



UNIVERSITA' DEGLI STUDI DI MILANO

SCUOLA DI DOTTORATO IN SCIENZE BIOMEDICHE CLINICHE E
SPERIMENTALI

DIPARTIMENTO DI SCIENZE CLINICHE E DI COMUNITA'

DOTTORATO DI RICERCA IN STATISTICA BIOMEDICA

XXV CICLO

TESI DI DOTTORATO DI RICERCA

Statistical Methods for the Assessment of Within-Platform
Repeatability and Between-Platform Agreement in microRNA
Microarray Technologies

Settore Scientifico Disciplinare MED/01

DOTTORANDO: Dr. Niccolo' Bassani
Matricola: R08561

Tutor: Prof. Elia Biganzoli

Coordinatore del Corso: Prof. Adriano DeCarli

Anno Accademico 2011/2012

*"Each day a new reason to give up
Each day another reason to sigh
A hundred thousand ways to live up
A hundred thousand ways to try"*

(Gammaray - Heaven Can Wait)

Contents

1	Introduction	1
1.1	New frontiers in biological research...	1
1.2	...and related statistical issues	3
1.3	Thesis organization	5
2	MicroRNA, Microarrays and Reliability	7
2.1	MicroRNAs: a short identikit	7
2.2	Microarray technology	9
2.3	Statistical issues related to reliability and agreement	14
2.3.1	Reliability	14
2.3.2	Agreement	17
2.4	Literature review	21
2.5	Limitations and proposed solutions	28
3	Platforms features and data processing	31
3.1	A terminology caveat	31
3.2	Samples and experimental design	32
3.3	The platforms: features and data processing	33
3.3.1	Affymetrix	33
3.3.2	Agilent	34
3.3.3	Illumina	36
3.4	miRNA selection and Normalization	38
3.4.1	Automatic filtering or manual selection?	38
3.4.2	Normalization: to do or not to do?	40
3.4.2.1	Quantile Normalization	43
3.4.2.2	Loess Normalization	44
4	Statistical Methods for Repeatability and Agreement	46
4.1	Within-platform Repeatability	46

4.1.1	A random effects model	46
4.1.2	CCC and Overall CCC	54
4.1.2.1	The standard CCC from Lin (1989)	54
4.1.2.2	The CCC for repeated measures from King <i>et al.</i> (2007) . .	58
4.1.2.3	The Overall CCC from Barnhart <i>et al.</i> (2002)	59
4.2	Between-platforms agreement	60
4.2.1	The Bland-Altman method	60
4.2.2	Measurement error and agreement assessment	63
4.3	Software	69
5	Results	70
5.1	Data description	70
5.1.1	Affymetrix	70
5.1.2	Agilent	73
5.1.3	Illumina	75
5.2	Repeatability analysis	76
5.2.1	Random effects models	76
5.2.1.1	Affymetrix	76
5.2.1.2	Agilent	85
5.2.1.3	Illumina	89
5.2.2	CCC and Overall CCC	94
5.2.2.1	Affymetrix	94
5.2.2.2	Agilent	97
5.2.2.3	Illumina	100
5.3	Agreement Assessment	103
5.3.1	Raw data	103
5.3.2	Normalized data	111
6	Discussion	116
6.1	Within-Platform Repeatability	116
6.2	Between-Platforms Agreement	120

7	Conclusions and Future Developments	123
A	R functions for Repeatability Analysis	125
A.1	ANOVA table for nested-factorial random effects models	125
A.2	Overall CCC - point estimate	130
B	R functions for Concordance Assessment	132
B.1	Liao-Capen modified Bland-Altman approach	132
	Bibliography	137

List of Tables

1.1	Type of measure and data dimensionality in genomics	3
2.1	Sequences for miRNAs belonging to let-7 family	13
2.2	Simulation results for non concordant data	20
3.1	Probeset replication for Agilent platform	34
3.2	ProcessedSignal for miRNA hsa-let-7a*	35
3.3	TotalProbeSignal for miRNA hsa-let-7a*	35
3.4	TotalGeneSignal for miRNA hsa-let-7a*	36
3.5	Pearson's ρ between Affymetrix replicates (hREF) - variance filtering	39
3.6	Pearson's ρ between Affymetrix replicates (hREF) - mean filtering	39
3.7	Pearson's ρ between platforms (hREF) - variance filtering	40
3.8	Pearson's ρ between platforms (hREF) - mean filtering	40
4.1	Experimental factors	48
4.2	3-way Nested-Factorial ANOVA table	50
4.3	Variance Components estimates - 3-way model	51
4.4	2-way factorial ANOVA table	53
4.5	Variance Components estimates - 2-way model	53
4.6	Estimates of ρ and ρ_c , simulated data	55
5.1	ANOVA table for line A498 - Affymetrix	80
5.2	ANOVA table for sample hREF - Affymetrix	82
5.3	Point and interval estimates (95% confidence level) for variance components and ICC of the different sources of variability - Affymetrix platform.	83
5.4	Point estimates for variance components and point/interval estimates (95% confidence level) for the ICC of different sources of variability - filtered data, Affymetrix platform	84
5.5	ANOVA table for line A498 - Agilent	86

5.6	ANOVA table for sample hREF - Agilent	87
5.7	Point and interval estimates (95% confidence level) for variance components and ICC of the different sources of variability - Agilent Platform.	87
5.8	Point estimates for variance components and point/interval estimates (95% confidence level) for the ICC of different sources of variability - filtered data, Agilent platform	88
5.9	ANOVA table for line A498 - Illumina	90
5.10	ANOVA table for sample hREF - Illumina	90
5.11	Point and interval estimates (95% confidence level) for variance components and ICC of the different sources of variability - Illumina Platform.	91
5.12	Point estimates and interval estimates (95% confidence level) for the ICC of different sources of variability - filtered and normalized data, Illumina platform	92
5.13	Point estimates and interval estimates (95% confidence level) for the ICC of different sources of variability - normalized data, Illumina platform	92
5.14	Point estimates and interval estimates (95% confidence level) for the ICC of different sources of variability - filtered data, Illumina platform	93
5.15	King's $\rho_{c,rm}$ and Overall CCC - Affymetrix	94
5.16	Lin's ρ_c and Overall CCC - Affymetrix	95
5.17	Lin's ρ_c and Overall CCC for normalized data - Affymetrix	96
5.18	King's $\rho_{c,rm}$ and Overall CCC - Agilent	97
5.19	Lin's ρ_c and Overall CCC - Agilent	98
5.20	Lin's ρ_c and Overall CCC for normalized data - Agilent	99
5.21	Lin's ρ_c and Overall CCC - Illumina	100
5.22	Lin's ρ_c and Overall CCC for normalized data - Illumina	100
5.23	ME model estimates, $\lambda = 1$	103
5.24	Bland-Altman method estimates	105
5.25	Point and interval estimates of λ	107
5.26	ME model parameter, λ estimated	107
5.27	Number of miRNA in agreement between platforms.	109
5.28	Number of miRNA in agreement between platforms - quantile normalization.	112
5.29	Number of miRNA in agreement between platforms - loess normalization. .	112

List of Figures

1.1	Number of DNA microarray publications per year from 1995 to 2006	2
2.1	MicroRNA mechanism of action	8
2.2	Number of miRNA microarray related publications per year	9
2.3	In-situ oligonucleotide arrays	11
2.4	Gene length in the Human Genome	13
2.5	Decision tree for the choice of ICC estimate	17
2.6	Bland-Altman plots for simulated data according to different patters of poor agreement	19
3.1	Kernel densities of n° of Bead on Illumina platform	37
3.2	Agreement and Filtering: Cross-platform Pearson's ρ	41
3.3	Common human miRNAs	41
3.4	Scatterplots of differently normalized data	42
3.5	Quantile normalization	44
4.1	Experimental design	49
4.2	Differences between the CCC ρ_c and Pearson's ρ	55
4.3	Violation of the Bland-Altman method assumptions	62
4.4	Attenuation in linear regression	64
5.1	Probe intensities, Affymetrix	71
5.2	Scatterplots between replicates, Affymetrix	72
5.3	Batch effect for line A498, Affymetrix	73
5.4	Probe intensities, Agilent	74
5.5	Scatterplots between replicates, Agilent	75
5.6	Probeset intensities, Illumina	76
5.7	Scatterplots between replicates, Illumina	76
5.8	Diagnostic plots for A498 line (Affymetrix)	77

5.9	Square root of absolute residuals vs fitted (A498), filtered data - Affymetrix	78
5.10	Diagnostic plots for intensity data transformations	79
5.11	Diagnostic plots for hREF line (Affymetrix)	81
5.12	ICC comparison - Affymetrix	83
5.13	Diagnostic plots for line A498 (Agilent)	86
5.14	Diagnostic plots for line A498 (Illumina)	90
5.15	ICC and CI95% for different pre-processing strategies	93
5.16	OCCC for differently pre-processed data - Affymetrix.	96
5.17	OCCC for differently pre-processed data - Agilent.	99
5.18	OCCC for differently pre-processed data - Illumina.	101
5.19	OCCC for all platforms	102
5.20	Agreement intervals, $\lambda = 1$	104
5.21	Bland-Altman plots	106
5.22	Agreement intervals, λ estimated.	108
5.23	CI95% for proportion of concordant miRNAs.	110
5.24	Pair-plots of intensity values for all platform comparisons	111
5.25	Proportion of miRNA within agreement intervals	114

Introduction

Contents

1.1	New frontiers in biological research...	1
1.2	...and related statistical issues	3
1.3	Thesis organization	5

1.1 New frontiers in biological research...

In the past thirty years there has been a considerable development in the technology available for biological research. From the first appearance of qPCR in 1983 [1] to the outbreak of microarrays [2] and of the more recent RNAseq [3] the amount of data generated by biomolecular experiments has rapidly grown, bringing the promise of a deeper knowledge of biological mechanisms underlying onset and development of diseases as well as their possible therapies. This technological advancements have made it possible to uncover many "secrets" hidden in the human genome, thus opening the doors to the era of personalized medicine, and to the possibility of managing healthcare in a different way by exploiting the genetic/genomic information produced by these technologies.

With respect to microarrays, many things have changed from 1995 (first microarray study): from the 45 genes profiled by Schena *et al.* [2] we have now arrived to the thousand of genes which are commonly measured on the most modern arrays. Additionally, not only gene expression is now measured using these technologies, but also microRNA [4], Single Nucleotide Polymorphisms or SNPs [5] and methylation [6] are commonly evaluated using microarrays. The bar chart of figure 1.1 reproduces the number of research papers devoted to DNA microarray from 1995 to 2006, from where it can be clearly seen the steep increase, starting around year 2000, in the amount of research involving microarrays.

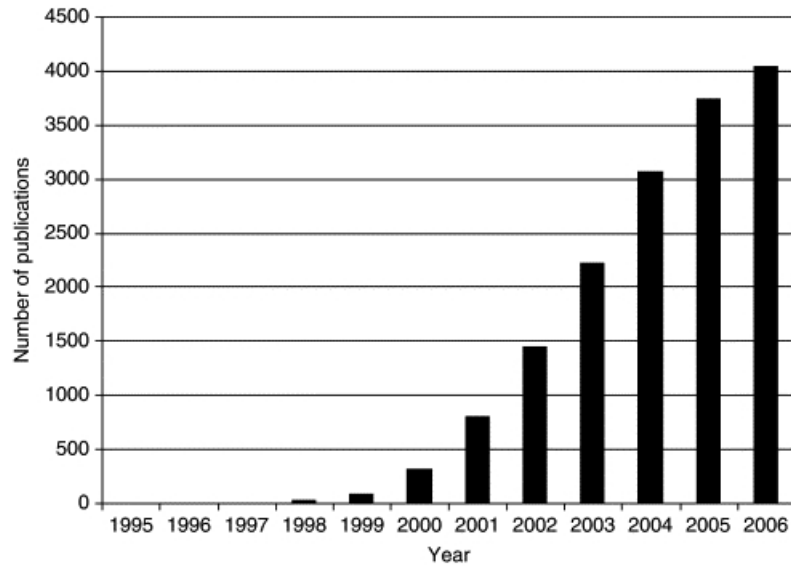


Figure 1.1: Number of DNA microarray publications per year from 1995 to 2006. Articles were identified using the search string "microarray AND gene AND expression", extracted from [7]

In the seminal paper from Golub *et al.* [8], for instance, a class predictor able to determine the class of new leukemia cases by considering the expression profiles of those genes which were most correlated with the outcome of interest, i.e. the leukemia classification, was developed, gaining an extremely satisfactory performance. The authors then suggest that *"the results demonstrate the feasibility of cancer classification based solely on gene expression monitoring and suggest a general strategy for discovering and predicting cancer classes for other types of cancer, independent of previous biological knowledge"*.

These potentialities have lead to a proliferation of studies attempting to discover the so-called *gene signatures*, i.e. panel of genes that could be able to discriminate between different phenotypes and, eventually, predict classification of a new sample. Despite all these efforts, results have been far below the original expectations, and to-date only Oncotype DX [9], a signature based on 21 genes derived from a qPCR experiment, has reached a level II of evidence to support tumor prognosis.

Contrasting results, however, have not generated a loss of interest in microarrays but, on the other hand, motivated an even greater effort in order to make the best out of them, both in terms of analytical properties and biological information. The Microarray Quality Control International Consortium [10], for instance, was designed to evaluate reliability of

expression measurements both within and between platforms, in order *to avoid procedural failures and to develop guidelines for microarray data analysis by providing the public with large reference datasets along with readily accessible reference RNA samples.*

With the recent advent of Next Generation Sequencing technologies, microarrays haven't lost their key role in biomolecular research but, conversely, have become the reference method with which these new technologies have to be compared. In this perspective, assessment of their reliability is a substantial feature, that needs to be evaluated by making use of proper statistical methods.

1.2 ...and related statistical issues

The main problem posed by recent biomolecular data lies in their dimensionality, since typically sample sizes are quite low (a few tens, often). But, whereas qPCR experiments generally are performed on a pre-selected number of transcripts which can range from some tens to at most some hundreds, a standard gene expression microarray experiment produces more than 40000 expression measures for each sample, and the numbers are much similar when it comes to RNASeq (see table 1.1).

Technology	What measures	Dimensionality
RT-qPCR [1]	number of cycles at which the transcript exceeds a threshold (above background)	$p > n$
Microarrays [2]	relative intensity of fluorescence of hybridized transcripts	$p \gg n$
NGS [3]	number of copies of a sequence of nucleotides (e.g. ATTCTTC) in a region of the genome	$p \gg n$

Table 1.1: *Type of measure and data dimensionality in different biomolecular platforms*

This situation, which is referred to as High Dimension Low Sample Size (HDLSS) data, makes it difficult to apply the standard statistical methods developed for the usual $n > p$ data framework, and requires the development of suitable methodologies capable of "recognizing" the relevant information and to extract it from such a huge amount of data while, at the same time, discarding the "noise", i.e. information attributable to technical features of the platform and/or of the experiment rather than to real biological characteristics.

The issues that arise in microarray research are many and various, and in many cases are strictly dependent on the specific platform used for the experiment. Between-array normalization [11, 12], batch-effect estimation and removal [13, 14], differential expression between clinical or treatment groups [15, 16] and multiple testing correction [17] are only some of the area of research that have received attention in the past years.

Another active field of research is the one related to platform reliability, meant as the combination of repeatability and reproducibility, as defined by the International Organization for Standardization (ISO) [18]. That is, just like any other laboratory technique, microarrays should satisfy some requirements: for instance, technical replicate arrays (i.e. arrays on which the same biological matter has been evaluated) within the same lab should produce results as much similar as possible. Parallel to these problems, researchers faced the issue of comparing results between platforms, which raised a fundamental question that has so far received a lot of attention: are different microarray platforms comparable in terms of expression measurements?

A lot of studies have focused on these questions (see [19] for a review), yet results have not only been contrasting, but often also based on incorrect statistical methods. For instance, many studies evaluated within-platform repeatability by computing correlation coefficients, claiming that high levels of correlation imply good platform reliability; additionally, sometimes also a statistical testing was performed, testing the null hypothesis that the correlation was different from 0, and interpreting a p-value less than a threshold α as an evidence for a good correlation between replicates. This approach, which is quite spread throughout literature, has been seriously criticized [20], yet is still used in not so few recent studies.

Because of this, and due to an initial lack of standardized protocols for microarray research, the first studies that attempted at performing some assessment on platform reliability and cross-platform "correlation" provided inconsistent results, somehow discouraging the spread of microarrays, and advising against the first optimistic claims about their great potentialities. The birth of great international projects (as the MAQC previously cited) were helpful in solving the latter issue, i.e. the deficiency of some consensus on how to deal with pre-processing of microarray data, but still some work remains to be done in terms of proper statistical methods for practical reliability assessment.

However, whereas great efforts have been devoted to understand patterns of performance of gene expression microarrays, the same can not be said for other array technologies, such as microRNA microarrays [4]. The particular relevance of microRNA microarrays lies in the biological importance of these small strands of RNA, that have been seen to play a key role in the regulation of genes, mostly by suppressing their signal. The possibility to profile a large number of microRNA simultaneously using a well established technology like that of hybridization microarrays have contributed to the initial success of this platform, which has provided many insights on the regulatory mechanisms underlying serious pathologies like cancer [21].

Nonetheless, to date, the issue of reliability and cross-platform comparison for microRNA arrays has been evaluated only in few studies, that suffered from the same limitations described above for gene expression arrays. This could possibly be due by the potentialities and advantages offered by Next Generation Sequencing platforms in terms, for instance, of new microRNAs discovery [22]. However, since reliability studies of these new technologies have provided contrasting results when compared to microarrays, miRNA arrays still remain an effective and useful technology, whose characteristics need to be properly assessed both in terms of within-platform reliability and of between-platforms comparison and concordance.

In this perspective, goal of this thesis is to highlight the pitfalls of current research and to illustrate the approaches developed for the assessment of within-platform reliability and between-platform comparison in the field of microRNA microarrays. In particular, the issue of comparing different platforms will be faced from an "agreement" perspective, by making use of the most recent methods for concordance assessment between different measurement methods.

1.3 Thesis organization

This dissertation is organized as follows. In Chapter 2 the main biological aspects related to microRNA and hybridization microarrays will be presented. Subsequently, the major statistical problem related to reliability and agreement will be described, together with a deep review of existing literature on this theme in both DNA and microRNA microarrays. Some limitations of current literature on the topic will be highlighted and solutions pro-

posed according to the methods developed in the present project. In Chapter 3 we will give an overview of the experimental platforms involved in the project and will describe the step performed in the data pre-processing phase. In Chapter 4 the statistical methods adopted for within-platform repeatability and between-platform agreement assessment will be described: the Concordance Correlation Coefficient (CCC) from Lin [23] and its generalization to the case with more than two measurements from Barnhard *et al.* [24] to evaluate reliability of a measurement method, and a modified version of Bland-Altman plot to assess concordance between different measurement methods [25]. In Chapter 5 results obtained from an application of the methods on 3 popular microRNA microarray platforms (Affymetrix, Agilent, Illumina) will be showed, with a comparison with results obtained from "standard" techniques currently used in literature. In Chapter 6 the results will be discussed alongside the possible limitation and drawbacks of the proposed approach. In Chapter 7 we present the conclusions and possibilities for future applications of the proposed method, and in the final Appendix we report commented R functions written for the computation of the measures of interest.

MicroRNA, Microarrays and Reliability

Contents

2.1	MicroRNAs: a short identikit	7
2.2	Microarray technology	9
2.3	Statistical issues related to reliability and agreement	14
2.3.1	Reliability	14
2.3.2	Agreement	17
2.4	Literature review	21
2.5	Limitations and proposed solutions	28

2.1 MicroRNAs: a short identikit

MicroRNAs (miRNAs) are small single-strand non-protein-coding RNA molecules, of 17-25 nucleotides in length. miRNAs function as negative post-transcriptional gene regulators, regulating mRNA translation and degradation in animals, plants and viruses [26, 27], and are believed to account for more than 3% of all human genes [28]. Recent evidences demonstrated that miRNAs play a critical role in many biological processes, also including haematopoietic cell differentiation, apoptosis, cell proliferation, organ development and stress response. They also have been implicated in tumorigenesis: in fact, recent evidences showed that the expression profiles of relatively few miRNAs (about 200 genes) were required to accurately classify human cancers, allowing their diagnosis and prognosis [29, 21]. Additionally, they have been proved to act in several other pathologies: Alzheimer's disease

[30], diabetes [31], multiple sclerosis [32] and ischemic diseases [33]. Their mechanism of action is described in figure 2.1.

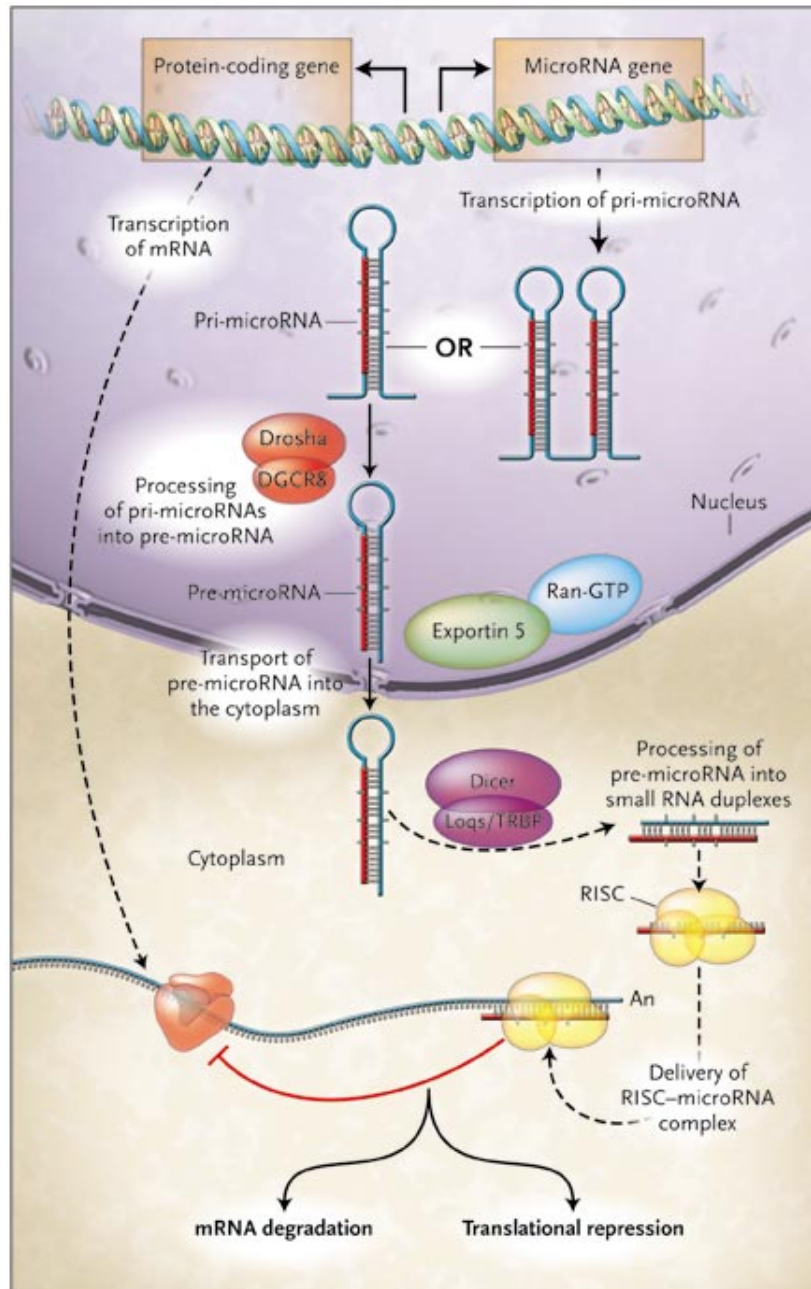


Figure 2.1: *MicroRNA mechanism of action, extracted from [34]*

Ever since their discovery back in 1993 [35] researchers have attempted at quantifying

level of expression of several microRNAs simultaneously. Methods based on RT-qPCR have been developed [36], yet hybridization-based techniques such as microarrays have rapidly become the assay of choice for this kind of experimentations. A search in PubMed using the keywords "miRNA" and "microarrays" in the title/abstract retrieves a growing number of contributions, suggesting the relevance of this kind of platforms for the analysis of microRNA expression levels (see figure 2.2). Adopting this type of assay however, though bringing alongside all the advantages derived from the deep knowledge of the hybridization techniques underlying microarrays, has posed to the researchers all the issues that had been already found out for gene expression microarrays, sometimes even exacerbated because of the reduced length of the transcripts that had to be analyzed. In section 2.2 the biology underlying microarray technology will be briefly described, and its application to the field of microRNA will be outlined together with the most critical issues related.

