Chapter 1
Functional genomics and molecular networks
Gene expression regulations
in complex diseases:
Down syndrome as a case study

Marie-Claude Potier[1] and Isabelle Rivals[2]
[1]CRICM, CNRS UMR7225, INSERM U975 and
UPMC, Hôpital de la Pitié-Salpêtrière, 47 Bd de
l'Hôpital, 75013 Paris, France
marie-claude.potier@upmc.fr
[2]ESA, ESPCI ParisTech,
10 rue Vauquelin, 75005 Paris, France
isabelle.rivals@espci.fr

Abstract

The goal of functional genomics is to understand the relationship between whole genomes and phenotypes through a dynamic approach. It requires high throughput technologies such as microarrays and data analysis. The power of this approach allowed to study complex biological functions as well as diseases. In this chapter, we introduce functional genomics and describe the statistical methods that are used to find differentially expressed genes. We analyze a large number of data sets produced on a complex disease, namely Down syndrome, in different models. We show that, whatever the model, genes that are in three copies are globally overexpressed. However, we failed to identify a set of two-copy genes that would be dysregulated in all studies. It either suggests that studies are incomplete, or that this set of genes does not exist and that overexpression of the three-copy genes impacts on the whole transcriptome in a “stochastic” way.

1.1 Introduction

Functional genomics has been or is being applied to complex diseases in the hope of finding molecular networks that are altered, as well as gene targets for treatment. The experiments were initiated as soon as tools were available, and this field of research has exploded with the commercialization of DNA microarrays and their relative affordability. Since the first development of DNA microarrays more than ten years ago (Schena *et al*, 1995), the technology has improved in many aspects. Genome annotations are being updated and the probes associated to individual genes have been optimized for their selectivity and sensitivity. Although probe collections covering all the genes corresponding to various genomes are not fully optimized (Golfier *et al*, 2009) the data are improving and becoming consistent for powerful statistical analysis. The initial studies aiming at defining lists of differentially expressed genes have been disappointing and revealed that data analysis had to be extended using other tools than statistical tests. Many clustering methods and network analysis have been applied since. In parallel, gene ontology categorization has allowed a more functional view on the list of differentially expressed genes. Gene ontology now groups 28 154 terms with 19 913 for biological_process, 2 775 for cellular_component and 8 908 for molecular_function (www.geneontology.org 11/30/2010).

Nowadays, researchers who envision gene expression studies have always the same question in mind which is: what are the genes differentially expressed between various samples? But they want to

know the answer beyond the list, meaning that they want to know what are the functions of these genes, and if they belong to a particular network or pathway. Knowing this pathway will eventually give them the key for tuning it. Of course, the first question to ask is: has it been done, published and deposited in public databases (GEO www.ncbi.nlm.nih.gov/geo/ or Arrayexpress www.ebi.ac.uk/microarray-as/ae/)? If the answer is yes, there are data available on the subject; one should then plan to produce a different set of data, keeping in mind that this new set will need to be integrated with data sets available for the ultimate meta-analysis. It is just not possible to ignore other data sets since the power of analysis will be increased along with the size of samples. Then, of course, data need to be comparable, meaning preferably performed on the same type of microarray and possibly on the same platform. If not, then microarray annotation becomes a real issue that will have to be improved in the future. Isn't it surprising that with 497 398 samples (19 918 series) in GEO the number of meta-analyses is so low? There are nowadays 44 datasets with 100 to 200 samples, 22 with 1 000 to 7 000 samples but none with more than 7 000 samples.

1.1.1 Alzheimer's disease (AD)

Let us take the example of AD, a neurodegenerative disease which affects 25 millions of individuals worldwide, and which is becoming a real societal problem. Many gene profiling studies have been performed on AD patient samples (brain, peripheral cells) but no coher-

ent picture of gene expression regulation in AD was obtained (Maes *et al*, 2007) (Nagasaka *et al*, 2005) (Blalock *et al*, 2004) (Emilsson *et al*, 2006) (Lu *et al*, 2004). One could argue that small sample size together with manipulating human tissues with artifacts associated to post-mortem delay have minimized the power of analysis because of high variability. Also analysis of brain samples with very heterogeneous cell composition brings another level of variability. One way around would be to analyze gene expression at the single cell level. Such analyses are still under development and will bring answers to this major problem (Bontoux *et al*, 2008). It might be though that the control of cellular function has both deterministic and stochastic elements: complex regulatory networks define stable states that drive individual cells, whereas stochastic fluctuations in gene expression ensure transitions and coherence at the population level (Macarthur *et al*, 2009). Stochastic "noise" arises from random fluctuations within the cell and is an unavoidable aspect of life at the single-cell level. Evidence is accumulating that this noise crucially influences cellular auto-regulatory circuits and can "flip" genetic switches to drive probabilistic fate decisions (Singh and Weinberger, 2009). Stochastic noise in gene expression propagates through active, but not inactive, regulatory links and it was recently shown that extrinsic noise sources generate correlations even without direct regulatory links (Dunlop *et al*, 2008). In bacteria, it was shown that noise in expression of specific genes selects cells for competence, and experimental reduction of this noise decreases the number of competent cells (Maamar *et al*, 2007). This stochastic noise could

have an impact on cell fate either during development but also during disease progression. It is assumed that during development, cells acquire their fate by virtue of their lineage or their proximity to an inductive signal from another cell. However, cells can choose one or another pathway of differentiation stochastically, without apparent regard to environment or history, and this stochastic character could be critical for the maintenance of species (Losick and Desplan, 2008). Although these aspects have been studied in bacteria and yeasts, it is still particularly difficult to explore in multicellular organisms and in diseases.

The experimental design applied to complex human diseases has focused on gene expression regulation in tissues or cultured cells, thus excluding the single cell resolution. Although stochastic gene expression was mentioned, it is not possible to differentiate single cell level noise from tissue complexity, cellular heterogeneity and inter-individual variability.

Recently, with the use of systems biology approaches, two studies have revealed new interesting molecular networks related to AD. The first study applied weighted gene coexpression network analysis (WGCNA) to microarray datasets analyzing brain samples (the CA1 region of the hippocampus) from AD patients and comparing to brain samples (frontal lobe) from normal elderly people (Miller *et al*, 2008). This analysis produced modules of co expressed genes that are functionally related with some relevant to disease progression and others conserved between AD and normal aging. In the second study, gene profiling of laser microdissected samples from the en-

torhinal cortex were analyzed slightly differently. Modules of highly correlated genes were constructed and among these genes regulatory cis elements were identified. New links have been identified between cardiovascular diseases, AD and diabetes (Ray *et al*, 2008).
Genome wide association studies (GWAS) have recently revealed the power of analyzing a very large number of samples (>1000) (Harold *et al*, 2009; Lambert *et al*, 2009). Although getting genomic DNA samples is far much easier than getting brain samples, one would imagine that larger sample gene profiling datasets with less heterogeneous samples will improve the readout of the analysis.

