilities of investigating more complex aspects of living organisms. Among the high-throughput technologies based on the genomic sequences, the DNA microarrays are widely used for many purposes, including the measurement of the relative quantity of the messenger RNAs. However, the reliability of microarrays has been strongly doubted as robust analysis of the complex microarray output data has been developed only after the technology had already been spread in the community. An objective of this study consisted of increasing the performance of microarrays, and was measured by the successful validation of the results by independent techniques. To this end, emphasis has been given to the possibility of selecting candidate genes with remarkable biological significance within specific experimental design. Along with literature evidence, the re-annotation of the probes and model-based normalization algorithms were found to be beneficial when analyzing Affymetrix GeneChip data. Typically, the analysis of microarrays aims at selecting genes whose expression is significantly different in different conditions followed by grouping them in functional categories, enabling a biological interpretation of the results. Another approach investigates the global differences in the expression of functionally related groups of genes. Here, this technique has been effective in discovering patterns related to temporal changes during infection of human cells.

Another aspect explored in this thesis is related to the possibility of combining independent gene expression data for creating a catalog of genes that are selectively expressed in healthy human tissues. Not all the genes present in human cells are active; some involved in basic activities (named housekeeping genes) are expressed ubiquitously. Other genes (named tissue-selective genes) provide more specific functions and they are expressed preferably in certain cell types or tissues. Defining the tissue-selective genes is also important as these genes can cause disease with phenotype in the tissues where they are expressed. The hypothesis that gene expression could be used as a measure of the relatedness of the tissues has been also proved.

Microarray experiments provide long lists of candidate genes that are often difficult to interpret and prioritize. Extending the power of microarray results is possible by inferring the relationships of genes under certain conditions. Gene transcription is constantly regulated by the coordinated binding of proteins, named transcription factors, to specific portions of the its promoter sequence. In this study, the analysis of promoters from groups of candidate genes has been utilized for predicting gene networks and highlighting modules of transcription factors playing a central role in the regulation of their transcription. Specific modules have been found regulating the expression of genes selectively expressed in the hippocampus, an area of the brain having a central role in the Major Depression Disorder. Similarly, gene networks derived from microarray results have elucidated aspects of the development of the mesencephalon, another region of the brain involved in Parkinson Disease.

ABBREVIATIONS

ANOVA	analysis of variance
BAC	bacterial artificial chromosome
cDNA	complementary deoxyribonucleic acid
CGH	comparative genomic hybridization
ChIP	chromatin immuno-precipitation
CNS	central nervous system
cRNA	complementary ribonucleic acid
Cy3	cyanine 3
Cy5	cyanine 5
DLP	digital light processor
DMD	digital micromirror device
DNA	deoxyribonucleic acid
FDR	false discovery rate
GABA	gamma-amminobutyric acid
GEO	gene expression omnibus
GO	gene ontology
KEGG	Kioto encyclopedia of genes and genomes
MBEI	model-based expression index
MIAME	minimum information about a microarray experiment
MM	mismatch
NCBI	national center for biotechnology information
MesE11	mesencephalon at embryonic stage E11
MesPC	mesencephalon neuronal primary culture
mRNA	messenger ribonucleic acid
PCR	polymerase chain reaction
PDNN	position-dependent nearest neighbor
PM	perfect match
PPi	inorganic pyrophosphate
qPCR	quantitative polymerase chain reaction
RMA	robust multiarray average
RNA	ribonucleic acid
SAGE	serial analysis of gene expression
TF	transcription factor
TFBS	transcription factor binding site
TIGR	The Institute for Genomic Research
UCSC	University of California Santa Cruz

1. INTRODUCTION

All living organisms carry precise instructions in their genome concerning how they grow and function. Genomics is the field of biological sciences that aims to study and decode this genetic information. The birth of genomics is generally thought to coincide with the completion of the first entire genome, the 5,375 base pairs long Phage PHI-X174 genome sequence in 1977 (Sanger et al. 1977). By January 13th 2009, the genome sequence of about 1,400 prokaryotes, about 200 eukaryotes, and 31 mammals has been completed or drafted, as reported by the NCBI genome sequencing project statistics (<http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html>). However, the only function of the genomic DNA is storing the information and ensuring its accurate delivery from one generation to another. Complexity arises primarily from more intricate regulatory interactions among genes, their products, and the environment.

1.1 Functional Genomics

Making use of the vast amount of data produced by genomics is the main task of the functional genomics. While genomics, proteomics, and structural biology focus on static aspects of the molecules of life (e.g. sequences and structures of DNA or proteins), functional genomics attempts to study dynamic aspects such as gene transcription and its regulation, as well as the interaction of genes and their products.

Each inheritable unit of DNA, usually referred to as a gene, contains the information required to make RNAs and proteins; such molecules constitute each and every cell, determining their functionality as well as their ability to

survive. The access to the information stored in the DNA is constantly modulated by dynamic processes that influence the amount of RNA and proteins present in the cells. Both the RNA-coding and the protein-coding genes are used as a template for the synthesis of RNA molecules by a process named transcription; similarly, the protein-coding RNAs are used as templates for synthesizing proteins during translation. The new proteins are thereafter folded, chemically modified, and delivered to the cellular compartment where they function; alternatively, they are secreted outside from the cells. Secreted proteins can act on the same cells where they are produced, or on neighbor cells, or on very distant cells by traveling within the blood stream.

1.2 Methods to analyze gene expression

Classical low-throughput techniques for quantifying the products of gene transcription, the messenger RNAs (mRNAs), include northern blotting and Polymerase Chain Reaction (PCR) (Saiki et al. 1988). In the mid-1990s, high-throughput technologies allowed many genes to be assayed within the same experiment. It is possible to divide these techniques into hybridization-based and sequencing-based methods. To the first class belong the microarrays, where target cDNA or cRNA is hybridized to complementary probes of the genes of interest and the abundance of a given transcript is estimated from the hybridization intensity of the corresponding probes.

In the family of the sequencing-based methods, the Serial Analysis of Gene Expression (SAGE), and the so named

next generation sequencing methods are among the most popular ones. In SAGE, short fragments of 14-17 bp length (usually referred to as tags) obtained from the 3' end of RNA molecules are concatenated and sequenced to quantify the expression levels of the corresponding transcripts (Velculescu et al. 1995). More recently, new ultra-high-throughput sequencing technologies have become available, including the Roche 454 GS FLX (<http://www.454.com>), the Illumina/Solexa Genome Analyzer (<http://www.illumina.com>), and the Applied Biosystems SOLiD (<http://www.appliedbiosystems.com>) technologies. The 454 technology uses emulsion PCR for producing beads-linked individual DNA fragments (Tawfik and Griffiths 1998). After transferring the beads into a multi-well picotiter plate, a sequencing-by-synthesis pyrosequencing approach is used, in which the release of inorganic pyrophosphate (PPi) is measured by chemiluminescence (Ronaghi et al. 1996). In the Illumina Solexa system, single-stranded DNA fragments are attached to a solid surface at one end by the use of adapters; next, the molecules bend, hybridizing to complementary adapters and are bridge-amplified to produce large amounts of clonal copies. The templates are sequenced using a sequencing-by-synthesis procedure, in which reversible terminators with removable fluorescent moieties and special DNA polymerases are used. ABI SOLiD technology is based on the polony technique (Shendure et al. 2005) and sequencing-by-ligation approach. Similar to the Roche 454 system, the emulsion PCR amplification products (on small beads) are transferred onto a glass support where sequencing occurs by multiple rounds of hybridization and ligation of fluorescently marked dinucleotides.

1.3 Regulation of gene expression

Gene expression is accomplished by modulating the accessibility of the genomic DNA, transcription, and the stability of messenger RNAs. Some long-term regulations involve chemical (eg. methylation) and steric (supercoiling) modification of the DNA molecules (van der Maarel 2008). Other levels of regulation might involve a variety of modifications of the proteins that are constitutively bound to the genomic DNA molecules, such as histones (Svaren and Hörz 1996). Each transcriptional unit (may be formed by a single gene or groups of related genes) is surrounded by regulatory DNA sequences, enhancers and promoter sequences (Sipos and Gyurkovics 2005). Once a promoter is available for binding the RNA polymerase, transcription is primarily regulated by the binding of transcription factors (TF) to their specific binding sites (TFBSs). Usually, multiple TFs and co-factors bind simultaneously to the promoter, recruiting or enforcing the binding of the RNA polymerase at the start site of the transcription (TSS) (Ross and Gourse 2009). The relative order and spacing of these TFBSs within a module are often highly conserved through evolution, highlighting their importance in regulation (Seifert et al. 2005). This conservation can allow the usage of computational tools for identifying clusters of known TFBS rather than specific nucleotide sequences.

1.4 Gene expression in complex organisms

While a copy of the same DNA molecule carrying the information for all the RNAs and proteins is present in each cell of the multi-cellular organisms, only some genes (called housekeeping genes) are active

in all the cells, as they are essential for the basic cellular functions. Other genes, providing more specialized molecular functions, are expressed selectively in particular tissues or cell types, or for example, at a particular moment of the development. Tissue-selective gene expression can be addressed in the strict terms of genes whose expression is limited to one tissue or cell type, but there is evidence indicating that functionally related tissues share many expression patterns (Liang et al. 2006). Compared to the housekeeping genes, the tissue-selective genes are thought to be longer (Vinogradov 2004), to have a more complex structure (Castillo-Davis et al. 2002), a different nucleotidic composition (Vinogradov 2003), and lower substitution rates at non-synonymous sites (Duret and Mouchiroud 2000). In addition, the tissue-selective genes show faster evolution rates and they are more likely to be mutated in genetic diseases with Mendelian inheritance (Winter et al. 2004).

The identification of tissue-selective genes sharing coordinate regulation can provide hints about the mechanisms governing development, the maintenance of the physiological state, and the establishment of pathological conditions. Table 1 summarizes the results of several studies where microarrays have been used for investigating the selective expression patterns in healthy human tissues (Hsiao et al. 2001, Saito-Hisaminato et al. 2002, Shyamsundar et al. 2005, Yanai et al. 2005, Liang et al. 2006).

1.5 DNA microarrays

Since their first description (Schena et al. 1995), DNA microarrays have become a routine tool in many laboratories worldwide. DNA microarrays can be defined as ordered and large series of known nucleic acid fragments that are placed on a solid support and that can function as molecular detectors. Through

Table 1. Microarrays in tissue-selectivity studies.

Each column represents a study where microarrays have been used for investigating tissue-specific or tissue-selective expression patterns. In rows, information concerning: the number of tissues and genes analyzed; the percentage of genes found specific or selective calculated as (selective genes / tot genes) * 100; the microarray platform, the preprocessing algorithm, and the method for the selection of genes utilized.

	Hsiao et al. 2001	Saito-Hisaminato et al. 2002	Yanai et al. 2005	Shyamsundar et al. 2005	Liang et al. 2006
n. tissues	19	29	12	35	97
n. of genes analyzed	7,000	27,000	23,000	26,000	27,000
% of specific genes	21 %	17 %	35 %	15 %	14 %
Microarray platform	Affy HuGeneFL	cDNA-MA	Affy HGU95A-E	cDNA-MA	Affy HGU-133A
Pre-processing method	MAS5	bg-correction	MAS5	bg-correction	MAS5
Identification of tissue-specific pickup	Student's t-test	fold-change	ANOVA + tissue-specificity index	fold-change	Tukey-Kramer's HSD

hybridization, it is possible to identify and quantify many labeled RNA or DNA species at a time. Nowadays, the microarrays are used for a variety of different purposes including comparative genomics hybridization (CGH) (Oostlander et al. 2004), ChIP-on-CHIP (Nègre et al. 2006), genotyping (Hacia 1999), and microRNA quantification (Yin et al. 2008). However, their most popular application is still the large-scale gene expression analysis. Profiling gene expression in human samples has been important for defining the functional identity of the tissues and, consequently, for uncovering the genomic signatures in many pathological conditions. Moreover, the microarrays and other high-throughput approaches are also potentially very useful in studying human complex diseases in an unbiased (i.e. hypothesis-free) manner. The number of publications tagged by the word “microarray” according to PubMed was 411 in the period spanning from 1995 to 2000, compared to 27,926 from 2001 to 2008. However, as the number of publications reporting microarray experiments has constantly grown, their reliability has also been questioned (Kothapalli et al. 2002, Draghici et al. 2006). Similar to other high throughput technologies, microarrays are prone to many uncontrolled and unknown sources of variability affecting their reproducibility. A general lack of standardization can also represent obstacles towards full comparability of independent experiments. In order to address these issues, the Microarray Gene Expression Data Group (MGED group) proposed in 2001 (Brazma et al. 2001) guidelines referred to as MIAME (Minimum Information About a Microarray Experiment). It defines three levels of microarray data: i) the scanned images (raw data); ii) the quantitative

outputs from the image analysis; and iii) the quantitative output from the preprocessing. The minimum information about a published microarray experiment should always include information concerning: i) the experimental design; ii) the array design; iii) the samples used; iv) the hybridization procedures and parameters; v) the measurements; and vi) the normalization specification.