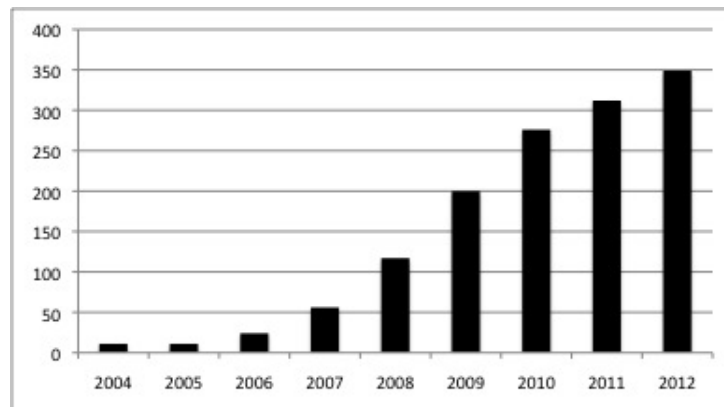


Figure 2.2: *Number of publications per year in which the keywords "miRNA" and "microarray" are cited in the abstract or in the title*

2.2 Microarray technology

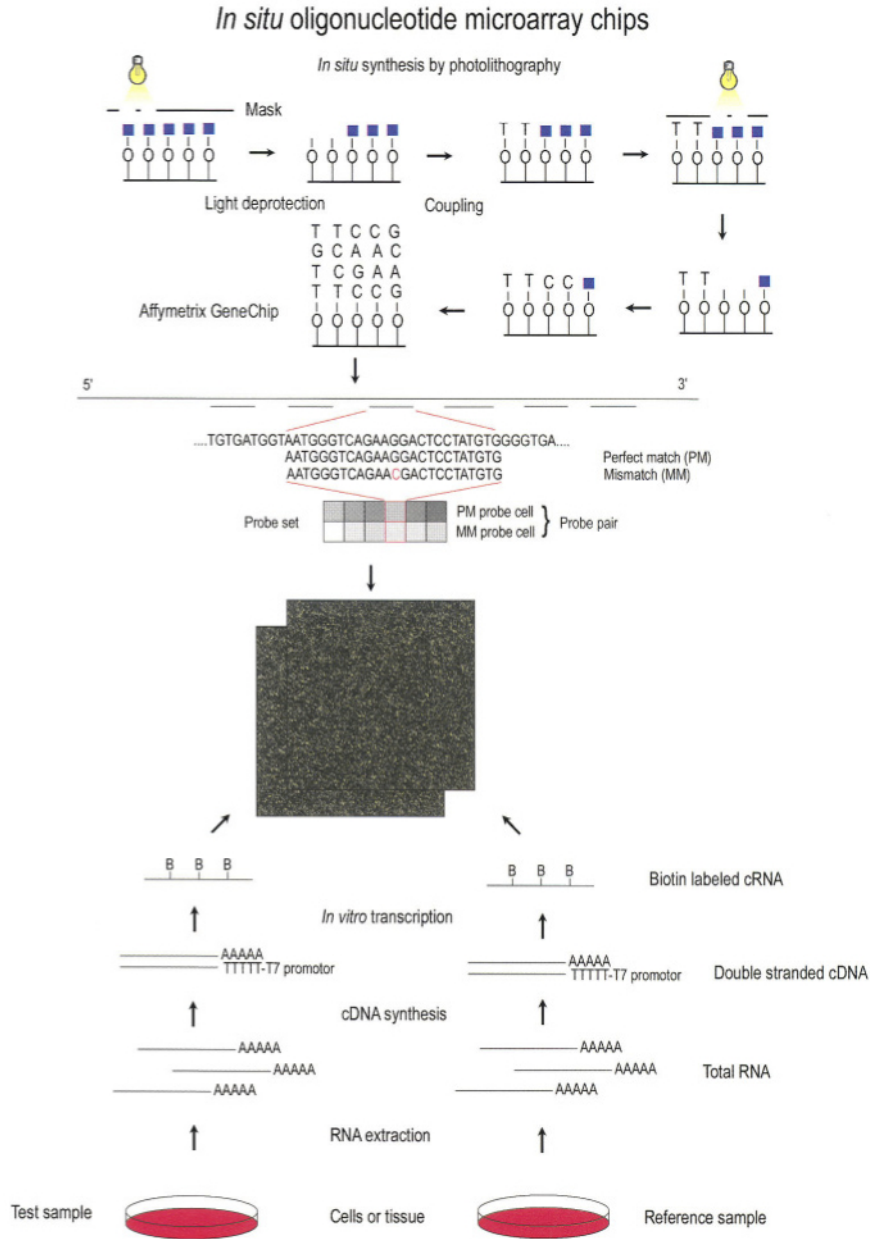
Microarray technology is fundamentally based on transcript hybridization, a phenomenon which occurs when two complementary nucleotides pair with each other by forming hydrogen bonds. Basically, a large number of probes (nowadays up to 50000 in gene expression microarrays) whose nucleotidic sequence is known and for whom we are able to determine which particular gene they refer to, is immobilized on a slide, and then the RNA target,

i.e. the biological matter that we want to know the content of, is fluorescently labeled and then gets hybridized with the probes on the slide. The more a specific sequence is present in the RNA target, the more the related fluorescent signal will be strong after the hybridization process, and the higher will be the quantitative measurement extracted by the image scanner.

One type of microarrays that was early developed were the so-called "two-color" microarrays, or "spotted" microarrays. These arrays were characterized by the fact that *test* and *control* sample had to be co-hybridized, i.e. hybridized on the same slide, in order to have a measure of how much a gene was expressed in a target sample with respect to a control one, where it was expected to be not expressed. To do this, different biological matters were labelled with different fluorochromes, typically red and green, and then the intensity measure from both *channels* had to be combined to have a relative measure of expression. The more recently developed oligonucleotide arrays (such as Affymetrix GeneChip, for instance), on the other hand, allow test and control sample to be processed separately, and then comparatively analyzed to understand whether gene A presents patterns of up-(down-) regulation in test vs control sample.

In figure 2.3 is described the pipeline for the *in-situ* preparation of the probes on the array slide. In general, given a mRNA sequence there are a number of probes 25-bp long which are designed to cover different regions of the target sequence. For instance, on Affymetrix Genechip for each mRNA sequence there are from 11 to 16 probes mapping to regions belonging to that sequence. This set of probes is called probeset. After hybridization has occurred, probes intensities within a probeset are summarized to obtain a single intensity measure for each target sequence whose mapping to a specific gene is known. For a short review of possible ways of summarization, see [37].

A point which deserves some attention is the issue of Cross-Hybridization (CH) or Multiple Targeting (MT), which may occur when a probeset specifically designed to match with a complementary mRNA sequence from gene A matches also with a complementary sequence from gene B, net of small differences between the complementary sequences from the two genes. For example, the probe TCGATCCTA, designed to match the transcript sequence AGCTAGGAT (according to the standard binding Adenine-Thymine and Guanine-Cytosine), could also hybridize with GCTAGG, a sequence which maps to another gene and

Figure 2.3: *In-situ oligonucleotide arrays*, extracted from [38]

whose only differences with the target transcript is the deletion of the first T and of the last two bases (see figure 1 in reference [39] for details).

Such a phenomenon has been object of debate in terms of how much it could influence statistical analysis and results, since it was argued that MT could concern up to half of the probe sets on the arrays [40]. In response to this work, Klebanov *et al.* [41] warned

that though the effect of cross-hybridization could lead to "strong positive correlations in families of those probesets that are known to have the potential for MT", yet this does not necessarily mean that such an effect would manifest itself in a tangible proportion of gene pairs.

This issue is likely to be much more evident in miRNA microarrays, mostly because of the different length of the transcripts involved. That is, genes have a large length variability, as shown in figure 2.4, which ranges from a few hundreds of base-pairs to some millions. This means that to capture a relevant amount of information on the expression levels of genes a single probeset will often be not sufficient, since a probe is generally 25-bp long. Thus, it is not uncommon to have more than one probeset mapping to the same gene, in order to cover a wider portion of the gene itself. For instance, on platform Affymetrix HG-U133Plus2 [42], out of the 54675 probesets available on the array only for 41290 of them there is an available Official Gene Symbol, i.e. an official gene identification code, and of these only 9350 map to a gene which is mapped only by that specific probe, whereas the remaining map to genes that are mapped also by a varying number of probesets, between 2 and 15. Notably, for each of these probesets it is possible to know the value of sensitivity and specificity, defined as follows for the gene they map to (see [43] for an example of calculation):

- **sensitivity:** fraction of probes in a probe set that match a respective gene. Namely, it is the number of probes in the given probe-set matching to a certain gene, divided by the total number of probes in the probe-set
- **specificity:** indicates to what extent probes of a probe-set bind to genes. It sums up the number of matching probes while giving lower weight to probes that match additional genes, and eventually divided by the total number of probes that matched any gene

These values are useful in that they allow to evaluate how much one should expect a problem of cross-hybridization for a given probeset. Additionally, this also means that for most genes there is a possibility, at least, to choose to consider only the probesets for which we can reasonably exclude that a wrong hybridization has taken place. For instance, one possible choice could be to consider only probesets with a sensitivity and a specificity score of 1 [44].

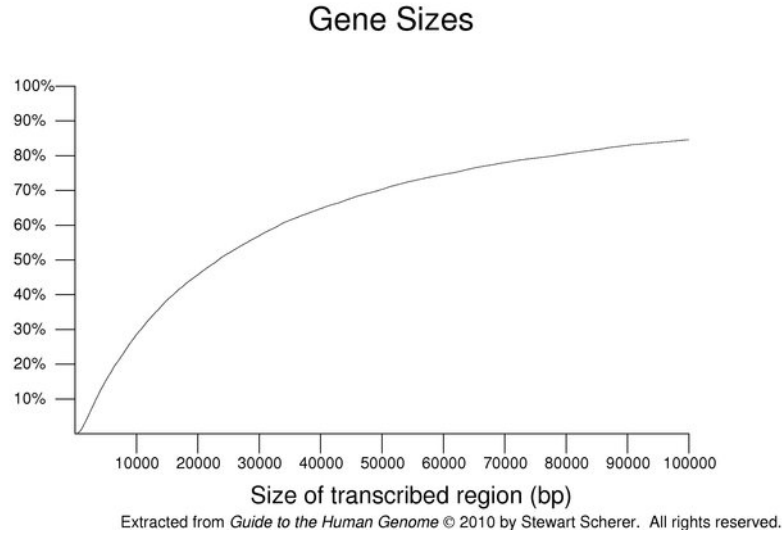


Figure 2.4: *Gene length in the Human Genome, extracted from [45]*

On the other hand, microRNAs are very short molecules, so that probes within a probe-set can cover the whole sequence of the transcript and be substantially identical (Affymetrix GeneChip, for instance). Additionally, it has been reported that microRNAs belonging to the same family show a high degree of similarity, i.e. sometimes only 1 base of difference, an issue which poses relevant problems when it comes to hybridization [46, 47]. As an example, in table 2.1 we report the sequences of four microRNAs all belonging to the let-7 family, which show only minor differences, and which are thus likely to show patterns of cross-hybridization (example taken from [48]).

let7-a	U	G	A	G	G	U	A	G	U	A	G	G	U	U	G	U	A	U	A	G	U	U
let7-b	U	G	A	G	G	U	A	G	U	A	G	G	U	U	G	U	G	U	G	G	U	U
let7-d	A	G	A	G	G	U	A	G	U	A	G	G	U	U	G	C	A	U	A	G	U	U
let7-f	U	G	A	G	G	U	A	G	U	A	G	A	U	U	G	U	A	U	A	G	U	U

Table 2.1: *Sequences for miRNAs belonging to let-7 family. Red highlighted bases in miRNAs let-7b, let-7d and let-7f represent differences from sequence of miRNA let-7a*

Because of these problems, several attempts have been made to increase platforms specificity, i.e. to reduce the number of MT probesets, mostly by evaluating different experimental conditions and their impact on CH (melting temperature of transcripts, etc) [47]. So far, though cross-hybridization is still an issue involving probesets whose differing bases are

located in the ends of the sequence, miRNA microarrays have reached a good performance level, in terms of what is actually measured and have become a *de facto* standard in many bio-molecular laboratories.

2.3 Statistical issues related to reliability and agreement

2.3.1 Reliability

Measurements reliability of miRNA microarrays is a relevant issue which has only poorly been addressed so far. Notably, this is contrasting with what has happened for gene expression microarrays, where a large number of studies have focused on the topic, though obtaining somehow inconsistent results. As we shall see in the next sections, the main problem related to this issue lies in the often incorrect statistical methods that have been used to address it, though analysis of reliability has a long history that dates back to 1962, when the method GAGE R&R was first introduced as a way to evaluate different components of variability in a measurement system at General Motors. Hence, several improvements have been suggested to the original framework, and now there is a wide literature focusing on the assessment of repeatability and reproducibility also in the context of clinical and biological measurement systems, and several methods have been proposed to address this issue [49, 50].

When we refer to reliability in the context of clinical/biological measurement, we refer to a combination of repeatability and reproducibility, where these two components are defined as follows [18, 51]:

- **Repeatability:** closeness of the agreement between the results of successive measurements of the same measurand carried out using the same method on identical test items in the same laboratory by the same operator using the same equipment within short intervals of time
- **Reproducibility:** closeness of the agreement between the results of successive measurements of the same measurand carried out using the same method on identical test items in different laboratories with different operators using different equipments

In the context of microarrays, these two components have been often used interchangeably, though they represent well distinct characteristics. In general we could say that in order to assess repeatability one needs to have a series of *technical* replicates, i.e. expression profiles of the same sample processed on different arrays within the same lab [52], whereas for assessing reproducibility a multi-center study involving the same samples being processed in various sites is required, such as the MicroArrayQuality Control study (MAQC, see [10]). In general, reliability studies aim at assessing different sources of variability which could affect the final measurement of interest, in order to provide useful information to prevent or minimize the related error. For instance, in a microarray study the expression profile of a sample could be influenced by the total amount of RNA extracted to perform the experiment on the array itself, or by the temperature of the laboratory where the study has been performed.

One common way to perform this type of evaluation is by means of the so-called "generalizability studies" [53, 54], where the ANalysis Of VAriance (ANOVA) framework is used to estimate the different variance components related to all the error sources involved in the experiment, e.g. day of the experiment, operator, etc. This means that the different sources of error are considered as random effects in the ANOVA model, rather than fixed. Once the variance components have been estimated, it is then possible to assess how much does each component contribute to the total experimental variability, by simply computing the ratio of each variance component to the total variance of the experiment. Practically, for a microarray study aiming at quantifying platform reliability, one could have just three main components to estimate (assuming no interaction exists): one which refers to technical replicates of the sample on the same array, and one which refers to the different labs where the study has been performed. The model would then be of this type:

$$y_{ijks} = a_i + b_j + c_k + e_{ijks} \quad (2.1)$$

where y_{ijks} is the expression level of the s -th replicate measurement for gene/miRNA i on the j -th technical replicate performed in the k -th laboratory, a_i is the effect of gene/microRNA (whose variance component estimate the biological variability of the assay), b_j is the effect of the technical replicate (whose variance component estimates platform repeatability), c_k is

the effect of the laboratory (whose variance component estimates platform reproducibility) and e_{ijks} is the residual variance. This model can then be integrated with interaction terms as well as other main effects and nested/crossed effects, according to the specific experimental design.

The different quotas of variance explained by each variable can also be interpreted in terms of Intraclass Correlation Coefficient (ICC), a measure which is commonly used also in the context of methods agreement evaluation, and which should reflect how much experimental units belonging to the same group are similar to each other [55]. Notably, ICC is a very debated quantity in literature and some alternatives to its use in the context of reliability have been presented, which should circumvent the intrinsic limitations of the index, such as its dependence on the range of the measuring scale [56]. In this last reference, the authors present a sort of "decision tree" (see figure 2.5) that should guide the researchers and practitioners in choosing one of the different strategies for estimating the ICC according to the characteristics of the design and of the data, depending on the observers (or methods of measurements) being assumed as fixed or random and on the measurements from every subject being considered as interchangeable between observers. Let us consider, for instance, a single-lab experiment, where a sample is profiled on 10 replicate arrays (i.e. the observers), each with a large number of genes/miRNAs (i.e. the subjects) measured on each plate: if the arrays are randomly drawn from a "population" of arrays that could be used for the experiment (which is reasonable) and the measurements of each "transcript" are considered to be not interchangeable across replicate arrays (reasonable because of possible batch effects), than the tree of figure 2.5 would lead us to use, among the others, a two-way random effects ANOVA model (Model B in the figure notation), which is the one defined in equation (2.1), with the deletion of the term c_k referred to laboratory effect.

Such a general approach however has been hardly seen in bio-molecular research, where the use of simple correlation coefficients has been, and in many case still is, the method of election for assessing platform reliability. Notably, since reliability evaluation and agreement assessment are two very connected issues, these methodological errors have propagated also in the study of cross-platform analysis, leading to some relevant pitfalls that have negatively influenced quality of results obtained, as we shall see in section 2.4.

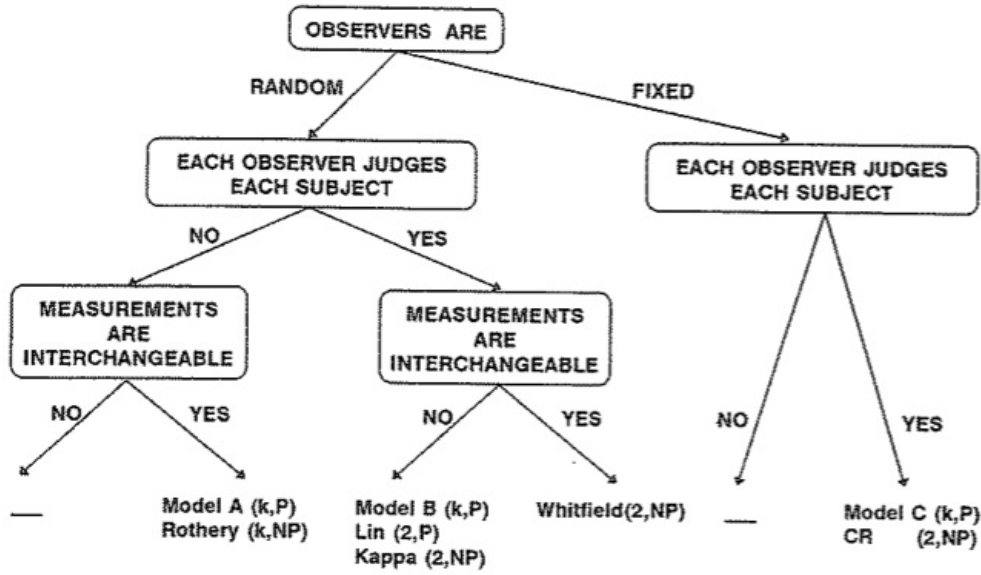


Figure 2.5: Decision tree for the choice of ICC estimate, extracted from [56]

2.3.2 Agreement

Another relevant topic which has become very urgent and actual, is the cross-platform comparison of results, which in the end turns into the comparison of measurements from different platforms: do different measurements actually convey the same information on any given microRNA? Literature on agreement of clinical measurement methods provides a great variety of approaches for dealing with this topic, yet in the field of bio-molecular research this has been often reduced to the comparison of list of genes differentially expressed between clinical/biological conditions or the computation of simple correlation coefficients [20] and, just as for the reliability issue, this topic is quite understudied for miRNA microarrays, in particular with respect to gene expression microarrays. Additionally, the few research studies which have been trying to address the problem of comparing different miRNA microarray platforms only tried to evaluate in which terms they could be said to be "comparable" or "concordant", and did not evaluate the reasons for the (possible) sources of disagreement. This is a very relevant lack, because, as stated in the Encyclopedia of Biostatistics, "*tracing the source and types (bias or random error) of the disagreements is the beginning of wisdom*" [57].

The agreement (or concordance) of measurement methods has its origins more than 100 years ago, when Karl Pearson introduced the correlation coefficient to assess existence of a linear relationship between two series of values. As already said, this index is now quite commonly used in many areas to assess agreement though it is well known that it may fail in detecting substantial disagreement [58] and many indexes have been suggested that have a close relationship with Pearson's ρ , such as the already discussed ICC and the Concordance Correlation Coefficient (CCC) proposed by Lin [23]. Despite its wide usage in many areas of biomedical science, the "index approach" has been seriously criticized because of its large sensitivity to sample heterogeneity [59]: heterogeneous samples, which are likely to occur in agreement studies because of the researcher's attempt to span across the whole range of possible measurements, are characterized by higher values of any correlation-related index (ρ , ICC, CCC, etc) than that obtained on more homogeneous samples with similar patterns of difference between methods but lower standard deviation. Moreover, the unsuitability of these kind of indices to provide information on sources of eventual disagreement could make them substantially useless in case a low value is found. That is, a $\text{CCC} < 0.5$, suggesting quite poor concordance between methods, could be due to, apart from sample homogeneity, to a substantial systematic bias between methods (i.e. the two methods actually give differing measurement over the same sample) or to a great random error. Since the two causes of disagreement are very different and can be handled with very different strategies (calibration of a method in case of systematic bias and replication in case of large random error), it is important that agreement analysis not only highlights patterns of concordance, but also sheds light over non-concordance sources. However, the CCC ρ_c can be obtained as the product of the standard correlation coefficient ρ (which defines the precision of the two methods) and of $c_b = \rho_c/\rho$ (which defines the accuracy of the two methods): depending on the sizes of these two quantities one is able to determine which is the main reason for eventual poor concordance [60], yet it is not so unusual to see only the CCC reported and discussed without further investigations.

Another approach which has been introduced to circumvent all the described issues is the Bland-Altman method, which relies on visualization of paired data average versus their differences [61, 62] to identify the so-called "Limits of Agreement" (LoA). This approach is very spread in biomedical research, mostly because of its simplicity of interpretation and

its direct connection with the problem at hand: in fact, it lets the clinician/biologist/etc. to be the ones who decide whether the LoA are satisfactorily narrow to conclude that the two competitive methods agree well. Moreover, and this is one the reasons for which it is often preferred to the computation of a simple concordance index, by looking at the Bland-Altman plot it is easy to ascertain for which reason the methods disagree: whether their differences are systematically different from 0 (systematic bias) or show a large degree of variability though randomly scattering around 0 (random error).

To compare the two approaches here discussed, we simulated two different situations where the two methods show poor patterns of concordance because of different reasons. In the first scenario (Bland-Altman plot shown in figure 2.6a) we generated a series of 100 normally distributed measurements ($\sim N(2, 1)$), whereas the second series was obtained by adding a random gaussian shift $\sim N(2, 0.5)$ to the first series to obtain a systematic bias. As for the second scenario (Bland-Altman plot shown in figure 2.6b), the first series was obtained by generating a series of 100 normally distributed values ($\sim N(1, 2)$) where the second series was obtained by adding a random gaussian shift $\sim N(0, 2)$ to obtain a mean difference close to 0 but a high standard deviation of the differences themselves.

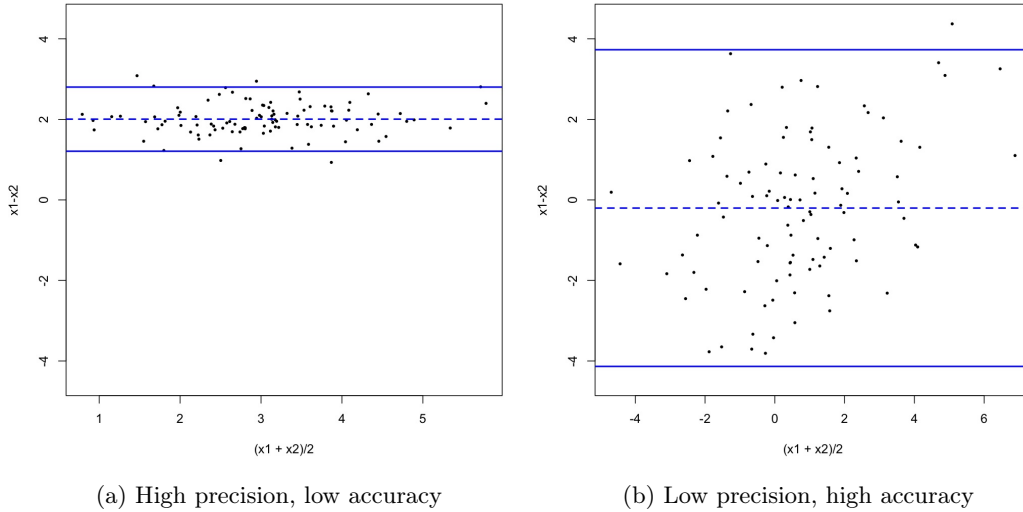


Figure 2.6: *Bland-Altman plots for data simulated according to the models described in the text. The LoA have been computed as $\bar{D} \pm 2\sigma_D$. Detailed results for both scenarios are described in the text and fully reported in table 2.2*

The full results of the simulation are shown in table 2.2. When the two methods actually differ mainly by a fixed shift the precision component of the CCC (ρ) is quite close to 1 (the maximum value) whereas the accuracy component (c_b) is substantially low; on the other hand, when the two method seem not to agree well because of a large random error (large LoA in figure 2.6b) the precision component is somehow lower than the accuracy component. In view of these results, the use of the CCC as a complementary tool to the Bland-Altman visualization can provide useful and practical information in that it can help in quantifying the degree to which the systematic bias or the random error affect the level of concordance of the two methods.

<i>Scenario</i>	$\bar{D}$	σ_D	<i>LoA</i>	ρ	c_b	ρ_c
Setting 1 (fig. 2.6a)	2.0056	0.3981	(1.2094, 2.8018)	0.9259	0.3441	0.3186
Setting 2 (fig. 2.6b)	-0.2024	1.9659	(-4.1341, 3.7293)	0.6989	0.9461	0.6612

Table 2.2: *Detailed results for the simulation scenarios presented in figure 2.6.*

The Bland-Altman method has however been criticized since it does not provide a direct evidence of agreement but leaves the final decision to the subjectivity of the researcher, as previously described. Other disadvantages lie in the underlying assumptions of constant variance and bias across the range of measurement values which are seldom met in common practice [25]. With respect to the constant bias issue, the standard approach of Altman and Bland has been recently generalized to more complex situations where the difference between measurements is not constant across the range of values, to derive *prediction intervals* for the measurement of one method given the measurement of the other method [63, 64].

As for the reliability evaluation, however, bio-molecular research is still a step behind these considerations. As a matter of fact, the few studies which have tried to evaluate cross-platform concordance so-far mainly focused more on the agreement of down-stream analysis results or, often, simply used correlation coefficients to assess whether two platforms were somehow "telling the same thing". Though it is reasonable to expect that a full "agreement" approach may not be the best strategy since the different array platforms are expected to provide very different values in terms of analytical range and properties, this approach is severely biased by the fact that the correlation only can provide some information on the presence of some intra-sample random error, and no information on the eventual existence of

a bias between arrays is provided. In this perspective, considering the described approaches, both the CCC and the Bland-Altman-like visualization method could certainly add a lot of information as well as provide the researcher with useful parameter to assess "where" do the differences between platforms lies.

2.4 Literature review

As said previously, evaluation of within-platform reliability and between-platform agreement has been much studied in the context of gene expression microarrays, whereas much less research focused on these topics for microRNA microarrays. Here we discuss some of the most relevant works which has focused on the issues deriving from the analysis of both reliability and agreement. However, these topics are substantially connected in a framework of global platform performance evaluation, and many studies concerning concordance between platforms also included some considerations on repeatability and reproducibility. One major stand-point for articles retrieval is the review from Yauk *et al.* [19], which only focused on cross-platform comparison but included many studies which devoted themselves also the evaluation of each platform behaviour in terms of internal consistency. This review somehow "splits" platform comparison literature in two distinct periods, 2002-2003 and 2004-present, claiming that advances in annotation, analytical technologies and methods of analysis have led to better results in latter works. In our review we tried not to focus on such a classification, which could be partially due to an over-interpretation of results, but evaluated critically all the papers discussed.

A study related to the MAQC project [65] aimed at evaluating within-platform reproducibility, using Agilent oligonucleotide platforms and two two-color arrays platform (CapitalBio and TeleChem), by computing Pearson correlation coefficients on $\log_2$ -scaled data both within different laboratories and between different laboratories for all possible replicate pairs, and then averaging over to have a single measure. All the coefficients are reported to be larger than 0.9, tending to decrease when computation are performed across laboratories, thus suggesting a lower degree of reproducibility and a better level of repeatability (according to the definition provided in section 2.3.1). Additionally, platform reproducibility was addressed in terms of how many genes were concordantly declared as significant

across labs, according to some criteria based on p-value (corrected for multiplicity with the False Discovery Rate method [66]) or on Fold-Change threshold. The authors report that these gene lists showed an overlap higher than 80% for Agilent platform, $> 70\%$ for CapitalBio and somewhat lower ($< 50\%$) for TeleChem platforms, thus suggesting that *"these results indicate that data generated from both one- and two-color assays are approximately equivalent and provide similar levels of biological insight"*. As for previous works, all platforms underwent some kind of normalization and of data-filtering, which were quite specific for each different type of array.