1.1.2 Down syndrome (DS)

We have been interested in another complex disease, namely Down syndrome. DS results from the presence in three copies of human chromosome 21, the smallest human autosome containing about 350 known protein-coding genes (Antonarakis *et al*, 2004; Epstein, 1990; Lejeune *et al*, 1959) The mechanisms by which this aneuploidy produces the complex and variable phenotype observed in DS patients are still under discussion. The use of large scale gene expression methods such as microarrays were expected to shed light on which genes (within or outside chromosome 21) contribute to the DS phenotype as well as to the phenotypic variability. For the genes on chromosome 21, all studies have confirmed a general increase of transcription following the chromosomal imbalance, the “primary gene dosage effect”. RNA samples prepared from cells or tissues of

DS patients or mouse models showed a global over-expression of the three-copy genes (Ait Yahya-Graison *et al*, 2007; Amano *et al*, 2004; Dauphinot *et al*, 2005; FitzPatrick *et al*, 2002; Giannone *et al*, 2004; Lockstone *et al*, 2007; Mao *et al*, 2005; Mao *et al*, 2003; Potier *et al*, 2006; Saran *et al*, 2003). However, even if the mean over-expression we and others reported to be close to the expected value of 1.5, recent studies in DS cell lines have reported that about 70% of the three-copy genes were significantly below the 1.5 ratio. In these particular cell lines at least, a large proportion of the chromosome 21 transcripts were compensated for the primary gene dosage effect (Ait Yahya-Graison *et al*, 2007; Prandini *et al*, 2007).

As for non-chromosome 21 genes, results are less consistent. The aneuploidy of an entire chromosome could affect the expression of either a limited number of genes, or a large number in a more random and extensive way (Mao *et al*, 2005; Saran *et al*, 2003). Conversely classification of samples on the basis of their whole transcriptome has not been applied systematically in the published gene expression studies of DS. Rather it was unfortunately wrongly applied such as in Slonim *et al*. (Slonim *et al*, 2009). In this study they conclude to a widespread differential expression between trisomic and euploid samples based on clustering of genes differentially expressed between trisomic and euploid, excluding the chromosome 21 genes. It seems obvious that differentially expressed genes between two conditions would be able to differentiate the two conditions. Nevertheless this question regarding the regulation of gene expression for non-chromosome 21 genes is still debated, and more com-

prehensive studies assessing the variability among samples, tissues and development stages are needed.

We have designed several large scale gene expression studies in which we could measure the effects of trisomy 21 on a large number of samples in tissues or cells that are affected in DS (Dauphinot *et al*, 2005; Laffaire *et al*, 2009; Moldrich *et al*, 2009). All were performed with the Ts1Cje mouse model of DS which is a segmental trisomy of mouse chromosome 16 (MMU16) with many genes orthologous to human chromosome 21 (HSA21) present in three copies (about 95). This mouse model has the advantage of being available as large colonies of mice on B6C3SnF1/Orl mixed genetic background and rapidly screened (Sago *et al*, 2000). Experiments were designed in order to correlate gene expression changes with the phenotype observed. Two data sets focused on cerebellum since adult Ts1Cje mice show a reduction in cerebellar volume that parallel the observations in DS patients and in another mouse model of DS (Ts65Dn mice) (Baxter *et al*, 2000; Olson *et al*, 2004). The reduced size of the cerebellum and the reduced cerebellar granule cell number in Ts65Dn adults originate around birth because of a defect in granule cell precursor proliferation (Roper *et al*, 2006). In our studies, three early postnatal time points that are crucial for cerebellar development were investigated which could provide a read-out of genes involved in cerebellar hypoplasia in DS. These three time points correspond to birth (P0) and postnatal days 15 (P15) and 30 (P30). During the P0-P10 time period granule cells proliferate and migrate from the external to the internal granule cell layer and Purk-

inje cells start differentiating and growing their highly dense dendritic tree. We quantified the proliferation of granule cell precursors on fixed cerebellum slices of Ts1Cje and euploid mice at P0, P3 and P7 using immunohistochemistry and histology. A significant 30% decrease of their mitotic index was observed at P0 but not at P3 and P7, in agreement with the results obtained in Ts65Dn mice (Roper *et al*, 2006). Finally and in order to find gene expression variations in cerebellar regions rich in granule cell precursors, external granule cell layers of newborn Ts1Cje and euploid mice were dissected and analyzed on microarrays.

We also integrated data sets that contained a number of samples that was sufficient for statistical analysis ($n \geq 4$). These included the studies of Mao et al. and Saran et al. from 2003 (Mao *et al*, 2003; Saran *et al*, 2003). The first dataset contains gene expression profiles of human fetal cortex and cultured astrocytes from 4 Down syndrome cases and 4 controls. The second study produced gene expression profiles of the adult cerebellum from the Down syndrome mouse model Ts65Dn.

We included in the meta analysis the data set from Amano *et al.* 2004 from whole brain of newborn Ts1Cje mice (Amano *et al*, 2004), the one from 2007 of Lockstone *et al*. (Lockstone *et al*, 2007) and Pevsner *et al*. (unpublished GEO GSE9762) on adult cortex and cultured fibroblasts respectively, from DS patients and controls. Finally, we failed to analyze the data set from Slonim *et al*. 2009 on uncultured amniotic fluid supernatants from DS and euploid fetuses (Slonim *et al*, 2009). Indeed, from all the samples published, less

than 1000 genes were expressed in all experiments, which were not representative enough for the analysis to be meaningful.

1.2 Elements of microarray statistical analysis

The aim of this section is not to propose an exhaustive panorama of the existing methods for the analysis of microarray data, but rather to give the necessary and sufficient technical elements needed in order to understand and to reproduce the statistical treatments that we or the authors we cite have applied to the various data sets surveyed in this chapter.

1.2.1 Data normalization

In addition to the variability of interest that is due to the difference between diseased (here DS) and normal tissue, observed expression levels are also subject to the variability introduced during sample preparation, the manufacture and the processing of the arrays (labeling, hybridization and scan). Even if some of this unwanted variability can be controlled using appropriate experimental design and procedures, for example by having all experiments performed at a single time point by a single operator, some of it can not be controlled, but still needs to be corrected. The most famous of these sources is perhaps the dye bias for cDNA microarray experiments, where the efficiency, heat and light sensitivities differ for Cy3 and Cy5, resulting in a systematically lower signal for Cy3. For cDNA microarrays, the normalization procedure proposed in (Dudoit S,

2002) was shown to be efficient. It is based on Cleveland's robust locally weighted regression for smoothing scatterplots (Cleveland, 1979), and consists in fitting a *lowess curve* to the MA plot of log intensities[1] of the red and green labels and considering the residuals as the normalized log ratios.

This approach is not directly applicable to single color arrays, such as the Affymetrix or Illumina arrays considered in this chapter. However, contrarily to the current perception that the lowess normalization is only suited for normalizing two single color arrays at a time, (Sasik CH, 2004) showed that lowess can indeed be applied across $n > 2$ arrays, assuming that most genes expressions do not change notably across the n experiments.

In practice, multiple lowess proves quite similar to *quantile normalization*, which is a much lighter procedure. The principle of quantile normalization is to make the distribution of the probe intensities equal to a reference distribution for each of the n arrays. This reference distribution is the mean distribution of the n arrays, computed by sorting all p probe intensities of each array in increasing order, and computing the i^{th} reference intensity value as the mean of the i^{th} intensity values of the n arrays. Boldstad *et al.* showed the efficiency of the method, which is commonly used for the normalization of Affymetrix data (Bolstad *et al*, 2003).

[1] Transforming expression data to a log scale (any base) reduces the asymmetry of the distribution of the intensities and homogenizes their variance. Here, probe intensities are systematically $\log_2$ values.