1.6 Experimental design

The design of microarray experiments is done, as for any other scientific experiment, balancing considerations such as skill, cost, equipment, and accuracy. The objective of experimental design is to make the analysis of the data and the interpretation of the results as simple and as powerful as possible. Several issues affect the microarray experimental design: i) the biological questions that the experiment is supposed to answer; ii) the meaning of the experiment with respect of the whole scientific project; iii) type of samples, amount, and complexity of the biological material; iv) the number of microarrays utilized for the experiment; v) the microarray platform utilized (Yang and Speed, 2002, Simon et al. 2002). As a general rule, a microarray experiment should be carried out only if it is feasible, given the type and the amount of resources available. It is also important to prioritize the biological objectives, as a design is usually able to answer only a limited number of questions with reasonable precision. A sensitive aspect of the experimental design is the number and the type of replicates used. The number of replicates largely depends on the desired magnitude of the gene expression differences as well as the noise level in the system. Different microarray technologies,

in fact, have different noise levels, and the only way to estimate the noise is to do adequate replicate hybridizations. There is substantial disagreement about whether to pool individual samples. In theory, if the gene expression variation among individuals is normally distributed, pooling individual samples results in smaller variance. In practice, the expression of most of the genes among individuals is not normal for a variety of biological and technical reasons (Pritchard et al. 2001). It has been argued that in small experiments, the inference for most genes is not adversely affected by pooling. On the other hand, pooling does not increase precision in larger experiments (Kendzioriski et al. 2005).

1.7 Microarray platforms

In gene expression microarrays, either synthetic oligonucleotides or cDNA fragments have been used as probes. Especially in the early years, cDNA libraries and Bacterial Artificial Chromosomes (BAC) sets have been the principal source of probe fragments (Holloway et al. 2002). Later, they have been almost completely replaced by oligonucleotides corresponding to known genes or transcripts. Because the oligonucleotides are much shorter than cDNAs, they allow more specificity but their base composition is likely to influence their performance (Kreil et al. 2006). Hence, an effective design is needed (Kreil et al. 2006). Probes are typically printed or synthesized on glass to allow visualization of the bound, fluorescently labeled targets. Glass slides have continued to be the favored solid support for immobilizing probes for reasons of availability, low fluorescence, transparency, high temperature resistance,

physical rigidity and the variety of surface chemical modifications possible (Affara 2003, Petersen and Kawasaki 2007).

The market of microarrays has changed markedly in the past few years as the price of commercial arrays has rapidly fallen. Affymetrix GeneChip arrays were increased in complexity and in the number of species represented. NimbleGen have described a technology for synthesizing microarrays containing about 200,000 features using a digital micromirror device (DMD or digital light processor – DLP) that creates digital masks to synthesize specific polymers (Nuwaysir et al. 2002). Febit has introduced a method that generates microarrays within a three-dimensional microstructure (Obermeier et al. 2003). Oligonucleotide probes are synthesized *in situ* via a light-activated process using a digital projector within the channels of a three-dimensional microfluidic reaction carrier. The three-dimensional microstructure contains, in total, four individual channel-like chambers or arrays, allowing eight array experiments to be run on a single carrier. Illumina introduced the BeadArray technology based on the random self-positioning of bead pools onto a patterned substrate (Michael et al. 1998). A decoding process is used for mapping the location of a specific bead type on the array. This is determined by serially hybridizing with fluorescently labeled complementary oligonucleotides. In this technology, the miniaturization is secured by adjusting the size of the beads and the pattern of the substrate; randomly assembled 300-nm diameter bead array is about 40,000 times higher than a typical spotted microarray.

1.7.1 Agilent microarray technology

Agilent produces microarrays by *in situ* inkjet printing of 60 nucleotides probes

(Hughes et al. 2001). The probe design relies on multiple up-to-date and publicly available sequence databases for a variety of organisms. For the *Homo sapiens* whole genome chipset, the probe design starts with the sequence comparison and the genome mapping of very well annotated sequences found in RefSeq (Pruitt et al. 2007), Ensembl (Flicek et al. 2008), UCSC GoldenPath (Kuhn et al. 2008) known genes and Incyte Foundation Full Length databases (Kronick 2004). Clusters of transcript sequences having sequence and genome overlap, namely GeneBins, are formed by using BLAT metrics (Kuhn et al. 2008). Additionally, a second GeneBin set is generated from more poorly annotated sequences from a variety of databases including Unigene (Sayers et al. 2009), the TIGR Tentative Human Consensus (Lee et al. 2005), Incyte Foundation partial transcripts and other GeneBank (Sayers et al. 2009) accessions. Any transcript sequences not mapping to the first set are included in the second round of GeneBins and additional consensus regions are defined. Once the final set of GeneBins is defined, the repetitive sequences are eliminated and a reference homology database is created, against which the probe sequences are compared to insure uniqueness.

Agilent technology also represents a versatile and budget choice as it allows production of custom arrays starting from any set of probes, the customization of the sample preparation protocols as well as the scanning and image analysis procedures. More recently, Agilent has also introduced the multiplex technology, where multiple sets of probes printed onto the same slide can be independently assayed (Wolber et al. 2006). The Agilent sample preparation protocol relies on direct labeling; one (Cy3-labeled) or two (Cy3- and Cy5-

labeled) samples are usually hybridized at a time (Wolber et al. 2006). Alternatively, indirect labeling techniques can also be successfully used. The electronic images produced during the scanning can be analyzed by the use of different algorithms and software. Agilent feature extraction methods aim at quantifying the feature signals and the background, performing the background subtraction, normalizing the dye effect, and computing the log ratios and their error estimates. Image segmentation and extraction of the feature intensities can also be performed with other software such as Axon GenePix (Paper II for an example). More recently, evidence supporting a simpler pre-processing strategy has been described, whereby the background correction step is skipped and intensity-dependent normalization is applied to the log-transformed signal intensities (Zahurak et al. 2007).

1.7.2 Affymetrix GeneChip technology

In the Affymetrix GeneChip technology, 25mer oligonucleotides probes are directly synthesized on the surface of the arrays by the use of photolithography technology (Lockhart et al. 1996). Multiple independent oligonucleotides (20, 16, or 11 couples according to the chipset) are designed *in silico*, from available sequence databases, to hybridize to different regions of the same transcript. In addition to each perfect match (PM) probes, oligonucleotides having a different base in the 13th position are also designed. This second type of probes, called mismatch (MM) probes, in principle, serve as controls for specific hybridization and they should facilitate the direct subtraction of background and cross-hybridization signals. All the probes for one transcript are referred to as probe set. Each probe set is formed by probe pairs, constituted by a PM probe with its own MM partner.

1.7.2.1 The mismatch probes

The mismatch probes should provide a way to quantify the hybridization noise of the PM partners, as the mutation in the 13th base should decrease their affinity to the target. However, about 30% of the MM probes show bigger signals than their respective PM partners suggesting that the measure obtained as the difference of the PM and MM is not reliable for many of the probes (Naef et al. 2002a, b). Moreover, the difference between the PM and MM intensities is affected by the nucleotide composition of the probes (Naef and Magnasco 2003). MM probes also introduce a systematic variability, which decreases the precision of expression measures (Binder and Preibish 2005). This suggests that subtracting the MM intensity from PM signal represents a major source of error, leading to fewer potentially biologically important candidate genes (Wang et al. 2007).

1.7.2.2 The annotation of the probes

In Affymetrix GeneChips, all the probes within a probe set should estimate the expression of the same gene. In recent years, however, evidence has shown that large portions of Affymetrix probes cross-hybridizing to multiple genes are non-specific or mis-targeted (Gautier et al. 2004b). Many probes do not even recognize their appropriate mRNA reference sequence (Mecham et al. 2004, Harbig et al. 2005). On the other hand, re-annotating the Affymetrix probes according to the RefSeq database improves the precision in estimating gene expression (Mecham et al. 2004). The Affymetrix probes have been aligned to different genomic databases such as UniGene, Refseq and Entrez Gene, and it was discovered that many probes are prone to mis-annotation issues (Dai et al.

2005). In addition, the genes identified as differentially expressed using the original and updated probe definition show only 50% overlap (Dai et al. 2005). More recently, it has been shown that updated definitions of the Affymetrix probes lead to more precise and accurate results as compared with the original annotations provided by the manufacturer (Sandberg and Larsson 2007). Several re-annotation methods are available allowing the probes to be mapped to genes, transcripts, or even exons sequences stored in public databases. However, exon-based re-annotation leads to decreased precision and increased variance in estimating gene expression, probably due to the smaller number of probes that map to each exon (Sandberg and Larsson 2007).

1.7.2.3 Preprocessing of Affymetrix GeneChips

The first task of the computational analysis of Affymetrix GeneChips is referred to as preprocessing and it consists of five main components: image analysis, background adjustment, normalization, summarization, and quality assessment. Image analysis allows converting the pixel intensities in the scanned images into the probe-level data. This process assigns one number to each probe cell (PM and MM). Background adjustment is essential, as part of the measured probe intensities is due to non-specific hybridization and the noise in the optical detection system. Observed intensities need to be adjusted to give accurate measurements of specific hybridizations. Without proper normalization, it is impossible to compare measurements from different arrays due to many sources of variation. These include sampling, different efficiencies of reverse transcription, labeling, hybridization reactions, physical problems of the arrays,

reagent batch effects, scanning, and laboratory conditions. Summarization is performed in order to obtain one number (usually referred to as the expression value) from the whole set of probes assayed for each transcript. At the end of preprocessing, an expression matrix carrying numerical information about the expression values per each gene/transcript (rows of the matrix) in each array (columns of the matrix) of the data set is obtained (Figure 1).

Affymetrix has developed a computational method for preprocessing, named MAS5 (<http://www.affymetrix.com>). First, the expression values are computed by averaging the PM-MM differences for all the probe pairs of the same probe set. Then, the expression

values are normalized by a scaling method. Already in 2001, Li and Wong (Li and Wong 2001a and b) reported that variation of a specific probe across the arrays is considerably smaller than the variance across probes within a probe set. Therefore, they concluded that one of the most critical issues in the analysis of the GeneChips is the way probe-specific effects are handled. They proposed a linear model, named Model-Based Expression Index (MBEI), where the probe-specific and the array-specific effect are estimated and used to calculate the expression values. In 2003, the robust multi-array average method (RMA) was also described (Irizarry et al. 2003). The RMA method allows robust estimation

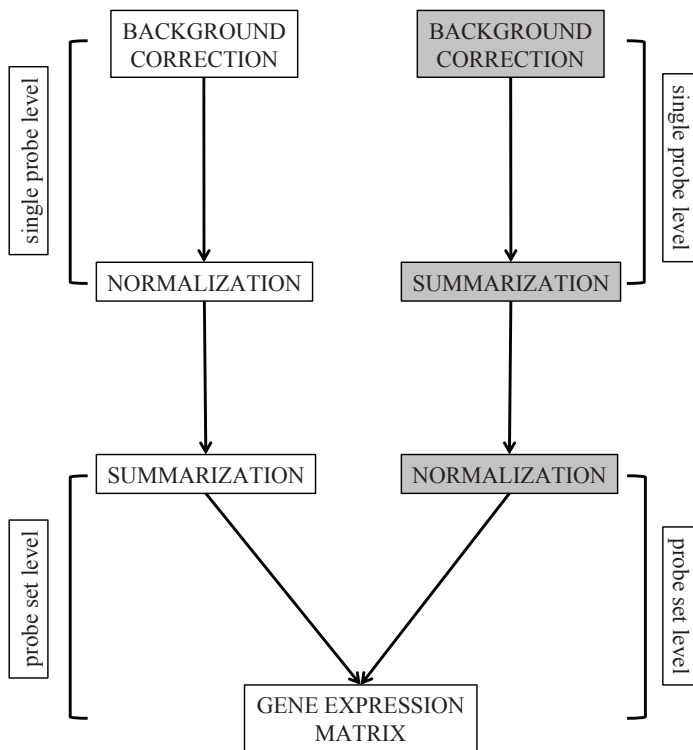


Figure 1. Affymetrix GeneChip preprocessing.

A schematic summary of the main steps of Affymetrix GeneChips preprocessing is shown. In some methods, such as RMA, the background correction and normalization are carried out at the single probe level; in other methods, such as MAS 5, the probes are summarized before the normalization step.

of inter-array variability. Similar to the MBEI, it uses information from multiple arrays for normalizing the dataset (through quantile normalization, the data are forced to have the same distribution) and fitting a linear model for each probe set across all the arrays of the dataset. RMA uses only the intensities from the PM probes for computing gene expression. Within the last few years, a multitude of model-based methods have been proposed. For instance, in the GCRMA algorithm, which is a direct evolution of the RMA, the nucleotide composition of the probes is taken into account (Wu and Irizarry 2004). Similarly, the PDNN algorithm estimates gene expression by using a free energy position-dependent nearest neighbor model based on PM sequences within each probe set (Zhang et al. 2003). Table 2 summarizes the features of the most popular methods.

This research field is still evolving and it is imaginable that new algorithms will allow more accurate gene expression estimations in the future. Several studies have compared the most popular preprocessing algorithms for Affymetrix GeneChips by using spike-in or dilution datasets, reporting that the model-based algorithms perform generally better than MAS5 (Irizarry et al. 2006). Elsewhere, the performance of preprocessing methodologies has been investigated in terms of the PCR validation rate (Qin et al. 2006).

1.7.2.4 Complex tissues and probe pre-filtering

Affymetrix GeneChips can detect cRNA species at very small concentrations. However, this has little value in gene expression detection in complex tissues, like the brain, which consists of specialized

Table 2. Affymetrix GeneChip preprocessing methods.

Each row summarizes the main features of the MAS5, MBEI, RMA, GCRMA, and PDNN preprocessing methods respectively.