The analysis from Tan *et al.* [67] aimed at comparing three commercial platforms (Affymetrix, Agilent and Amersham) both in terms of within-platform reliability and between-platform comparison. To assess reliability, the authors computed Pearson's and Spearman's correlation coefficients between time points of measurements, where to obtain measures at each time point data from biological replicates were averaged. All the intra-platform correlation were reported to be higher than 0.9, suggesting a good level of repeatability for all the platforms considered. Also to assess cross-platform "concordance", the authors compute correlation coefficients, but additionally they evaluate similarity of results in terms of differential gene expression by fitting a mixed model to each platform, then for each pair of platform a contingency table for the event *The gene is differentially expressed* is created and a test (McNemar's or Fisher's exact) is performed to evaluate a sort of cross-platform agreement. The analysis was performed only on 2009 genes which were found to be in common across all the platforms according to the sequences GenBank ID, whereas considering genes matching by UniGene ID, though leading to a higher degree of overlapping (4012 genes) would have introduced additional confounding platforms that could prevent from comparing properly different technologies. Overall, correlation coefficient show poor values (around 0.5), and also concordance of differentially expressed genes was quite low, leading the authors to say that *"the results of this study suggest that cross-platform differences arise from the intrinsic properties of the microarrays themselves, and/or the processing and analytical steps of these microarrays"*.

Wang *et al.* [68] evaluated agreement between different platforms (2 Affymetrix platforms and cDNA) and different labs by means of a simple two tissue experiment involving three different platforms and three different laboratories. The measures used to assess

"agreement" were the common Pearson correlation coefficient, the intraclass correlation coefficient (both within sample replicates of the same platform and between different platforms and labs) and Cohen's kappa for a fold-change of gene expression greater than two. The authors report good results both in term of within platform consistency both in term of between platform agreement, saying that different platforms in the same lab were in larger agreement than the same platform in different labs. In terms of within-platform reliability the authors report Pearson's ρ larger than 0.8, ICC larger than 0.77 and Cohen's kappa larger than 0.43, suggesting that *"the replicates were quite consistent within each technology"*. Additionally, in the discussion section they state that *"results indicate that the agreement of different technologies is decent"* according to an ICC of 0.662 for overall agreement assessment. However, some pre-processing was performed, that is data were filtered according to different criteria related to each single platform and normalization was applied. This could be a source of major bias in their results because different methods were applied also to the same platform: for instance, three different normalization methods (Robust Multiarray Average, Average difference between PM and MM, Li and Wong model-based method [69]) were applied to different datasets coming from the same Affymetrix platform, so differences in expression level could have been induced by the choice of normalization methods adopted.

Reproducibility of three microarray platforms was addressed also by de Reynies and colleagues [70], by using correlation coefficients between replicates (all yielding values > 0.9) computed within quartiles of expression values, finding that also for poorly expressed genes the degree of within-platform consistency was moderately good. Additionally, they found that oligo-nucleotide arrays were substantially better than two-color in that the amount of genes showing a log2-ratio between replicates different from 0 appeared much higher in the Agilent two-color arrays. Between-platforms comparison was performed on 3400 targets *chosen for their unique correspondence to a single transcript* (thus known not to be involved in Multiple Targeting or Cross Hybridization, see section 2.2), and returned inter-platform correlation values for each of the three samples involved within the range 0.72-0.85, thus substantially lower than the intra-platform ones (all > 0.96). The authors also report that of all the genes analyzed, 965 showed patterns of discordance between platforms, and for each platform they were able to identify in which range of expression

values lied the non concordant genes. Differential expression analysis resulted in a good overlap of differentially expressed genes when an ANOVA model was performed using the cell line type as the comparison group. The authors thus conclude that all array types show good degree of both within-platform reliability and between-platform concordance, but also suggest an increased precision of detection for oligonucleotide arrays when compared to two-color arrays.

A comparison between cDNA microarrays and Affymetrix GeneChip was carried out in a paper by Kuo *et al.* [71]. The authors used 56 of the 60 tumor cell lines described in the NCI60 panel of lines, considering information on 2895 genes with known matching sequences between platforms. Regardless of the intensity measure considered for the cDNA probes (whereas only Average Difference between Partial Match (PM) and MisMatch (MM) was considered for Affymetrix), the authors found poor correlation (both Spearman and Pearson coefficients not larger than 0.5) between both genes and lines, and results were similar even after some data filtering which possibly resulted in removing noisy genes. Moreover, the authors stated that *"as many microarray studies draw their conclusions from clustering analyses, cross-technology preservation of clustering results may be more interesting than preservation of the raw measurements"*, and evaluated whether samples and/or genes grouped in the same cluster on both platforms by means of Cohen's Kappa and Jaccard index. Since also this analysis gave poor results, authors warn on considering cross-platform dataset comparable, suggesting that more research on the hybridization dynamics is needed.

A quite different strategy has been chosen by Yuen *et al.* [72], which performed a sort of "indirect" comparison between Affymetrix and custom cDNA microarrays, that is they compare both platform with a reference method (quantitative Real-Time PCR), estimating measurement bias for both platforms and empirically comparing them through graphical visualization. Moreover, they suggest a formula for calibrating microarray data in order to estimate unbiased measures of expression starting from the ratio of the fold-change of each platform to the fold-change of the PCR assay (details of the procedure can be found in the original paper). The authors conclude that microarray platforms underestimate fold-change of expression with respect to PCR, and show that after calibrating the assays using a general equation the bias is relevantly reduced, resulting in more reliable differential expression results. Though interesting, the paper suffers from some major limitations, which

are quite similar to those seen for previous reports, i.e. the analysis of comparison to PCR and the calibration process has been developed and tested only on 47 genes/transcripts present on both platforms, known to be regulated or unregulated, that were not chosen at random but specified on the basis of previous knowledge.

The point of view of Lee and colleagues [73] is different, in that they assess that *"our focus in this study was less on the comparison of technologies than on the generation of mutually validated, and therefore robust, datasets"*. Their approach is based on a filtering procedure which relies on estimating the correlation of expression patterns for genes whose sequences had been identified as belonging to the same UniGene cluster for both platforms compared and choosing a cut-off for gene filtering. After this step, they propose to compute the correlation of correlations, that is they compute pairwise correlation for all samples within each platform, and then compute correlation of these platform specific correlation coefficients. Higher values indicate similar pattern of regulation in the same samples, leading thus to conclude for good degree of concordance. The authors apply the proposed methods to the NCI60 dataset, profiled with Affymetrix and cDNA arrays, by considering whether different filtering thresholds result in better "presume" concordance. For a cut-off of 0 the correlation of correlations was found to be very poor, whereas increasing it leads to better results, as expected. The main point here is that good results are only seen when removing large part of the genes on the platform, and this is a common point between all studies previously described, and that no consideration on reason for which poor "concordance" is seen are discussed, with the exception of the well known annotation problems.

Comparing three different platforms, one oligonucleotide array (Affymetrix) and two cDNA arrays (Agilent and custom) was the focus of Järvinen *et al.* [74], whose conclusions were based on correlation analysis of 5 samples (4 breast cancer cells and 1 human reference) on which different normalizations were applied. To assess reliability of custom arrays the authors computed correlation coefficients and averaged them over all the possible pair-wise comparisons, obtaining a correlation of 0.9 (but using different numbers of probes, ranging from 7412 to 9982). All concordance results indicated poor correlation between platforms, particularly from commercial and custom arrays, for both filtered and unfiltered data. For instance, Pearson correlation coefficient between Agilent and Affymetrix ranged from 0.7 to 0.79 for the cell lines considered, whereas Spearman coefficient range from 0.7 to 0.76,

with slightly higher values for filtered data. The authors suggest that *"the most significant impact on gene expression levels resulted from placement of the wrong clones on the arrays"*, i.e. errors in sequences, and highlight the need for optimization of the data pre-processing workflow in order to have some standardized benchmark to use in microarray data analysis.

Mah *et al.* compared cDNA and oligonucleotide based systems [75], obtaining disheartening results: the Affymetrix platform and the cDNA based platform showed discordant results both in term of present/absent calls and of rank order correlations between the platforms. Actually, the analysis was performed only on subsets of the whole platforms, i.e. probes for which there was concordance in annotations, but the conclusions of the authors are very clear: *"in summary, the results of this study indicate that oligonucleotide-based arrays, such as those produced by Affymetrix, and full-length clone-based arrays may be too different in experimental design to be expected to give global expression results that can be directly correlated"*.

Park *et al.* also compared performance of several microarrays (both oligonucleotide and cDNA) with PCR assay [76], using classical methods such as those seen previously for other similar works. They explicitly considered the possible effect of expression levels on the cross-platform correlation of matching genes, by dividing genes on each platforms in quantile of expression depending on averaged intensity ranks across platforms, evaluating Pearson correlation within each of these artificially created bins. Pearson's ρ show a rather clear increasing pattern, since they tend to be higher for genes in the upper quantile of the distribution. Comparison with PCR confirms what has been described in the work by Yuen *et al.* [72], i.e. that microarrays tend to underestimate differential expression with respect to PCR, which is here considered as a gold-standard. Actually, validation analysis was performed only on 23 genes for which sequences had been verified, so the sample for which these conclusions can be drawn is very small. With respect to previous works, the authors conclude that *"the agreement between Affymetrix and cDNA arrays seems to be fairly good and reproducible"*, whereas *"the oligo arrays in many cases do not give a concordant result"*. These conclusions appear a little bit "overfitting", in the sense that they tend to over-interpret actual results, which talk about correlation coefficient hardly above 0.8.

Comparison of cDNA and oligo microarrays is also the focus of a work by Larkin *et al.*

[77], which however implements different methods to come to a conclusion on microarrays concordance. The experiment consists of mice treated with acute (24 hours) or chronic (14 days) angiotensin II or with acute/chronic saline control: after data processing has been performed according to strict protocols, log2 ratios between treatment and control are calculated for both the acute setting and the chronic setting, since *"knowing how a transcript level changes in response to a particular stimulus often is the more relevant parameter to use in comparing platforms"*. Concordance analysis is then performed only on genes with matching sequence and detected in at least half of the samples on both platforms, resulting in 5853 'good' genes (as the authors calls them), and consists of fitting a two-way ANOVA model (including treatment and platform) for each gene, evaluating for which genes there is a significant effect of platform, implying that genes for which no platform-effect is seen can be considered concordant between platforms. 90% of genes shows no effect of platform and a visualization via a PCA-based biplot showed that these genes were helpful in discriminating chronic treatment from acute treatment, whereas the remaining 10% of genes mainly discriminated different platforms. So, somehow in contrast with what has been previously described in other studies, *"our analysis of cardiac gene expression yielded consistent results for greater than 90% of genes in common between the Affymetrix GeneChip and TIGR cDNA arrays"*.

These works represent only a subset, though quite relevant and representative, of the complete amount of literature available on this topic. Since most work are substantially very similar both conceptually and analytically, not all of them are reported here. A review on the topic from Draghici *et al.* [78] illustrate common drawbacks in platform comparison studies, including annotation issues and noise filtering procedures, highlighting that *"accurate measurements of absolute expression levels and the reliable detection of low abundance genes are currently beyond the reach of microarray technology"*. Another critical view of commonly used methods for assessing concordance has been presented by Chen *et al.* [79], whose paper underlines that *"a correlation coefficient of 0.90 or higher does not necessarily imply reproducibility between two measurements"*, whereas they suggest to use ANOVA models to determine whether there's a relevant/significant effect of platform in determining differences between different arrays, as suggested in other works (for instance [77]).

Notably, all the discussed work refer to gene expression microarrays, but similar meth-

ods have been used also for comparing different microRNA platforms, coming to results which show higher consistency seen in the previous ones, also because of less problems in terms of annotation and sequence matching. Sato *et al.* [80] found a good correlation between log-ratios of expression both when using expression values and when using their ranks, and a good overlap between differentially expressed miRNAs, at least for three of the platforms considered. To assess intra-platform repeatability the authors also evaluated distribution of Coefficient of Variation computed for each gene across replicates available for each sample and graphically compared them between platforms and by using t-test and Mann-Whitney test. Additionally, they computed log-ratios of sample-to-sample expression (only two sample types involved) for all pair-wise comparisons of three technical replicates, resulting in a total of 9 log-ratio intensity profiles for each of the six platform, and visualized all these data with a heatmap, where data from the same platform was seen to correlate very well and seemed to be poorly correlated with the remaining platforms. Cross-platform comparison was performed by means of Spearman rank correlation on 309 miRNAs with concordant detection calls between replicates on different platforms, and pointed out poor level of concordance, with a maximum correlation of 0.8717. Also, overlap of differentially expressed genes was found to be substantially low.

Yauk *et al.* [81] found moderate correlation between the six arrays considered for analysis (Spearman's index between 0.663 and 0.949), and when concordance with a PCR assay was considered, the Concordance Correlation Coefficient was never found to be above 0.532. Curiously, the authors conclude that *the platforms examined in the present study show high levels of both within- and between-platform reproducibility. The platforms also was found to correlate well with TaqMan arrays.*

2.5 Limitations and proposed solutions

Two main limitations can be seen in the works here described: the first is about statistical methods and the second strictly relates to data analysis and workflows implemented in practice. With respect to statistical methods, correlation or overlap of list of differently expressed genes are not the proper measures to use when the goal is describing patterns of reliability or agreement and this has already been outlined also in microarray literature [20],

yet still they are commonly used. The more general concern with the statistical methods summed up in this review refers to what we have described in section 2.3, that is no clue, other than hybridization problems, is generally given on which are the causes for poor levels of repeatability/reproducibility and concordance. Additionally, also the interpretation of many metrics considered is quite difficult, since no reference value exists to say that two platforms actually agree. The sometimes reported p-values for the hypothesis of null correlation are meaningless, since it is well expected that some correlation between array must exist. One possible approach for within-platform reproducibility would be to work in a non-inferiority setting, defining a sufficiently stringent threshold and then properly computing confidence intervals for the CCC, evaluating whether the interval lies above or includes the threshold [82].

The other major concerns with this type of study relates to the type of data processing which is performed. We could restrict to two problems that often occur:

- **data filtering:** in this step genes with low signal compared to variance are often filtered out, and thus are not considered in the analysis. As in Kuo *et al.* [71], for instance, this is done in order to evaluate whether filtering is able to "improve the correlation" between platforms. Of course, to perform platform comparison one has to filter out the same genes on all platforms to be compared, so one may exclude a gene in one platform with low signal and high noise whereas on another platform this gene could be at a top-level intensity, thus leading to a severely biased estimate of correlation between platforms. This issue poses several questions on whether it is correct to do this type of action in this context.
- **normalization:** such a step is probably the most critical of all, because of what it implies. Suppose we normalize data from two platform using the same method (e.g. quantile normalization): then, when we compare normalized data actually we are not simply comparing platforms, but a sort of platform/normalization interaction to evaluate whether that kind of normalization leads to concordant data. Things may become even more complicated if we use different normalization methods for different platforms. If normalization is performed, we may lose some patterns of discordance which are relevant to describe the relationship between platforms, but if it is not, we

may highlight some discordance which can be possibly due to experimental factors different from platform itself (i.e. the batch effect). A possible strategy could be to compare platforms before and after normalization, to understand whether a specific procedure could determine a general increase in concordance.

The main point is that though these steps are possibly relevant when one performs downstream analysis such as clustering or differential expression, when the focus is on analytical properties of the assay itself, all the data should be considered as relevant and should not be excluded or manipulated, unless one wishes to study how do these methods affect expression measurements. However, since this is not the case in most of the studies presented above, these two points appear as major flaws which need to be properly addressed.

In this project we propose to assess within-platform reliability of three major microRNA microarray platforms (Affymetrix, Agilent and Illumina) by means of a random-effects model to estimate variability contributed by different experimental factors and, additionally, to evaluate platform repeatability using the Overall CCC, a weighted mean of pair-wise CCC, in a non-inferiority setting, as proposed in [82]. In this way we will be able to quantify the technical and the biological variability connected to the experimental setting as well as providing a proper measure for evaluating agreement rather than a simple mean of correlation coefficients. To address the issue of between-platform comparison we suggest to use a modified version of the classical Bland-Altman visualization method described in section 2.3.2: this modified version substantially allows to incorporate a linear measurement error model in the definition of the Limits of Agreement, thus enabling us to evaluate possible source of disagreement between different platforms and to provide a direct information on the concordance of the measurement methods. Moreover, to cope with the problems deriving from data filtering and normalization, all the analysis have been performed on unfiltered and un-normalized data, and results were compared with those obtained with data normalized using standard procedures such as quantile normalization and loess normalization.

Platforms features and data processing

Contents

3.1	A terminology caveat	31
3.2	Samples and experimental design	32
3.3	The platforms: features and data processing	33
3.3.1	Affymetrix	33
3.3.2	Agilent	34
3.3.3	Illumina	36
3.4	miRNA selection and Normalization	38
3.4.1	Automatic filtering or manual selection?	38
3.4.2	Normalization: to do or not to do?	40
3.4.2.1	Quantile Normalization	43
3.4.2.2	Loess Normalization	44

3.1 A terminology caveat

Since the terms reliability, its components reproducibility and repeatability, and agreement are used in microarray literature quite interchangeably, it is important to clearly define what we are referring to when we use one term or the other, provided that in a certain sense reproducibility and agreement refer to a similar phenomenon, that is closeness of measures obtained under different conditions (reproducibility) or with different instruments (agreement). In the laboratoristic context however, reproducibility and agreement are two

very distinct issues, since the first necessarily requires that the performance of a single instrument methods is evaluated across multiple sites whereas the second can be addressed within a single lab using different measuring instruments.

In the present study all of the platforms analyzed were studied within a single lab, so that no reproducibility but only repeatability of measurements can be assessed. For simplicity, however, we will use the general term reliability to refer to within-platform repeatability.

3.2 Samples and experimental design

This study involved two samples: a renal tumor cell line named A498 (ATCC, Manassas, VA, USA) [83] and a pool of twenty different human normal tissues (namely hREF), obtained from the First Choice[©] Human Total RNA Survey Panel, (Ambion Inc, Austin, TX, USA). RNA material was analyzed in different labs, as follows: Affymetrix processing took place at the Biomedical Technologies Institute of the University of Milan (Segrate, Italy), Illumina and Agilent processing was performed at the Department of Experimental Oncology of the National Cancer Institute (Milan, Italy). For both samples three technical replicates for each platform were performed, leading to a total of 18 arrays, 6 for each different microarray platform. Details on each platform features are described in section 3.3.

The cell lines were cultured according to the corresponding ATCC datasheets, at cell confluency, total RNA samples were extracted using the miRNeasy kit (Qiagen Spa) and quantified by ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). Using Agilent 2100 BioAnalyzer (Agilent Technologies, Santa Clara, CA, USA), RNA integrity was assessed on the basis of the RIN (RNA Integrity Number) factor and presence of low molecular weight RNA (5S) was verified. Labeling of total RNA samples was performed using FlashTagTM Biotin RNA labeling Kit (Genisphere Inc., Hatfield, PA) starting from 1 μ g of total RNA. Briefly, the tailing reaction is followed by ligation of the biotinylated signal molecule to the target RNA sample. The labelling reaction is based on Genisphere's proprietary 3DNA dendrimer signal amplification technology. Prior to array hybridization, total RNA labeling was verified using Enzyme Linked Oligosorbent Assay (ELOSA). After that, biotin-labeled samples were hybridized onto the arrays at 48°C, then washed and stained using the Affymetrix Fluidics Station 450. Arrays were scanned with the GeneChip[©] Scan-

ner 3000 7G to acquire fluorescent images of each array and analyzed by use of GeneChip Operating Software (GCOS).

3.3 The platforms: features and data processing

3.3.1 Affymetrix

The Affymetrix GeneChip[®] miRNA Array (Affymetrix Santa Clara, CA, USA) includes 46227 perfect match (PM) probes, representing 7815 probesets, of which only 847 (10.83%) represents human microRNA; the remaining act as control probe sets which are expected not to be expressed when human samples are hybridized. This annotation refers however to an old version of the array, since, at present time, more than 1000 microRNAs have been discovered and characterized [84]. The array contains 4 identical probes for each miRNA, located in specific spots on the array, with a length between 16 and 25-mer.

Intensity .CEL files were obtained from the scan images and imported to Affymetrix[®] miRNA QC Tool software, Version 1.0.33.0 [85] to quantify the signal value. We assessed QC by plotting the average intensity of the oligo spike-in and background probe sets (included in Control target content) across all the arrays. The miRNA arrays were detected using Affymetrix detection algorithm, based on non parametric Wilcoxon Rank-Sum test, applied independently on each array and probe/probe set; a p-value greater than 0.06 stands for "not detected above background". For data normalization, we chose "default" method obtaining log₂ expression values (expression values data matrix) from the raw data (intensity values data matrix). Briefly, this method involves the following three steps: grouping the background probes intensities, i.e. probes that are supposed to reflect noise of the array, based on Guanine-Cytosine (GC) content, where the median intensity of each bin was the correction value for each probe with the same GC content; a quantile normalization and, finally, a median polish summarization. Quantile normalization will be further described in section 3.4.2.1 with reference to probeset level intensity data; in this phase, normalization on probe-level data is necessary to obtain a single intensity measure for each probeset. Median polish summarization [12, 86] is an algorithm that for each probeset i fits an additive model defined as

$$y_{ijkl} = \mu_i + A_j + P_{ik} + \varepsilon_{ijkl} \quad (3.1)$$

where y_{ijkl} is the intensity value of probe i referred to probe set k on array j , and l is the number of replicate spots for each probe, μ_i is the general mean of probeset i , A_j is the effect of the j -th array, P_{ik} is the effect of probe k referred to probeset i and ε_{ijkl} reflects residual variability. Then the fitted expression levels for probeset i on the j -th array is obtained by summing the general μ_i to a specific normalized array effect NA_{ik} which is obtained with a non-parametric smoothing that should remove effect of extreme intensities (see [86] for details).

3.3.2 Agilent

The Agilent Human miRNA Microarray (V1) platform contains information on 961 microRNAs, of which 851 (88.55%) are human microRNAs, for each of which there's a variable number of replicated probes mapping to it. Notably, these probes are not all different, and there are sub-levels of replication within each set of replicated probes: that is, out of 16 replicated probes (constituting a probeset) for miRNA A there could be four groups of 4 probes with the same nucleotide sequence, for instance. In table 3.1 the various levels of probeset replication are reported. Of all the microRNAs, both human and non-human, 97.7% have 16 replicates, and this includes all the human microRNAs. The miRNA with 182 replicates is labelled as **NegativeControl** on the array, and represents the level of background noise. In terms of probe replication within each probeset, if we consider only human microRNA we find that 308 miRNAs (36.2%) are "interrogated" by 4 different sequences, 56 (6.6%) by 3 sequences, 483 (56.8%) by 2 sequences and 4 (0.5%) by a single sequence.

N. of miRNAs	1	1	939	12	1	6	1
N. of Replicates	5	10	16	20	26	42	182

Table 3.1: *Levels of probeset replication for Agilent Human miRNA Microarray (V1) platform*

Images were scanned using the Agilent Feature Extraction (AFE) software (Version 9.5.3.1) [87], obtaining the background-subtracted **ProcessedSignal** (PS) for all the probes, the **TotalProbeSignal** (TPS) for all replicated probes within a probeset and the **TotalGeneSignal** (TGS) for all probesets on the array. The TPS is obtained as the robust average of all the

PS for each replicated probe (i.e. probes mapping to same miRNA which share the same sequence) multiplied by the total number of probe replicates. The TGS is then obtained as the sum of the unique values of the TPS. As a practical example, in table 3.2 we report PS data for miRNA hsa-let-7a* over our experimental samples, where the rows have been highlighted according to different probe sequence.

miRNA	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
hsa-let-7a*	-3.3299	0.4680	-2.0335	-1.3901	-3.7643	-2.5698
hsa-let-7a*	1.5148	-3.1829	-2.2008	-0.3657	-1.4792	-5.8714
hsa-let-7a*	2.9001	5.0021	7.7388	3.8242	3.9510	2.1170
hsa-let-7a*	0.7713	-0.2400	1.8596	-4.9641	-1.7849	0.4060
hsa-let-7a*	1.9520	-1.2598	0.9588	-6.2737	-3.2143	-2.8047
hsa-let-7a*	0.3236	-2.1647	2.1497	-0.2515	0.7653	-3.9289
hsa-let-7a*	-0.8569	-4.6409	-0.1646	-1.0033	-0.6864	0.6811
hsa-let-7a*	1.5214	1.4006	2.2075	-3.3012	-1.8923	-7.9353
hsa-let-7a*	0.9325	4.1192	2.6478	-0.3589	4.3844	2.1793
hsa-let-7a*	-2.3443	-1.2741	-0.1310	0.4935	5.4329	0.3316
hsa-let-7a*	1.8203	1.6870	-0.2460	0.4119	1.1411	-2.0285
hsa-let-7a*	-1.9471	1.4837	-2.3905	1.0477	-0.5786	-5.9691
hsa-let-7a*	2.1316	1.4491	4.0628	-2.1970	2.8992	1.5832
hsa-let-7a*	1.2334	1.6761	1.9301	2.7012	8.5070	2.4994
hsa-let-7a*	-0.3531	-0.8127	-4.1391	1.8035	1.4809	-4.9203
hsa-let-7a*	13.2036	-2.0677	1.3615	2.2118	4.7325	-0.6009

Table 3.2: *ProcessedSignal* values for miRNA hsa-let-7a* on 6 experimental samples

By using a proprietary algorithm described in [87], these values are then summarized into 4 different TPS, that are reported in table 3.3. The rows have been highlighted accordingly to what can be seen in table 3.2.

miRNA	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
hsa-let-7a*	-0.0894	-0.0282	0.2178	-0.5038	-0.1810	-0.2531
hsa-let-7a*	0.2159	-0.0818	-0.1435	0.0987	-0.2760	-1.1449
hsa-let-7a*	0.4652	0.7878	1.0508	0.2686	1.3188	0.5753
hsa-let-7a*	-0.2088	-0.5721	0.1586	0.0705	0.1389	-0.5381

Table 3.3: *TotalProbeSignal* values for miRNA hsa-let-7a* on 6 experimental samples

By summing up the values reported in table 3.3 the TGS value for miRNA hsa-let-7a* is then obtained, as in table 3.4. Notably, one can see that for all quantities described some

negative values are present. Since log-transformation is not possible for negative values, to handle this issue with TGS it has been suggested to add, to all intensities, the absolute value of the minimum TGS intensity experiment-wise as extracted by the $AFE + 2$ before log2 transformation [88]. In our analysis we have performed the same transformation also for the PS values. Data extracted from AFE were imported in the R environment [89] and processed using the AgiMicroRna package, available in Bioconductor [90, 91]. Notably, the PS has been used as starting data for performing repeatability analysis (see Chapter 4 for details), whereas assessment agreement has been performed on the summarized TGS.

miRNA	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
hsa-let-7a*	0.3828	0.1057	1.2837	-0.0659	1.0008	-1.3607

Table 3.4: *TotalGeneSignal* values for miRNA hsa-let-7a* on 6 experimental samples

3.3.3 Illumina

Illumina humanMI_V2 platform contains 1145 miRNAs of which 858 (74.93%) are human miRNAs. Each miRNA is quantified by a large series of copies of the same oligonucleotides which are attached to a "bead". This bead-level information is then collapsed into a single intensity measure for each microRNA by means of a proprietary algorithm by Illumina, Inc. This means that there is no way to access bead-level information and that in this case we only have, differently from Affymetrix and Agilent, summarized information at the probeset level.