Let us illustrate this efficiency with an example exhibiting a known undesirable effect. Gene expression was measured twice on cell lines from 12 DS patients at a two month interval on Illumina chips with 48 701 probes, the labeling being the same for the two hybridizations. Figure 1a shows the raw intensity values for the 24 arrays, those of the first hybridization in black, those of the second in grey: the two groups differ visibly. One also notices differences among the first and second hybridization, the arrays being located on two different Illumina chips (there were up to 6 arrays on the considered Illumina chips). Figure 1b shows the mean distribution used for quantile normalization.

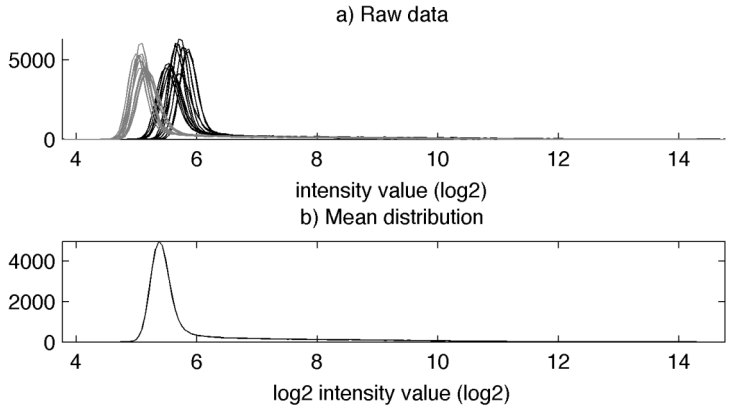

Figure 1.1 a) Distributions of the probe intensities for the 12 DS patients in black for the first hybridization, and in grey for the second one.

In order to demonstrate the efficiency of quantile normalization, we performed a PCA (see next section for further technical details about PCA) of the raw and of the quantile normalized data. Both are

shown in Figure 1.2. Whereas the arrays are grouped according to the hybridization when considering the raw data (Figure 1.2a), they are clearly grouped two by two when using the normalized data (Figure 1.2b), i.e. two arrays corresponding to the same tissue are now very close. Furthermore, the markers used for the arrays correspond to the chip they belong to. With the raw data, a chip effect can be noted (for arrays 1, 2 and 3 for example), which lessens considerably after normalization.

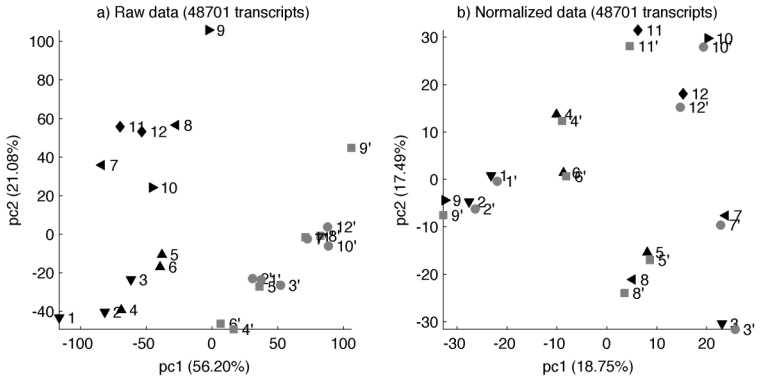

Figure 1.2 PCA of the 24 arrays along the first two principal axes, each sample being originally represented by the intensities of 48 701 probes. The arrays that where first hybridized are shown in black, the second in grey. Identical markers denote arrays located on the same Illumina chip.

This illustration using PCA leads us to the second part of the analysis, that of data visualization prior to differential analysis.

1.2.2 Dimensionality reduction and data visualization

The result of a microarray experiment involving *n* arrays with *p* probes presents itself as a *nxp* matrix of – now normalized – intensi-

ties, which can be viewed as the representation of *n* tissues by the intensities of their *p* genes or probes (typically hundreds or thousands), or conversely, as the representation of the *p* probes by their expression in *n* tissues (typically tens or even less). In this chapter, we will focus on the first view, which raises the problem of visualizing objects in a high dimension space, see McLachlan *et al*. for an exhaustive analysis of both views (McLachlan GJ, 2004).

A common way mean of reducing dimensionality is to carry out a principal component analysis (PCA): the principle of PCA is precisely to project multidimensional data to a lower dimension space retaining as much as possible of the variability of the data.

A first purpose of such a PCA prior to differential analysis is to detect outliers and possible biases, as well as to validate their correction by a proper normalization: in the previous example, PCA showed the reduction of the effect of having different hybridizations by quantile normalization.

A second goal may be to exhibit groups of tissues, especially according to the known differences between them, such as normal and DS tissues. In this context, we must insist that PCA is an unsupervised procedure, whose only property is that the projection in the *d*-dimension space generated by the *d* first principal axes has the highest variance among all possible projections in a *d*-dimension space. The direction of maximum variance being composed of variance *within* the groups and variance *between* the groups, the first principal components need not necessarily reflect the direction in the probe space that is best for revealing the group structure of the tissues.

However, conversely, if PCA indeed reveals clusters, it implies a large variance between the groups, i.e. the presence of many differentially expressed probes. In this chapter, whose main object is the characterization of DS versus normal tissues, we will systematically present three different PCA of the data: the PCA on all transcripts, the PCA on the HSA21 chromosome (or the equivalent part of MMU16 chromosome in three copies in the case of mouse models) transcripts, and the PCA on the remaining euploid transcripts. Because of the gene dosage effect, PCA on the three-copy transcripts systematically separates normal from DS tissues. If PCA without the three-copy transcripts does, it means that the whole transcriptome is affected by DS. This might be a useful and complementary information to differential analysis, especially in the case of less powerful experiments (i.e. with too few samples) where only a few genes can be determined as significantly differentially expressed. We could have completed the PCA with a cluster analysis, however for all the data sets presented in the next section, hierarchical clustering never exhibits two separate clusters of DS and euploid samples when PCA does not (while the opposite case often appears).

Now, a few technical details need to be clarified. The lower dimensional space used for the PCA projection is the space generated by the eigenvectors of the feature (probe) correlation matrix corresponding to its largest eigenvalues, called principal axes, see for example Johnson & Wichern (DW, 2002). In many applications, it happens that some features have completely different scalings. For example, one of the features may have been measured in meters and

another one, by design or accident, in micrometers. Since eigenvalues are scale dependent, it might be appropriate in such cases to rescale all features to the same scale, which amounts to use the correlation matrix of the features, instead of their covariance matrix. In the case of gene expression, rescaling leads to give low- or unexpressed genes (the variance of which corresponds to noise) the same importance as highly expressed genes, which is indeed not desirable.

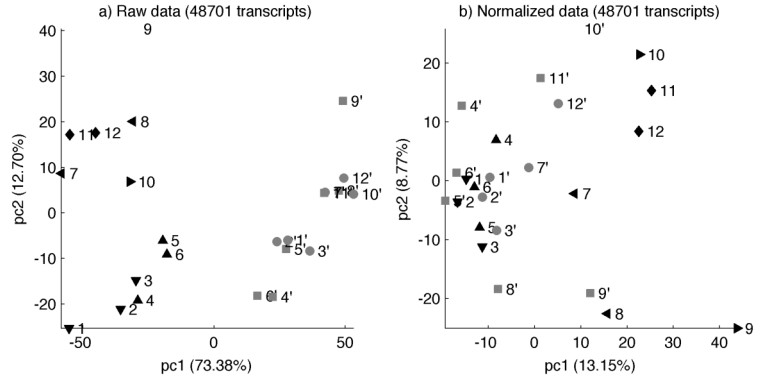

Figure 1.3 Same as in Figure 1.2, except that the PCA is performed on the correlation matrix (i.e. with rescaled probe intensities).