Method	Citation	Background correction	Normalization	Summarization
MAS 5	Affymetrix 2002	Spatial background and MM are subtracted	Scale normalization	Robust average (Tukey biweight)
MBEI	Li and Wong 2001	MM are subtracted	Splines from a reference array and invariant set	Model assuming multiplicative probe-effect and additive error
RMA	Irizarry et al. 2003	Global correction from posterior mean given the observed PM	Quantile	Linear model including array and probe effects using median polish
GCRMA	Wu and Irizarry 2004	Probe specific correction using posterior mean of PM and MM; probe sequence used to predict model parameters	Quantile	Linear model including array and probe effects using median polish
PDNN	Zhang et al. 2003	Model with optical background, non-specific binding, and specific binding as additive components		

cells with variant transcriptional profiles. In practice, relatively high-abundance transcripts are reliably detected by GeneChips but a significant percentage of low-abundance transcripts are undetected or, in most of the cases, unreliably detected. As a result, the magnitude of expression changes found with microarrays is often modest and hard to separate from the experimental noise. In addition to producing normalized expression values, the preprocessing could also consider whether all the hybridizations of a single experiment are reliable. Methods that eliminate potentially unreliable data can help, beginning from the assumption that not all genes are expressed at levels that are either biologically significant or detectable by the Affymetrix technology in a particular tissue. Pre-filtering based on hybridization quality before the statistical evaluation of each transcript can aid in reducing the noise. Different methods have been used to pre-filter data to remove probe sets that are believed to be less reliable but the effects of such pre-filtering have rarely been analyzed (Wildhaber et al. 2003, Ryan et al. 2004, Stossi et al. 2004). Filtering by expression level (Modlich et al. 2004) aims to eliminate probe sets with signal close to background; the choice of how close to background is arbitrary. Removal of probe sets that are called “Absent” on all arrays has been reported (Ryan et al. 2004). Some use post-hoc methods by eliminating significant probe sets with low fold changes (Wildhaber et al. 2003). McClintick and Edenberg have filtered out probe sets that were not called Present by the MAS5 detection call in at least 50% of the samples in one treatment group (McClintick and Edenberg 2006). Others use combinations of these strategies (Perrier et al. 2004, Stossi et al. 2004, Aston et al. 2005, Tang et al. 2004).

1.8 Microarray analysis of differential gene expression

A microarray experiment typically aims to identify the relative differences between the biological conditions examined. The first computational techniques utilized for inferring the differential expression relied on the simple assumption that the reliability and, consequently, the significance would increase together with the magnitude in the gene expression. Accordingly, the fold changes calculated between samples served also as a significance cut-off. More strict statistical evaluation has been established and the number of methodological papers introducing novel statistical approaches has been increasing as the biological papers presenting microarray results. Usually, in gene-wise analyses, p-values are calculated for each gene present on the microarray by using the t-test or some other analytical strategies such as the ANOVA, which helps to estimate the contribution of experimental factors to the distribution of the measured gene expression. Next, a cut-off is found to separate the differentially expressed genes from the genes whose expression is not changed. This cut-off is usually based on a multiple testing criterion such as the Bonferroni or the false discovery rate (Benjamini and Hochberg 1995). Post-hoc corrections are also recommended because the number of genes tested is much bigger than the amount of samples replicated across two or more biological conditions.

1.8.1 Microarray functional analysis

A typical microarray experiment results in lists of differentially expressed genes. Long gene lists, however, cannot be considered the end point of the analysis. Rather, they have to be regarded as the starting point of a more meaningful interpretation,

whereby biological patterns are typically highlighted. By taking advantage of the increasing knowledge about the functions of the genes within the cells, it is also possible to infer the overall changes in terms of functions and processes. This essentially shifts the level of analysis from individual genes to sets of biologically related genes. The annotation terms are usually obtained from libraries such as Gene Ontology (Ashburner et al. 2000) or KEGG (Ogata et al. 1999). Metabolic pathways, though, are controlled to a large extent by protein-based events, having no direct implication to the levels of mRNA measured by microarray assays. Similarly, one can test whether the expression of genes sitting in specific portions of chromatin (i.e. cytobands or entire chromosome) are involved in certain experimental conditions. For any of the annotations used for grouping the genes, the terms are defined *a priori* and constructed independently from the experimental data. The most popular method starts from a list of differentially expressed genes and assesses whether a given gene set is overrepresented by using a test for independence in a contingency matrix (Khatri and Draghici 2005 for an overview). These methods imply the use of a strict significance cut-off for the differential expression of individual genes. Alternatively, one can test whether the ranked list of genes annotated in a given gene set differs from a uniform distribution by using the Kolmogorov-Smirnov test (Mootha et al. 2003). Other approaches do not compute the p-values per each gene, but start the analysis directly from the raw expression data. It has been proposed to test whether samples with similar expression profiles have similar class labels. This can be achieved by using logistic regression models (Goeman et al. 2004), ANOVA models (Mansmann and

Meister 2005), or a t-test after reducing the gene set to its first principal component (Tomfohr et al. 2005).

1.8.2 Gene regulatory networks

Increasing attention is being oriented to the inference of transcriptional regulatory networks based on high throughput gene expression screenings (Lee 2005, Sivachenko et al. 2007, Wang et al. 2007). These approaches aim to link gene expression data to the activity of transcription factors in cause-effect models (Goutsias and Lee 2007, Babu 2008). Fundamental to the idea of a gene network is the notion of modularity, according to which a complex system is built by combining simpler parts (Alon 2007). Modularity exists in a variety of biological contexts, including protein complexes, metabolic pathways, signaling pathways and transcriptional programs (Wagner et al. 2007). For transcriptional programs, for instance, modules are defined as sets of genes controlled by the same set of transcription factors under certain conditions. Learning the structures of networks based on biological data and estimating their parameters is a crucial step. This is accomplished by integrating *a priori* knowledge about the network structure based on assumptions about the function of a gene (Schlitt and Brazma 2006). Co-regulation of mammalian genes usually depends on sets of transcription factors that coordinately bind the promoter sequences and interact with each other (Werner 2007). Regulatory motif sequences within the promoter regions are organized into defined frameworks or modules of two or more transcription factor binding sites. Subsequent to the definition of frameworks, it is possible to scan large promoter sequences repositories for matches of such predefined modules.

1.9 Microarray meta-analysis

Despite their broad use, microarrays are still suffering a substantial lack of standardization levels that would easily allow a combination of independent experiments (Kuo et al. 2002, Järvinen et al. 2004). There is anyway an increasing need for integrating the massive amount of gene expression data that are continuously produced worldwide. This kind of integration would sensitively improve our knowledge of the complex events that take place during the embryonic development of tissues, during the genesis of diseases, or the mechanisms that modulate the response to drugs. In recent years, several attempts have been made in comparing and integrating high throughput gene expression experiments. Wang et al. observed that different microarray

platforms show good agreement both within and across laboratories when using the same RNA samples (Wang et al. 2005). On the other hand, the laboratory effect plays a more significant role than the platform effect (Wang et al. 2005). Severgnini et al. effectively compared gene expression data from similar microarray technologies, using identical sample preparation protocols and identical statistical analysis (Severgnini et al. 2006). Microarrays have also been collected for studying gene expression in human cancers (Kilpinen et al. 2008). There is evidence that one way to reliably combine microarray data is by matching the probes from different chipsets or platforms on the sequence base (Hwang et al. 2004, Carter et al. 2005, Stec et al. 2005, Ji et al. 2006).

2. AIMS OF THE STUDY

Due to the multi- and inter-disciplinary nature of this thesis, it is possible to divide its objectives in two orders: methodological and biological.

Methodological objectives:

- Establishing statistical frameworks for increasing the reproducibility of Affymetrix GeneChip experiments;
- Defining methods for reliably meta-analyzing independent Affymetrix GeneChip data sets;
- Extending microarray results to regulatory gene networks.

Biological objectives:

- Exploring gene expression patterns in human tissues and cell lines;
- Investigating the relationships of human tissues based on gene expression information;
- Evaluating gene expression in neuronal primary cultures and brain tissues for studying the developing brain.

3. METHODS

An overview of the methods used in the publications included in the thesis is shown in Table 3.

Table 3. Summary of the methods used in this thesis.

Each row corresponds to a particular method. The paper (I – IV) where the method is used is also reported. Each method is described in details in the following paragraphs.

Method	Paper
Microarray data collection from public repositories	III
Microarray quality control	I, II, III, IV
Affymetrix probes re-annotation	III, IV
Affymetrix GeneChips preprocessing	I, III, IV
Affymetrix GeneChips pre-filtering	I
Agilent microarray preprocessing	II
Differential gene expression analysis	I, II, IV
Tissue-selective gene selection	III
Microarray results functional analysis	I, III, IV
Microarray functional global-testing	II
Literature-based gene network analysis	III, IV
Promoter computational analysis	III, IV

3.1 Microarray data collection from public repositories (III)

Affymetrix (<http://www.affymetrix.com>) GeneChip raw data files (CEL files) were collected from the Gene Expression for Omnibus (GEO) public database (Edgar et al. 2002). Strict criteria for the data selection were applied: i) the experiments had been documented according to the MIAME protocol (Brazma et al. 2001); ii) the arrays had been hybridized to normal fetal or adult human tissues or cell types; iii) the specimens had been obtained from healthy subjects or from reference RNA samples; iv) the raw data files had been made available for download; v) all the samples had been hybridized to Affymetrix GeneChips chipset HGU-133A.

3.2 Microarray quality control (I, II, III, IV)

Affymetrix data (I, III, IV) were checked for quality by using the package *affy* (Gautier et al. 2004a) and *affyQCReport* (Parman and Halling 2008) for R (R Development Core Team 2008). Agilent (<http://www.agilent.com>) data (II) were checked for quality by using the R package *limma* (Smyth 2005).

3.3 Affymetrix probes re-annotation (III, IV)

Sequence-based re-annotation of the Affymetrix probes was applied. Each single oligonucleotide probe was re-annotated according to the *Homo sapiens*

release March 3, 2006 (III) and the *Rattus norvegicus* release June 28, 2006 (IV) Entrez Gene databases (Maglott et al. 2007). In Paper III, the probes were also re-annotated according to the RefSeq version 24 (Pruitt et al. 2007) and Ensemble version 42 gene databases (Flicek et al. 2008). R packages for the re-annotated Affymetrix chipset are available for download at http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/CDF_download.asp.

3.4 Affymetrix GeneChips preprocessing (I, III, IV)

CEL files were imported into R (R Development Core Team 2008) and preprocessed using the algorithm RMA (Irizarry et al. 2003) implemented in the BioConductor (Gentleman et al. 2004) package *affy*.

3.5 Affymetrix GeneChips pre-filtering (I)

Three different pre-filtering methods were applied to normalized Affymetrix GeneChip data. Pre-filtering based on the Affymetrix detection call (Liu et al. 2002): probe sets were retained if its detection call was equal to “Present” in at least 50% + 1 arrays in at least one group of biologically replicated arrays. Detection calls “Marginal” were converted to “Absent”. Pre-filtering based on the MBEI standard error (Li and Wong 2001a and b): probe sets were kept if its MBEI standard error was falling below the 95th percentile of the distribution of all the standard errors computed for each probe set across all the arrays of the experiment. Combinational pre-filter: both the detection call-based and the MBEI standard error-based pre-filters were applied.

3.6 Agilent microarray preprocessing (II)

Image segmentation as well as estimation of foreground and local background intensities for each feature was performed using Axon Genepix Pro version 6.0 (http://www.moleculardevices.com/pages/software/gn_genepix_pro.html). The data were then imported into R (R Development Core Team 2008) by using methods implemented in the package *limma* (Smyth 2005). Background-corrected intensities were normalized using the variance stabilization normalization (VSN) method (Huber et al. 2002).

3.7 Differential gene expression analysis (I, II, IV)

In paper I, a permutation-corrected t-test (Tusher et al. 2001) was used; probe sets with p-value < 0.01 after false discovery rate FDR correction were selected as differentially expressed. In paper II, genes with analysis of variance (ANOVA) p-value < 0.01 were considered. In paper IV, a moderated t-test and p-value cut-off of 0.001 after Benjamini Hockberg post-hoc correction were applied.

3.8 Tissue-selective gene selection (III)

RMA-normalized expression values were transformed so that the maximum value was set to 1 for each gene across the tissues; the method proposed by Yanai and collaborators (Yanai et al. 2005) is used as a gene-specific weight; the tissue-selectivity score per gene per tissue is then computed for each gene in each tissue separately as the transformed expression value by its specific weight.

3.9 Microarray results: functional analysis (I, III, IV)

In paper I and III, Fisher's exact test was used for screening the over-representation of gene ontology categories (Ashburner et al. 2000); p-value cut-offs of 0.05 and 0.01 were applied respectively for selecting significant families. In paper IV, the methods implemented in the DAVID gene annotation system were utilized with default parameters (Huang et al. 2007).

3.10 Microarray functional global-testing (II)

Global statistics implemented in the R package global test (Goeman et al. 2004) for R (R Development Core Team 2008) were applied to the normalized expression matrix in order to find gene ontology categories affected during *Chlamydia pneumoniae* infection. Gene ontology families showing a p-value < 0.01 after permutation correction were considered to be significant; for each of these, the genes showing the most significant differential expression were selected for further investigation.

3.11 Literature-based gene network analysis (III, IV)

Lists of candidate genes were imported into the software Genomatix Bibliosphere (<http://www.genomatix.de/products/BiblioSphere/>) in order to build networks. Two genes were connected in the graph if they appeared to be co-cited in the PubMed literature database (Wheeler et al. 2008), or if the consensus for a known transcription factor family was present in their promoter regions. In Bibliosphere, it is possible to highlight both consensus-based connections between the candidate genes, as well as the connection of the input genes with other transcription factors.

3.12 Promoter computational analysis (III, IV)

The transcription factors presenting an interesting topology within the literature-based gene network were selected; promoter regions of candidate genes presenting specific consensus sequences were retrieved using the software Genomatix Gene2Promoter (http://www.genomatix.de/online_help/helpeldorado/Gene2Promoter_Intro.html) and screened with the methods implemented in Genomatix FrameWorker (http://www.genomatix.de/online_help/help_gems/FrameWorker.html) in order to find common regulatory modules containing at least two transcription factor binding sites.