In figure 3.1 we report the estimated distribution of the number of beads for each of the microRNA measured by the platform. On the left panel (figure 3.1a) all 1145 microRNAs are considered, while in the right panel (figure 3.1b) only human miRNAs are considered. No relevant differences are actually seen. Of note, sample 1 appears to be characterized by a consistently lower number of beads contributing to miRNA intensities with respect to all the remaining samples, that on the other hand show very similar patterns. As a matter of fact, mean number of beads for each microRNA being profiled is ≈ 338 for Sample 1 and higher than 370 for all the remaining samples when we consider all miRNAs. Substantially identical results are seen if only human miRNAs are taken into account.

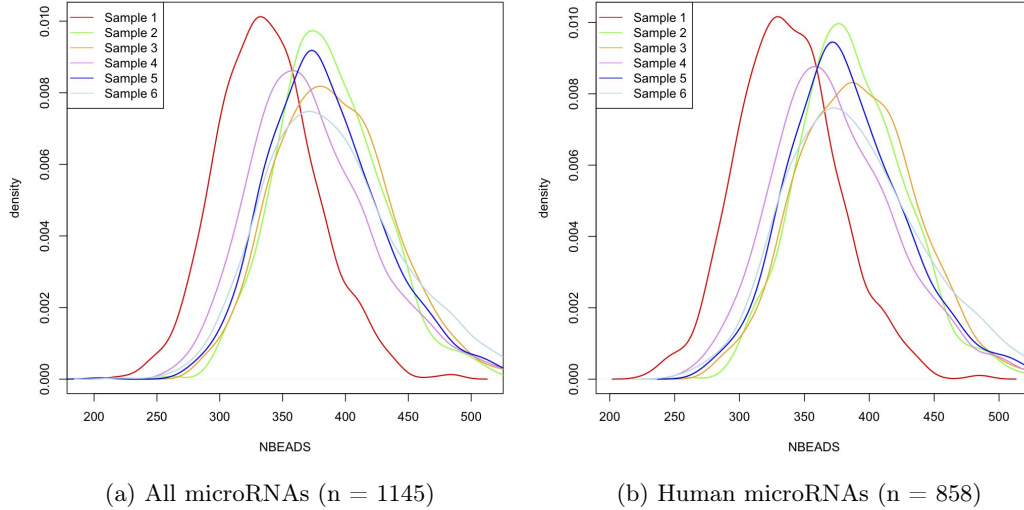


Figure 3.1: *Estimated kernel densities of number of Beads for the microRNA on the Illumina platform*

One relevant thing concerning pre-processing of Illumina data is that background subtraction, that in the end should remove the technical noise from the measurements by subtracting intensities of the negative control probes from the intensities of "regular" probes, has not been performed using the standard Illumina method, as suggested for instance in [92]. The reason for this is that simple background subtraction is likely to introduce negative estimates of intensity which are not allowed on the Illumina platform (differently from what happens, for instance, on the Agilent array), thus being automatically changed to missing value. Also, this method has been seen to introduce conservative estimates of expression levels, which turns into highly variable statistics for differential expression [93]. However, since some background subtraction has been performed for all platforms, we applied the method developed by Irizarry et al. for Affymetrix microarrays [12].

3.4 miRNA selection and Normalization

3.4.1 Automatic filtering or manual selection?

Automatic filtering is a very commonly used technique in genomic analysis which aims at removing those genes that are assumed not to be detected above a certain threshold. The key idea under this point is that if a gene shows poor level of expression than it is likely that the measured intensity values may reflect experimental noise rather than true biological knowledge. Proper filtering has been shown to increase power in detection of truly differentially expressed genes irrespective of the proportion of genes that are regulated [94] and so far many different strategies have been proposed that involve, among the others, detection p-values (used in Illumina BeadArray [95]), detection call as Absent/Present (specific of Affymetrix arrays [96]), filtering on the empirical gene variance distribution [97] and model-driven filtering [98].

However, these type of procedures are useful only if one is interested in the so-called "down-stream" analysis, i.e. evaluating differences in gene regulation between groups, unsupervised classification of samples, gene set enrichment and so on. That is, analysis where noisy data do not add information to answer our questions but increase complexity. When the purpose of an analysis is evaluate an analytical property of a platform such as its reliability or agreement between-platforms, filtering genes no matter which is the threshold considered is likely to artificially "change" results both by improving them or by worsening them. This artifact is due to the fact that assessing if two technical replicates actually "say the same thing" should be done by taking into account also the "shady" zone of the arrays themselves, i.e. those genes for which levels of expression float on the boundary between detection and non-detection. Depending on the type of the analysis to perform, filtering can thus have different effects which may lead to an under-(over-) estimation of the parameter of interest

As an exemplification, let us consider our experimental data on the hREF sample profiled on the Affymetrix platform. We filter out all microRNAs that show an empirical variance of intensity below a quantile q_α of the distribution of miRNA-wise variance of intensities, where α is set to vary between 0 and 0.5 with a step of 0.05, and for all pair-wise comparison between technical replicates we compute the simple Pearson correlation coefficient. Results

are displayed in table 3.5.

Threshold	$\geq q_0$	$\geq q_{0.05}$	$\geq q_{0.1}$	$\geq q_{0.15}$	$\geq q_{0.2}$	$\geq q_{0.25}$	$\geq q_{0.3}$	$\geq q_{0.35}$	$\geq q_{0.4}$	$\geq q_{0.45}$	$\geq q_{0.5}$
# of miRNAs	847	804	762	720	677	635	593	550	508	466	424
Repl. 1 vs Repl. 2	0.9921	0.9920	0.9920	0.9918	0.9917	0.9916	0.9915	0.9913	0.9912	0.9910	0.9907
Repl. 1 vs Repl. 3	0.9941	0.9941	0.9940	0.9939	0.9938	0.9937	0.9937	0.9935	0.9934	0.9932	0.9930
Repl. 2 vs Repl. 3	0.9956	0.9955	0.9955	0.9955	0.9954	0.9954	0.9953	0.9953	0.9953	0.9953	0.9952

Table 3.5: *Pearson correlation coefficients between technical replicates of the Affymetrix platform for line hREF according to different degrees of variance-based filtering*

Notably, there seem to be not a relevant difference between different thresholds, meaning that correlation patterns are substantially identical for the majority of microRNAs detected on the arrays. Actually, we used Pearson's ρ simply for consistency with data reported in literature, but, as we will see in chapter 5, values of proper concordance indices will not differ too much from these results. We have also evaluated a different strategy which makes use of the miRNA-wise mean intensity level as filtering statistic and uses the same quantile-approach described above to determine which microRNAs are to be filtered out. Results for the Affymetrix hREF replicates are reported in table 3.6, and resemble very much those presented in table 3.5.

Threshold	$\geq q_0$	$\geq q_{0.05}$	$\geq q_{0.1}$	$\geq q_{0.15}$	$\geq q_{0.2}$	$\geq q_{0.25}$	$\geq q_{0.3}$	$\geq q_{0.35}$	$\geq q_{0.4}$	$\geq q_{0.45}$	$\geq q_{0.5}$
# of miRNAs	847	804	763	722	677	636	594	550	508	466	424
Repl. 1 vs Repl. 2	0.9921	0.9921	0.9921	0.9920	0.9919	0.9917	0.9916	0.9915	0.9913	0.9911	0.9909
Repl. 1 vs Repl. 3	0.9941	0.9941	0.9941	0.9940	0.9939	0.9938	0.9938	0.9936	0.9935	0.9935	0.9934
Repl. 2 vs Repl. 3	0.9956	0.9955	0.9955	0.9955	0.9955	0.9955	0.9955	0.9954	0.9954	0.9953	0.9953

Table 3.6: *Pearson correlation coefficients between technical replicates of the Affymetrix platform for line hREF according to different degrees of mean-based filtering*

When it comes to cross-platform evaluations, however, things may work a little bit differently. Let us consider the same filtering strategy applied above and expression data for line hREF for all the three platforms considered. Once filtering has been performed according to a specific threshold q_α for all platforms, common unfiltered microRNAs across the three platforms have been selected, data were averaged for each miRNA within each platform to have a single intensity measure for each microRNA for each technology and simple Pearson's ρ correlation coefficient were then computed for all pair-wise platform comparisons (see table 3.7).

It is easy to see that for each platform performing some filtering can increase correlation levels, yet some caution is to be taken, since for too much higher thresholds the gain in the correlation levels is "reversed" (see the Affymetrix-Agilent comparison, for instance). If

Threshold	$\geq q_0$	$\geq q_{0.05}$	$\geq q_{0.1}$	$\geq q_{0.15}$	$\geq q_{0.2}$	$\geq q_{0.25}$	$\geq q_{0.3}$	$\geq q_{0.35}$	$\geq q_{0.4}$	$\geq q_{0.45}$	$\geq q_{0.5}$
# of miRNAs	813	694	588	502	429	362	301	251	217	182	155
Affymetrix vs Agilent	0.667	0.680	0.674	0.679	0.693	0.691	0.680	0.657	0.635	0.604	0.615
Affymetrix vs Illumina	0.538	0.557	0.570	0.594	0.604	0.613	0.619	0.607	0.601	0.584	0.548
Agilent vs Illumina	0.623	0.637	0.646	0.663	0.677	0.731	0.734	0.722	0.721	0.717	0.701

Table 3.7: *Pearson correlation coefficients according to different degrees of variance-based filtering*

we perform filtering using the mean intensity as filter statistic, on the other hand, results are substantially very different (see table 3.8): when we compare Affymetrix and Agilent platform the the level of correlation is almost constantly decreasing with the increasing of the stringency of the filtering threshold, whereas the opposite pattern can be seen for the comparison between Agilent and Illumina platforms.

Threshold	$\geq q_0$	$\geq q_{0.05}$	$\geq q_{0.1}$	$\geq q_{0.15}$	$\geq q_{0.2}$	$\geq q_{0.25}$	$\geq q_{0.3}$	$\geq q_{0.35}$	$\geq q_{0.4}$	$\geq q_{0.45}$	$\geq q_{0.5}$
# of miRNAs	813	698	605	534	471	414	364	318	284	250	215
Affymetrix vs Agilent	0.667	0.663	0.652	0.645	0.635	0.631	0.613	0.588	0.591	0.582	0.568
Affymetrix vs Illumina	0.538	0.545	0.550	0.549	0.563	0.559	0.563	0.550	0.528	0.526	0.507
Agilent vs Illumina	0.623	0.624	0.649	0.648	0.652	0.665	0.679	0.664	0.673	0.697	0.703

Table 3.8: *Pearson correlation coefficients according to different degrees of intensity-based filtering*

In figure 3.2 we have plotted the values of correlation coefficients reported in tables 3.7 and 3.8 to better compare the effects of different filtering strategies. Notably, filtering using miRNA variance is characterized by a quite similar trend for all comparison, with a peak in the correlation level at a threshold around 0.3, whereas no similar behavior is seen when filtering is performed on mean-intensity.

In view of these results, we have chosen not to perform any filtering, and to rely on the whole profile of expression of human microRNAs. Repeatability analysis was thus performed on all the human microRNA available for each platform, whereas cross-platform comparison made use of the common humanmiRNAs across platforms, resulting in a total of 813 variables, i.e. 95.99% of Affymetrix human miRNAs, 95.53% of Agilent and 94.76% of Illumina. Levels of overlap between different platforms are visualized via Venn diagram in figure 3.3.

3.4.2 Normalization: to do or not to do?

Normalization of microarray data is one of the most discussed topics throughout all genomic literature. So far, a large variety of methods exist that aim at removing array and

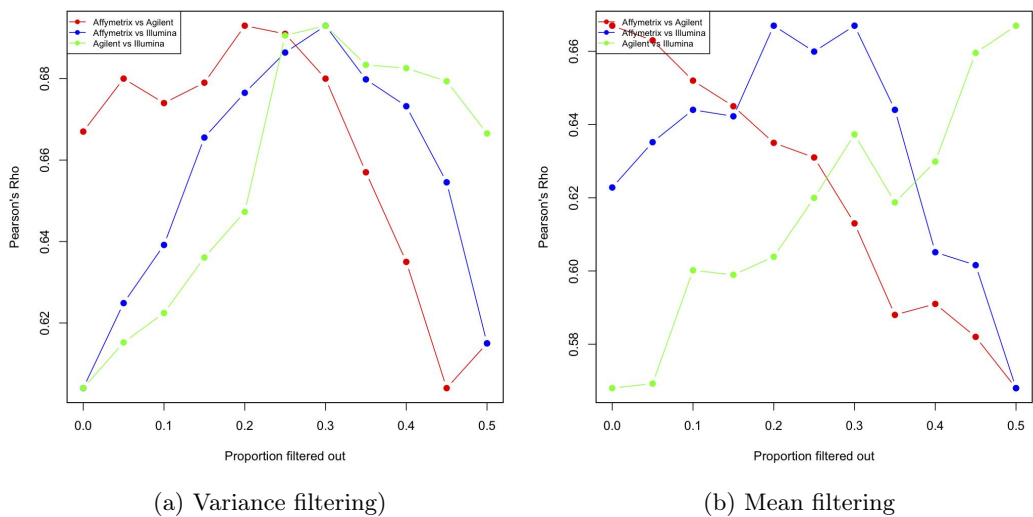


Figure 3.2: Agreement and Filtering: Cross-platform Pearson correlation coefficients according to different filtering strategies

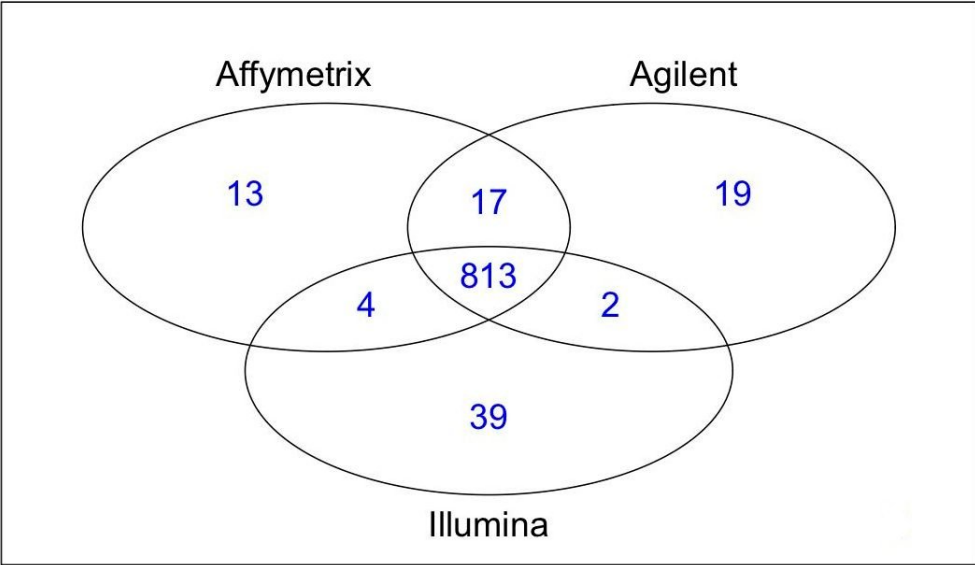


Figure 3.3: Overlap of human miRNAs on the different platforms considered.

experiment-specific effects in order to maximize the true biological information contained in expression measures, and no consensus exists about which procedure performs best. In

the field of microRNA microarrays many works have devoted to modify methods commonly applied to gene expression or attempted to developing news ones [99, 100, 101, 102, 103], yet no "gold-standard" methods exists. In figure 3.4 we show scatterplots of normalized expression values according to the different methods compared in [101], with the addition of the loess method, for one random sample processed using all the platforms considered. In the lower triangles we have reported the values of the Concordance Correlation Coefficient [23] to have an indication of the level of concordance between differently normalized data. Interestingly, the Variance Stabilizing Normalization (VSN) method [104] is the one which provides the more different results with respect to other methods, with the exception of the Illumina data, where it is seen to be mainly concordant with both loess and quantile methods (see sections 3.4.2.1 and 3.4.2.2 for details on the algorithms).

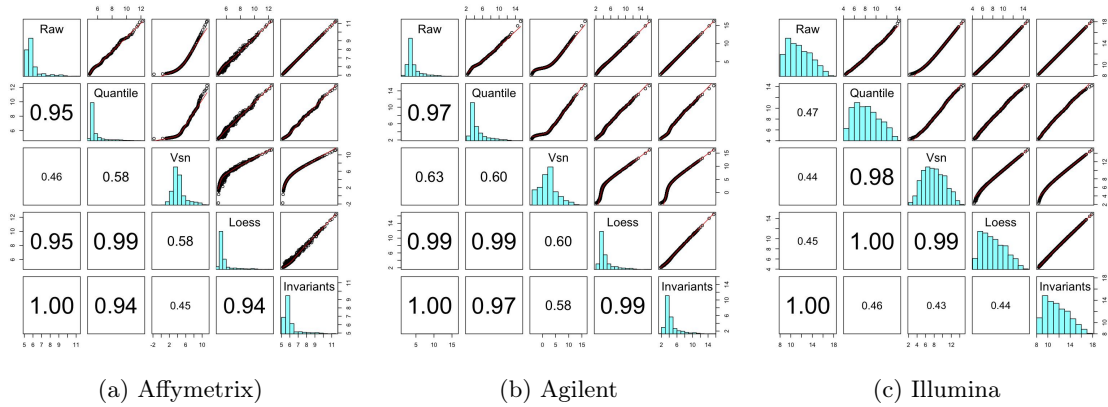


Figure 3.4: *Scatterplots of expression measures of one sample randomly chosen from the experimental data. The values in the lower triangle of the plot are the Concordance Correlation Coefficient [23] between pairs of expression measurement vectors.*

In terms of within-platform repeatability, the only effect that should be expected is that eventual platform-specific bias could be removed by the normalization procedure, thus resulting in an under-estimation of the technical effects. Concerning cross-platform comparison however, normalization is a much more relevant issue, since each algorithm is clearly affecting expression measurements in a way that is strictly platform dependent. Practically, if we normalize data we could add an unwanted and uncontrollable effect that could mask the real levels of concordance between platforms, so that we could find high level of between-platforms agreement due not to the platforms themselves but to the normalization

used or, on the other hand, the same normalization on different platforms could highlight patterns of discordance that can not be ascribed to the platforms. Nonetheless, comparing un-normalized data exposes to the risk of finding poor concordance because of incidental batch effects occurred in the experiment which may lead to an under-estimate of the "real" agreement between platforms. Actually, all the studies described in section 2.4 normalized their data, and this is certainly one of the major limitations to their findings. Moreover, in [41] the authors say that *"the idea of normalization was initially offered as an ad hoc expedient to improve significance testing for differentially expressed genes in two-sample comparisons"*, warning against the use of normalization when the focus is on the correlation structure of microarray data. Their concern is about gene-to-gene correlation, yet their point can be surely extended to the context described in this work.

In this perspective, in the present study we have chosen to use non-normalized data, in order to avoid spurious findings. However, results were also compared, when feasible, with those obtained on data normalized using the quantile and the loess normalization methods.

3.4.2.1 Quantile Normalization

Quantile normalization was first introduced for gene expression microarrays [11], and aims at making the distribution of expression measures on each array the same as the other arrays included in the experiment. Briefly, the algorithm works as follows:

1. consider the X matrix of expression values of dimension $p \times n$, where p is the number of genes/miRNAs and n is the number of subjects
2. sort each column (i.e. array/sample) independently in an increasing order to obtain X_{order}
3. compute the row-means of matrix X_{order} and assign these means to each element in the row as to get X'_{order}
4. re-arrange each column back to the original ordering of X to obtain the quantile normalized data matrix

One possible issue with this type of normalization is that it forces the distribution of possibly very different arrays to be the same (see the boxplots of figure 3.5), yet this method

has been seen to work well with microRNA microarray when compared to other methods, so it's actually much used throughout literature [101].

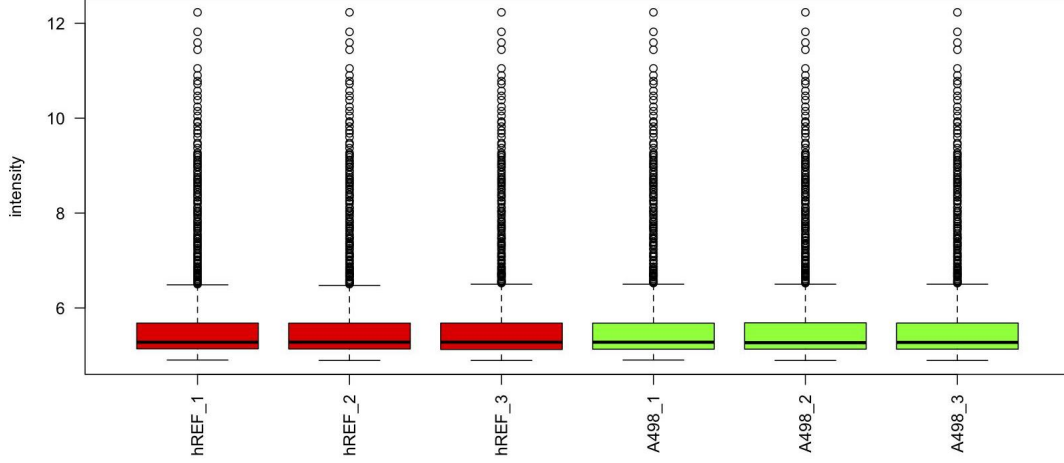


Figure 3.5: *Box-plots of quantile normalized data for Affymetrix platform.*

3.4.2.2 Loess Normalization

This method was first proposed for two-color cDNA microarrays, but has been modified to handle also oligonucleotide arrays [11]. It makes use of the well known M-A plot [105], where M is the difference in log-expression values between two probesets and A represents the average of the log-expression values. Formally, given two arrays i, j on which $k = 1, \dots, p$ probesets expression levels are measured, the algorithm follows these steps:

1. compute $M_k = \log_2(x_{ki}/x_{kj})$ and $A_k = \frac{1}{2}\log_2(x_{ki}x_{kj})$ for every probeset
2. fit a normalization curve to the M-A plot using the loess method [106]
3. estimate the adjustments $M'_k = M_k - \hat{M}_k$
4. obtain normalized probeset intensities as $x'_{ki} = 2^{A_k + \frac{M'_k}{2}}$ and $x'_{kj} = 2^{A_k - \frac{M'_k}{2}}$

When more than two arrays are involved, the algorithm simply performs this adjustment procedure for all possible pairwise combination of arrays, store all the estimated adjustments $M'_k = M_k - \hat{M}_k$ and then averages them using equal weights to obtain a single adjustment which is then applied to all the arrays. Usually, Step 2 is performed by considering only a set

of invariant genes properly defined, to avoid influence of extremely high or low expression values on the loess fit. The method, which has been seen to perform well in terms of reduction of both bias and variance has a limitation due to its computational costs, since when the number of n arrays is large $\frac{(n-1) \times n}{2}$ loess fit are to be done. A microarray experiment processed by our lab consisted of 57 arrays that were loess normalized through $\frac{(56) \times 57}{2} = 1596$ repetitions of step 2, and it took ≈ 10 minutes to do the normalization, whereas the quantile method required less than 1 second to run.

Statistical Methods for Repeatability and Agreement

Contents

4.1 Within-platform Repeatability	46
4.1.1 A random effects model	46
4.1.2 CCC and Overall CCC	54
4.1.2.1 The standard CCC from Lin (1989)	54
4.1.2.2 The CCC for repeated measures from King <i>et al.</i> (2007) . .	58
4.1.2.3 The Overall CCC from Barnhart <i>et al.</i> (2002)	59
4.2 Between-platforms agreement	60
4.2.1 The Bland-Altman method	60
4.2.2 Measurement error and agreement assessment	63
4.3 Software	69

4.1 Within-platform Repeatability

4.1.1 A random effects model

Random (and mixed) effects model are a widely used method for the analysis of longitudinal and grouped data, where it is assumed that some relationship (i.e. correlation) exists between measurements on the same unit or belonging to the same grouping unit [107]. This approach has rapidly become very popular and is now very used in the row-by-row modeling approach of microarray data to assess differential expression between experimental or clinical conditions.

Consider the variable Y which represents expression values for gene A, and let X be a categorical variable indicating a treatment group with k possible values. If our goal is to model only the k experimental groups considered than a simple linear model can be written as

$$y_{ij} = \beta_i + \varepsilon_{ij} \quad i = 1, \dots, k \quad j = 1, \dots, n_i \quad (4.1)$$

where β_i represents mean expression level of gene A for the generic i -th condition and ε_{ij} are the errors assumed to be normally distributed as $\sim N(0, \sigma^2)$. In this case the treatment is handled as a fixed effect. However, suppose that the k treatments considered are only a random sample from all the possible treatments available; if that is the case, then we can consider the factor treatment as a random effect, and thus rewrite the model in 4.1 as

$$y_{ij} = \beta + b_i + \varepsilon_{ij} \quad (4.2)$$

Here β is the overall mean expression value for gene A across sampled treatment groups and b_i is a random variable representing the deviation from the population mean of the mean expression value for the i -th treatment group. To complete the model we assume that b_i follows a Gaussian distribution $N(0, \sigma_b^2)$. The parameter of interest in this case is σ_b^2 , which tells us how much of the variation in y_{ij} is due to the treatment group. The model in (4.2) is called a *random effects model*. In a more general framework, we can define the so-called *mixed effects models*, i.e. models where both fixed and random effects are included, and define them as

$$y_i = X_i\beta + Z_ib_i + \varepsilon_i \quad (4.3)$$

In this model, β is the p -dimensional vector of fixed effects, b_i is the q -dimensional vector of random effects which follows a Normal distribution $N(0, \Sigma)$, X_i is a $n_i \times p$ regressor matrix of the known fixed effects, Z_i is a $n_i \times q$ regressor matrix of random effects and ε_i , which follows a Normal distribution $N(0, \sigma^2 I)$ is the n_i -dimensional *within-group error* vector.

As described in section 2.3.1, random-effects model are much used in different areas to assess contribution to global experimental variability from different sources. Given the

experimental design described in section 3.2 and the platform features seen in section 3.3, we can define three models, one for each array technology, that may allow us to estimate different quotas of variability for each source of variation to assess eventual technical issues related to each array. Actually, since data provided from Affymetrix and Agilent are quite similar net of the degree of probe-level data replication, the model to be defined are only two. The experimental factors involved in the study according to different technologies are reported in table 4.1.

	microRNA	Replicate	Probe
Affymetrix	yes	yes	yes
Agilent	yes	yes	yes
Illumina	yes	yes	no

Table 4.1: *Experimental factors involved for each of the platforms considered*

The experimental design for Affymetrix and Agilent can be graphically visualized as in figure 4.1. The main effects to be included in the model are related to the miRNAs, to the technical array replicates and to the replicated probes within each probeset, 4 for Affymetrix and 16 for Agilent (thus resulting in a balanced experiment). The effect of microRNA is the biological component of the assay, i.e. the one that should account for most of the total variation in the experiment, whereas the effect of the replicate arrays and of the different probes can be thought of as technical effects that, ideally, should be close to 0. MicroRNAs and Array effects can thus be considered as crossed effects, since the same microRNA is measured on each platform; opposite to this, the replicated probes can be seen as *nested* within the microRNA, since the different replicated probes are completely interchangeable for the Affymetrix platform (exactly the same sequence) and at least partially for the Agilent platform (from 1 to 4 different sequences for 16 different probes, see section 3.3.2), and thus no clear link between probe 1 for miRNA A and probe 1 for miRNA B can be established. Notably, if we consider probe-level data for Affymetrix and `ProcessedSignal` for Agilent we have a perfectly balanced experiment, resulting in a computational simplification.