To illustrate this, Figure 1.3 shows the two PCA of the 12 DS tissues hybridized twice, this time with rescaled intensities. On the raw data, the main variability being due to the different hybridizations, the projection is quite similar as when performed on the un-rescaled data. But on the normalized data, where this effect has been removed, we see that we have lost the close neighborhood of the couples of arrays corresponding to the same tissue. Thus, all PCA presented in this chapter are performed on the normalized, un-rescaled

probe intensities. On PCA figures, the percentage indicated in parentheses in a principal axis label corresponds to the proportion of the variance explained by this axis.

Finally, whatever the platform, intensity values are usually provided with "calls" (present, absent, marginal) and/or detection *p*-values. The PCA shown in the next section have been performed on the transcripts considered present or with expression *p*-values lower than 5% or 1% for all *n* arrays (when the *p*–values were not available, we chose a cutoff on the probe intensity so as to obtain the same proportion of expressed transcripts). For our example, 10 626 transcripts are considered present on all arrays with a threshold of 5% on the detection *p*-value, and the PCA on these 10 626 transcripts is shown on Figure 1.4. The benefit of removing the non-expressed genes is especially noticeable on the raw data, where the couples are now visible (though still much less than on the normalized data).

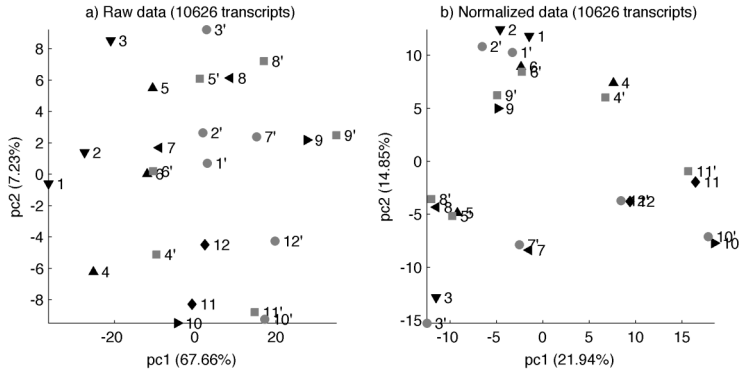

Figure 1.4 Same as in Figure 1.2, but only for the 10 626 transcripts considered present on all 24 arrays.

1.2.3 Differential analysis at the gene level

The purpose of differential analysis at the gene (or transcript) level is to identify genes whose expression level differs from a condition to another, based on statistical hypothesis testing. Almost all experiments analyzed in this chapter involve two groups of tissues, normal and DS tissues, usually unpaired. Thus, the traditional *t*-test is relevant for our purpose, which reformulates the question of differential expression of gene *i* in terms of a null hypothesis H_{0i} "there is no difference of mean expression for the transcript *i* between the normal and the DS tissues". Student's *t*-test is indeed the test that was used in almost all reviewed papers, and that we used for the experiments for which no analysis was published. Once the *t*-statistic is computed, the classical decision rule to accept or reject H_{0i} consisting in controlling the type I error probability can be applied for declaring each gene differentially expressed (DE) or not.

However, the specificity of microarray differential analysis lies in the large number of tests to be performed: as many as probes on the array, or at least, as expressed transcripts. The question of differential expression must hence be restated as a multiple testing problem. The first attempts to solve this problem aimed at controlling the Family Wise Error Rate (FWER), that is the probability to have at least one false positive, and the procedures of Bonferroni and Sidak are the most widely used to this end. An alternative approach has been proposed in Benjamini & Hochberg, based on the principle that

the designer of a microarray experiment is ready to accept some type I errors, provided that their number is small as compared to the total number of rejected hypotheses (i.e. of genes decided DE) (Benjamini, 1995). This approach aims at controlling the False Discovery Rate (FDR), i.e. the expected proportion of false positives among the total number of positives. Storey & Tibshirani proposed to define an equivalent of the *p*-value for the control of the FDR, the *q*-value (Storey and Tibshirani, 2003). If genes with *q*-values smaller than 5% are decided DE, then there is a FDR of 5% among the DE genes. In practice, the *q*-values can be computed from the *p*-values, and are often called "adjusted *p*-values". Most papers reviewed here use the *q*-values corresponding to Benjamini and Hochberg's rule to control the FDR, possibly with an estimation of the number m_0 of true null hypotheses H_{0i} (i.e. the number of not DE genes), see (Storey *et al*, 2003).

Let us take the example of Pevsner's data available on GEO without published analysis. We analyze the expression of human skin fibroblasts, from five normal and five from DS individuals, as measured by Affymetrix arrays involving 54 675 probes (we use the calls and normalized intensity values calculated by the MAS 5 or GCOS software as available on GEO).

We performed the *t*-test for the transcripts which were considered present at least three times in both conditions, i.e. for 22 606 transcripts. The histogram of the corresponding *p*-values is shown in Figure 1.5a. Their distribution is far from being uniform, which means that many genes are differentially expressed. As a matter of

fact, when controlling only an individual type I error risk of 5% using the *p*-values, 2 938 transcripts are decided DE.

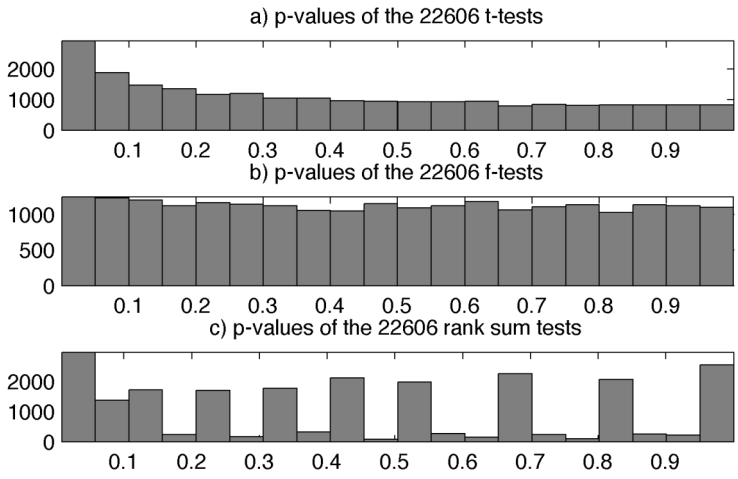

Figure 1.5 Histograms of the *p*-values of three statistical hypothesis tests: a) Student's test of equality of the means, b) Fisher's test of equality of the variances, c) Wilcoxon's non parametric rank sum test.

The number of true null hypotheses m_0 is roughly given the number of *p*-values in the flat part of the histogram (the on which would correspond to the uniform distribution). It can be estimated at 17 108 according to Storey & Tibshirani (Storey *et al*, 2003) (with the tuning parameter $\lambda = 0.5$). Using this estimate for the computation of the *q*-values, only 11 transcripts are decided DE when imposing a FDR of 5% (76 with a FDR of 10%).