4. RESULTS

4.1 Pre-filtering improves the reliability of Affymetrix GeneChip experiments in complex tissues as tested by qPCR (I).

The effect of the treatment with the psycho-stimulant drug methylphenidate (MPH) was evaluated in male rats. Gene expression screening was carried out on the striatum of these animals by using Affymetrix GeneChips RAE-230A. Several pre-filtering methods of the normalized expression values were applied (Paper I, Figure 1) and a set of 85 biologically relevant genes were tested by qPCR. In particular, the genes chosen included those encoding post-synaptic density proteins (Yao et al. 2004, Elkins et al. 2003), neurotransmitter receptors (Sari 2004, Heidbreder et al. 2005), transcription factors (Guerriero et al. 2005), trophic factors (Castrén 2004), extra-cellular matrix proteins (McCracken et al. 2005), and synaptic vesicle release proteins (Kahlig et al. 2005), for their expression had already known to be related to drug abuse (Paper I, Table 2). The qPCR validation showed large agreement (~ 98%) with the microarray predictions after the detection call and MBEI standard error pre-filters, with exception of the gene *Bmpr1a* (qPCR-based t-test p-value = 0.31). None of the genes from the other analyses were validated (Paper I, Table 3).

4.2 Integrating global testing and gene-wise analysis in gene expression data (II).

Global testing was used for finding gene ontology classes containing at least 3 genes that significantly associated (p-value < 0.01) with *Chlamydia pneumoniae*

infection at different temporal stages. In this analysis, the p-value represented the probability of the differential global expression of all the genes associated to a given GO term at each time point as compared to all the others (Paper II, Table 2). The GO-wise and the gene-wise analyses were combined in this study for determining the candidate genes to be considered for further investigation (Paper II, Figure 1). At 12 hours time point the GO term “DNA modification”, possibly related to the manipulation of gene expression of the host by the *Chlamydia pneumoniae*, was globally induced; from this group, the gene *vFOS* was selected. During all the stages of the experiment, the expression of several steroid-related categories went through an overall modification; the gene *NR4A1* was chosen from the “steroid hormone receptor activity”. Similarly, the gene *DKK1* was picked up as a member of the GO class “negative regulation of the WNT signaling pathway”, which was drastically induced after 12 hours and repressed after 72 hours of infection. Finally, the gene *CYR61* was selected from the functional group “Insulin-like growth factor binding activity”. In addition, 6 genes, namely *EGR1*, *FLJ32065*, *EMP1*, *IGFBP1*, *ACHE*, *FLJ23356*, were also selected as showing notable induction in the gene-wise analysis, creating a group of 10 candidate genes (Paper II, Table 3). After qPCR validation of the selected genes, 4 of them were successfully silenced with corresponding siRNAs (Paper II, Table 4). The silencing of the genes *EGR1* or *DKK1* was capable of reducing the amount of *Chlamydia pneumoniae* by more than 25% (Paper II, Table 5).

4.3 Building a catalog of tissue-selective genes (III).

The pipeline designed for identifying the tissue-selective genes (Paper III) consists of several consecutive steps (Figure 2).

A total of 4,985 gene-tissue pairs, corresponding to 1,601 unique genes, were considered as expressed in a tissue-selective manner after permutation testing (Paper III, File S1, Table 0.1). Significant gene-tissue pairs were found in 77 out of 78 tissues analyzed, with the exception of the superior cervical ganglion. About 35% of the 1,601 genes were selectively expressed in one tissue, 20% in two, 13% in three; 10% of the tissue selective genes were expressed in six or more tissues (Figure 3).

The majority of the tissue-selective genes shared by ten or more tissues were expressed in neural system tissues. The greatest part of the tissue-selective genes were found in the immune system (32%), followed by central and peripheral nervous system (17%), muscles (15%), and reproductive organs (9%); altogether, the other categories accounted for 27% of the selective genes (Figure 4).

By using the normalized expression of the 1,601 genes, the tissues could be successfully segregated by hierarchical clustering (Paper III, File S3, Figure 2), principal component analysis (Paper III, File S3, Figure 4), and curvilinear component analysis (Paper III, S3, Figure 6).

The tissue-selective genes represented many biological and molecular themes, as they could be significantly annotated in many gene ontology terms (Paper III, File S1, Tables 0.2, 0.3, and 0.4). Nineteen percent of the tissue selective genes were involved in signal transduction, 16% in development, and 14% in immune response. Moreover, about 18% of these genes coded for secreted proteins, and 8% for receptors. When the selective genes in each tissue were annotated, they were able to depict the main known physiological traits, for instance, the liver-selective genes (Paper III, File S1, Table 44.2) or the testis-selective genes (Paper III, File S1, Table 55.2). The 1,601 tissue-selective genes were enriched in disease genes, for they were associated with 361 human Mendelian disorders (Paper III, File S1, Table 0.5). In many cases, tissue-

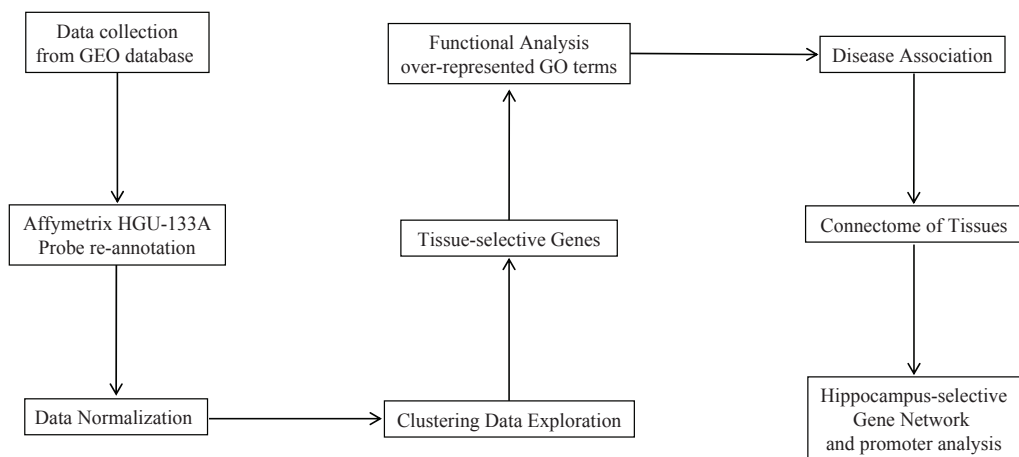


Figure 2. Analytical flowchart of paper III.

Each box represents an analytical step used in the paper III.

selective genes were found to be related to pathologies having strong impact on the tissues from where they were found to be selectively expressed. This was, for instance, the case for numerous muscle-selective genes linked to myopathies, or gland-selective genes linked to endocrine system and metabolic disorders. The fetal

heart-selective GATA4 and NKX2.5 had been associated with heart malformations, such as tetralogy of Fallot and atrial septal defects (Goldmuntz et al. 2001, Hirayama-Yamada et al. 2005).

About 65% of the 1,601 tissue selective genes were found in two or more tissues. Hence, investigating the

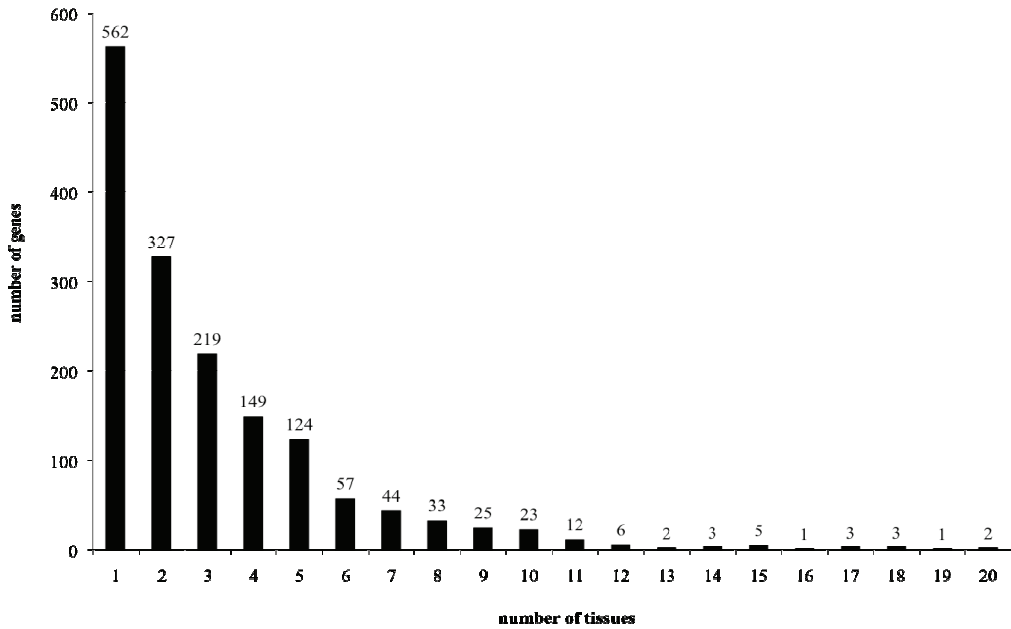


Figure 3. Distribution of the tissue-selective genes.

In x axis, the number of tissues sharing the expression of selective genes; in y axis, the number of genes in each category.

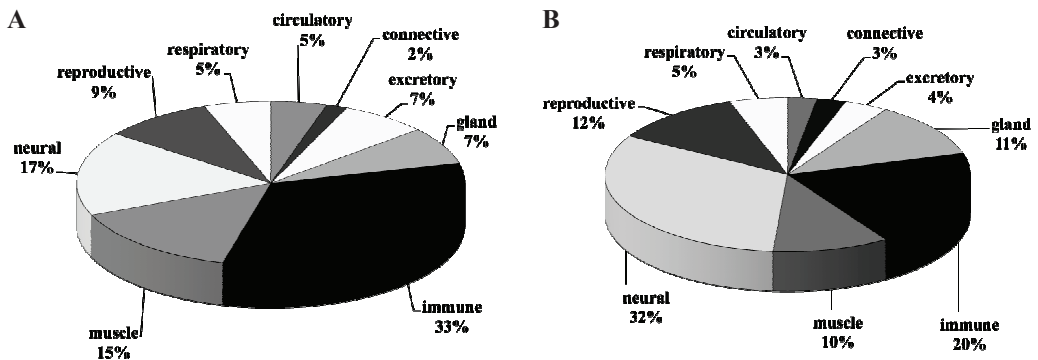


Figure 4. Tissue representation.

In A: the distribution of tissue-selective genes in groups of related tissues. In B: the groups of tissues analyzed in paper III.

relationship of tissues based on the amount of tissue selective genes shared was relevant. Networks of tissues (namely connectome, in Paper III) were built, where each node represented an analyzed tissue and the number of selective genes shared by two or more tissues formed the edges. The number of edges in the graphs was computed as the function of the number of shared genes. Thus, three cut-off values of at least 30, at least 20, and at least 5 shared genes were selected as a representation of different degrees of relatedness of tissues (Paper III, Figure 1). Four main results were obtained from the connectome analysis: i) central nervous system, immune, and testis tissues showed very distinct expression signatures, as they formed tight intra-connections already at 30 sharing genes cut-off. Testis tissues did not join any other tissues until the connectivity cut-off was lowered to 5 genes; ii) amygdala, which is thought to be a collection of adjacent cell groups within the forebrain, was located in the center of a CNS tissues network, sharing expression patterns with anatomically neighbor areas; iii) tonsil, an immune organ with a myoepithelial histological structure, bridged the networks of the immune tissues and the muscles; iv) the hippocampus, the fetal brain, and the olfactory bulb showed an interesting topology for they were placed at the interface between the nervous tissues and other tissues where active cell replication is known to take place. Neurogenesis is present in these three neural tissues. Therefore, the hippocampus-selective genes were further investigated in search of genetic networks that would underlie physiological functions of the hippocampus.

A network of the hippocampus-selective genes was modeled based on

their literature co-citation as well as the presence of consensus sequence motifs for specific transcription factor families within their regulatory regions. The transcription factor NF κ B, which was not selectively expressed in hippocampus, was found to potentially bind the promoters of several hippocampus selective genes (Paper III, Figure S1). Extensive analysis of the regulatory regions of the NF κ B interactors revealed the presence of a conserved transcriptional module composed by E2F and NF κ B transcription factor families (Paper III, Figure S2). Hence, the E2F-NF κ B module was found in an independent set of 1,901 human promoter sequences. The functional annotation of these genes, according to the gene ontology system, highlighted biological themes such as nervous system development, cell adhesion, and tyrosine kinase receptor signaling (Paper III, File S2).