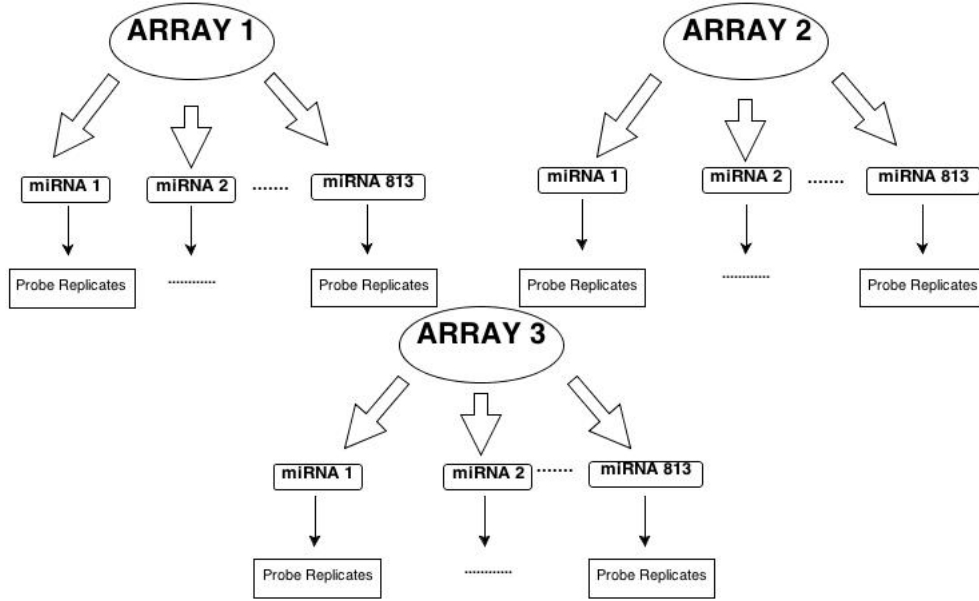


Figure 4.1: *Experimental design for the Affymetrix and Agilent platforms.*

The model can then be completely specified as

$$y_{ijkl} = \beta + a_i + b_j + (ab)_{ij} + c_{k(i)} + (bc)_{jk(i)} + e_{ijkl} \quad \begin{cases} i = 1, \dots, m \text{ (microRNA)} \\ j = 1, \dots, r \text{ (array)} \\ k = 1, \dots, s \text{ (probe)} \\ l = 1. \end{cases} \quad (4.4)$$

where greek letters represent fixed effect and latin letters represent random effects. In particular, β is the fixed intercept, or grand-mean, of the model, a_i is the effect of the i -th miRNA, b_j is the effect of the j -th replicate array, $(ab)_{ij}$ is the interaction between miRNA and array, $c_{k(i)}$ is effect of the k -th replicate miRNA probe nested within the i -th miRNA, $bc_{jk(i)}$ is the interaction between array and spot nested in miRNA and e_{ijkl} is the error component of the model. The model specified in this way leaves no degrees of freedom ($l = 1$) for the estimation of the error component since no replicates for the highest-order interaction are available: that is, if we sum degrees of freedom for the different effects in the model we obtain $N - 1$, that is the total number of intensity measurements available for each

cell line. Such a specification is however useful to have an explicit reference to all source of technical variation, a matter of fundamental importance in this study. It is thus possible, starting from the model in (4.4), to compute a classical ANOVA table in order to estimate variance components for the model effects, as presented in table 4.2. Note that the nested effect Spot in miRNA and the crossed-nested interactions Array*Spot in miRNA sum of squares are obtained as sum of simple crossed effects.

<i>Effect</i>	<i>SS</i>	<i>df</i>	<i>MS</i>
miRNA	$rs \sum_{i=1}^m (\bar{Y}_i - \bar{Y})^2$	$m - 1$	SS_a/df
Array	$ms \sum_{j=1}^r (\bar{Y}_j - \bar{Y})^2$	$r - 1$	SS_b/df
miRNA*Array	$s \sum_{i=1}^m \sum_{j=1}^r (\bar{Y}_{ij} - \bar{Y}_i - \bar{Y}_j + \bar{Y})^2$	$(m - 1)(r - 1)$	SS_{ab}/df
Probe in miRNA	$SS_{Probe} + SS_{Probe*miRNA}$	$m * (s - 1)$	SS_c/df
(probe in miRNA)*Array	$SS_{Array*Probe} + SS_{Array*miRNA*Probe}$	$m * (r - 1) * (s - 1)$	SS_{bc}/df

Table 4.2: ANOVA table for general 3-way nested factorial design.

where

$$SS_{Probe} = mr \sum_{k=1}^s (\bar{Y}_k - \bar{Y})^2 \quad (4.5)$$

$$SS_{miRNA*Probe} = r \sum_{i=1}^m \sum_{k=1}^s (\bar{Y}_{ik} - \bar{Y}_i - \bar{Y}_k + \bar{Y})^2 \quad (4.6)$$

$$SS_{Array*Probe} = m \sum_{j=1}^r \sum_{k=1}^s (\bar{Y}_{jk} - \bar{Y}_j - \bar{Y}_k + \bar{Y})^2 \quad (4.7)$$

$$SS_{Array*miRNA*Probe} = \sum_{i=1}^m \sum_{j=1}^r \sum_{k=1}^s (\bar{Y}_{ijk} - \bar{Y}_{ij} - \bar{Y}_{ik} - \bar{Y}_{jk} + \bar{Y}_i + \bar{Y}_j + \bar{Y}_k + \bar{Y})^2 \quad (4.8)$$

Note that $\bar{Y}$ is the grand mean of the intensity values, and that indexed mean are the group-specific means of miRNA intensity. Starting from table 4.2, it is possible to estimate the variance components for each factor included in the model by using the method of moments (see table 4.3).

Estimation of confidence intervals for these variance components is based on Satterthwaite's approximation [108], as suggested also by Montgomery in [109]. Given σ the generic

<i>Effect</i>	<i>Expected MS</i>	σ^2
miRNA	$rs\sigma_a^2 + r\sigma_c^2 + s\sigma_{ab}^2 + \sigma_{bc}^2$	$(MS_a - MS_c - MS_{ab} + MS_{bc})/rs$
Array	$ms\sigma_b^2 + s\sigma_{ab}^2 + \sigma_{bc}^2$	$(MS_b - MS_{ab})/ms$
miRNA*Array	$s\sigma_{ab}^2 + \sigma_{bc}^2$	$(MS_{ab} - MS_{bc})s$
Probe in miRNA	$r\sigma_c^2 + \sigma_{bc}^2$	$(MS_c - MS_{bc})/r$
(Probe in miRNA)*Array	σ_{bc}^2	MS_{bc}

Table 4.3: *Expected Mean Squares and variance components estimates for the effects in model (4.4). MS_a = Mean Square for miRNA; MS_b = Mean Square for Array; MS_{ab} = Mean Square for miRNA*Array; MS_c = Mean Square for Probe in miRNA; MS_{bc} = Mean Square for (Probe in miRNA)*Array*

variance component, its confidence interval at level α can be obtained as

$$\left(\frac{f\hat{\sigma}^2}{\chi_{1-\alpha/2, f}^2}, \frac{f\hat{\sigma}^2}{\chi_{\alpha/2, f}^2} \right) \quad (4.9)$$

To compute the quantity f , which depends on the specific component for which the interval is required, one has to look at table 4.3 and make a ratio between the squared linear combination of mean squares in third column (excluding eventual denominators) and the sum of the squared mean squares in the numerator divided by the degrees of freedom of the related effect which can be derived from table 4.2. Equations from 4.10 to 4.14 report formulas for calculating the values of f for all the variance components estimated in model (4.4).

$$f_a = \frac{(MS_a - MS_c - MS_{ab} + MS_{bc})^2}{\frac{MS_a^2}{(m-1)} + \frac{MS_c^2}{m(s-1)} + \frac{MS_{ab}^2}{(m-1)(r-1)} + \frac{MS_{bc}^2}{mr(s-1)}} \quad (miRNA) \quad (4.10)$$

$$f_b = \frac{(MS_b - MS_{ab})^2}{\frac{MS_b^2}{(r-1)} + \frac{MS_{ab}^2}{(m-1)(r-1)}} \quad (Array) \quad (4.11)$$

$$f_{ab} = \frac{(MS_{ab} - MS_{bc})^2}{\frac{MS_{ab}^2}{(m-1)(r-1)} + \frac{MS_{bc}^2}{m(r-1)(s-1)}} \quad (miRNA * Array) \quad (4.12)$$

$$f_c = \frac{(MS_c - MS_{bc})^2}{\frac{MS_c^2}{m(s-1)} + \frac{MS_{bc}^2}{m(r-1)(s-1)}} \quad (Probe in miRNA) \quad (4.13)$$

$$f_{bc} = \frac{(MS_{bc}^2)}{\frac{MS_{bc}^2}{m(r-1)(s-1)}} \quad ((Probe \text{ in } miRNA)*Array) \quad (4.14)$$

Actually, the issue of estimating confidence intervals for variance components is a very discussed one, and also a bootstrap-oriented approach has been proposed [110], that however was not pursued in this work.

Relevance of the single variability sources was also assessed by computing ICC-like measures as a ratio of each factor related variance component and the total variability, as follows:

$$ICC_{\sigma_{Effect}^2} = \frac{\sigma_{Effect}^2}{\sigma_{Total}^2} \quad (4.15)$$

To make inference on these measures, we have implemented a parametric bootstrap procedure [111] using 1000 simulations from the model fitted in equation (4.4) and estimating the CI as

$$CI \left(ICC_{\sigma_{Effect}^2} \right) = \hat{ICC}_{\sigma_{Effect}^2} \pm t_{B-1, 1-\alpha/2} * se \left(\hat{ICC}_{\sigma_{Effect}^2} \right) \quad (4.16)$$

where $t_{n, 1-\alpha/2}$ is the quantile of a student's T r.v. whose degrees of freedom are defined as the number of bootstrap replicates B -1, and $se \left(\hat{ICC}_{\sigma_{Effect}^2} \right)$ is the standard error of the bootstrap distribution of the 1000 $\hat{ICC}_{\sigma_{Effect}^2}$ bootstrap estimates of $ICC_{\sigma_{Effect}^2}$.

In the case of the Illumina platform, the model in (4.4) is simplified by the absence of the replicated probes within each miRNA so that it actually becomes as follows:

$$y_{ijl} = \beta + a_i + b_j + (ab)_{ij} + e_{ijl} \quad \begin{cases} i = 1, \dots, m \text{ (microRNA)} \\ j = 1, \dots, r \text{ (array)} \\ l = 1. \end{cases} \quad (4.17)$$

i.e. a standard two-way ANOVA random effects model with crossed interaction. Notably, also in this case the model is saturated in the sense that there is no replicate to estimate the error component, which in the end is estimated by the random effect $(ab)_{ij}$ (see table

4.4 for relevant formulas).

<i>Effect</i>	<i>SS</i>	<i>df</i>	<i>MS</i>
miRNA	$r \sum_{i=1}^m (\bar{Y}_i - \bar{Y})^2$	$m - 1$	SS_a/df
Array	$m \sum_{j=1}^r (\bar{Y}_j - \bar{Y})^2$	$r - 1$	SS_b/df
miRNA*Array	$\sum_{i=1}^m \sum_{j=1}^r (\bar{Y}_{ij} - \bar{Y}_i - \bar{Y}_j + \bar{Y})^2$	$(m - 1)(r - 1)$	SS_{ab}/df

Table 4.4: ANOVA table for general 2-way full factorial design.

Estimates of the variance components for each of the factors involved in model (4.17) are estimated accordingly, similarly to the model for the Affymetrix and Agilent platforms (see table 4.5).

<i>Effect</i>	<i>Expected MS</i>	σ^2
miRNA	$\sigma_{ab}^2 + r\sigma_a^2$	$(MS_a - MS_{ab})/r$
Array	$\sigma_{ab}^2 + m\sigma_b^2$	$(MS_b - MS_{ab})/m$
miRNA*Array	σ_{ab}^2	MS_{ab}

Table 4.5: Expected Mean Squares and variance components estimates for the effects in model (4.17). MS_a = Mean Square for miRNA; MS_b = Mean Square for Array; MS_{ab} = Mean Square for miRNA*Array.

Estimation of confidence intervals relies on the estimation of the f values just as for the previous model, see formulas from 4.18 to 4.20 for their computation. For the computation of point and interval estimates for the ICC-like measure defined in equation (4.15) we have considered the same strategy described for model (4.4), i.e. ratio of variance components for the point estimates and the standard t-approach in a parametric bootstrap framework with 1000 simulated samples.

$$f_a = \frac{(MS_a - MS_{ab})^2}{\frac{MS_a^2}{m-1} + \frac{MS_{ab}^2}{(m-1)(r-1)}} \quad (miRNA) \quad (4.18)$$

$$f_b = \frac{(MS_b - MS_{ab})^2}{\frac{MS_b^2}{r-1} + \frac{MS_{ab}^2}{(m-1)(r-1)}} \quad (Array) \quad (4.19)$$

$$f_{ab} = \frac{MS_{ab}^2}{\frac{MS_{ab}^2}{(m-1)(r-1)}} \quad (miRNA*Array) \quad (4.20)$$

Model adequacy has been checked via quantile-quantile plots and fitted vs residuals

plots, to assess eventual departure from the underlying assumptions of the model.

4.1.2 CCC and Overall CCC

4.1.2.1 The standard CCC from Lin (1989)

Within platform repeatability was assessed also by means of pairwise Concordance Correlation Coefficient by Lin [23]. The CCC ρ_c between two series of n measurements x and y is defined as

$$\rho_c = \frac{2\sigma_{xy}}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (4.21)$$

where $\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$ is the Pearson correlation coefficient between x and y , $\mu_x = \frac{\sum_{i=1}^n x_i}{n}$ and $\mu_y = \frac{\sum_{i=1}^n y_i}{n}$ are the sample means of x and y and $\sigma_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$ and $\sigma_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$ are the sample variances of x and y . Notably, in our study we have treated the miRNAs as the subjects of the study and the technical replicates as the different methods whose repeatability has to be assessed.

Whereas ρ gives an indication about precision of the relationship between x and y by telling us how close are the two series to a linear relationship (i.e. precision), c_b tells us how much do the two series depart from the 45° line of perfect agreement (i.e. accuracy). This index varies from -1 and +1, like ρ does, where -1 suggest perfect disagreement and +1 perfect agreement. Despite the range of possible values being common with that of Pearson's ρ , the CCC properly addresses the question of whether two measurements series are actually repeatable/reproducible/concordant, whereas ρ is unable to distinguish agreement from disagreement. To highlight this, we have simulated two series of measurements according to four different scenarios, and have computed both ρ and ρ_c . The different scenarios were simulated as follows: 1) $y = x$, 2) $y = x/2$, 3) $y = 2x$ and 4) $y = x - 2$, and results have been plotted in figure 4.2.

Estimates of ρ and ρ_c have been reported in table 4.6. Whereas the two estimates are quite close when the two methods actually agree quite well (scenario 1, red dots in the scatterplot), when the two methods clearly do not agree (e.g. scenario 2, green squares) Pearson's ρ is even larger than the one obtained in presence of good reproducibility between

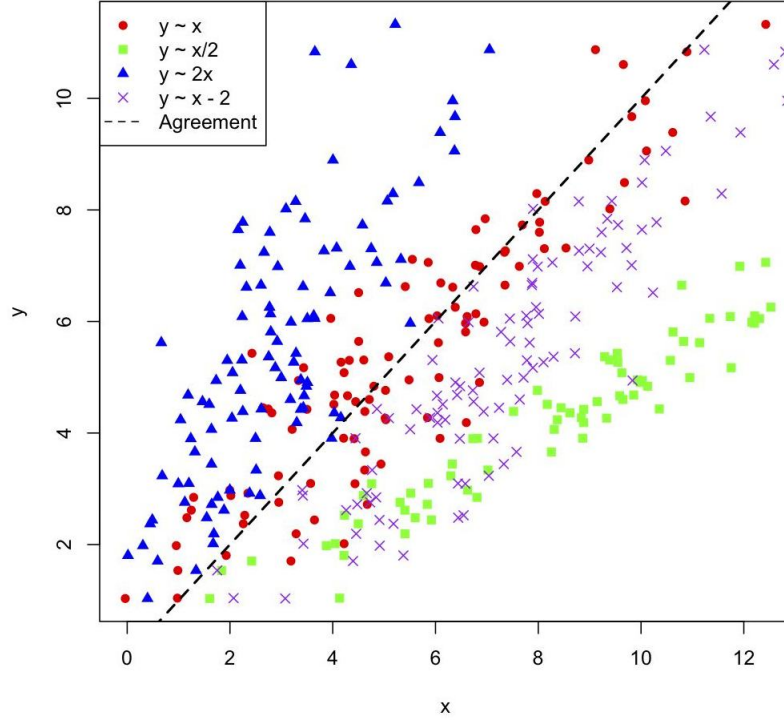


Figure 4.2: *Toy example for highlighting differences between the CCC ρ_c and Pearson's ρ .*

methods, giving a clear indication that Pearson's correlation coefficient is not a suitable measure for assessing repeatability and/or reproducibility between measurements.

Scenario	ρ_c	ρ
$y \sim x$	0.900	0.905
$y \sim x/2$	0.387	0.983
$y \sim 2x$	0.403	0.787
$y \sim x - 2$	0.673	0.911

Table 4.6: *CCC and Pearson's Rho for different x/y settings*

The CCC can also be written (see section 2.3.2) as a product of ρ and c_b , where

$$c_b = \frac{2\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (4.22)$$

Notably, c_b can also be rewritten as

$$c_b = \frac{2}{v + 1/v + u^2} \quad (4.23)$$

where:

- $v = \sigma_1/\sigma_2$ represents the scale shift between the two measurements series
- $u = (\mu_1 - \mu_2)/\sqrt{\sigma_1\sigma_2}$ is the location shift relative to the scale

By computing all these quantities (ρ, c_b, u and v) one can thus both evaluate the levels of concordance between replicate arrays for each single platform, and additionally assess which are the reasons for possible sources of poor repeatability. Notably, the CCC was computed for all possible pair-wise combination of the three technical replicates, separately for each of the two samples on each of the three platforms. Then, inference can be made on ρ_c by computing 95% confidence intervals using the inverse hyperbolic tangent transformation (or Z-transformation):

$$\hat{Z} = \tanh^{-1}(\hat{\rho}_c) = \frac{1}{2} \ln \frac{1 + \hat{\rho}_c}{1 - \hat{\rho}_c} \quad (4.24)$$

which provides better asymptotic normality with mean $Z = \tanh^{-1}(\rho^c) = \frac{1}{2} \ln 1 + \rho_c / 1 - \rho_c$ and variance

$$\sigma_{\hat{Z}}^2 = \frac{1}{n-2} \left[\frac{(1-\rho^2) \rho_c^2}{(1-\rho_c^2) \rho^2} + \frac{4\rho_c^3 (1-\rho_c) u^2}{\rho (1-\rho_c^2)^2} - \frac{2\rho_c^4 u^4}{\rho^2 (1-\rho_c^2)^2} \right] \quad (4.25)$$

The parameters in (4.25) can be replaced by their sample estimates and hyperbolic tangent transformation can be applied to the lower and upper limit of the interval computed for $\hat{Z}$ to have the limits for ρ_c . One other possible strategy is to use bootstrap methodology [111], i.e. to resample B times with replacement the original dataset and to compute for each resampled dataset the statistic of interest. Then the confidence intervals can be built using percentiles of the empirical distribution defined by the B bootstrap values of the statistic itself. Here we have used the *bias-corrected and accelerated* (BC_a) bootstrap intervals with an *intended coverage* of 95%, thus a value of $\alpha = 0.025$. Suppose that $\hat{\theta}^{*(\alpha)}$ is the 100 $\cdot\alpha$ th percentile of the B bootstrap replications of the parameter of interest θ , then the BC_a of *intended coverage* $1-2\alpha$ is obtained as

$$BC_a = \left(\hat{\theta}^{*(\alpha_1)}, \hat{\theta}^{*(\alpha_2)} \right) \quad (4.26)$$

where

$$\begin{aligned}\alpha_1 &= \Phi \left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(\alpha)})} \right) \\ \alpha_2 &= \Phi \left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(1-\alpha)})} \right)\end{aligned}\tag{4.27}$$

are the *corrected* percentiles to be used when calculating the confidence limits. $\hat{z}_0$ reflects the bias in the bootstrap distribution intended as the discrepancy between the median of the B bootstrap estimates $\hat{\theta}^*$ and the point estimate $\hat{\theta}$, and $\hat{a}$ is the *acceleration* with which the standard error of $\hat{\theta}$ changes with respect to the true value θ . In this study we have used 1000 bootstrap replicates to estimate the confidence intervals, as suggested also by Efron (see [111], Chapter 14, Section 4).

Then, to evaluate whether pairs of technical replicates are actually in agreement we follow the non-inferiority approach proposed by Liao and colleagues [82], which mainly consists in defining a threshold, or lower-bound, $\rho_{c(CL)}$ reflecting the minimal value that the CCC should assume to conclude that two methods agree, and then testing the following hypothesis:

$$H_0 : \rho_c \leq \rho_{c(CL)} \quad \text{vs} \quad H_1 : \rho_c > \rho_{c(CL)}\tag{4.28}$$

This can be done using the confidence intervals described above, interpreting the results as follows: if the lower confidence bound falls below $\rho_{c(CL)}$ than the null hypothesis can not be rejected and the two replicates can not said to be in agreement, otherwise the two replicate are in agreement. To assess the value of ρ_{CL} the authors define the minimum thresholds of precision ρ_{CL} and accuracy $(c_b)_{CL}$, and then compute it as the product of these two thresholds. In their example they propose ρ_{CL} to be 0.95 and estimate $(c_b)_{CL}$ as a combination of $u_{CL} = 0.25$ and $v_{CL} = 0.8$, which results in a threshold $\rho_{c(CL)} \approx 0.9$ (see formula (4.23)). In this paper we have chosen to use a more stringent threshold, obtained as follows:

$$\rho_{c(CL)} = \frac{2\rho_{CL}}{(v_{CL} + v_{CL}^{-1} + u_{CL}^2)} = \frac{2 * 0.98}{0.9 + 0.9^{-1} + 0.15^2} = 0.9638 \approx 0.96\tag{4.29}$$

The reason for the choice of these values is extremely subjective, but in this case we have also tried to be as much conservative as possible: a higher value for ρ_{CL} means that we want a relationship between technical replicates to be as much linear as possible, though leaving space for small departures due to ineffective probes or small experimental effects, whereas increasing v_{CL} to 0.9 is due to the fact that we assume microRNA measurements to be less variable than gene expression and so we believe that technical replicates show very similar patterns of variability. Only u_{CL} is unchanged, because the value proposed in [82] appeared reasonable also for microRNA microarrays.

4.1.2.2 The CCC for repeated measures from King *et al.* (2007)

The measures defined in previous section assume that all the measurements from the two methods to be compared are independent, which in our case is actually true only for the Illumina data, where no repeated measures is available. To make the data from the Agilent and Affymetrix platform suitable for using the classical CCC we should summarize probe-level data to have a single intensity measure for each microRNA, but this level of data is crucial for assessing the amount of variability related to probe replication in the assay. To estimate concordance between technical replicates we have thus used the modified version of the CCC from King *et al.*, that takes into account the presence of repeated measures in the data [112].

To define the index, assume that X_{ij} and Y_{ij} are the measurements of a certain quantity on the i -th subject and on the j -th repeated measure, with $i = 1, \dots, n$ and $j = 1, \dots, p$. Let D be a $p \times p$ non-negative definite matrix whose elements d_{jk} are weights that describe the "strength" of the correlation between repeated measures within a subject. Then the RM-CCC is defined as

$$\rho_{c,rm} = \frac{\sum_{j=1}^p \sum_{k=1}^p d_{jk} (\sigma_{X_j Y_k} + \sigma_{Y_j X_k})}{\sum_{j=1}^p \sum_{k=1}^p d_{jk} (\sigma_{X_j Y_k} + \sigma_{Y_j X_k}) + \sum_{j=1}^p \sum_{k=1}^p d_{jk} (\mu_{X_j} - \mu_{Y_j})(\mu_{X_k} - \mu_{Y_k})} \quad (4.30)$$

which reduces to the standard CCC in (4.21) when $p = 1$. To estimate $\rho_{c,rm}$ the sample estimates of the parameters involved in (4.30) are substituted to the true values. The authors suggest 4 different types of weight matrix D , depending on the relevance one wishes to give

to the between replicates concordance, i.e. the off-diagonal elements d_{jk} where $j \neq k$. Here we have assumed that replicate probes within both the Affymetrix and the Agilent platforms are completely interchangeable, which is reasonable for the first one but questionable for the second, and choose to assign weight 1 to every replicate pair, so that $d_{jk} = 1, \forall (j, k)$.

As for Lin's coefficient, confidence intervals can be built for this measure by resorting to the inverse hyperbolic tangent transformation, so that we have $\hat{Z}_{rm} \sim N(Z, V(\hat{\rho}_{c,rm})/(1 - \rho_{c,rm}^2)^2)$, where $V(\hat{\rho}_{c,rm})$ is the variance of $\rho_{c,rm}^2$ and is estimated using the delta method [112]. Another alternative approach that we have applied in this study is the bootstrap, which is applied as described in section 4.1.2.1 using 1000 bootstrap samples. To preserve the grouped data structure resampling was performed on miRNAs rather than on simple intensity measures.

Starting from these interval estimates we perform repeatability assessment between technical replicates by using the non-inferiority approach described above. Defining threshold values for the measure described in equation (4.30) is however non trivial, so for simplicity we consider the same threshold value described in formula (4.29).

4.1.2.3 The Overall CCC from Barnhart *et al.* (2002)

Once the CCC in (4.21) for Illumina and the RM-CCC in (4.30) for Agilent and Affymetrix have been computed alongside their confidence intervals for all possible pairwise comparisons for each of the three platforms, a global assessment of within-platform repeatability for a specific sample can be performed by evaluating whether the null hypothesis described in (4.28) is rejected for all pairs of replicates or not. If this happens, the platform show good patterns of repeatability, otherwise not. Another strategy which has been proposed is to consider a weighted average of pairwise CCC, to obtain an Overall CCC (OCCC), that can provide information on global reproducibility patterns [24]. This measure is defined as

$$\rho_c^0 = \frac{\sum_{j=1}^{J-1} \sum_{k=j+1}^J \xi_{jk} \rho_c^{jk}}{\sum_{j=1}^{J-1} \sum_{k=j+1}^J \xi_{jk}} \quad (4.31)$$

where ρ_c^{jk} is the standard Lin's CCC between j -th and k -th replicate measurement series, (in our study these are the replicate arrays), and ξ_{jk} are the weights, specific of each paired

comparison:

$$\xi_{jk} = \sigma_j^2 + \sigma_k^2 + (\mu_j - \mu_k)^2 \quad (4.32)$$

These weights substantially tend to disadvantage replicate pairs characterized by large differences in means and with large variance, thus reflecting the bias and the imprecision in the replication process. This measure has been proposed for data where no repeated measures for each individual subject are available, yet the weights defined in (4.32) can be reasonably adopted also in situations where multiple measurements are made on every subjects.