Let us now discuss the relevance of the *t*-tests. For a *t*-test to be valid, in addition to the absence of correlation of the measurements,

two assumptions are supposed to be true: the normality of the data, and the equality of the variance in the two conditions. If, like here, the number of measurements is small, the normality can hardly be tested efficiently. But (assuming normality), the Fisher test of the equality of the variances can indeed be performed. For our example, the histogram of its *p*-values is shown on Figure 1.5b. The distribution being uniform, as would be the case if all null hypotheses were true, we can conclude that the variances are equal in the two conditions, and this justifies the use of the *t*-test.

If the previous Fisher test establishes that many variances are different, or if non-normality is suspected, a solution could be to use Wilcoxon's non parametric rank sum test. A problem then arises with small samples that is clearly visible on Figure 1.5c: the Wilcoxon statistic being discrete, so are the *p*-values and hence the *q*-values. Here, the smallest *q*-value equals 0.26, one cannot impose the FDR to be smaller than 26% (627 transcripts are DE with a FDR of 26%).

Thus, in the situations where the assumptions that the data is normal and/or that the variances are unequal are really unsuitable, the best alternative is to estimate the empirical distribution of the *t*-statistics using permutation methods such as bootstrap or permutations, see Westfall & Young (SS, 1992). A particularly popular and efficient permutation method is the Significance Analysis of Microarrays (SAM) proposed by Tusher *et al*. (Tusher *et al*, 2001).

1.2.4 Differential analysis at the gene set level

In order to take full advantage of the differential analysis at the gene level, which merely provides an unstructured list of DE genes, an integration at a higher level is necessary. Thus, the identification of predefined sets of biologically related genes enriched or depleted with DE genes has become a routine part of the analysis and of the interpretation of microarray data.

Gene sets can be built on several criteria. These criteria can be based on the available annotation sources such as GO, the Gene Ontology project, KEGG , the Kyoto Encyclopedia of Genes and Genomes, or GenMAPP for example. In the case of DS studies, other gene sets of interest are the HSA21 genes, or even genes belonging to the specific bands of HSA21, as analyzed in Slonim *et al*. (Slonim *et al*, 2009).

The first and most common approaches used to identify gene sets enriched or depleted in DE genes are based on the two-by-two contingency table obtained by classifying the genes into "being DE or not DE" on one hand, and "belonging to the gene set *S* of interest or not" on the other hand. The statistical significance of the overlap between being DE and belonging to *S* can be established more or less equivalently using the hypergeometric test, Fisher's exact test or chi-square tests, as proposed by many GO processing tools, see Rivals *et al*. (Rivals *et al*, 2007) for a review. Though these approaches are quite efficient, their limitation is to require a preliminary categorization into DE and not DE genes, i.e. they necessitate the choice of a significance cutoff (be it in the form an individual type I error risk, or of a FDR), which is always arbitrary.

More recently, several methods have been proposed that avoid categorizing the genes into DE and not DE, by simply using the *t*-statistics or the associated *p*-values. For example, Sartor *et al*. (Sartor *et al*, 2009) propose a very intuitive logistic regression approach, *LRpath*. Given a gene set S of interest, a target variable y is defined as having value 1 for the genes in S, and value 0 for the others. The –log(*p*-value) is used as explanatory variable x, and y is modeled by a logistic function of x, $1/(1+\exp(-(ax+b))$. If the slope a is found significant according to a classic Wald test, the subset is decided significantly enriched ($a > 0$) or depleted ($a < 0$) in DE genes.

Let us illustrate the enrichment/depletion analysis using the hypergeometric test and LRpath on the example of Pevsner's data, simply defining the gene subsets of interest according to the chromosomes they belong to. For the hypergeometric test, we define a threshold of 5% for the *p*-values, i.e. genes with $p \leq 5\%$ are considered DE. The percentage of DE transcripts for each chromosome is shown on Figure 1.6: with 49% of DE transcripts, chromosome 21 appears clearly enriched. But is chromosome 3 significantly depleted with 10.6%?

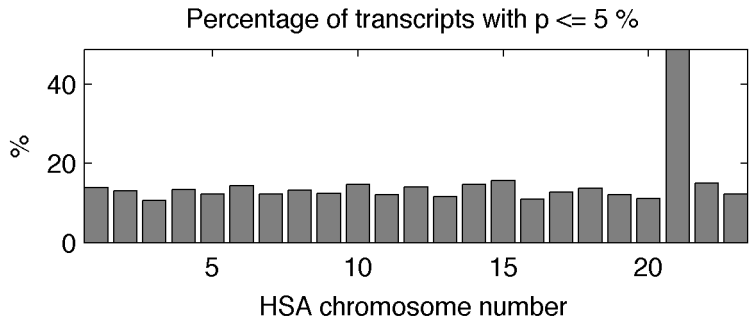

Figure 1.6 Percentage of DE transcripts for each chromosome (the 23th is the X).

The *p*-values of the hypergeometric test and of LRpath are shown in Table 1. It is interesting to note that they often disagree (Spearman's $\rho = 0.34$), hence the interest for the recent approaches avoiding categorization into DE/not DE.

			p		
chromosome	transcripts	DE transcripts	Hypergeometric	LRpath	Wilcoxon
1	2087	291	0.44	0.63	0.058
2	1461	190	0.7	0.36	0.36
3	1184	125	0.003	$7.2\ 10^{-4}$	0.0054
4	788	105	1	0.031	0.015
5	1054	130	0.33	0.32	0.12
6	1079	155	0.35	0.18	0.2
7	954	117	0.32	0.42	0.87
8	712	94	0.94	0.56	0.3
9	823	102	0.43	0.99	0.65
10	839	123	0.29	0.14	0.18
11	1103	134	0.23	0.28	0.92
12	1095	154	0.52	0.072	0.061
13	430	50	0.31	0.59	0.5
14	700	103	0.32	0.042	0.073
15	675	106	0.085	0.92	0.78
16	839	92	0.037	0.0045	0.035
17	1117	143	0.6	0.44	0.34
18	329	45	0.92	0.67	0.35

19	1109	135	0.24	0.029	0.058
20	561	62	0.11	0.2	0.13
21	219	107	7.9e-11	0	0
22	427	64	0.36	0.93	0.86
X	665	82	0.46	0.037	0.038

Table 1.1. Enrichment/depletion tests for the chromosomes (except chromosome Y with too few transcripts).

On the other hand, we have computed Wilcoxon's rank sum statistic for the *p*-values (one group being the set *S* corresponding to one chromosome, the second all the other transcripts), which is in good agreement with LRpath (Spearman's $\rho = 0.80$): we see that this simple test is a good indicator for enrichment/depletion.

For the chromosomes for which the three tests agree (chromosomes 3, 1 and 21), Figure 1.7 shows the results of the logistic regression for LRpath. Chromosomes 3 and 15 are depleted in DE genes, whereas chromosome 21 is enriched, as expected.