4.4 Gene expression screening for characterizing embryonic mesencephalon and neuronal primary cultures (IV).

Three sequence-based re-annotations, based on the Entrez Gene (Maglott et al. 2007), RefSeq (Pruitt et al. 2007), and Ensembl gene (Flicek et al. 2008) databases, were used for alternatively grouping the probes of the *Rattus norvegicus* Affymetrix ChipSet RAE230A. These re-annotated independent data sets were then normalized and differential expression was assayed by moderated t-test (Paper IV, Table 1). The lists of differentially expressed genes were then analyzed using the DAVID database (Huang et al. 2007) and it was found that 425 differentially expressed unique genes were shared between the three annotations

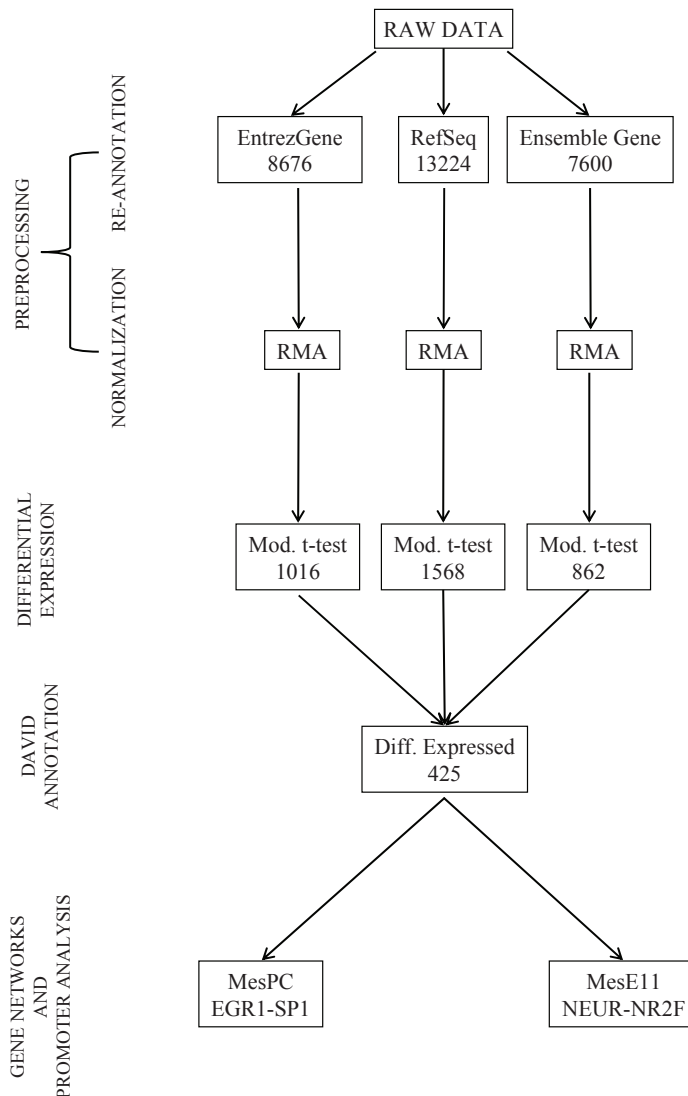


Figure 5. Analytical flowchart of paper IV.

Each box represents an analytical step used in the paper IV. When reported, the numbers represent the number of genes present at the specific step.

(Figure 5, and Paper IV, Figure 1 and Table ST1).

A total of 268 genes were found to be over-expressed in mesencephalon primary cultures (MesPC), representing the functional families of development, lipid metabolism, extracellular matrix, and mitochondrion (Paper IV, Table ST2). Analogously, 157 genes were significantly

associated with E11.5 mesencephalon (MesE11), covering functions such as synaptic transmission, nervous system development, neurogenesis, and ion channels (Paper IV, Table ST3).

A number of promoters of MesPC genes showed a binding site for the transcription factor Egr1. Further investigation highlighted the module

composed of binding sites for Egr1 and Sp1 as potentially involved in their expression regulation. The Egr1-Sp1 module was also found in promoters of other *Rattus norvegicus* genes involved in neuron differentiation and neurogenesis (Paper IV, File SR1).

As for the MesE11-genes, co-citation based network analysis and promoter analysis revealed a role of the transcription factor families Neur and

Nr2f in modulating the expression of many MesE11-genes. The genes Neurod3 (NEUR family) and Nr2f2 (Nr2f family) were found to be significantly over-expressed in MesE11. Other *Rattus norvegicus* promoters, covering functional families such as dopamine metabolism, synaptic transmission, and development, showed consensus for the Neur-Nr2f module.

5. DISCUSSION

This is the 15th year since the first microarray publication (Schena et al. 1995). Since then, the microarrays became a technique for large gene expression screening worldwide; increasing number of articles report gene expression microarray results in a variety of organisms and a multitude of experimental conditions. Despite their popularity, there has been increasing skepticism concerning the microarrays, determining the paradoxical situation that more scientists choose microarrays for their projects and at the same time more scientists struggle for publishing their microarray results. More specifically, several studies have highlighted major issues of reproducibility and predictability of microarray experiments (Tan et al. 2003, Kawasaki 2006, Walker and Hughes 2008, Ioannidis et al. 2009). Efforts have been made for increasing the reliability of microarray experiments, including increasing the sample size (Ein-Dor et al. 2006), improving the computational analysis of the data (Tilstone 2003, Allison et al. 2006, Jafari and Azuaje 2006), updating the probe annotation (Taylor et al. 2001, Dai et al. 2005, Carter et al. 2005) and design (Mecham et al. 2004), and standardizing manufacturing processes (Tan et al. 2003) as well as the sample preparation and hybridization procedures (Vartanian et al. 2009).

Microarrays have developed very rapidly along with the progressive completion of the genome sequencing projects but only recently the community focused on questions inherent to the computational analysis of the microarray data. When statistics entered into the microarray field, a peculiar split in the literature has happened: statisticians,

mathematicians, engineer and computer scientists have been proposing better and more sophisticated methods for analyzing microarray data, but their job has rarely been able to influence the way of working of the biologists. Rather, the microarray users have often entrusted their data to easy-to-use graphic-interfaced software that not always ensure adequate levels of strictness and customizability. Alternatively, some biologists have assigned the responsibility of the analysis to theoretical mathematicians or statisticians, assuring very high numerical reliability but scarce biological interpretability. Another reason for the increased skepticism towards microarrays is related to the strategic management of projects where microarrays have been used. Because of their complexity and costs, microarrays have been often regarded as the final step of longer projects. However, the real essence of large-scale and non-quantitative screening methods would naturally place the microarray experiments in the beginning of more articulated research projects. Whenever microarrays are used, problems also arise from the fact that usually too big and too much complex output is generated, making the interpretation of the general picture appearing in the results very difficult (Slonim 2002). For these and more practical motives related to budget issues, post-array work plan is extremely complicated, and often it abruptly stops at the validation of a handful of genes by the means of other methods such as PCR (Holland 2002, Czechowski et al. 2004). Owing to the costs associated with independent verifications, in most papers only a few genes (typically less than 20) are validated, including the

differentially expressed genes that are also widely studied and those with well-agreed sequences (Larkin et al. 2005, Paper II, Paper IV). When working with standardized microarray protocols and platforms, the validation rate can also be very high (Paper I). Regardless of the method used for testing the differential expression, the genes to be further investigated should be selected by prioritizing their potential biological meaning (Paper I and Paper II). In this thesis, much effort has been given to combining rigorous statistics with high biological interpretability of the results. In the papers presented here, people with different expertise and disciplinary backgrounds have independently screened the lists of differentially expressed genes, ensuring unbiased selection of genes to be validated by PCR. Additionally, when designing PCR primers, taking into account the position of the microarray probes onto the gene sequence has proven to be beneficial in terms of validation rates, despite the microarray platform utilized (Paper I, Paper II, Adriani et al. 2006, Volpicelli et al. 2007, Consales et al. 2007, Kivi et al. 2008). Post-array work is not always limited to the independent validation of the microarray results. In paper II, a gene silencing approach has been utilized in order to find genes whose expression would be essential for the replication of *Chlamydia pneumoniae*. Elsewhere, the information concerning the transcripts levels has been integrated with protein quantification and functional assays (Kivi et al. 2008).

Microarray data are expensive and time consuming to generate, nonetheless, they are rarely fully mined for their information content. Soon after the introduction of microarrays, it has been clear of the importance of sharing data within the community (Stoeckert et al.

2002), as well as the need of standardizing their description (Brazma et al. 2001). Additional motivations for public archiving of the microarray data has been to avoid duplicating experiments, as well as to re-analyze using new and more efficient algorithms. Several repositories offer the possibility to store and consult microarray data, including Gene Expression Omnibus (Barrett et al. 2005), ArrayExpress (Parkinson et al. 2005), Stanford Microarray Database (Ball et al. 2005), oncoMine (Rhodes et al. 2004), Celsius (Day et al. 2007), and ArrayWiki (Stokes et al. 2008). GEO and ArrayExpress contain data of about 280,000 and about 235,000 individual hybridizations respectively.

In Paper III, gene expression obtained by using human HGU-133A Affymetrix GeneChips was retrieved from the GEO database. Only a small part of the available data has been included in the analysis, as many experiments were stored in the SOFT format (Barrett et al. 2005), where only the normalized gene expression matrix had been made available. For these, the re-annotation of the probes is not possible; in addition, many experiments had been normalized using outdated methods no longer considered reliable. The quality of any meta-analysis depends on the underlying data. For this reason, strict pre-selection criteria of the experiments as well as rigorous quality assessment has been preferred over collecting as many hybridizations as possible to include in the study. In fact, 20-50% of all the HGU-133A arrays present in GEO and ArrayExpress databases have been reported to be of insufficient quality and should not be considered for inclusion in any meta-data set (Larsson et al. 2006). Finally, 195 arrays from six data sets have been selected for further investigation.

There are two general approaches for integrating microarray studies: i) the comparative analysis of the results published in different studies; and ii) the comprehensive re-analysis by merging primary data from different experiments. Comparison of aging and cellular senescence microarray studies has demonstrated that the expression patterns of genes involved in cellular senescence were similar to those of aging in mice but not in humans (Wennmalm et al. 2005). Similarly, microarray results from several studies have been integrated to identify common host transcriptional responses to pathogens (Jenner and Young 2005). The comprehensive re-analysis approach can be conceptually divided into three main groups, according to their primary purpose: i) summarizing meta-analysis, where the investigators try to identify more accurate results associated to a given phenotype; ii) hypothesis-driven meta-analysis, needed when a biological question can be answered with the integration of data; and iii) exploratory meta-analysis, where large sets of data are used to find previously uncharacterized gene expression patterns (Larsson et al. 2006). Additional biases in microarray meta-analysis are represented by the laboratory-effect (Irizarry et al. 2005), the platform-effect (Carter et al. 2005) and the species-effect (McCarroll et al. 2004). In this study, the platform-effect and the species-effect have not influenced the analysis, for only arrays from the same chipset have been considered. In Affymetrix GeneChips, the hybridization thermodynamics has been modeled considering some probe-specific and array-specific effects (Li and Wong 2001a and b, Irizarry et al. 2003). The probe-effect is balanced by re-annotating the probes, (Dai et al. 2005), as not only are they re-arranged within more appropriate probe

sets, but all the probes that potentially bias the hybridization are eliminated. On the other hand, the standardization of the GeneChip manufacturing and the consistency of the laboratory procedures have drastically diminished the array-specific effect. Therefore, the model-based algorithms typically used to normalize single experiments can largely deal with the laboratory-effect, similar to what they do with the array-specific effect. In these studies, the RMA algorithm (Irizarry et al. 2003) has been successfully utilized for normalizing arrays from different experiments. The RMA particularly allows robust estimation of inter-array variability. It uses information from multiple arrays for normalizing, through the quantile method (Bolstad et al. 2003), gene expression. More recently, large groups of independent GeneChips have been used as training sets for increasing the accuracy of the RMA-based gene expression estimation (Katz et al. 2006).

A group of genes, called housekeeping genes, are virtually expressed in all tissues to maintain the basic cellular functions, whereas the tissue-selective genes show differential expression patterns among the tissues and provide specialized functions that distinguish the tissues from each other. Although largely ubiquitous, the housekeeping genes are far from being constantly expressed, their expression levels may vary significantly between different tissues (Thellin et al. 1999, Lee et al. 2002, Barber et al. 2005). The estimation of the number of housekeeping genes is still debated (Warrington et al. 2000, Hsiao et al. 2001, Zhu et al. 2008). On the other hand, big interest is also given to the definition of the tissue-selective group of genes, and the new high-throughput technologies have represented in recent years valuable

tools for large-scale screening of gene expression across multiple tissues and cell types (Hsiao et al. 2001, Saito-Hisaminato et al. 2002, Shyamsundar et al. 2005, Yanai et al. 2005, Liang et al. 2006, Paper III). Housekeeping genes are thought to evolve slowly (Winter et al. 2004), and the slowly evolving genes are generally highly expressed (Drummond et al. 2005). Therefore, the importance of focusing on genes with middle-range and low expression has been pointed out (Yanai et al. 2005, Paper III). The housekeeping genes have also been reported to be under-represented among the disease genes due to a higher chance of embryonic lethality when mutated (Winter et al. 2004). Likewise, many human diseases caused by genetic defects are usually highly tissue-selective as well as disease genes are inclined to encode non-hub proteins in the protein network (Goh et al. 2007). In this study, the tissue-selective genes were found to be highly associated with human diseases. Additionally, an evident tendency emerged for disease genes to be selectively expressed in the tissues where their defects are described to cause pathology. Similar observations have been made elsewhere (Lage et al. 2008). Tissue-selectivity can be addressed in strict terms of considering only the genes whose expression is limited to a specific tissue, in which case it is referred to as tissue-specificity. However, tissues involved in the same functions or having common developmental origin share also numerous expression patterns.

Here, attention has been given to those genes whose expression is enriched in one or more similar tissues (Liang et al. 2006, Zhu et al. 2008, Paper III). Further investigation has been concentrated on the relatedness of tissues that shared selective expression patterns. The central nervous

system, testis, immune, and muscle tissues, showed a high grade of intra-relatedness, as these groups were connected already at very high degrees in the tissue network. However, the testis tissues have showed the most peculiar expression patterns, as their sub-network has remained isolated to all the other tissues also at medium and low degrees. Within the sub-network of CNS tissues, the amygdala, which originates from neurons migrating from different portions of the brain (Swanson and Petrovich 1998), presented high relatedness and central position. At a lower degree of relatedness, the hippocampus, together with a few other CNS structures, shared numerous expression patterns with other tissues where active cell replication is reported under certain conditions such as the liver and some fetal structures. Production of new neurons is thought to be possible in the adult hippocampus (Jacobs et al. 2000).