Confidence intervals are then computed using the bootstrap procedure already described for both the CCC and the RM-CCC, accounting for the within-subject (i.e. the miRNAs) replication by cluster resampling, when necessary. 1000 bootstrap replicates were considered for consistency with previously discussed methods, and the BC_a bootstrap confidence intervals were considered. Since the OCCC is an average of CCC, repeatability assessment is then performed following the non-inferiority approach described before, using the threshold used for the standard ρ_c of 0.96.

4.2 Between-platforms agreement

4.2.1 The Bland-Altman method

As described in section 2.3.2 the Bland-Altman approach is a widely used method which is commonly used to assess agreement between different measurement methods [61, 62]. Basically, given two series of normally distributed measurements X and Y on $i = 1, \dots, n$ subjects then $D = Y - X$ is also normally distributed so that

$$D \sim N(\mu_D, \sigma_D) \quad (4.33)$$

Thus, Bland and Altman suggest to use the following interval to assess whether two methods can be considered in agreement:

$$LoA = [\bar{D} - 2 S_D, \bar{D} + 2 S_D] \quad (4.34)$$

These are called Limits of Agreement (LoA), as briefly introduced in section 2.3.2, and identify the range of values where 95% of all the differences, including not observed ones, between measurements from method Y and measurements from method X are expected to lie if normality described in equation (4.33) holds and for n large. Here $\bar{D} = \frac{1}{n} \sum_{i=1}^n D_i$ is the estimate of μ_D and $S_D^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2$ is the estimate of σ_D^2 .

Once the limits are computed, they are compared to a scientifically acceptable boundaries to assess whether the two methods can be said to agree or not. In practice, this methodology is commonly used to assess whether two different methods, providing measurements of the same quantity, are interchangeable in the sense that *one might replace the other with sufficient accuracy for the intended purpose of measurement* [61]. Additionally, one can visualize the LoA as described in section 2.3.2, to have a better characterization of the patterns of agreement and, more important in some situations, of disagreement. Precision of the LoA can be assessed by estimating their $(1-\alpha)\%$ confidence intervals, which can be obtained as follows:

$$CI95\%_{LoA_L} = \hat{LoA}_L \pm t_{n-1, 1-\alpha/2} 1.71 S_D / \sqrt{n} \quad (4.35)$$

$$CI95\%_{LoA_U} = \hat{LoA}_U \pm t_{n-1, 1-\alpha/2} 1.71 S_D / \sqrt{n}$$

which hold, once again, if normality assumption introduced in equation (4.33) is satisfied. This method is very powerful for evaluating patterns of reproducibility and/or agreement, however, only if two important assumptions other than normality of differences are met, which is not always the case: 1) the variation of differences is constant across all the range of measurements, and 2) the difference between the methods is constant across all the range of measurements. Practically, assumptions 1) and 2) are met if:

- 1) the points scatter around the mean difference with no increasing (or decreasing) pattern
- 2) the points scatter along a line parallel to the x axis

In figure 4.3 we show situations where these assumptions are violated. The red line is the fit of a standard linear model of the type $D = \alpha + \beta A + \varepsilon_i$, that we use to

assess assumptions violations. In the left panel the estimate of the intercept α is 0.245 (CI95%: -3.69 - 4.179) and that of the slope β is -0.013 (CI95%: -0.08 - 0.055), yet we can see that the differences show a scattering pattern around the 0 horizontal line which tends to increase with the value of the average measurement between the methods, possibly suggesting a mean-variance dependence. On the other hand, in the right panel variability is constant across the range of measurements yet the regression of differences vs mean returns an intercept of 1.724 (CI95%: 1.174 - 2.274) and a slope of -0.583 (CI95%: -0.0752 - 0.414).

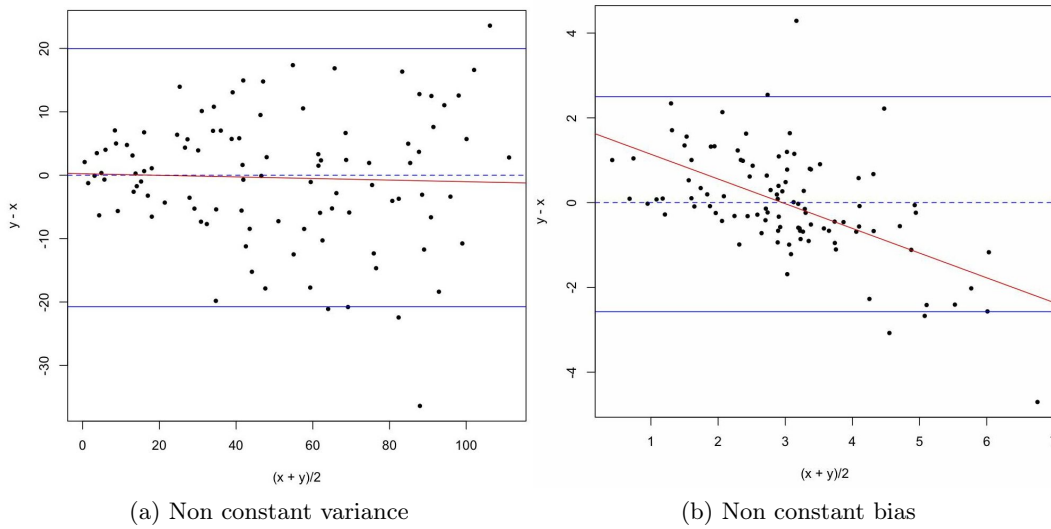


Figure 4.3: *Scenarios where the assumptions of the Bland-Altman method are violated. Solid blue line represents $\bar{D}$, dashed blue lines are the LoA (see equation (4.34)) and red line is the fit of a linear model of the differences of the measurements vs their average.*

To circumvent the issue of non constant bias some modifications to the standard approach have been proposed [63, 64, 113], which rely on fitting a regression model such as the one described above and to use parameter estimates to derive proper LoA, whereas for removing non constant variability it is suggested that some data transformation is performed. All the proposed modifications however, still do not provide a conclusion on the reproducibility/interchangeability/agreement between the methods being compared, since the LoA discussed so far are intervals that should contain 95% of all the differences between the two methods that are likely to arise if we measured them on an infinite population, yet this information still has to be compared with some knowledge which is to be assumed as a

gold-standard. The main issue is that when no previous knowledge exist, we have nothing to use as reference to assess methods agreement, and in the field of microarrays this is a quite relevant problem which has to be taken in consideration.

4.2.2 Measurement error and agreement assessment

Another problem which is often underevaluated in microarrays is the issue of measurement error. All the studies that focused on assessing platform performance both in terms of reliability and concordance did not consider the intrinsic nature of measurement assay of microarray platforms, and simply performed normalization "hoping" that this would have removed all the technical artifacts occurred during the experiments, measurement error as well. Actually, this is likely to produce severe bias to the obtained results, in that not considering it is likely to *attenuate* the real relationship which could be observed if we had had "error-free" data, which in practice are not available [114].

In the context of regression, this means that both the response Y and the covariates X have to be assumed to be affected by some measurement error, whereas in the classical linear model theory the X are assumed to be "gold" in the sense that their value is not affected by any type of error in the measuring phase. The previously introduced concept of attenuation of effect due to measurement error is illustrated in linear regression in figure 4.4: the small "x" on the plot represent the real values of 200 (X, Y) pairs, simulated according to the model $Y = \alpha + \beta X + \varepsilon$ with $\alpha = 2$, $\beta = 2$ and $\varepsilon \sim N(0, 1)$, and the small triangles represent the actual "observed" values, that is the ones we do "read in" when we perform statistical analysis, obtained by adding Gaussian measurement error to X (normally distributed with 0 mean and $\sigma = 3$). The OLS estimates on the true values are $\alpha = 1.694$ and $\beta = 2.027$, whereas the OLS on the observed values are $\alpha = 10.36$ and $\beta = 1.164$, which substantially means an attenuation of the relationship between X and Y [115]. Fitting a model that takes into account the existence of measurement error to the "observed" data leads to an estimate of the intercept of 4.546 and to a slope of 1.745, which are much closer to the "real" estimates, though still showing some patterns of attenuation.

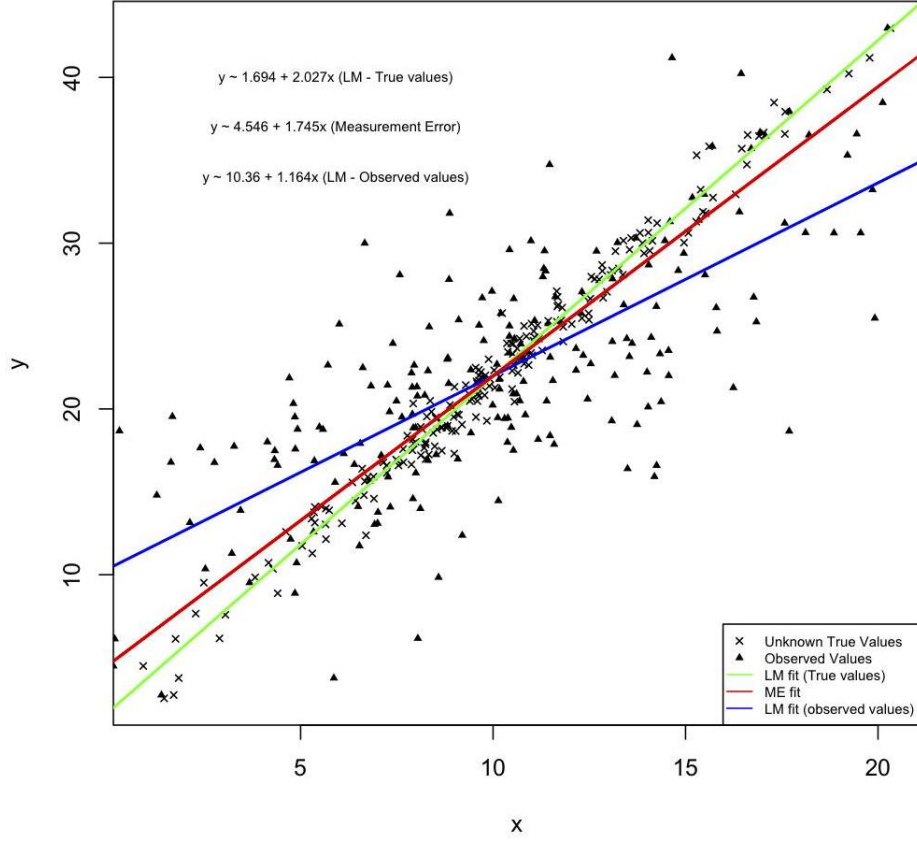


Figure 4.4: *Example of attenuation in linear regression due to measurement error.*

In general, a linear measurement error model has the following form:

$$Y_i = a_0 + b_o X_i^0 + \varepsilon_i \quad (4.36)$$

$$X_i = X_i^0 + \delta_i \quad (4.37)$$

where $(X_i^0, Y_i^0), i = 1, \dots, n$ are the unobserved true values of the two variables X and Y, and ε_i and δ_i are the i.i.d. error components of the model, which follow a normal distribution with zero mean and variances equal to σ_ε^2 and σ_δ^2 , respectively [116, 117]. This kind of models can be used to assess patterns of agreement between measurement methods by incorporating its parameter estimates in the building of the Limits of Agreement seen for the Bland-Altman approach [25, 118].

With reference to the model in (4.36) and (4.37), the parameters a_0 and b_0 are estimated as follows:

$$a_0 = \bar{Y} - \hat{b}_0 \times \bar{X} \quad (4.38)$$

$$b_0 = \frac{-(\lambda S_{xx} - S_{yy}) + \sqrt{(S_{yy} - \lambda S_{xx})^2 + 4\lambda S_{xy}^2}}{2S_{xy}} \quad (4.39)$$

where:

- $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ is the sample mean for method X measurements
- $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ is the sample mean for method Y measurements
- $S_{xx} = \sum_{i=1}^n (X_i - \bar{X})^2$ is the sample deviance for method X measurements
- $S_{yy} = \sum_{i=1}^n (Y_i - \bar{Y})^2$ is the sample deviance for method Y measurements
- $S_{xy} = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$ is the sample codeviance
- $\lambda = \sigma_\varepsilon^2 / \sigma_\delta^2$ is the ratio of the error variances of methods Y and X

The key of model parameters estimation is the value of λ , the ratio of the error variances of methods Y and X. In many applications it is common practice to set it equal to 1, thus assuming that the two measurement methods have the same, though unknown, error variance. This assumption is reasonable in many situations, and does not require previous knowledge of the actual values of σ_ε^2 and σ_δ^2 . Alternative approaches for choosing the values of λ imply its estimation, by means of replication. In this work we use the technical replicates available for each microarray platform and for each sample to estimate the error variances and then directly compute the value of ratio of error variances for each platform pair. That is, within each platform and sample we fit the following random effects model:

$$y_{ijks} = a_k + \varepsilon_{ijks} \quad (4.40)$$

where $i = 1, 2, 3$ represents the platform, $j = 1, 2$ the sample, $k = 1, 2, 3$ the technical replicate and $s = 1, 2, \dots, 813$ the microRNA. From this model we take ε_{ijk} to be the error

variance for platform i and sample j and values of λ are estimated as the ratio of the residual variances for all platform pairs for both samples involved in the experiment. To assess whether these estimated values differ from 1 we compute 95% confidence intervals as follows:

$$\frac{\hat{\lambda}}{F_{(n_1-1, n_2-1), 1-\alpha/2}} < \lambda < \frac{\hat{\lambda}}{F_{(n_1-1, n_2-1), \alpha/2}} \quad (4.41)$$

where $n_1 = n_2 = 813$ and $\alpha = 0.05$. Once the point estimates of the model parameters are obtained, one can compute confidence intervals for both a_0 and b_0 :

$$\hat{a}_0 - \frac{z_{\alpha/2} \hat{\sigma}_a^2}{\sqrt{(n)}} \leq a \leq \hat{a}_0 + \frac{z_{\alpha/2} \hat{\sigma}_a^2}{\sqrt{(n)}} \quad (4.42)$$

$$\hat{b}_0 - \frac{z_{\alpha/2} \hat{\sigma}_b^2}{\sqrt{(n)}} \leq b \leq \hat{b}_0 + \frac{z_{\alpha/2} \hat{\sigma}_b^2}{\sqrt{(n)}} \quad (4.43)$$

where:

- $\hat{\sigma}_a^2 = \frac{\hat{\sigma}_\delta^2}{n} + \bar{X}^2 \hat{\sigma}_b^2$ is the estimated variance of a_0
- $\hat{\sigma}_\delta^2 = \frac{1}{n} \left(S_{xx} - \frac{S_{xy}^2}{b_0} \right)$ is the estimate of the error variance of Y .
- $\hat{\sigma}_b^2 = \frac{(1 + \lambda b_0^2)(S_{xx} S_{yy} - S_{xy}^2)}{(S_{xx} - \lambda S_{yy})^2 + 4\lambda S_{xy}^2}$ is the estimated variance of b_0

Starting from point and interval estimates of the intercept and slope of the model in (4.36) and (4.37) we build the modified versions of the agreement interval for $Y - X$ that was originally proposed by Bland & Altman [61, 62]. That is, $Y - X$ is assumed to be normally distributed

$$Y - X \sim N(a_0 + (b_0 - 1) \times X, \text{Var}(Y - X)) \quad (4.44)$$

and the variance of this Gaussian distribution is obtained as follows:

$$\text{Var}(Y - X) = \text{Var}(Y) + \text{Var}(X) = \sigma_\epsilon^2 + \sigma_\delta^2 = \lambda \sigma_\delta^2 + \sigma_\delta^2 = (1 + \lambda) \sigma_\delta^2 \quad (4.45)$$

Under condition of perfect agreement (i.e. $a = 0$ and $b = 1$), then the new agreement interval has the following form:

$$\Delta = \left(-t_{1-\alpha/2, n-1} \sqrt{1 + \lambda \hat{\sigma}_\delta}, +t_{1-\alpha/2, n-1} \sqrt{1 + \lambda \hat{\sigma}_\delta} \right) \quad (4.46)$$

This interval, though somehow resembling the classical LoA reported in equation (4.34) has a crucial difference: this interval does not suggest the limits within which 95% of all of the differences between Y and X are expected to lie, but tells us that if perfect agreement exists, i.e. if the measurement error model is a 45° straight line through the origin, this is the *possible range for the difference if there is no real "changes" occurred* [25]. So, depending on the scattering of the observed points around the line estimated by model (4.36)-(4.37), a certain percentage of subjects will lie within this interval and the remaining will not. The other difference with the Bland-Altman LoA lies in the choice of the critical value: in Bland-Altman the value 2 was used as an approximation of $Z_{1-\alpha/2} = 1.96$, here we use the $t_{1-\alpha/2, n-1}$ that for large n converges to 1.96.

Equation (4.46) is reasonable only if inference on parameters a_0 and b_0 based on the confidence intervals (see equations (4.42)-(4.43)) suggest that a perfect agreement situation should be expected. However, in some cases it is known that some bias exists between the two methods, both fixed and/or proportional. In the first case, i.e. when only a fixed bias between the two methods is present, this is equivalent to a situation when the intercept in the model is different from 0 (i.e. the confidence interval in (4.42) does not contain the value 0) and the slope b_0 of the model is not different from 1 (i.e. the value 1 is included in the confidence interval in (4.43)). This means that the two methods possibly differ by a fixed shift, which is represented by the estimate of a_0 , which leads to the following modified agreement interval:

$$\Delta = \left(a_0 - t_{1-\alpha/2, n-1} \sqrt{1 + \lambda \hat{\sigma}_\delta}, a_0 + t_{1-\alpha/2, n-1} \sqrt{1 + \lambda \hat{\sigma}_\delta} \right) \quad (4.47)$$

Similarly, the two methods could differ only by a proportional bias described by the slope of the model (i.e. $a_0 = 0$ and $b_0 \neq 1$), meaning that their difference tend to increase proportionally to the increase of the X measurements values, leading to a modified agreement

interval which has the following form:

$$\Delta = \left((b_0 - 1) X_i - t_{1-\alpha/2, n-1} \sqrt{1 + \lambda} \hat{\sigma}_\delta, (b_0 - 1) X_i + t_{1-\alpha/2, n-1} \sqrt{1 + \lambda} \hat{\sigma}_\delta \right) \quad (4.48)$$

Finally, the two methods could differ both by a fixed and a proportional bias (i.e. $a_0 \neq 0$ and $b_0 \neq 1$), which leads to this interval:

$$\Delta = \left(a_0 + (b_0 - 1) X_i - t_{1-\alpha/2, n-1} \sqrt{1 + \lambda} \hat{\sigma}_\delta, a_0 + (b_0 - 1) X_i + t_{1-\alpha/2, n-1} \sqrt{1 + \lambda} \hat{\sigma}_\delta \right) \quad (4.49)$$

Notably, the X_i are the actual measured values of the "reference" method X. Once the proper interval has been computed, depending on the inference on the parameters, one can visualize the results by plotting the differences between measurement methods versus the sample labels (when no bias or only fixed bias is present) or versus the sample labels ordered according to the values of the X method (when only proportional or both biases are present), then adding the lines representing the agreement interval and evaluating how many subjects lie within the interval. If no more than a predefined number of k subjects show $Y - X$ differences that lie outside these intervals than the two methods are said to be in agreement. The choice of the threshold k depends on the tolerance one wishes to accept: the lower the tolerance for disagreeing subjects, the lower the value of k . In this work, we have chosen a sufficiently stringent threshold for agreement, allowing no more than 5% of microRNA to lie outside the agreement interval defined by the measurement error model parameters, which means that $k = 0.05 * 813 \approx 41$. Alternatively, this means that for two platforms to be concordant we request that 95% of microRNA lie within the agreement interval defined according to formulations (4.46) - (4.49).

To perform cross-platform analysis we averaged miRNA intensities across technical replicates for each sample and platforms, and applied the proposed method for all the pair-wise combinations of platforms, separately for line A498 and hREF. Such a procedure was undertaken only after having summarized data at a probeset level on all platforms, which consisted in averaging replicated probes for Affymetrix and in extracting the `TotalGeneSignal` for the Agilent arrays. Additionally, similarly to repeatability analysis, we have also studied concordance on normalized data, to assess impact of pre-processing on final results.

4.3 Software

All the pre-processing of data discussed in chapter 3 and the statistical analysis described in this chapter have been performed in the R/Bioconductor [89, 91] environment using built-in packages as well as *ad hoc* written functions. In particular:

- import and signal extraction for the Agilent arrays was performed using package `AgiMicroRna` [119]
- import and normalization of Illumina arrays was performed using package `lumi` [120]
- quantile normalization was performed using the function `normalize.quantiles` in the package `PreProcessCore` [121] (only for Affymetrix and Agilent arrays)
- loess normalization was performed using the function `lumiN` in package `lumi`
- the nested-factorial random effects model described in section 4.1.1 was fitted using function `lmer` in package `lme4` [122]
- the ANOVA table described in table 4.2 was obtained using function `ANOVA.crossed.3ways`, see Appendix A for commented R code
- Lin's CCC and 95% confidence intervals according to inverse hyperbolic transformation were computed using function `epi.ccc` of package `epiR` [123]
- King's RM-CCC and 95% confidence intervals according to inverse hyperbolic transformation were computed using function `cccUst` in package `cccrm` [124]
- BC_a bootstrap confidence intervals for both the CCC and the RM-CCC were computed using both routines included in package `boot` [125]
- the OCCC from Barnhart *et al.* were estimated using function `occc`, and the relative confidence interval were estimated using the routines included in package `boot`
- modified Bland-Altman approach was implemented by writing a comprehensive function `liao.capen` which both estimates model parameters and graphical visualization of the proper agreement interval, according to formulas (4.46) - (4.49), see Appendix B for commented R code

Results

Contents

5.1 Data description	70
5.1.1 Affymetrix	70
5.1.2 Agilent	73
5.1.3 Illumina	75
5.2 Repeatability analysis	76
5.2.1 Random effects models	76
5.2.1.1 Affymetrix	76
5.2.1.2 Agilent	85
5.2.1.3 Illumina	89
5.2.2 CCC and Overall CCC	94
5.2.2.1 Affymetrix	94
5.2.2.2 Agilent	97
5.2.2.3 Illumina	100
5.3 Agreement Assessment	103
5.3.1 Raw data	103
5.3.2 Normalized data	111

5.1 Data description

5.1.1 Affymetrix

In figure 5.1 we report the distributional features of probe-level data for all human microRNAs ($n = 847$) available on the Affymetrix array. Since the equally colored array are

technical replicates of the same biological matter, we would expect both the boxes and the density lines to be almost identical within each sample, showing possible differences between samples. Notably, remember that hREF is a pool of tissues, so that some biological variability is likely to be present in there.

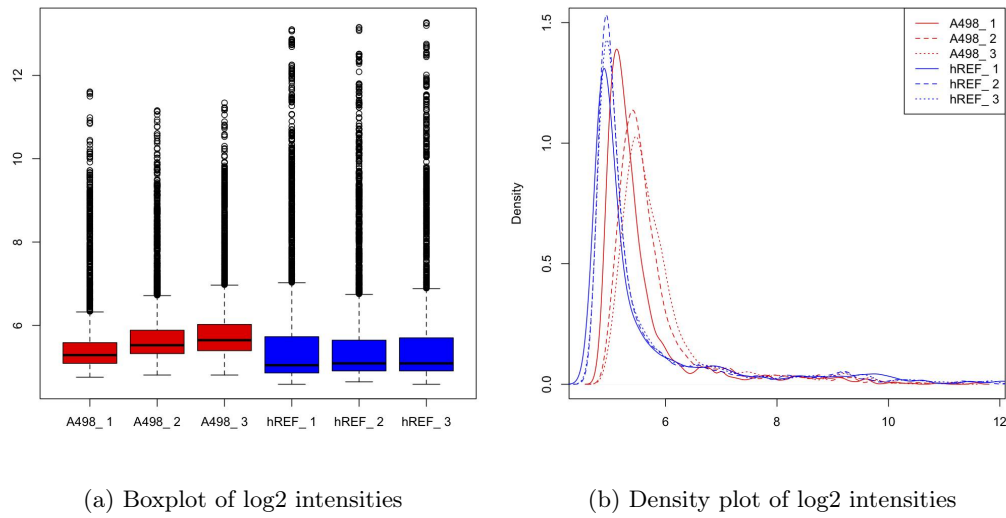


Figure 5.1: *Distribution of intensities at probe-level for log2 transformed Affymetrix data*

It is immediate to see that the first replicate on the A498 sample shows a distribution which is markedly different from the other technical replicates, with a substantially lower level of expression. Notably, the intensity peak is much more similar to that of hREF than that of the A498 replicates, possibly suggesting that some problem has occurred during the experiment which may bias results. To have a better view of possible issues in platform repeatability, we have visualized pairs of technical replicates in the scatterplots of figure 5.2, from where one can better see that in particular for line A498 there is a bulk of probes in the poorly-expressed region which shows different patterns on the different replicates, as suggested also by the red smoother. Actually, a similar yet less relevant trend can be seen for the hREF sample, in particular for Array 1, whereas Array 2 and 3 seem to be largely "concordant" with each other, at least at a first glance. Notably, however, for this samples the smoother suggests an overall good level of repeatability, being quite closer to the 45° line for all pairwise plots.