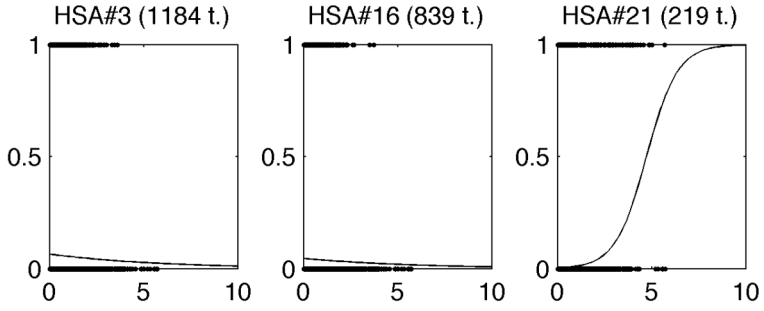

Figure 1.7 Logistic regression on $-\log_{10}(p)$ for the chromosoms dignificantly enriched or depleted in DE transcrips.

Gene set enrichment analysis (GSEA), in the version proposed by Subramanian *et al.* (Subramanian *et al*, 2005), is used for example in Slonim *et al*. (Slonim *et al*, 2009) in order to detect enriched bands

on chromosome 21. Like the Wilcoxon test, GSEA uses the complete distribution of the *p*-values, divides the genes into the set *S* of interest and the rest, and ranks them according to the *p*-value. But the enrichment score is computed by walking down the list, increasing a running sum statistic when a gene in *S* is encountered, and decreasing it when the gene is outside *S*; the enrichment score is the maximum deviation from zero encountered during the walk, and its significance is evaluated by estimating the null distribution through permutations (i.e. the correlation structure of the gene expression is taken into account, what the simple Wilcoxon test does not).

Finally, let us mention ProbeCD, the method proposed by Vencio & Schmulevitch (Vencio and Shmulevich, 2007). ProbCD not only presents the advantage of not requiring the choice of a significance cutoff, but it is also able to take the uncertainty in the gene annotation into account.

1.3 Results

Table 1.2 summarizes the data sets considered in this study. As detailed in the previous section, PCA was applied to the normalized datasets, on the transcripts expressed across all arrays. Three different PCA were systematically performed: one with all expressed transcripts, another with the expressed three-copy transcripts only (those of HSA21 or of the triplicated part of MMU16), and the last with the remaining euploid transcripts. The ultimate goal of this analysis was to visualize whether samples would be grouped according to their

genotype (DS or control) and in which conditions (with all genes, with triplicated genes only and/or with the euploid genes only). Figure 1.8 to Figure 1.14 show the results from these PCA applied to the data sets. On all of them, DS samples are shown in black, control samples in white.

From the analysis including the three copy genes only, samples from DS models are very clearly separated from samples from euploid controls. This is due to the global overexpression of the three copy genes that has been largely described previously. Indeed, in DS, three-copy genes are globally over-expressed by a mean factor of 1.5. However at the single gene resolution, this 1.5 overexpression does not strictly apply and several comprehensive studies have shown that compensation and amplification mechanisms do exist. Compensated three-copy genes will not be over-expressed while amplified three-copy genes will be over-expressed by a factor significantly higher than 1.5 (Ait Yahya-Graison *et al*, 2007; Prandini *et al*, 2007).

Authors	Sample type	Number of expressed genes	Statistical test	Differentially expressed genes	Differentially expressed genes from Hsa21
Mao et al.	Human fetal cortex from	15106	ANOVA (5%)	725	
Mao et al.	Human fetal cultured astrocytes	15106	ANOVA (5%)	679	
Saran et al.	Cerebellum from adult Ts65Dn mice				
Amano et al.	Whole brain of new born Ts1Cje mice	10602			
Dauphinot et al.	Cerebellum from P0, P15 and P30 Ts1Cje mice	8287			
Lockstone et al.	Human adult prefrontal and dorsolateral cortex				
Pevsner et al.	Human cultured fibroblasts		t-test FDR 10%	11	
Pevsner et al.	Human cultured fibroblasts		t-test 1%	873	
Slonim et al.	Human uncultured amniotic fluid				

Table 1.2: list of datasets used in this study.

When comparing the PCA performed on all expressed genes and applied to the various sets of data, the results are quite different. With three sets of data (Figure 1.8, Figure 1.9, Figure 1.10), samples from DS models are separated from samples from euploid controls, although comparatively less than when the analysis is applied to three copy genes only. With one set of data (Figure 1.11), the separation is present in a lesser extent. Finally, with the last three sets of data all samples are mixed and no separation is clearly depicted (Figure 1.11, Figure 1.12, Figure 1.13 and Figure 1.14).

For the datasets with a clear separation, we tested the influence of the three-copy genes. We removed them and run the PCA on all expressed genes except the three-copy genes. The right panels of Figure 1.8, Figure 1.9, and Figure 1.10 show the same projections than the right panels, thus suggesting that the categorization into normal and DS samples is not due to the overexpression of the three-copy genes only but rather to a modification of the whole transcriptome.

We tried to analyze the reasons why datasets would behave differently towards PCA. One obvious reason would be that there is a factor which is stronger than the genotype (DS or control) that drives the separation of samples. This is the case for samples that include different time points during development in the same analysis (Figure 1.13 and Figure 1.14). On Figure 1.13, samples segregate with the litter. In this particular analysis the external granular layer of the

cerebellum was dissected at birth (P0) from the Ts1Cje mice. What is called P0 can in fact be between birth and P1 depending on the time of birth during the day or during the night. According to the PCA, samples were separated according to the litter, indicating that the up to 12-24 hours can impact seriously on the transcriptome of this particular cell type. On Figure 1.14, it is clear that the impact of development on gene expression is much bigger than the impact of trisomy 21, as was discussed previously (Dauphinot *et al*, 2005).

In the case of the data set from Amano *et al*., again whole brains were obtained at birth with possibly an up to 24hours difference between litters and even between pups. It is known that the embryos from a litter are not totally equivalent in term of development depending on their position in the uterus.

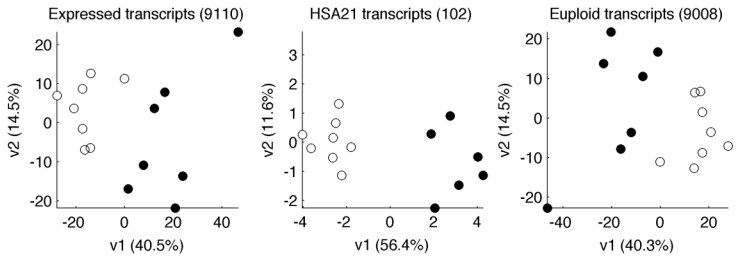

Figure 1.8 PCA of the data described in Lockstone *et al*. (Lockstone *et al*, 2007).

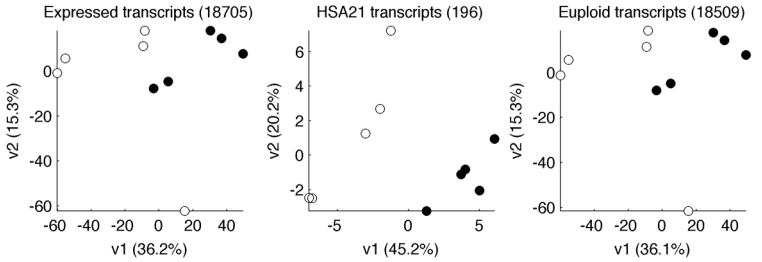

Figure 1.9 PCA of the data described in Pevsner (GEO GSE9762).