High-throughput technologies are only merely capable of cataloguing biological events in a relatively flat format, rather than providing a frame of deeper understanding or a broad key of interpretation. Alongside, microarray experiments provide only estimation of the transcript levels without directly revealing any possible regulatory mechanisms (Werner 2007). Regulatory networks provide a good way to represent the apparatuses that control the molecular processes of a living cell (Werner and Nelson 2006). Therefore, it has been proven that coordinated sets of transcription factor binding sites retrieved from the promoter sequences of related genes can provide an interpretative framework for the changes observed in gene expression (Pilpel et al. 2001). The relationships of genes can be inferred from the gene expression data in combination

with multiple independent evidence (Seifert et al. 2005). A similar approach has been considered in Papers III and IV of this thesis. In Paper III, the hippocampus-selective genes have been interconnected based on literature-based evidence, whilst in Paper IV, the genes found over-expressed in embryo mesencephalon or in the primary cell cultures derived from it have been the object of investigation. In both cases, additional links between the genes have been drawn according to the presence of TFBS for specific transcription factor families, enabling the discovery of regulatory relationships not described before. Visual inspection of the gene networks allows selecting sub-networks of genes whose promoters show possible co-regulation. Next, promoter sequence alignment can help in finding common modules constituted of multiple TFBSs (Cohen et al. 2006). In Paper III, the E2F-NFKB module has emerged as possibly regulating the expression of several hippocampus-selective genes as well as additional genes found selectively expressed in other CNS regions extensively studied for neurogenesis (Gould 2007). Transcription factors of the family E2F have a well-established role regulating gene expression during the cell cycle (Attwooll et al. 2004) and silencing several S-phase genes in differentiated neurons (Liu et al. 2005). Additionally, E2F binding sites have been identified in promoters of genes involved

in hippocampal development (Dabrowski et al. 2006). Within the CNS, NFKB genes have been reported to play a crucial role in synaptic plasticity, neuroprotection, as well as in learning and memory (Mémét 2006). They have also reported to be expressed in areas of active neurogenesis (Denis-Donini et al. 2005). Protein members of the E2F and NFKB families have been also reported to physically interact and cooperate in regulating the expression of common effectors (Lim et al. 2007). Likewise, in Paper IV the novel modules EGRF-SP1F and NEUR-NR2F have been identified in *Rattus norvegicus* dopaminergic neuronal primary cultures and embryonic mesencephalon respectively. EGR transcription factors have been described as having a role in chronic CNS diseases (Beckmann and Wilce 1997), as well as being involved in several neurophysiological aspects (Swiatek and Gridley 1993, Topiko et al. 1994, Jones et al. 2001, Li et al. 2007). Depolarization positively regulates the transcription of *Egr1*, indicating that this gene functions in neuronal differentiation following electrical stimuli. Neurogenic helix-loop-helix transcription factor family NEUR play a pivotal role in Nurr1-induced dopaminergic neuronal differentiation (Park et al. 2006). NR2F genes seem also involved in GABAergic interneurons migration during the development of the brain (Tripodi et al. 2004).

6. CONCLUSIONS

The completion of the human genome sequence and several model organisms has shown that complexity arises primarily from more complex regulatory interactions among genes, their products, and the environment. At the same time, high-throughput technologies delineate the location, abundance, state and interactions of genes and their products with increasing resolution in terms of cell types, time points and conditions. Among them, the microarrays allow investigating the function of many genes at once, thereby providing an assay of the transcriptional status of cells or tissues in a wide variety of physiological or pathological situations. At the same time, new ultra-high-throughput sequencing methods enable measurement of gene expression and genomic variations within the same experiments, at unparalleled sensitivity. These methods will not replace the microarrays, but will drive their evolution towards more powerful and more affordable levels. Investigating

gene expression and its regulation is of big impact in defining the identity of organisms, tissues, and cells, hence allowing the characterization of disease conditions and their etiology as well as the evaluation of new treatments. Organizing, combining, and reanalyzing the massive information already obtained are a central need as well as an unprecedented resource for the scientific community.

Biomedical sciences are going through a profound revolution, as important as the one that invested all the fields of knowledge in XVI and XVII centuries. Under the thrust of more and more capable technologies, they are moving from the reductionist approach to a powerful integrated approach. Along with their internal metamorphosis, biosciences are also challenging the whole society, as the fast progress especially in genetics have profound implications and consequences also on politics, economy, and ethics. Indeed, the very concepts of life and death are being forever modified.

7. ACKNOWLEDGEMENTS

It is impossible to properly acknowledge all the people who have inspired me as a person and a scientist; only some of them are mentioned in this book. Many of the others are not scientists: some are musicians, writers, or painters, but they are not less important. All of them formed in many ways the foundation of my personality and my way of thinking.

This study has been carried out at the DNA Sequencing and Genomics Laboratory of the Institute of Biotechnology, University of Helsinki, with financial support from the University of Helsinki, Academy of Finland, Finnish Cultural Foundation, and Ehnrooth Foundation.

All this would not have been possible without my supervisor Dr. Petri Auvinen. Besides being one of the most brilliant minds I have encountered in my life, he is a nice and sensitive person. He has well tolerated all my crazy moments and worn his headphones only when really exhausted. Among many other things, Petri taught me that extraordinary goals can be achieved simply and quietly... and that bosses can ask things so gently that you don't realize it is something you "have" to do (and even fast)!

I wish to thank Prof. Mart Saarma, Director of the Institute of Biotechnology, for giving us all at the Institute an excellent work environment and facilities. It is also because of my great admiration for him that I moved to Finland.

I am grateful to the members of my follow up group, Professors Eero Castren and Jukka Corander, for advising and guiding my work. I am also grateful to Jukka and Dr. Iris Hovatta for critically reviewing my thesis. It has been my great pleasure and honor to collaborate with them. I express my gratitude to Dr. Christopher Carroll for the language revision of this thesis.

I wish to thank to Prof. Tapio Palva and Dr. Pekka Heino, at the Division of Genetics in the Department of Biological and Environmental Sciences, for their help during my whole curriculum. Dr. Eeva Sievi and Dr. Sandra Falck of the Viikki Graduate School in Biosciences are also acknowledged.

My time in the lab has been nicer thanks to the people who have populated it or that currently do so: Panu, Tuomas, Eeva-Marja, Lasse, Kui, Jenni, Olli, Jarmo, Miia, Kaisa, Anu S. & P., Pia, Matias, Eetu, Rashi, Paula, Anna-Liisa, Päivi, Kirsi, Pasi, Ritu, Tuuli, Mira, Suvi, Lea, Ritva, Tuula, Riikka, Noora, Robert, Ari-Matti, Hannu, Markku, Janne, Tanja, Matthew, Juha. Thank you for enduring my up-n-down days, usually starting at your lunchtime.

During these years in Finland, I have had the chance to work with many talented scientists. I wish to mention those who have challenged and influenced me most: Dr. Eeva Auvinen, Dr. Mikko Frilander, M.Sc. Nina Kivi, M.Sc. Heli Pessa, M.Sc. Maria Sundvik, M.Sc. Helena Kilpinen, Prof. Arto Urtti, Prof. Pertti Panula, Dr. Tapio Heino, Dr. Johan Peränen, Prof. Anu Wartiovaara, Dr. Joni Alvesalo, Prof. Pia Vuorela, Prof. Kristina Lindström, Dr. Jarno Tuimala, Dr. Eija Korpelainen, Dr. Saara Laitinen, Dr. Oscar Puig, Dr. Zewdu Terefework, Dr. Ettore Tiraboschi, Dr. Petri Törönen, M.Sc. Kati-Sisko Vellonen, Prof. Hannele Yki-Järvinen, Dr. Anna Kotronen, M.Sc. Mari Palgi.

My adventure in science began in 1997, when I was a first year student at the Faculty of Medicine in Naples. Prof. Corrado Garbi, my first mentor, transferred to me his love and enthusiasm for science and disciplined me to the rigorous way of thinking.

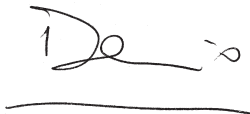
I will owe him a lot for the rest of my life. Prof. Lucio Nitsch has supported me even when I moved from cell biology to bioinformatics. I am thankful to Pasquale De Luca at BioGeM for tutoring and becoming a good friend. My last stage in Italy was at the Developmental Neurobiology lab, CNR, directed by Dr. Umberto di Porzio. He introduced me to neurobiology, which I have not abandoned anymore. He also allowed me to visit Helsinki in March 2003, which changed my life forever.

The following are only a few of the many friends and colleagues whose support and ideas were vital also to my work: Antonello, Damiana, Salvatore, Massimiliano, Carla, Floriana, Luigi, Claudia, Simone, Ombretta, Eleonora, Cinzia, Gaetano, Flaviana, Anna, Paolo, Remo, Pandelis, Rosario, Irene, Michele, Alessandro, Laura.

I am also enormously indebted to my family for the endless support and encouragement.

My greatest thank goes to Leena for making my life so colorful. I love her more than microarrays.

Helsinki, April 2009



A handwritten signature in black ink, appearing to read 'U. Porzio', is written above a horizontal line.

8. REFERENCES

- Adriani W, Leo D, Greco D, Rea M, di Porzio U, Laviola G, Perrone-Capano C. 2006. Methylphenidate administration to adolescent rats determines plastic changes on reward-related behavior and striatal gene expression. *Neuropsychopharmacology*. 31(9):1946-56.
- Affara NA. 2003. Resource and hardware options for microarray-based experimentation. *Brief Funct Genomic Proteomic*. 2(1):7-20.
- Allison DB, Cui X, Page GP, Sabripour M. 2006. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet*. 7(1):55-65.
- Alon U. 2007. Network motifs: theory and experimental approaches. *Nat Rev Genet*. 8(6):450-61.
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 25(1):25-9.
- Aston C, Jiang L, Sokolov BP. 2004. Transcriptional profiling reveals evidence for signaling and oligodendroglial abnormalities in the temporal cortex from patients with major depressive disorder. *Mol Psychiatry*. 10(3):309-22.
- Attwooll C, Lazzarini Denchi E, Helin K. 2004. The E2F family: specific functions and overlapping interests. *EMBO J*. 23(24):4709-16.
- Babu MM. 2008. Computational approaches to study transcriptional regulation. *Biochem Soc Trans*. 36(Pt 4):758-65.
- Ball CA, Awad IA, Demeter J, Gollub J, Hebert JM, Hernandez-Boussard T, Jin H, Matese JC, Nitzberg M, Wymore F, Zachariah ZK, Brown PO, Sherlock G. 2005. The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Res*. 33:D580-2.
- Barber RD, Harmer DW, Coleman RA, Clark BJ. 2005. GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. *Physiol Genomics*. 21(3):389-95.
- 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*. 33:D562-6.
- Beckmann AM, Wilce PA. 1997. Egr transcription factors in the nervous system. *Neurochem Int*. 31(4):477-510.
- Benjamini Y and Hochberg Y. 1995. Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc*. 57(1):289-300.
- Binder H, Preibisch S. 2005. Specific and nonspecific hybridization of oligonucleotide probes on microarrays. *Biophys J*. 89(1):337-52.
- Bolstad BM, Irizarry RA, Astrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. 19(2):185-93.
- Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansong W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*. 29(4):365-71.
- Carter SL, Eklund AC, Mecham BH, Kohane IS, Szallasi Z. 2005. Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements. *BMC Bioinformatics*. 6:107.
- Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. *Nat Genet*. 31(4):415-8.
- Castr n E. 2004. Neurotrophic effects of antidepressant drugs. *Curr Opin Pharmacol*. 4(1):58-64.

- Cohen CD, Klingenhoff A, Boucherot A, Nitsche A, Henger A, Brunner B, Schmid H, Merkle M, Saleem MA, Koller KP, Werner T, Gröne HJ, Nelson PJ, Kretzler M. 2006. Comparative promoter analysis allows de novo identification of specialized cell junction-associated proteins. *Proc Natl Acad Sci U S A*. 103(15):5682-7.
- Consales C, Volpicelli F, Greco D, Leone L, Colucci-D'Amato L, Perrone-Capano C, di Porzio U. 2007. GDNF signaling in embryonic midbrain neurons in vitro. *Brain Res*. 2007 1159:28-39.
- Czechowski T, Bari RP, Stitt M, Scheible WR, Udvardi MK. 2004. Real-time RT-PCR profiling of over 1400 Arabidopsis transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. *Plant J*. 38(2):366-79.
- Dabrowski M, Aerts S, Kaminska B. 2006. Prediction of a key role of motifs binding E2F and NR2F in down-regulation of numerous genes during the development of the mouse hippocampus. *BMC Bioinformatics*. 7:367.
- Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson SJ, Meng F. 2005. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res*. 33(20):e175.
- Day A, Carlson MR, Dong J, O'Connor BD, Nelson SF. 2007. Celsius: a community resource for Affymetrix microarray data. *Genome Biol*. 8(6):R112.
- Denis-Donini S, Caprini A, Frassoni C, Grilli M. 2005. Members of the NF-kappaB family expressed in zones of active neurogenesis in the postnatal and adult mouse brain. *Brain Res Dev Brain Res*. 154(1):81-9.
- Draghici S, Khatri P, Eklund AC, Szallasi Z. 2006. Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet*. 22(2):101-9.
- Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A*. 102(40):14338-43.
- Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol*. 17(1):68-74
- Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 30(1):207-10.
- Ein-Dor L, Zuk O, Domany E. 2006. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci U S A*. 103(15):5923-8.
- Elkins RL, Orr TE, Rausch JL, Fei YJ, Carl GF, Hobbs SH, Buccafusco JJ and Edwards GL. 2003. Cocaine-induced expression differences in PSD-95/SAP-90-associated protein 4 and in Ca2+/calmodulin-dependent protein kinase subunits in amygdalae of taste aversion-prone and taste aversion-resistant rats. *Ann N Y Acad Sci* 1003:386-390.
- Flieck P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Eyre T, Fitzgerald S, Fernandez-Banet J, Gräf S, Haider S, Hammond M, Holland R, Howe KL, Howe K, Johnson N, Jenkinson A, Kähäri A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Slater G, Smedley D, Spudich G, Trevanion S, Vilella AJ, Vogel J, White S, Wood M, Birney E, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Hubbard TJ, Kasprzyk A, Proctor G, Smith J, Ureta-Vidal A, Searle S. 2008. Ensembl 2008. *Nucleic Acids Res*. 36:D707-14.
- Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. 2004a. affy---analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20:307-315.
- Gautier L, Møller M, Friis-Hansen L, Knudsen S. 2004b. Alternative mapping of probes to genes for Affymetrix chips. *BMC Bioinformatics*. 14;5:111.
- 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(10):R80.
- Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. 2004. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*. 20(1):93-9.

- Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. 2007. The human disease network. *Proc Natl Acad Sci U S A*. 104(21):8685-90.
- Goldmuntz E. 2001. The epidemiology and genetics of congenital heart disease. *Clin Perinatol*. 28(1):1-10.
- Gould E. 2007. How widespread is adult neurogenesis in mammals? *Nat Rev Neurosci*. 8(6):481-8.
- Goutsias J, Lee NH. 2007. Computational and experimental approaches for modeling gene regulatory networks. *Curr Pharm Des*. 13(14):1415-36.
- Guerriero RM, Rajadhyaksha A, Crozatier C, Giros B, Nosten-Bertrand M and Kosofsky BE. 2005. Augmented constitutive CREB expression in the nucleus accumbens and striatum may contribute to the altered behavioral response to cocaine of adult mice exposed to cocaine in utero. *Dev Neurosci* 27:235-248.
- Hacia JG. 1999. Resequencing and mutational analysis using oligonucleotide microarrays. *Nat Genet*. 21(1 Suppl):42-7.
- Harbig J, Sprinkle R, Enkemann SA. 2005. A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. *Nucleic Acids Res*. 33(3):e31.
- Heidbreder CA, Gardner EL, Xi ZX, Thanos PK, Mugnaini M, Hagan JJ and Ashby CR Jr. 2005. The role of central dopamine D3 receptors in drug addiction: a review of pharmacological evidence. *Brain Res Brain Res Rev*. 49:77-105.
- Hirayama-Yamada K, Kamisago M, Akimoto K, Aotsuka H, Nakamura Y, Tomita H, Furutani M, Imamura S, Takao A, Nakazawa M, Matsuoka R. 2005. Phenotypes with GATA4 or NKX2.5 mutations in familial atrial septal defect. *Am J Med Genet A*. 135(1):47-52.
- Holland MJ. 2002. Transcript abundance in yeast varies over six orders of magnitude. *J Biol Chem*. 277(17):14363-6.
- Holloway AJ, van Laar RK, Tothill RW, Bowtell DD. 2002. Options available--from start to finish--for obtaining data from DNA microarrays II. *Nat Genet*. 32 Suppl:481-9.
- Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen RV, Misra J, Dillon W, Lee KF, Clark KE, Haverty P, Weng Z, Mutter GL, Frosch MP, Macdonald ME, Milford EL, Crum CP, Bueno R, Pratt RE, Mahadevappa M, Warrington JA, Stephanopoulos G, Stephanopoulos G, Gullans SR. 2001. A compendium of gene expression in normal human tissues. *Physiol Genomics*. 7(2):97-104.
- Huang W, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA. 2007. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol*. 8(9):R183.
- Huber W, von Heydebreck A, Sueltmann H, Poustka A, Vingron M. 2002. Variance Stabilization Applied to Microarray Data Calibration and to the Quantification of Differential Expression. *Bioinformatics* 18:96-S104.
- Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, Kobayashi S, Davis C, Dai H, He YD, Stephanians SB, Cavet G, Walker WL, West A, Coffey E, Shoemaker DD, Stoughton R, Blanchard AP, Friend SH, Linsley PS. 2001. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol*. 19(4):342-7.
- Hwang KB, Kong SW, Greenberg SA, Park PJ. 2004. Combining gene expression data from different generations of oligonucleotide arrays. *BMC Bioinformatics*. 5:159.
- Ioannidis JP, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, Falchi M, Furlanello C, Game L, Jurman G, Mangion J, Mehta T, Nitzberg M, Page GP, Petretto E, van Noort V. 2009. Repeatability of published microarray gene expression analyses. *Nat Genet*. 41(2):149-55.
- Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U and Speed TP. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 4:249-264.
- Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JG, Geoghegan J, Germino G, Griffin C, Hilmer SC, Hoffman E, Jedlicka AE, Kawasaki E, Martinez-Murillo F, Morsberger L, Lee H, Petersen D, Quackenbush J, Scott A, Wilson M, Yang Y, Ye SQ, Yu W. 2005. Multiple-laboratory comparison of microarray platforms. *Nat Methods*. 2(5):345-50.

- Irizarry RA, Wu Z, Jaffee HA. 2006. Comparison of Affymetrix GeneChip expression measures. *Bioinformatics*. 22(7):789-94.
- Jacobs BL, Praag H, Gage FH. 2000. Adult brain neurogenesis and psychiatry: a novel theory of depression. *Mol Psychiatry*. 5(3):262-9.
- Jafari P, Azuaje F. 2006. An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. *BMC Med Inform Decis Mak*. 6:27.
- Järvinen AK, Hautaniemi S, Edgren H, Auvinen P, Saarela J, Kallioniemi OP, Monni O. 2004. Are data from different gene expression microarray platforms comparable? *Genomics*. 83(6):1164-8.
- Jenner RG, Young RA. 2005. Insights into host responses against pathogens from transcriptional profiling. *Nat Rev Microbiol*. 3(4):281-94.
- Ji Y, Coombes K, Zhang J, Wen S, Mitchell J, Puzstai L, Symmans WF, Wang J. 2006. RefSeq refinements of UniGene-based gene matching improve the correlation of expression measurements between two microarray platforms. *Appl Bioinformatics*. 5(2):89-98.
- Jones MW, Errington ML, French PJ, Fine A, Bliss TV, Garel S, Charnay P, Bozon B, Laroche S, Davis S. 2001. A requirement for the immediate early gene *Zif268* in the expression of late LTP and long-term memories. *Nat Neurosci*. 4(3):289-96.
- Kahlig KM, Binda F, Khoshbouei H, Blakely RD, McMahon DG, Javitch JA, Galli A. 2005. Amphetamine induces dopamine efflux through a dopamine transporter channel. *Proc Natl Acad Sci U S A*. 102(9):3495-500. Epub 2005 Feb 22.
- Katz S, Irizarry RA, Lin X, Tripputi M, Porter MW. 2006. A summarization approach for Affymetrix GeneChip data using a reference training set from a large, biologically diverse database. *BMC Bioinformatics*. 7:464.
- Kawasaki ES. 2006. The end of the microarray Tower of Babel: will universal standards lead the way? *J Biomol Tech*. 17(3):200-6.
- Kendzioriski C, Irizarry RA, Chen KS, Haag JD, Gould MN. 2005. On the utility of pooling biological samples in microarray experiments. *Proc Natl Acad Sci U S A*. 102(12):4252-7.
- Khatri P, Drăghici S. 2005. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*. 21(18):3587-95.
- Kilpinen S, Autio R, Ojala K, Iljin K, Bucher E, Sara H, Pisto T, Saarela M, Skotheim RI, Björkman M, Mpindi JP, Haapa-Paananen S, Vainio P, Edgren H, Wolf M, Astola J, Nees M, Hautaniemi S, Kallioniemi O. 2008. Systematic bioinformatic analysis of expression levels of 17,330 human genes across 9,783 samples from 175 types of healthy and pathological tissues. *Genome Biol*. 9(9):R139.
- Kivi N, Greco D, Auvinen P, Auvinen E. 2008. Genes involved in cell adhesion, cell motility and mitogenic signaling are altered due to HPV 16 E5 protein expression. *Oncogene*. 27(18):2532-41.
- Kothapalli R, Yoder SJ, Mane S, Loughran TP Jr. 2002. Microarray results: how accurate are they? *BMC Bioinformatics*. 3:22.
- Kreil DP, Russell RR, Russell S. 2006. Microarray oligonucleotide probes. *Methods Enzymol*. 410:73-98.
- Kronick MN. 2004. Creation of the whole human genome microarray. *Expert Rev Proteomics*. 1(1):19-28.
- Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, Meyer L, Hsu F, Hinrichs AS, Harte RA, Giardine B, Fujita P, Diekhans M, Dreszer T, Clawson H, Barber GP, Haussler D, Kent WJ. 2009. The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res*. 37:D755-61.
- Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS. 2002. Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*. 18(3):405-12.
- Lage K, Hansen NT, Karlberg EO, Eklund AC, Roque FS, Donahoe PK, Szallasi Z, Jensen TS, Brunak S. 2008. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc Natl Acad Sci U S A*. 105(52):20870-5.
- Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J. 2005. Independence and reproducibility across microarray platforms. *Nat Methods*. 2(5):337-44.
- Larsson O, Wennmalm K, Sandberg R. 2006. Comparative microarray analysis. *OMICS*. 10(3):381-97.
- Lee NH. 2005. Genomic approaches for reconstructing gene networks. *Pharmacogenomics*. 6(3):245-58.

- Lee PD, Sladek R, Greenwood CM, Hudson TJ. 2002. Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. *Genome Res.* 12(2):292-7.
- Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J. 2005. The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Res.* 33:D71-4.
- Li C and Wong WH. 2001a. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc Natl Acad Sci.* 98:31-36.
- Li C and Wong WH. 2001b. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.* 2:research0032.1-0032.11.
- Li L, Yun SH, Keblesh J, Trommer BL, Xiong H, Radulovic J, Tourtellotte WG. 2007. Egr3, a synaptic activity regulated transcription factor that is essential for learning and memory. *Mol Cell Neurosci.* 35(1):76-88.
- Liang S, Li Y, Be X, Howes S, Liu W. 2006. Detecting and profiling tissue-selective genes. *Physiol Genomics.* 26(2):158-62.
- Lim CA, Yao F, Wong JJ, George J, Xu H, Chiu KP, Sung WK, Lipovich L, Vega VB, Chen J, Shahab A, Zhao XD, Hibberd M, Wei CL, Lim B, Ng HH, Ruan Y, Chin KC. 2007. Genome-wide mapping of RELA(p65) binding identifies E2F1 as a transcriptional activator recruited by NF-kappaB upon TLR4 activation. *Mol Cell.* 27(4):622-35.
- Liu DX, Nath N, Chellappan SP, Greene LA. 2005. Regulation of neuron survival and death by p130 and associated chromatin modifiers. *Genes Dev.* 19(6):719-32.
- Liu WM, Mei R, Di X, Ryder TB, Hubbell E, Dee S, Webster TA, Harrington CA, Ho MH, Baid J, Smeekens SP. 2002. Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics.* 18(12):1593-9.
- Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol.* 14(13):1675-80.
- Maglott D, Ostell J, Pruitt KD, Tatusova T. 2007. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* 35:D26-31.
- Mansmann U, Meister R. 2005. Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach. *Methods Inf Med.* 44(3):449-53.
- McCarroll SA, Murphy CT, Zou S, Pletcher SD, Chin CS, Jan YN, Kenyon C, Bargmann CI, Li H. 2004. Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat Genet.* 36(2):197-204.
- McClintick JN, Edenberg HJ. 2006. Effects of filtering by Present call on analysis of microarray experiments. *BMC Bioinformatics.* 31:7:49.
- McCracken CB, Hamby SM, Patel KM, Morgan D, Vrana KE, Roberts DC. 2005. Extended cocaine self-administration and deprivation produces region-specific and time-dependent changes in connexin36 expression in rat brain. *Synapse.* 58(3):141-50.
- Mecham BH, Wetmore DZ, Szallasi Z, Sadovsky Y, Kohane I, Mariani TJ. 2004. Increased measurement accuracy for sequence-verified microarray probes. *Physiol Genomics.* 18(3):308-15.
- Mémet S. 2006. NF-kappaB functions in the nervous system: from development to disease. *Biochem Pharmacol.* 72(9):1180-95.
- Michael KL, Taylor LC, Schultz SL, Walt DR. 1998. Randomly ordered addressable high-density optical sensor arrays. *Anal Chem.* 70(7):1242-8.
- Modlich O, Prisack HB, Munnes M, Audretsch W, Bojar H. 2004. Immediate gene expression changes after the first course of neoadjuvant chemotherapy in patients with primary breast cancer disease. *Clin Cancer Res.* 10:6418-6431.
- Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC. 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet.* 34(3):267-73.