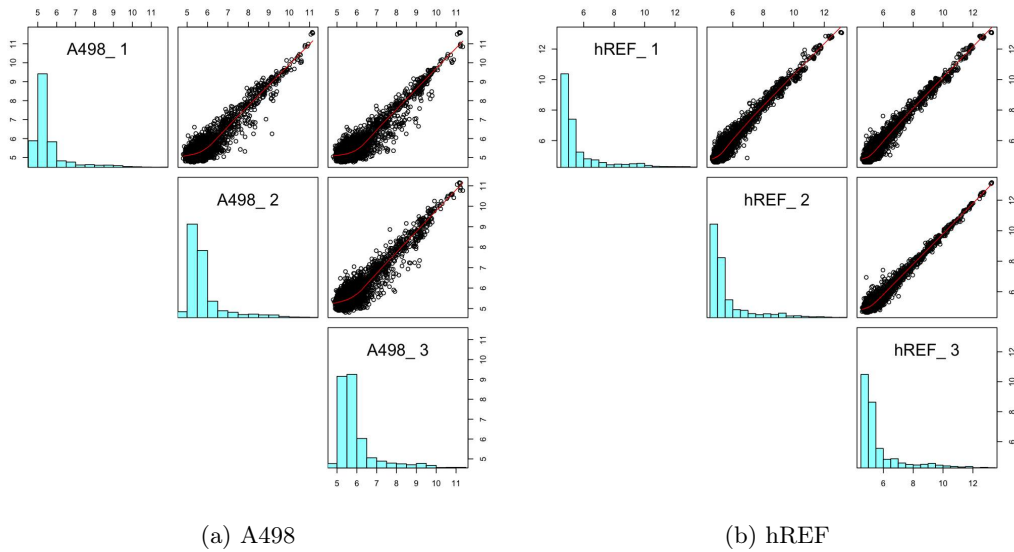


Figure 5.2: *Scatterplots of pairwise replicate arrays for Affymetrix.*

Because of these strange features, we investigated probe replicates across arrays, to assess the existence of possible biases due to some systematic effect. We extracted 9 random microRNAs (`set.seed(447)` in R) and evaluated each probe pattern across different array replicates by plotting their intensity values in a Trellis-like graphical device (see figure 5.3). Additionally to each probe "trend" we also plotted the mean trend, i.e. the trend of the mean expression across replicates for the selected microRNA. It is quite evident that line hREF show substantially mixed patterns, with microRNAs whose mean intensity value is largest on the first array (`hsa-miR-591_st`, for instance), on the second array (`hsa-miR-302b-star_st`) or on the third (`hsa-miR-136_st`). On the other hand, it is even clearer that line A498 is affected by a systematic bias, in that the pattern of the mean intensity values (solid red lines) is almost identical for all the selected miRNAs, with only `hsa-miR-218-1-star_st` showing a slightly different behavior. This has been seen to occur not only for this particular set of randomly chosen miRNAs, but for a large variety of subsamples. In particular, we have seen that $\approx 90\%$ of microRNAs has its lowest average intensity value on the first replicate array, whereas more than 70% has the largest intensity on array 3. Notably, almost 65% of the microRNAs shows the pattern described in figure 5.3b, with lowest value on array 1 and largest on array 3. Such an effect will much likely affect our

results, and will thus be carefully considered to avoid over-interpretation of the findings.

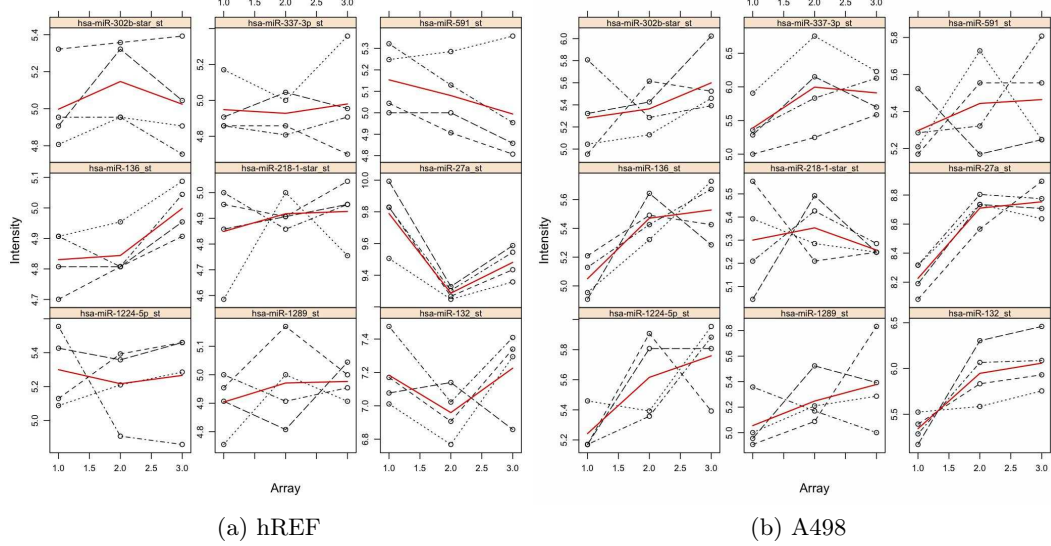


Figure 5.3: Line plot of the four spots of 9 randomly selected miRNAs. Dashed line correspond to the first spot, dotted line correspond to the second spot, dotdashed to the third and longdashed to the fourth. The solid red line represent the mean of the miRNA within each array.

5.1.2 Agilent

As discussed in section 3.3.2, repeatability analysis for Agilent platform has been performed on the ProcessedSignal, since it is the only measure which is peculiar of each probe, and is background subtracted, so that we have rid of a relevant noise source. Since the ProcessedSignal data type contains also some negative values because of features of the Agilent Feature Extraction software, we have added an offset to all signal to remove all negative intensities; such an offset was defined as suggested in [88, 90] as $|min(x)| + 2$, where $|min(x)|$ is the lowest intensity value across all arrays, and 2 is added so that the lowest intensity value after log2 transformation is 1. As in figure 5.1 we have plotted distribution of log-intensities by mean of boxplots and density plots, to highlight possible flaws in the data generating process (see figure 5.4). By looking at the distributions of technical replicates one can immediately see that the level of raw repeatability appears to be quite good, in particular for sample hREF. Notably, the third replicate for line A498 shows a discrete amount of low-intensity microRNAs. It has to be underlined that, just like the probe level

data of Affymetrix array, this intensity signal has been pre-processed, yet no "low-level" normalization has been done, differently from the Affymetrix Genechip, so that a larger degree of variability than what has been seen for previous technology should be expected for this particular type of data.

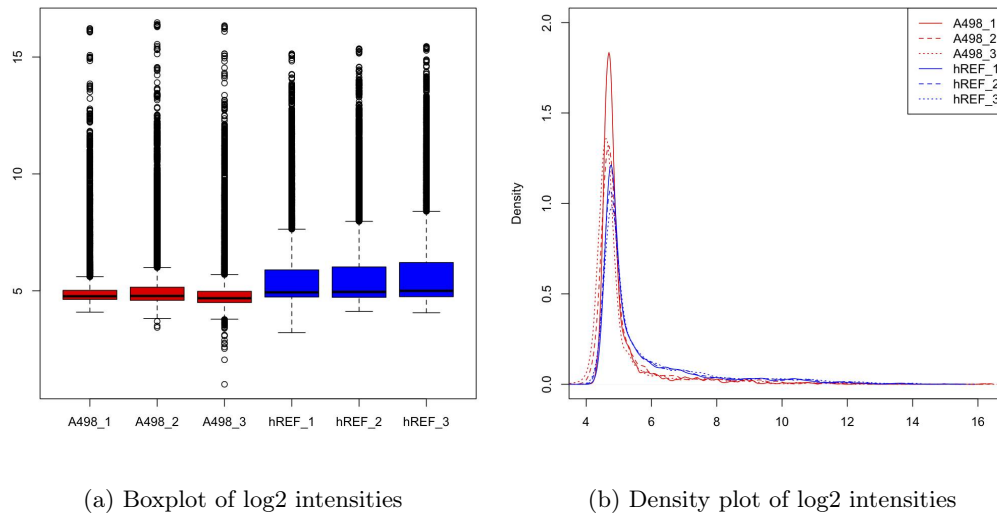


Figure 5.4: *Distribution of intensities at probe-level for log2 transformed Agilent data*

To have a first view of platform repeatability we have visualized, as for Affymetrix, scatterplots of all replicate pairs, to see whether individual points depart from the 45° line, and which is the eventual pattern of departure (see figure 5.5). The red line in the upper triangle is a loess smooth of $y \sim x$, using 2/3 of the data at each point to produce the smoothing. It is worthy to note that A498 replicates show a larger degree of variability in the lower range of values, whereas for middle-high intensities data appear quite concordant. Replicates from hREF show even better patterns, yet some scattering for low intensity miRNAs can be seen also in this case.

Though no evident systematic bias seems to affect results from this platform, yet the A498 line shows some interesting patterns on low intensity microRNAs, in particular replicate 3 is the whose smoothing curve shows the more marked departure from linearity on the left portions of its scatterplots, confirming what the box-plot had suggested, i.e. that tends to under-measure poorly expressed miRNAs. This bias appears to be more evident

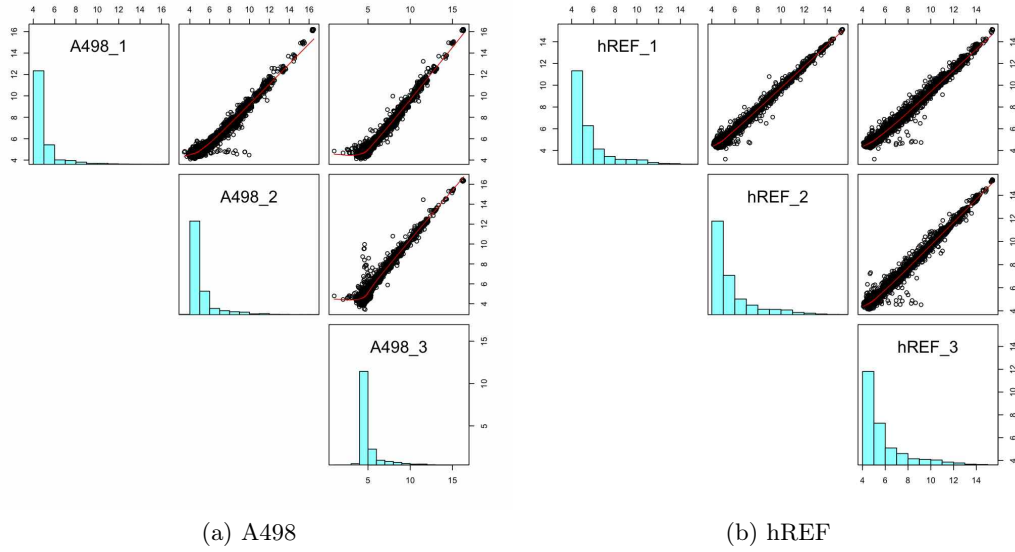


Figure 5.5: *Scatterplots of pairwise replicate arrays for Agilent.*

hen the comparison is made with array 2, for which there is a moderate bunch of miRNAs that are clearly over-measured.

5.1.3 Illumina

As outlined in Chapter 3, no probe level replicate is available for the Illumina data, so that we can only evaluate reliability of summarized probeset intensities. The boxplot and the density plot of figure 5.6 suggest a great degree of repeatability for this platform, yet this can not be directly compared to figures 5.1 and 5.4 because of the different pre-processing involved. Notably, the distribution appear to be slightly bi-modal for both samples, yet this is more evident for line A498, which shows a first peak at 2.5 (i.e. $\approx 2^{2.5} = 5.7$ on original scale) and a second approximately at 7 (i.e. $\approx 2^7 = 128$ on original scale). hREF, on the other hand, is not clearly bi-modal, yet a small peak can be detected at around 5 (i.e. $\approx 2^5 = 32$ on original scale) and the major one is seen around 10.5 (i.e. $\approx 2^{10.5} = 1448.15$ on original scale). Additionally, if we consider the scatterplots of figure 5.7, we can appreciate how all pairs of replicates appear to be very concordant, in that they lie on the 45° straight line through the origin.

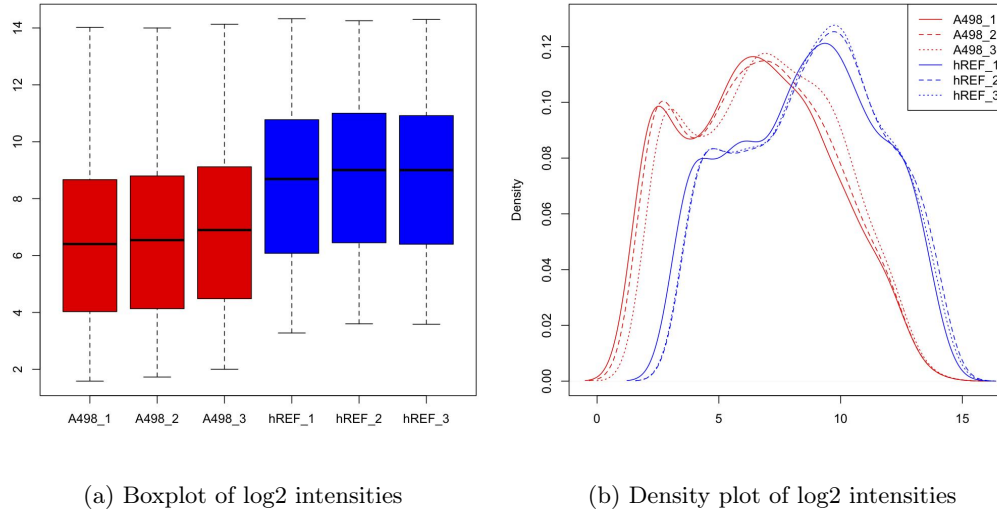


Figure 5.6: *Distribution of intensities at summarized probeset-level for log2 transformed Illumina data*

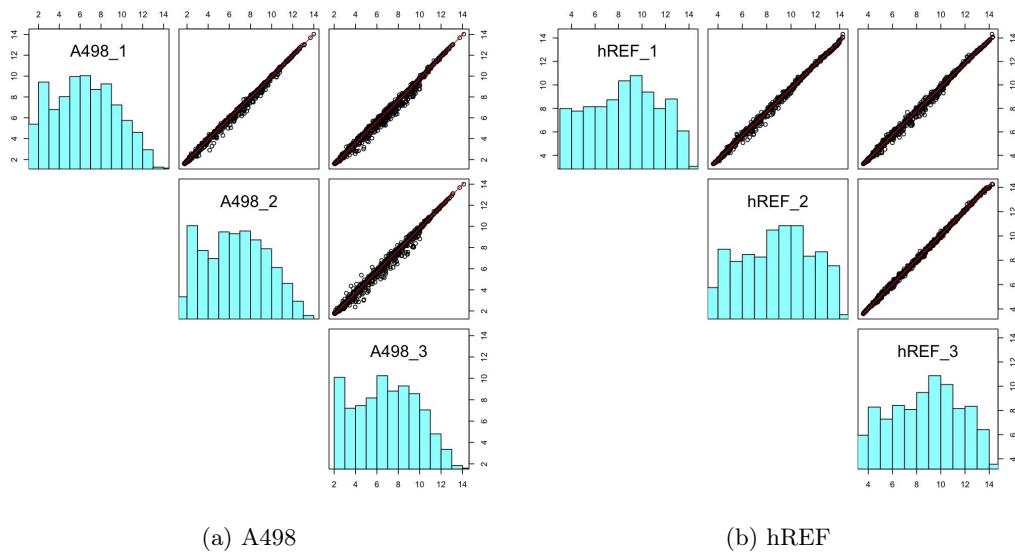


Figure 5.7: *Scatterplots of pairwise replicate arrays for Illumina platform.*

5.2 Repeatability analysis

5.2.1 Random effects models

5.2.1.1 Affymetrix

The model described in equation (4.4) was checked for adequacy by visualizing normal quantile-quantile plots of the residuals and scatterplots of residuals vs fitted values. Ac-

tually, the commonly used residuals are named *standardized residuals*, in that they are estimated as $\hat{\varepsilon}/\hat{\sigma}$, where $\hat{\varepsilon}$ is an estimate of the error component ε of the model, and assumed to be normally distributed with 0 mean and variance $\sigma^2\Sigma$, and $\hat{\sigma}$ is an estimate of the residual standard deviation [126]. Additionally, they are considered *conditional* residuals in that they are computed as $\hat{\varepsilon} = Y - (X\hat{\beta} + Z\hat{b})$, i.e. the random effects are taken into account in the estimation.

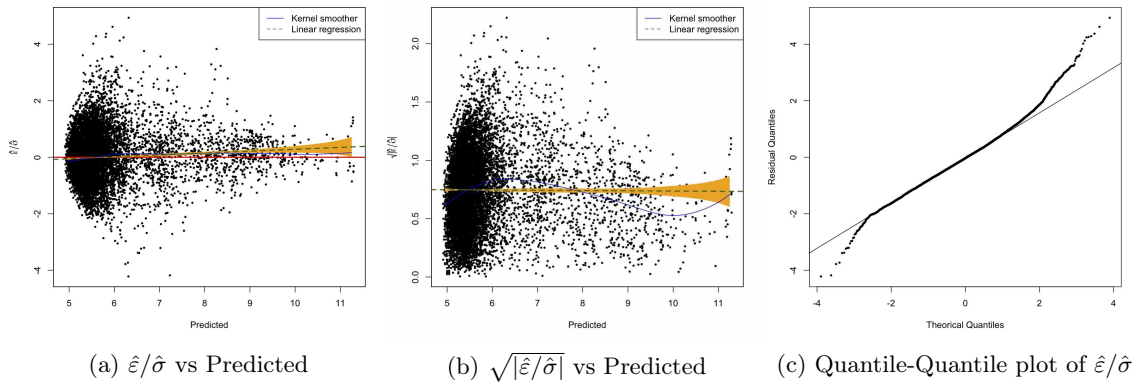


Figure 5.8: *Diagnostic plots for nested-factorial random effects, Affymetrix platform (line A498.)*

The plots of figure 5.8 refer to line A498 for Affymetrix platform. Additionally to the basic fitted vs residuals plot we have added a kernel smoother to assess eventual trends in the relationship (solid blue line in figure 5.8a), which is expected to produce points randomly scattered around 0 with a homoscedastic pattern. Additionally, we have also plotted the fit of a standard linear model, together with the confidence band around the fitted values. This linear model actually suggest a pattern different from the horizontal red line plotted as reference (Intercept = -0.388, CI95%: -0.493, -0.283; Slope = 0.067, CI95%: 0.049, 0.085), yet the difference is so small that it can be considered irrelevant. What actually is to be carefully considered is the evident heteroscedasticity pattern that can be devised, with a variability which seems to be clearly related to the predicted value. To better highlight this, it is useful to visualize, additionally to the scatter of figure 5.8a, the plot of the square root of the absolute residuals versus the predicted values (see figure 5.8b). From this plot, one should retrieve no specific pattern in the residuals, i.e. the points should scatter around some horizontal line with the same spread pattern, where this line

depends on the variability of the residuals themselves. To see this, we overlaid both a kernel smoother and a standard linear model, which return quite different results: whereas the linear model does not provide evidence that a slope is present in the relationship (Slope = -0.002, CI95%: -0.009, 0.005), the smoother seems to suggest a non-monotonous trend, which suggest that the model is possibly missing some pattern. Additionally, this plot provides further evidence against homoscedasticity of the residuals. About this issue, two points should be addressed: 1) why this is likely to occur, and 2) how could we possibly correct for this. With respect to problem 1), we must note that the major deviations from homoscedasticity occur for low intensity microRNA. That is, if we remove all those probes with an average intensity value across all replicate arrays < 7 , we are likely to remove a large part of possibly poorly hybridized probes which return the patterns seen in figure 5.9. Notably, this results however in a reduced amount of information for the estimation of the model, and thus gives larger weight to outliers, so that a scatterplot of residual vs fitted values is sufficiently homoscedastic, yet both the smoother and the linear model do suggest that scattering does not occur randomly around 0 (see figure 5.9).

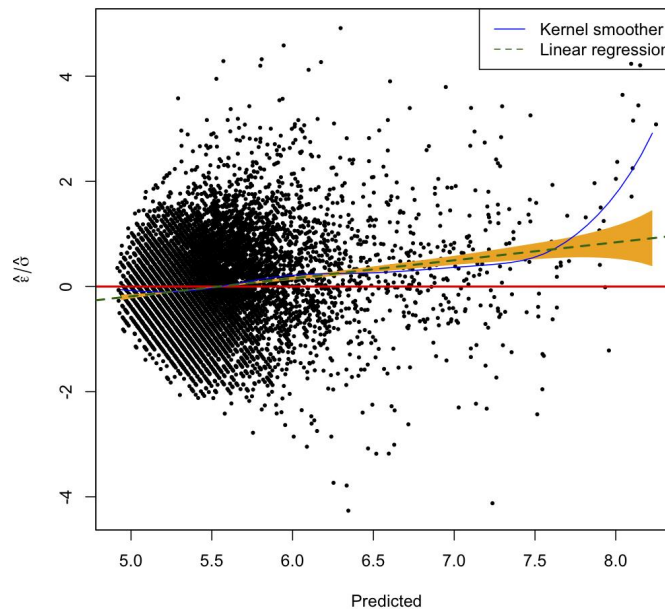


Figure 5.9: $\hat{\varepsilon}/\hat{\sigma}$ versus predicted of random effects model for line A498 on Affymetrix platform (filtered data).

As for problem 2), one possible way would be to apply some proper transformation to the data: natural logarithm, log10, square root, cubic root, variance stabilizing transformation (vst, see [104]) and Rocke-Durbin [127, 128] transformation have all been applied to these data, yet the results have not been much different, as the plots of figure 5.10 suggest. The overlaid linear models return slopes and intercepts which are all significantly different from 0, with the exception of the cubic root transformation, for which the confidence interval on the intercept includes 0 (Intercept = 0.015, CI95%: -0.032, 0.003).

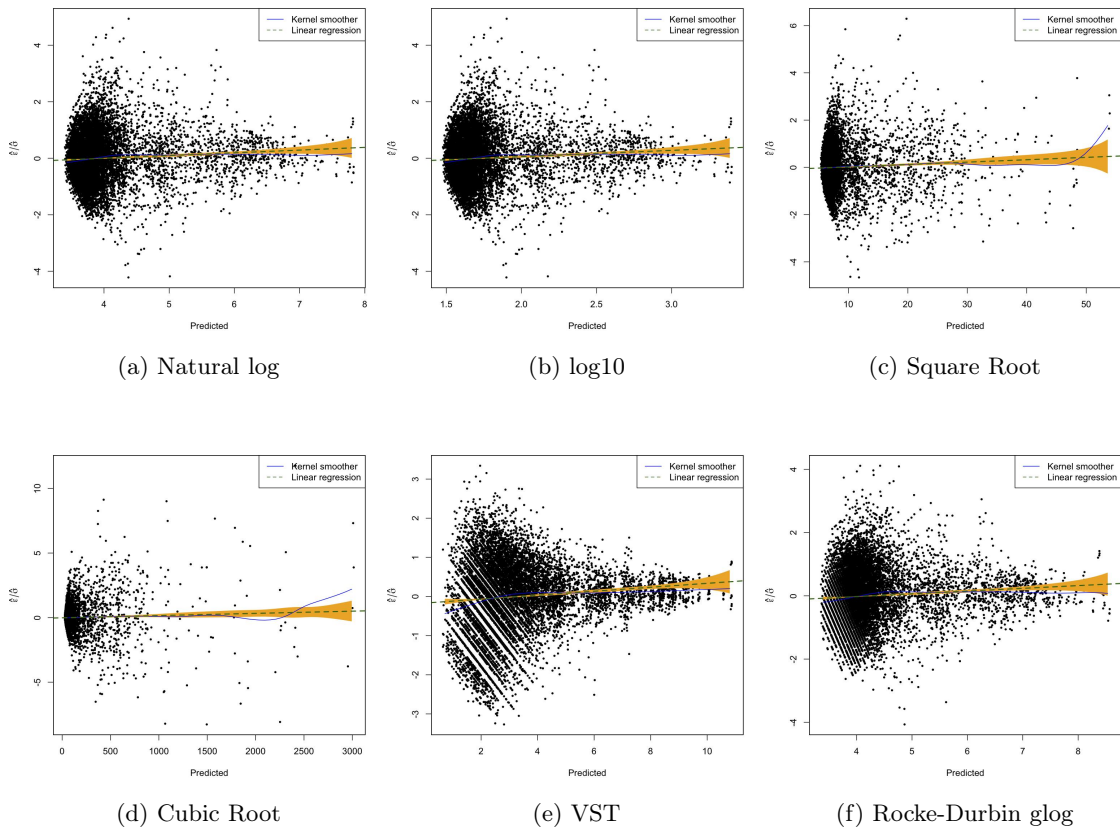


Figure 5.10: $\hat{\sigma}/\sigma$ vs fitted plots for nested-factorial random effects on Affymetrix platform, considering different types of data transformation (line A498.)

According to these plots, the issue of heteroscedasticity is hardly manageable, since 6 different transformations were not able to cope with it, and only by filtering out a large part of the miRNAs could somehow deal with it. However, since issues arose only for poorly expressed miRNAs, and given the considerations following figure 5.9, we feel confident in

that results from the model provide valid information on the variability structure of the experiment.

In table 5.1 we report the output for the nested-factorial ANOVA related to the A498 line. In the last column we have reported the quota of variability explained by each effect included in the model, which in the end can be interpreted as Intraclass Correlation Coefficients (see section 2.3.1). In this perspective, the lower the ICC the higher the "similarity" between units belonging to same grouping unit for a factor. So, the Array factor, which is one of the key effects under evaluation for it tells us how much the platform is repeatable i.e. returns similar values of expression across different experimental runs, has an ICC of 0.0370, less than 4% of total assay variability, thus suggesting a quite satisfactory degree of reliability, whereas the Probes within each miRNA show an even lower degree of variability, with a quota of explained variability $\approx 1\%$. We can see however that the largest effect is the one attributable to the microRNAs, which represent the biological variability of the experiment, which should be maximized with respect to the technical one.

Source	SS	df	MS	σ^2	Quota
miRNA	8734.7610	846	10.3248	0.8492	0.8826
Array	241.2549	2	120.6274	0.0356	0.0370
miRNA*Array	172.0814	1692	0.1017	0.0117	0.0121
Probe in miRNA	221.4426	2541	0.0871	0.0107	0.0111
(Probe in miRNA)*Array	279.8767	5082	0.0551	0.0551	0.0572

Table 5.1: ANOVA table for nested-factorial random effects model on Affymetrix platform (line A498)

Summing up, an ICC-like measure for all the technical effects involved in the study can be defined as

$$\frac{\sigma_A + \sigma_{M*A} + \sigma_{PinM} + \sigma_{(PinM)*A}}{(\sigma_A + \sigma_{M*A} + \sigma_{PinM} + \sigma_{(PinM)*A}) + \sigma_M} = \frac{\sigma_{Tech}}{\sigma_{Tech} + \sigma_{Biol}} \quad (5.1)$$

which returns a value of 0.1174, which suggests that almost 12% of total experimental variability is to be attributed to technical reasons other than just microRNA biological variability. Noteworthy, whereas the miRNA interaction with the array does not seem to account that much, the variability connected to interaction between the probe and the array

is the highest of all technical variability sources, with an explained variability $\approx 6\%$. Such a result can be explained by returning to the line plots of figure 5.3a, which show that whereas the average miRNA intensity value across arrays shows a degree of variability quite low, the single probes actually can vary quite a lot between arrays, mainly following the pattern described in section 5.1.1. These results however refer only to A498 line, a tumor line for which some miRNA regulation (both up- and down-) is to be expected, and for which we have noted that some "batch effect" [129] has occurred. Fitting the same model on sample hREF returns quite similar patterns in terms of both quantile-quantile plots of standardized residuals and scatterplots of residuals vs fitted values (see figure 5.11). Though the cone appears more narrow than that seen for line A498, actually this occurs because of two outliers (highlighted in bigger red points) which widen the plotting range of residuals, and refer to the 4th probe of microRNA hsa-miR-218_st and hsa-miR-139-5p_st, respectively on replicate array 2 and 3.