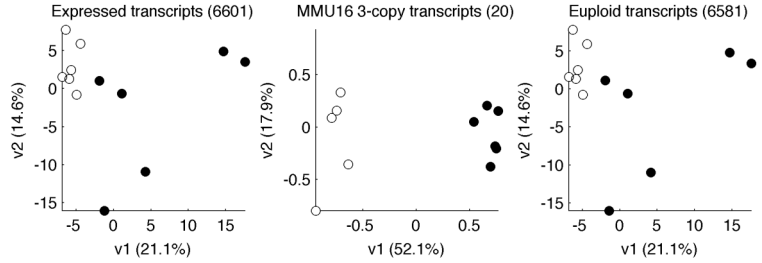

Figure 1.10 PCA of the data described in Saran *et al.* (Saran *et al*, 2003).

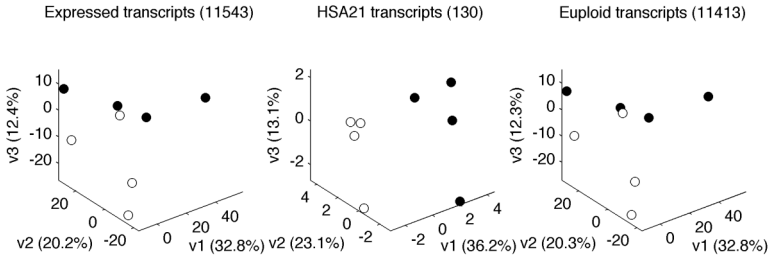

Figure 1.11 PCA of the data described in Mao *et al*. (Mao *et al*, 2003).

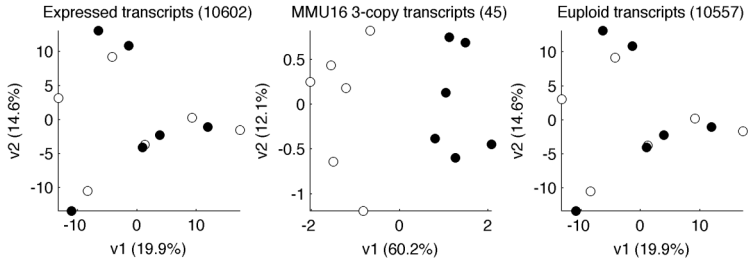

Figure 1.12 PCA of the data described in Amano *et al.* (Amano *et al*, 2004).

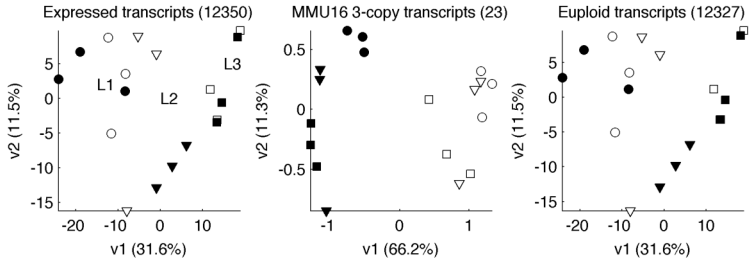

Figure 1.13 PCA of the data described in Laffaire *et al.* (Laffaire *et al*, 2009). The three markers correspond to three different litters.

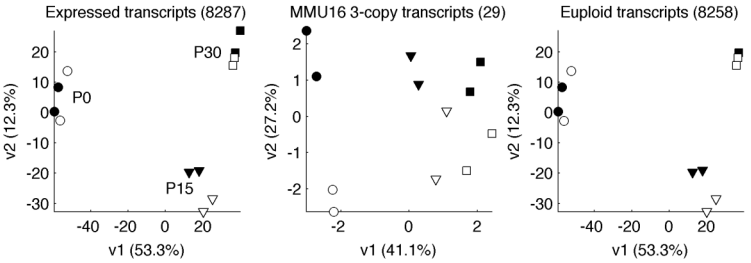

Figure 1.14 PCA of the data described in Dauphinot *et al*. (Dauphinot *et al*, 2005). The three markers correspond to three different developmental stages (P0, P15, P30).

1.4 Conclusion

Functional genomics needs to be applied to complex diseases. In the case of Down syndrome, we have used gene expression profiling in various human samples or in mouse models and shown that, when

we selected the three-copy genes for the analysis, samples were separated according to their genotype (DS or euploid) in all data sets. This is due to the global over-expression of the three-copy genes in DS or in mouse models. When using all expressed genes, samples were separated according to their genotype only in some datasets. This suggests that, in the datasets with no separation, there is a factor other than trisomy that strongly impacts on the transcriptome. We have shown that this factor can be the postnatal development of the cerebellum.

It now remains to be shown whether, beside the global over-expression of the three-copy genes, there will be a common set of genes that is modified in all samples analyzed. We and others have tried to search for this group of genes without any frank success. To get a more precise answer, very large sets of data will need to be generated, or alternatively, gene profiling should be obtained from single cells either trisomic or euploid. At the present time, gene expression profiles are obtained from samples that are too variable (different tissues or cells, different time points during development, different individuals with too many inter-individual variations and not enough samples). If the common set of dysregulated genes does not exist, it suggests that the most important trend is the overexpression of the three-copy genes themselves that secondarily impacts on the whole transcriptome in a “stochastic” way.

References

Ait Yahya-Graison E, Aubert J, Dauphinot L, Rivals I, Prieur M, Golfier G, et al (2007). Classification of human chromosome 21 gene-expression variations in Down syndrome: impact on disease phenotypes. Am J Hum Genet 81(3): 475-491.

Amano K, Sago H, Uchikawa C, Suzuki T, Kotliarova SE, Nukina N, et al (2004). Dosage-dependent over-expression of genes in the trisomic region of Ts1Cje mouse model for Down syndrome. Hum Mol Genet 13(13): 1333-1340.

Antonarakis SE, Lyle R, Dermitzakis ET, Reymond A, Deutsch S (2004). Chromosome 21 and down syndrome: from genomics to pathophysiology. Nat Rev Genet 5(10): 725-738.

Baxter LL, Moran TH, Richtsmeier JT, Troncoso J, Reeves RH (2000). Discovery and genetic localization of Down syndrome cerebellar phenotypes using the Ts65Dn mouse. Hum Mol Genet 9(2): 195-202.

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B57(1): 289-300.

Blalock EM, Geddes JW, Chen KC, Porter NM, Markesbery WR, Landfield PW (2004). Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proc Natl Acad Sci U S A 101(7): 2173-2178.

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): 185-193.

Bontoux N, Dauphinot L, Vitalis T, Studer V, Chen Y, Rossier J, et al (2008). Integrating whole transcriptome assays on a lab-on-a-chip for single cell gene profiling. Lab Chip 8(3): 443-450.

Cleveland WS (1979). Robust locally weighted regression and smoothing scatterplots. American Statistical Association 74: 829-836.

Dauphinot L, Lyle R, Rivals I, Dang MT, Moldrich RX, Golfier G, et al (2005). The cerebellar transcriptome during postnatal development of the Ts1Cje mouse, a segmental trisomy model for Down syndrome. Hum Mol Genet 14(3): 373-384.

Dudoit S YY, Callow MJ and Speed T (2002). Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. Statistical Sinica 12: 111-139.

Dunlop MJ, Cox RS, 3rd, Levine JH, Murray RM, Elowitz MB (2008). Regulatory activity revealed by dynamic correlations in gene expression noise. Nat Genet 40(12): 1493-1498.

Johnson RA, Wichern DW (2002). Applied multivariate statistical analysis. Prentice-Hall.