- Naef F, Hacker CR, Patil N, Magnasco M. 2002a. Characterization of the expression ratio noise structure in high-density oligonucleotide arrays. *Genome Biol.* 3:PREPRINT0001.
- Naef F, Hacker CR, Patil N, Magnasco M. 2002b. Empirical characterization of the expression ratio noise structure in high-density oligonucleotide arrays. *Genome Biol.* 3(4):RESEARCH0018.
- Naef, F., and M. O. Magnasco. 2003. Solving the riddle of the bright mismatches: hybridization in oligonucleotide arrays. *Phys. Rev. E.* 68:11906–11910.
- Nègre N, Lavrov S, Hennetin J, Bellis M, Cavalli G. 2006. Mapping the distribution of chromatin proteins by ChIP on chip. *Methods Enzymol.* 410:316-41.
- Nuwaysir EF, Huang W, Albert TJ, Singh J, Nuwaysir K, Pitas A, Richmond T, Gorski T, Berg JP, Ballin J, McCormick M, Norton J, Pollock T, Sumwalt T, Butcher L, Porter D, Molla M, Hall C, Blattner F, Sussman MR, Wallace RL, Cerrina F and Green RD. 2002. Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Res.* 12:1749-55.
- Obermeier F, Burgmaier J, Thome K, Weichert S, Hein S, Binnewies T, Foitzik V, Muller M, Stahler CF and Stahler PF. 2003. Validation of a novel, fully integrated and flexible microarray benchtop facility for gene expression profiling. *Nucleic Acids Res.* 31:e151.
- Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 1;27(1):29-34.
- Oostlander AE, Meijer GA, Ylstra B. 2004. Microarray-based comparative genomic hybridization and its applications in human genetics. *Clin Genet.* 66(6):488-95.
- Qin LX, Beyer RP, Hudson FN, Linford NJ, Morris DE, Kerr KF. 2006. Evaluation of methods for oligonucleotide array data via quantitative real-time PCR. *BMC Bioinformatics.* 7:23.
- Park CH, Kang JS, Shin YH, Chang MY, Chung S, Koh HC, Zhu MH, Oh SB, Lee YS, Panagiotakos G, Tabar V, Studer L, Lee SH. 2006. Acquisition of in vitro and in vivo functionality of Nurr1-induced dopamine neurons. *FASEB J.* 20(14):2553-5.
- Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, Coulson R, Farne A, Lara GG, Holloway E, Kapushesky M, Lilja P, Mukherjee G, Oezcimen A, Rayner T, Rocca-Serra P, Sharma A, Sansone S, Brazma A. 2005. ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* 33:D553-5.
- Parman C and Halling C. 2008. affyQCReport: A Package to Generate QC Reports for Affymetrix Array Data.
- Perrier P, Martinez FO, Locati M, Bianchi G, Nebuloni M, Vago G, Bazzoni F, Sozzani S, Allavena P, Mantovani A. 2004. Distinct transcriptional programs activated by interleukin-10 with or without lipopolysaccharide in dendritic cells: induction of the B cell-activating chemokine, CXC chemokine ligand 13. *J Immunol.* 172:7031–7042.
- Petersen DW, Kawasaki ES. 2007. Manufacturing of microarrays. *Adv Exp Med Biol.* 593:1-11.
- Pilpel Y, Sudarsanam P, Church GM. 2001. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet.* 29(2):153-9.
- Pritchard CC, Hsu L, Delrow J, Nelson PS. 2001. Project normal: defining normal variance in mouse gene expression. *Proc Natl Acad Sci U S A.* 98(23):13266-71.
- Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35:D61-5.
- R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM. 2004. ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia.* 6(1):1-6.
- Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyérén P. 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem.* 242(1):84-9.
- Ross W, Gourse RL. 2009. Analysis of RNA polymerase-promoter complex formation. *Methods.* 47(1):13-24.

- Ryan CA, Gildea LA, Hulette BC, Dearman RJ, Kimber I, Gerberick GF. 2004. Gene expression changes in peripheral blood-derived dendritic cells following exposure to a contact allergen. *Toxicol Lett.* 2;150(3):301-16.
- Saito-Hisaminato A, Katagiri T, Kakiuchi S, Nakamura T, Tsunoda T, Nakamura Y. 2002. Genome-wide profiling of gene expression in 29 normal human tissues with a cDNA microarray. *DNA Res.* 9(2):35-45.
- Sandberg R, Larsson O. 2007. Improved precision and accuracy for microarrays using updated probe set definitions. *BMC Bioinformatics.* 8:48.
- Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M. 1977. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature.* 265(5596):687-95.
- Sari Y. 2004. Serotonin1B receptors: from protein to physiological function and behavior. *Neurosci Biobehav Rev* 28:565-582.
- Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science.* 239(4839):487-91.
- Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Mizrachi I, Ostell J, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Yaschenko E, Ye J. 2009. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 37:D5-15.
- Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science.* 270(5235):467-70.
- Schlitt T, Brazma A. 2006. Modelling in molecular biology: describing transcription regulatory networks at different scales. *Philos Trans R Soc Lond B Biol Sci.* 361(1467):483-94.
- Seifert M, Scherf M, Eppl A, Werner T. 2005. Multievidence microarray mining. *Trends Genet.* 21(10):553-8.
- Severgnini M, Biccato S, Mangano E, Scarlatti F, Mezzelani A, Mattioli M, Ghidoni R, Peano C, Bonnal R, Viti F, Milanesi L, De Bellis G, Battaglia C. 2006. Strategies for comparing gene expression profiles from different microarray platforms: application to a case-control experiment. *Anal Biochem.* 353(1):43-56.
- Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science.* 309(5741):1728-32.
- Shyamsundar R, Kim YH, Higgins JP, Montgomery K, Jordan M, Sethuraman A, van de Rijn M, Botstein D, Brown PO, Pollack JR. 2005. A DNA microarray survey of gene expression in normal human tissues. *Genome Biol.* 6(3):R22.
- Simon R, Radmacher MD, Dobbin K. 2002. Design of studies using DNA microarrays. *Genet Epidemiol.* 23(1):21-36.
- Sipos L, Gyurkovics H. 2005. Long-distance interactions between enhancers and promoters. *FEBS J.* 272(13):3253-9.
- Sivachenko AY, Yuryev A, Daraselia N, Mazo I. 2007. Molecular networks in microarray analysis. *J Bioinform Comput Biol.* 5(2B):429-56.
- Slonim DK. 2002. From patterns to pathways: gene expression data analysis comes of age. *Nat Genet.* 32 Suppl:502-8.
- Smyth GK. 2005. Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W, editors. *Bioinformatics and Computational Biology Solutions using R and Bioconductor.* New York: Springer. pp 397-420.
- Stec J, Wang J, Coombes K, Ayers M, Hoersch S, Gold DL, Ross JS, Hess KR, Tirrell S, Linette G, Hortobagyi GN, Fraser Symmans W, Pusztai L. 2005. Comparison of the predictive accuracy of DNA array-based multigene classifiers across cDNA arrays and Affymetrix GeneChips. *J Mol Diagn.* 7(3):357-67.

- Stoeckert CJ Jr, Causton HC, Ball CA. 2002. Microarray databases: standards and ontologies. *Nat Genet.* 32 Suppl:469-73.
- Stokes TH, Torrance JT, Li H, Wang MD. 2008. ArrayWiki: an enabling technology for sharing public microarray data repositories and meta-analyses. *BMC Bioinformatics.* 9 Suppl 6:S18.
- Stossi F, Barnett DH, Frasor J, Komm B, Lyttle CR, Katzenellenbogen BS. 2004. Transcriptional profiling of estrogen-regulated gene expression via estrogen receptor (ER) alpha or ERbeta in human osteosarcoma cells: distinct and common target genes for these receptors. *Endocrinology.* 145(7):3473-86.
- Swaren J, Hörz W. 1996. Regulation of gene expression by nucleosomes. *Curr Opin Genet Dev.* 6(2):164-70.
- Swanson LW, Petrovich GD. 1998. What is the amygdala? *Trends Neurosci.* 21(8):323-31.
- Swiatek PJ, Gridley T. 1993. Perinatal lethality and defects in hindbrain development in mice homozygous for a targeted mutation of the zinc finger gene *Krox20*. *Genes Dev.* 7(11):2071-84.
- Tan PK, Downey TJ, Spitznagel EL Jr, Xu P, Fu D, Dimitrov DS, Lempicki RA, Raaka BM, Cam MC. 2003. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.* 31(19):5676-84.
- Tang Z, McGowan BS, Huber SA, McTiernan CF, Addya S, Surrey S, Kubota T, Fortina P, Higuchi Y, Diamond MA, Wyre DS, Feldman AM. 2004. Gene expression profiling during the transition to failure in TNF-alpha over-expressing mice demonstrates the development of autoimmune myocarditis. *J Mol Cell Cardiol.* 36:515-530.
- Tawfik DS, Griffiths AD. 1998. Man-made cell-like compartments for molecular evolution. *Nat Biotechnol.* 16(7):652-6.
- Taylor E, Cogdell D, Coombes K, Hu L, Ramdas L, Tabor A, Hamilton S, Zhang W. 2001. Sequence verification as quality-control step for production of cDNA microarrays. *Biotechniques.* 31(1):62-5.
- Thellin O, Zorzi W, Lakaye B, De Borman B, Coumans B, Hennen G, Grisar T, Igout A, Heinen E. 1999. Housekeeping genes as internal standards: use and limits. *J Biotechnol.* 75(2-3):291-5.
- Tilstone C. 2003. DNA microarrays: vital statistics. *Nature.* 424(6949):610-2.
- Tomfohr J, Lu J, Kepler TB. 2005. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics.* 6:225.
- Topilko P, Schneider-Maunoury S, Levi G, Baron-Van Evercooren A, Chennoufi AB, Seitanidou T, Babinet C, Charnay P. 1994. *Krox-20* controls myelination in the peripheral nervous system. *Nature.* 371(6500):796-9.
- Tripodi M, Filosa A, Armentano M, Studer M. 2004. The COUP-TF nuclear receptors regulate cell migration in the mammalian basal forebrain. *Development.* 131(24):6119-29.
- Tusher VG, Tibshirani R, Chu G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A.* 2001 98(9):5116-21.
- van der Maarel SM. 2008. Epigenetic mechanisms in health and disease. *Ann Rheum Dis.* 67 Suppl 3:iii97-100.
- Vartanian K, Slotke R, Johnstone T, Casale A, Planck SR, Choi D, Smith JR, Rosenbaum JT, Harrington CA. 2009. Gene expression profiling of whole blood: Comparison of target preparation methods for accurate and reproducible microarray analysis. *BMC Genomics.* 10(1):2.
- Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. 1995. Serial analysis of gene expression. *Science.* 270(5235):484-7.
- Vinogradov AE. 2003. Isochores and tissue-specificity. *Nucleic Acids Res.* 31(17):5212-20.
- Vinogradov AE. 2004. Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet.* 20(5):248-53.
- Volpicelli F, Caiazzo M, Greco D, Consales C, Leone L, Perrone-Capano C, Colucci D'Amato L, di Porzio U. 2007. *Bdnf* gene is a downstream target of *Nurr1* transcription factor in rat midbrain neurons in vitro. *J Neurochem.* 102(2):441-53.
- Wagner GP, Pavlicev M, Cheverud JM. 2007. The road to modularity. *Nat Rev Genet.* 8(12):921-31.

- Walker MS, Hughes TA. 2008. Messenger RNA expression profiling using DNA microarray technology: diagnostic tool, scientific analysis or un-interpretable data? *Int J Mol Med.* 21(1):13-7.
- Wang H, He X, Band M, Wilson C, Liu L. 2005. A study of inter-lab and inter-platform agreement of DNA microarray data. *BMC Genomics.* 6(1):71.
- Wang Y, Miao ZH, Pommier Y, Kawasaki ES, Player A. 2007. Characterization of mismatch and high-signal intensity probes associated with Affymetrix genechips. *Bioinformatics.* 23(16):2088-95.
- Wang RS, Zhang XS, Chen L. 2007. Inferring transcriptional interactions and regulator activities from experimental data. *Mol Cells.* 24(3):307-15.
- Wamat P, Eils R, Brors B. 2005. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics.* 6:265.
- Warrington JA, Nair A, Mahadevappa M, Tsyganskaya M. 2000. Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol Genomics.* 2(3):143-7.
- Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetverin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Ostell J, Pruitt KD, Schuler GD, Shumway M, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E. 2008. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 36:D13-21.
- Wennmalm K, Wahlestedt C, Larsson O. 2005. The expression signature of in vitro senescence resembles mouse but not human aging. *Genome Biol.* 6(13):R109.
- Werner T. 2007. Regulatory networks: linking microarray data to systems biology. *Mech Ageing Dev.* 128(1):168-72.
- Werner T, Nelson PJ. 2006. Joining high-throughput technology with in silico modelling advances genome-wide screening towards targeted discovery. *Brief Funct Genomic Proteomic.* 5(1):32-6.
- Wildhaber BE, Yang H, Tazuke Y, Teitelbaum DH. 2003. Gene alteration of intestinal intraepithelial lymphocytes with administration of total parenteral nutrition. *J Pediatr Surg.* 38(6):840-3.
- Winter EE, Goodstadt L, Ponting CP. 2004. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res.* 14(1):54-61.
- Wolber PK, Collins PJ, Lucas AB, De Witte A, and Shannon KW. 2006. The Agilent in situ-synthesized microarray platform. *Methods in enzymology* 410:28-57.
- Wu Z, Irizarry RA. 2004. Preprocessing of oligonucleotide array data. *Nature Biotechnology* 22:656-658
- Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, Lancet D, Shmueli O. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics.* 21(5):650-9.
- Yang YH, Speed T. 2002. Design issues for cDNA microarray experiments. *Nat Rev Genet.* 3(8):579-88.
- Yao WD, Gainetdinov RR, Arbuckle MI, Sotnikova TD, Cyr M, Beaulieu JM, Torres GE, Grant SG and Caron MG. 2004. Identification of PSD-95 as a regulator of dopamine-mediated synaptic and behavioral plasticity. *Neuron* 41:625-638.
- Yin JQ, Zhao RC, Morris KV. 2008. Profiling microRNA expression with microarrays. *Trends Biotechnol.* 26(2):70-6.
- Zahurak M, Parmigiani G, Yu W, Scharpf RB, Berman D, Schaeffer E, Shabbeer S, Cope L. 2007. Pre-processing Agilent microarray data. *BMC Bioinformatics.* 8:142.
- Zhang L, Miles MF, Aldape KD. 2003. A model of molecular interactions on short oligonucleotide microarrays. *Nature Biotechnology* 21:818-821.
- Zhu J, He F, Song S, Wang J, Yu J. 2008. How many human genes can be defined as housekeeping with current expression data? *BMC Genomics.* 9:172.