Emilsson L, Saetre P, Jazin E (2006). Alzheimer's disease: mRNA expression profiles of multiple patients show alterations of genes involved with calcium signaling. Neurobiol Dis 21(3): 618-625.

Epstein CJ (1990). The consequences of chromosome imbalance. Am J Med Genet Suppl 7: 31-37.

FitzPatrick DR, Ramsay J, McGill NI, Shade M, Carothers AD, Hastie ND (2002). Transcriptome analysis of human autosomal trisomy. Hum Mol Genet 11(26): 3249-3256.

Giannone S, Strippoli P, Vitale L, Casadei R, Canaider S, Lenzi L, et al (2004). Gene expression profile analysis in human T lymphocytes from patients with Down Syndrome. Ann Hum Genet 68(Pt 6): 546-554.

Golfier G, Lemoine S, van Miltenberg A, Bendjoudi A, Rossier J, Le Crom S, et al (2009). Selection of oligonucleotides for whole-

genome microarrays with semi-automatic update. Bioinformatics 25(1): 128-129.

Harold D, Abraham R, Hollingworth P, Sims R, Gerrish A, Hamshere ML, et al (2009). Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat Genet 41(10): 1088-1093.

Laffaire J, Rivals I, Dauphinot L, Pasteau F, Wehrle R, Larrat B, et al (2009). Gene expression signature of cerebellar hypoplasia in a mouse model of Down syndrome during postnatal development. BMC Genomics 10: 138.

Lambert JC, Heath S, Even G, Campion D, Sleegers K, Hiltunen M, et al (2009). Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat Genet 41(10): 1094-1099.

Lejeune J, Gautier M, Turpin R (1959). [Study of somatic chromosomes from 9 mongoloid children.]. C R Hebd Seances Acad Sci 248(11): 1721-1722.

Lockstone HE, Harris LW, Swatton JE, Wayland MT, Holland AJ, Bahn S (2007). Gene expression profiling in the adult Down syndrome brain. Genomics 90(6): 647-660.

Losick R, Desplan C (2008). Stochasticity and cell fate. Science 320(5872): 65-68.

Lu T, Pan Y, Kao SY, Li C, Kohane I, Chan J, et al (2004). Gene regulation and DNA damage in the ageing human brain. Nature 429(6994): 883-891.

Maamar H, Raj A, Dubnau D (2007). Noise in gene expression determines cell fate in Bacillus subtilis. Science 317(5837): 526-529.

Macarthur BD, Ma'ayan A, Lemischka IR (2009). Systems biology of stem cell fate and cellular reprogramming. Nat Rev Mol Cell Biol 10(10): 672-681.

Maes OC, Xu S, Yu B, Chertkow HM, Wang E, Schipper HM (2007). Transcriptional profiling of Alzheimer blood mononuclear cells by microarray. Neurobiol Aging 28(12): 1795-1809.

Mao R, Wang X, Spitznagel EL, Jr., Frelin LP, Ting JC, Ding H, et al (2005). Primary and secondary transcriptional effects in the developing human Down syndrome brain and heart. Genome Biol 6(13): R107.

Mao R, Zielke CL, Zielke HR, Pevsner J (2003). Global up-regulation of chromosome 21 gene expression in the developing Down syndrome brain. Genomics 81(5): 457-467.

McLachlan GJ DKaAC (2004). Analyzinf microarray gene expression data. Wiley.

Miller JA, Oldham MC, Geschwind DH (2008). A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging. J Neurosci 28(6): 1410-1420.

Moldrich RX, Dauphinot L, Laffaire J, Vitalis T, Herault Y, Beart PM, et al (2009). Proliferation deficits and gene expression dysregulation in Down's syndrome (Ts1Cje) neural progenitor cells cultured from neurospheres. J Neurosci Res.

Nagasaka Y, Dillner K, Ebise H, Teramoto R, Nakagawa H, Lilius L, et al (2005). A unique gene expression signature discriminates familial Alzheimer's disease mutation carriers from their wild-type siblings. Proc Natl Acad Sci U S A 102(41): 14854-14859.

Olson LE, Roper RJ, Baxter LL, Carlson EJ, Epstein CJ, Reeves RH (2004). Down syndrome mouse models Ts65Dn, Ts1Cje, and Ms1Cje/Ts65Dn exhibit variable severity of cerebellar phenotypes. Dev Dyn 230(3): 581-589.

Potier MC, Rivals I, Mercier G, Ettwiller L, Moldrich RX, Laffaire J, et al (2006). Transcriptional disruptions in Down syndrome: a

case study in the Ts1Cje mouse cerebellum during post-natal development. J Neurochem 97 Suppl 1: 104-109.

Prandini P, Deutsch S, Lyle R, Gagnebin M, Delucinge Vivier C, Delorenzi M, et al (2007). Natural gene-expression variation in Down syndrome modulates the outcome of gene-dosage imbalance. Am J Hum Genet 81(2): 252-263.

Ray M, Ruan J, Zhang W (2008). Variations in the transcriptome of Alzheimer's disease reveal molecular networks involved in cardiovascular diseases. Genome Biol 9(10): R148.

Rivals I, Personnaz L, Taing L, Potier MC (2007). Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics 23(4): 401-407.

Roper RJ, Baxter LL, Saran NG, Klinedinst DK, Beachy PA, Reeves RH (2006). Defective cerebellar response to mitogenic Hedgehog signaling in Down [corrected] syndrome mice. Proc Natl Acad Sci U S A 103(5): 1452-1456.

Sago H, Carlson EJ, Smith DJ, Rubin EM, Crnic LS, Huang TT, et al (2000). Genetic dissection of region associated with behavioral abnormalities in mouse models for Down syndrome. Pediatr Res 48(5): 606-613.

Saran NG, Pletcher MT, Natale JE, Cheng Y, Reeves RH (2003). Global disruption of the cerebellar transcriptome in a Down syndrome mouse model. Hum Mol Genet 12(16): 2013-2019.

Sartor MA, Leikauf GD, Medvedovic M (2009). LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. Bioinformatics 25(2): 211-217.

Sasik CH WRaCJ (2004). Microarray truths and consequences. Journal of Molecular Endocrinology 33: 1-9.

Schena M, Shalon D, Davis RW, Brown PO (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270(5235): 467-470.

Singh A, Weinberger LS (2009). Stochastic gene expression as a molecular switch for viral latency. Curr Opin Microbiol http://www.doodle.com/yix62vkik5gks8v612(4): 460-466.

Slonim DK, Koide K, Johnson KL, Tantravahi U, Cowan JM, Jarrah Z, et al (2009). Functional genomic analysis of amniotic fluid cell-free mRNA suggests that oxidative stress is significant in Down syndrome fetuses. Proc Natl Acad Sci U S A 106(23): 9425-9429.

Storey JD, Tibshirani R (2003). Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100(16): 9440-9445.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102(43): 15545-15550.

Tusher VG, Tibshirani R, Chu G (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98(9): 5116-5121.

Vencio RZ, Shmulevich I (2007). ProbCD: enrichment analysis accounting for categorization uncertainty. BMC Bioinformatics 8: 383.

Westfall PH, Young SS (1992). Resampling-based multiple testing. Wiley

Acknowledgements:

The authors wish to thank The European program AnEUploidy and the Fondation Jérôme Lejeune for their financial support.