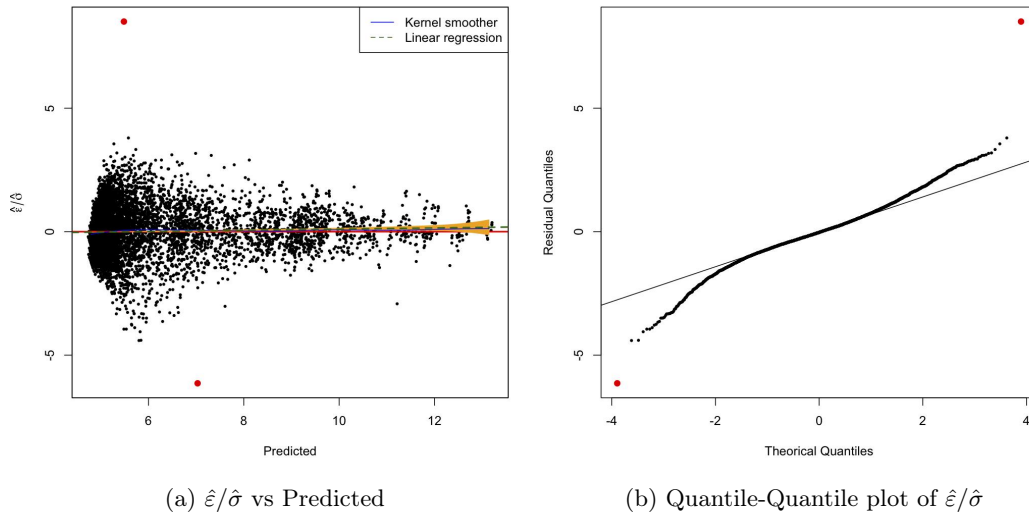


Figure 5.11: *Diagnostic plots for nested-factorial random effects, Affymetrix platform (sample hREF.)*

ANOVA results for the model are reported in table 5.2. Similarly as before, the quotas of variability may be interpreted as ICC, so that the lowest the largest within-group homogeneity. The technical components ICC is equal to $(0.0008 + 0.0093 + 0.0089 +$

$0.0294)/(2.1186 + 0.0008 + 0.0093 + 0.0089 + 0.0294) = 0.0224$, a value which is substantially lower than the 0.1174 seen for line A498. Notably, the array component here has an extremely low variability, accounting for less than 0.05%, suggesting an excellent overall platform repeatability. Similarly, the Probe effect is very low, accounting for $\approx 0.4\%$ of total variability; this has a relevant effect because it allows to perform summarization at the probeset level with a good reliability, an issue which has often been a problem in gene expression microarray data.

Source	SS	df	MS	σ^2	Quota
miRNA	21587.2617	846	25.5169	2.1186	0.9777
Array	5.6216	2	2.8108	0.0008	0.0004
miRNA*Array	112.5699	1692	0.0665	0.0093	0.0043
Probe in miRNA	142.4018	2541	0.0560	0.0089	0.0041
(Probe in miRNA)*Array	149.2798	5082	0.0294	0.0294	0.0136

Table 5.2: ANOVA table for nested-factorial random effects model on Affymetrix platform (line hREF).

Actually, the differences that we can highlight between tables 5.1 and 5.2 are substantial, but we can not conclude anything by simply relying on point estimates of the ICC-like measured introduced before. As described in section 4.1.1 a parametric bootstrap interval was built according to 1000 bootstrap simulations from the estimated model. The bootstrap standard error was estimated as described in [111], and results are reported in table 5.3 for both variance components and ICC-like measures. Notably, the bootstrap confidence intervals for the Array effect for both lines show a negative lower limit, an issue that has been pointed out also by Müller and colleagues [56] with reference to point estimate. The interpretation of such a result is substantially unclear, mainly because since the ICC here should estimate a ratio of variances, its resulting estimators, both point and interval, should be strictly positive.

To better compare results across lines we have plotted point and interval estimates for the ICC for both lines in figure 5.12, where we have splitted the technical and biological components in two sub panel to better highlight between-samples differences as well as to avoid shrinking of technical components plot because of the vary different scale of the y-axis. We can see that the hREF sample shows a substantially better pattern in terms of biological/technical effects, and that only the array component is somehow comparable

Sample	Source	σ^2	CI95% $_{\sigma^2}$	ICC	CI95% $_{ICC}$
A498	miRNA	0.8492	0.7730 - 0.9378	0.8826	0.8234 - 0.9417
	Array	0.0356	0.0096 - 1.4027	0.0370	-0.0268 - 0.1007
	miRNA*Array	0.0117	0.0101 - 0.0137	0.0121	0.0098 - 0.0144
	Probe in miRNA	0.0107	0.0091 - 0.0127	0.0111	0.0089 - 0.0133
	(Probe in miRNA)*Array	0.0551	0.0530 - 0.0573	0.0572	0.0507 - 0.0638
hREF	miRNA	2.1186	1.9294 - 2.3363	0.9777	0.9753 - 0.9801
	Array	0.0008	0.0002 - 0.0305	0.0004	-0.0004 - 0.0011
	miRNA*Array	0.0093	0.0082 - 0.0106	0.0043	0.0036 - 0.0050
	Probe in miRNA	0.0089	0.0790 - 0.0101	0.0041	0.0035 - 0.0048
	(Probe in miRNA)*Array	0.0294	0.0283 - 0.0306	0.0136	0.0122 - 0.0149

Table 5.3: Point and interval estimates (95% confidence level) for variance components and ICC of the different sources of variability for the Affymetrix platform. Intervals for variance components were estimated using the Satterthwaite approximation [108]; intervals for ICC were estimated using the standard t approach in a parametric bootstrap framework with 1000 simulations [111].

with the one from line A498. This, however, is likely to be due to the great uncertainty connected to the array effect for line A498 that we already discussed, but also to the very few degrees of freedom available to estimate this random effects (only 2, see ANOVA tables 5.1 and 5.2).

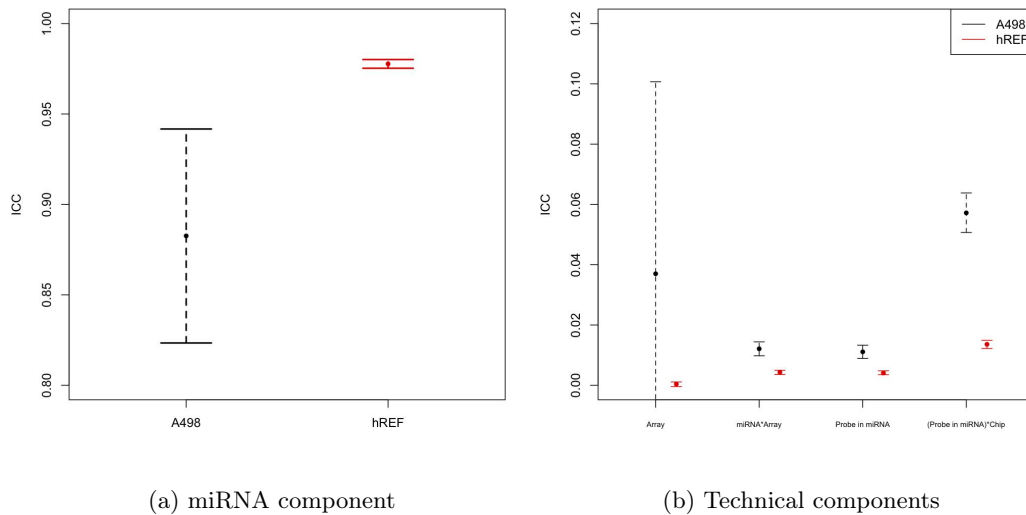


Figure 5.12: Parametric bootstrap confidence intervals for biological and technical effects of model (4.4).

In the end, we could conclude that the platform has a good level of repeatability, in that technical sources of variability generally account for a low amount of total experimental variation, despite some issues occurred for one of the processed samples. Additionally, it's very relevant to note that the replicated probes within the microRNA provide a very reliable amount of information which can then be summarized to obtain a single intensity measure for each microRNA.

These results have been compared to those obtained on filtered data, where filtering was performed as previously described, i.e. by removing all genes with an average expression level across all samples (both A498 and hREF replicates) < 5.5 on log2 scale. This procedure left us with a very unbalanced setting, since the filtering was performed at a probe level rather than at a microRNA level, and with a dataset containing information on 1158 probes (the original data consisted of 3388 probes, thus we have removed $\approx 66\%$ of data). In table 5.4 we report estimates for the variance components σ^2 and the ICC with their confidence intervals; given the unbalance in data replication it was not possible to estimate an ANOVA-like table, so we could not use formulas given in equations (4.9)-(4.14) for the intervals of σ^2 , and this could be estimated only for the ICC.

Sample	Source	σ^2	ICC	CI95% $_{ICC}$
A498	miRNA	1.3308	0.8987	0.8392 - 0.9582
	Array	0.0511	0.0345	-0.0282 - 0.0972
	miRNA*Array	0.0332	0.0224	0.0173 - 0.0275
	Probe in miRNA	0.0093	0.0063	0.0039 - 0.0087
	(Probe in miRNA)*Array	0.0564	0.0381	0.0319 - 0.0443
hREF	miRNA	3.1127	0.9772	0.9711 - 0.9833
	Array	0.0088	0.0028	-0.0022 - 0.0078
	miRNA*Array	0.0171	0.0054	0.0039 - 0.0069
	Probe in miRNA	0.0046	0.0014	0.0007 - 0.0021
	(Probe in miRNA)*Array	0.0422	0.0132	0.0110 - 0.0154

Table 5.4: *Point estimates for variance components and point/interval estimates (95% confidence level) for the ICC of different sources of variability on filtered data (1158 probes), Affymetrix platform. Confidence intervals for ICC were estimated using the standard t approach in a parametric bootstrap framework with 1000 simulations [111].*

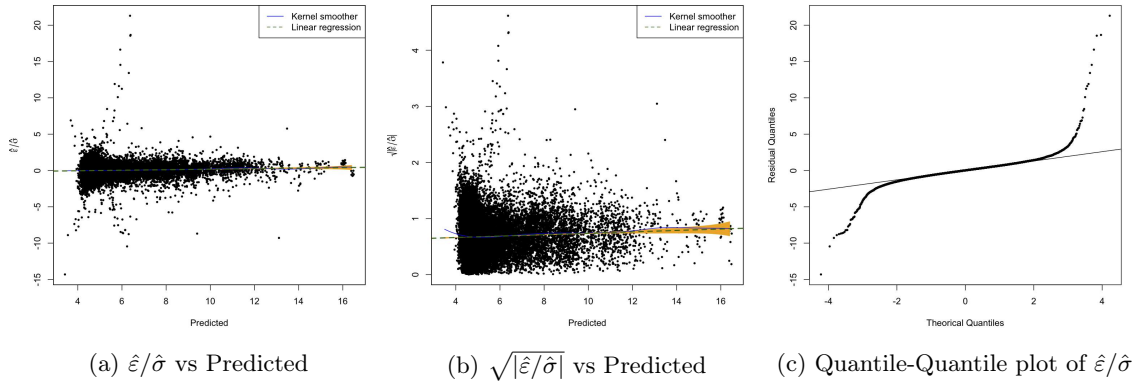
Filtering has a similar effect on the two samples: on A498 line it causes a slight increase in the variability explained by the biological components, coupled with a moderate decrease

in some technical components like the interaction between Probes and Array (ICC from 0.0572, CI95%: 0.0507 - 0.0638 to 0.0381, CI95%: 0.0319 - 0.0443), whereas for hREF the variability related to replicated Probes after filtering moderately decreases in terms of ICC (from 0.0041, CI95%: 0.0035 - 0.0048 to 0.0014, CI95%: 0.0007 - 0.0021). Thus, filtering seems to only moderately affect our results, thus suggesting a platform performance substantially good, irrespective of this pre-processing step.

5.2.1.2 Agilent

As for the Affymetrix arrays, we checked the model fit for departures from normality and heteroscedasticity of standardized conditional residuals by means of residuals vs fitted plots, square root of absolute residuals vs fitted plots and Normal quantile-quantile plot. The residuals vs fitted plots (figures 5.13a-5.13b) show a pattern which is quite similar to the one seen for Affymetrix in figure 5.8, yet we can note some outliers which show both largely positive and largely negative residuals that do induce a reduction in the scattering pattern around 0. Notably, the linear model fit and the smoother are almost completely overlapping in figure 5.13a, and the orange confidence band is extremely narrow, suggesting a good fit precision. If we examine the linear fit we find that both the intercept (-0.187, CI95%: -0.22, -0.155) and the slope (0.036, CI95%: 0.03, 0.042) suggest that a mild increasing pattern exists, yet since this model was fitted on an extremely large number of records, even a small difference is likely to return a significant difference from the null hypothesis being tested, even when this difference is practically irrelevant. The quantile-quantile plot suggests that on the extreme boundaries of measurements range some issues occur that lead to a deviation from the normality assumption. As an additional check, we have evaluated quantile-quantile plots using the Student's t distribution with degrees of freedom ranging from 1 to 10, but since no considerable results were obtained the plots are not shown here.

Results from the random effects models are reported in the ANOVA table 5.5. If we compare them with those obtained on the same sample with Affymetrix arrays (see table 5.1) we can see that there is good comparability in terms of biological variability explained (88.26% for Affymetrix vs 89.08% for Agilent), but that the "splitting" of the technical components is quite different, with a large decrease of the Array effect (from 3.70% to 0.42%) and a relevant increase in the Probe effect (from 1.11% to 7.85%). The lower Array

Figure 5.13: *Diagnostic plots for nested-factorial random effects, Agilent platform (line A498.)*

effect is easily explained by the consideration that on Affymetrix platform a relevant batch effect has possibly led to an increase that would not have occurred otherwise, whereas the increase in the Probe variability is related to the different level of probe sequences replication that we have discussed in section 3.3.2, since the 16 intensity values for a microRNA on Agilent are not generated by the same targeting nucleotide sequence, opposite to what happens with the 4 Affymetrix probes.

Source	SS	df	MS	σ^2	Quota
miRNA	61281.6309	850	72.0960	1.4880	0.8908
Array	192.0608	2	96.0304	0.0070	0.0042
miRNA*Array	475.9860	1700	0.2800	0.0157	0.0094
Probe in miRNA	5384.4194	12765	0.4218	0.1311	0.0785
(Probe in miRNA)*Array	728.7267	25530	0.0285	0.0285	0.0171

Table 5.5: *ANOVA table for nested-factorial random effects model on Agilent platform (line A498).*

We have fitted the same random-effects model also to data coming from hREF sample. Diagnostic checks return patterns which are substantially identical to those seen for line A498, even slightly better in terms of residuals homoscedasticity and thus are not reported here, for the sake of brevity. The ANOVA table of this model (see table 5.6) provides results which are very close to those obtained for line A498, with only some minor changes. For instance, the Array-related variability is halved (from 0.42% to 0.21%), but the absolute

values appear quite low as to state that actual replicate-related variability accounts for a minimal part of total assay variation. Similarly, the miRNA*Array interaction decreases with a three-fold pattern, yet the quota of variability connected to it still it's under 1%, suggesting that its contribution may be considered irrelevant.

Source	SS	df	MS	σ^2	Quota
miRNA	118364.1165	850	139.2519	2.8802	0.9031
Array	186.2643	2	93.1322	0.0068	0.0021
miRNA*Array	310.1646	1700	0.1824	0.0103	0.0032
Probe in miRNA	10719.0828	12765	0.8397	0.2738	0.0858
(Probe in miRNA)*Array	467.6406	25530	0.0183	0.0183	0.0057

Table 5.6: ANOVA table for nested-factorial random effects model on Agilent platform (line hREF).

The comparison of variance components σ^2 and ICC (i.e. the quotas of variability) between the two lines was thus performed via confidence intervals (see table 5.7). All the effects appear comparable between lines, with the only exception of the higher-level interaction and of the miRNA*Array interaction which are lower for the hREF sample. In this sense, the platform shows a satisfactory level of repeatability, yet somehow lower than that seen for Affymetrix if we consider sample hREF (A498 line is somehow affected by a bias so that estimated ICC are likely to under-estimate actual reliability levels).

Sample	Source	σ^2	CI95% σ^2	ICC	CI95% ICC
A498	miRNA	1.4880	1.3552 - 1.6422	0.8908	0.8789 - 0.9028
	Array	0.0070	0.0019 - 0.2761	0.0042	-0.0037 - 0.0122
	miRNA*Array	0.0157	0.0146 - 0.0170	0.0094	0.0083 - 0.0105
	Probe in miRNA	0.1311	0.1277 - 0.1346	0.0785	0.0713 - 0.0857
	(Probe in miRNA)*Array	0.0285	0.0281 - 0.0290	0.0171	0.0156 - 0.0186
hREF	miRNA	2.8802	2.6225 - 3.1763	0.9031	0.8939 - 0.9122
	Array	0.0068	0.0018 - 0.2686	0.0021	-0.0022 - 0.0065
	miRNA*Array	0.0103	0.0095 - 0.0111	0.0032	0.0028 - 0.0036
	Probe in miRNA	0.2738	0.2671 - 0.2808	0.0858	0.0783 - 0.0934
	(Probe in miRNA)*Array	0.0183	0.0180 - 0.0186	0.0057	0.0052 - 0.0063

Table 5.7: Point and interval estimates (95% confidence level) for variance components and ICC of the different sources of variability on the Agilent platform. Intervals for variance components were estimated using the Satterthwaite approximation [108]; intervals for ICC were estimated using the standard t approach in a parametric bootstrap framework with 1000 simulations [111].

Since normalization on the `ProcessedSignal` is not feasible, we only evaluated effect of filtering on the results, similarly to what has been done for Affymetrix arrays. We filtered out all miRNAs that showed an overall mean log2-intensity of 5, which resulted in excluding 7947 probes, i.e. $\approx 58\%$ of the total amount of probes. The choice of this threshold was substantially subjective, and was made to have results as much as possible comparable to those obtained with Affymetrix platform, which present a range of measurements quite similar to those seen for Agilent arrays. Estimates of σ^2 and ICC are reported in table 5.8.

Sample	Source	σ^2	ICC	CI95% _{ICC}
A498	miRNA	2.1851	0.8443	0.8184 - 0.8702
	Array	0.0270	0.0104	-0.0106 - 0.0314
	miRNA*Array	0.0571	0.0221	0.0188 - 0.0254
	Probe in miRNA	0.2923	0.1129	0.0989 - 0.1269
	(Probe in miRNA)*Array	0.0265	0.0102	0.0090 - 0.0114
hREF	miRNA	3.2959	0.8298	0.8059 - 0.8537
	Array	0.0316	0.0080	-0.0079 - 0.0239
	miRNA*Array	0.0393	0.0099	0.0085 - 0.0113
	Probe in miRNA	0.5868	0.1477	0.1300 - 0.1654
	(Probe in miRNA)*Array	0.0184	0.0046	0.0041 - 0.0051

Table 5.8: *Point estimates for variance components and point/interval estimates (95% confidence level) for the ICC of different sources of variability on filtered data (5669 probes), Agilent platform. Confidence intervals for ICC were estimated using the standard t approach in a parametric bootstrap framework with 1000 simulations [111].*

The decrease in platform performance after filtering is evident mainly for hREF sample, where Probe-related variability is almost twice what was found on unfiltered data (ICC from 0.0858 to 0.1477). Actually, this decrease in technical-related ICCs is possibly due to the removal of highly variable miRNAs, so that the biological effect has been largely reduced, yet should warn on the effect that filtering could have on interpretation of results. Overall, platform performance appears somehow lower than what seen for Affymetrix, though not unsatisfactory, and this could be due to the different strategy in the design of the replicated probes.

5.2.1.3 Illumina

As described previously, the model fitted for the Illumina arrays was simpler because of the reduced amount of information available. That is, since no bead-level data is provided, we have no way to evaluate accuracy of replicated probes for this particular platform. Diagnostic checks provide quite a different picture of model adequacy, since the scattering around the 0 line appears much more uniform than in figures 5.8a and 5.13a, with a panel of miRNAs showing moderately large (small) residuals, that however, differently from other platforms, are located in the center of the range of measurements. These molecules however do not seem to relevantly affect the overall model fit, since the smoother and the linear fit are well overlayed on each other, at least until a log2-expression value of ≈ 12 (which correspond to a value on the original scale $= 2^{12} = 4096$). Thus, in this case larger intensities are associated with higher uncertainty, represented also by the widening of the confidence band around the linear fit. Whereas homoscedasticity appears to be substantially satisfied, this is certainly less true for their normality, with a relevant departure from the 45° line of the quantile-quantile normal plots of figure 5.14a. The pattern we can see appears somehow half-way between Affymetrix (figure 5.8c) and Agilent (figure 5.13c), possibly suggesting that on the extreme boundaries of measurements range the assumption is far from being verified. Notably, this happens for very low and very large intensity values, i.e. values for which issues in pre-processing phase are known: intensities below the background signal for poorly expressed microRNAs and saturation of signal for highly expressed ones. Actually, an empirical evaluation of other possible distributions showed that a Normal distribution with $\mu = -0.2$ and $\sigma = 0.5$ fits quite well the histogram of density of expression values, indicating that it is reasonable to be confident in the adequacy of the fitted model.

The ANOVA table for this model (see table 5.9) indicates a preeminent role of the biological component in determining experimental variability, accounting for almost 99% of total assay variation. Yet, this result can not be directly compared with what has been found for Affymetrix and Agilent, because of the different data structure. Results, however, suggest a very good overall level of repeatability of the platform, with a variability for both technical components here considered falling below 1%. All these considerations concerning line A498 are valid also for sample hREF, for which adequacy check of the model

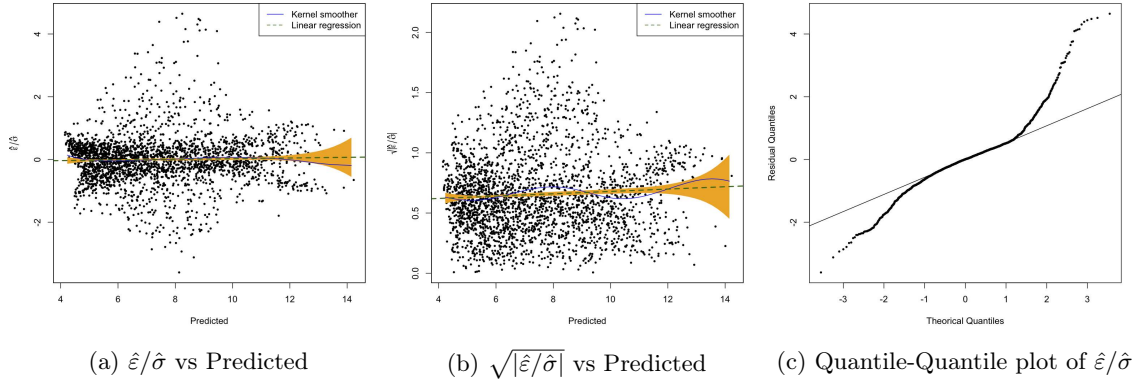


Figure 5.14: Diagnostic plots for nested-factorial random effects, Illumina platform (line A498.)

returns results which are slightly better than those obtained for A498 both in terms of heteroscedasticity and normality of residuals. ANOVA results for this samples are reported in table 5.10.

Source	SS	df	MS	σ^2	Quota
miRNA	12038.2024	857	14.0469	4.6751	0.9892
Array	50.8236	2	25.4118	0.0296	0.0063
miRNA*Array	37.1661	1714	0.0217	0.0217	0.0046

Table 5.9: ANOVA table for full factorial random effects model on Illumina platform (line A498).

Source	SS	df	MS	σ^2	Quota
miRNA	15275.3471	857	17.8242	5.9385	0.9943
Array	43.3812	2	21.6906	0.0253	0.0042
miRNA*Array	15.1326	1714	0.0088	0.0088	0.0015

Table 5.10: ANOVA table for full factorial random effects model on Illumina platform (sample hREF).

The 95%CI for both variance components σ^2 and ICCs are reported in table 5.11. Additionally to the issue of the lower limits falling below 0 for the Array component, we also face the problem of the upper limits falling above 1 for the miRNA component. In our case, this occurs for components whose point estimates are quite close to 1, so that

interpretation of results is sufficiently easy to make, yet it may not always be this simple. In terms of comparison between samples, we can conclude that the platform shows a very good reliability level, yet this conclusion can not be directly compared with Affymetrix and Agilent arrays processed here, because of the different type of data available (i.e. probe-level replication).

Sample	Source	σ^2	CI95% $_{\sigma^2}$	ICC	CI95% $_{ICC}$
A498	miRNA	4.6751	4.2616 - 5.1521	0.9892	0.9825 - 0.9958
	Array	0.0296	0.0270 - 0.0326	0.0063	-0.0004 - 0.0129
	miRNA*Array	0.0217	0.0203 - 0.0232	0.0046	0.0043 - 0.0049
hREF	miRNA	5.9385	5.4137 - 6.5437	0.9943	0.9881 - 1.0005
	Array	0.0253	0.0230 - 0.0278	0.0042	-0.0020 - 0.0104
	miRNA*Array	0.0088	0.0083 - 0.0095	0.0015	0.0013 - 0.0016

Table 5.11: *Point and interval estimates (95% confidence level) for variance components and ICC of the different sources of variability on the Illumina platform. Intervals for variance components were estimated using the Satterthwaite approximation [108]; intervals for ICC were estimated using the standard t approach in a parametric bootstrap framework with 1000 simulations [111].*

We compared these results with those obtained after filtering and normalizing data according to quantile and loess algorithms described in section 3.4.2. For filtering, we used the `detectionCall` method, by removing all the miRNAs with a detection p-value above 0.01 for more than 2 samples. In a nutshell, the detection p-value is a measure of how much it is likely that a gene is found to be expressed above the background by chance, so that the lower the p-value, the higher the probability that the gene is actually expressed. This resulted in keeping only 669 human microRNAs, i.e. $\approx 78\%$ of total human miRNAs on the array, and returned point estimates and confidence intervals for the ICC detailed in table 5.12. Also in this simulations we can see that some confidence intervals lower limits lie below 0, yet results appear substantially similar to those obtained without performing any filtering or normalization,

To evaluate separately the role of data filtering and normalization we have evaluated the ICC alongside their confidence intervals also on data that were in turn only normalized (both loess and quantile) or only filtered (more than 2 samples with detection p-value > 0.1). Results in tables 5.13 and 5.14 suggest however that in our study they have a very moderate, and somehow similar, impact on final results in terms of platform repeatability.

Sample	Source	Quantile		Loess	
		ICC	CI95% _{ICC}	ICC	CI95% _{ICC}
A498	miRNA	0.9864	0.9715 - 1.0013	0.9864	0.9718 - 1.0010
	Array	0.0078	-0.0071 - 0.0227	0.0078	-0.0068 - 0.0224
	miRNA*Array	0.0059	0.0051 - 0.0067	0.0059	0.0051 - 0.0067
hREF	miRNA	0.9918	0.9807 - 1.0029	0.9918	0.9316 - 1.0520
	Array	0.0059	-0.0052 - 0.0170	0.0059	-0.0545 - 0.0663
	miRNA*Array	0.0024	0.0021 - 0.0027	0.0024	0.0021 - 0.0027

Table 5.12: *Point estimates and interval estimates (95% confidence level) for the ICC of different sources of variability on filtered and normalized data (669 probes), Illumina platform. Confidence intervals for ICC were estimated using the standard t approach in a parametric bootstrap framework with 1000 simulations [111].*

Sample	Source	Quantile		Loess	
		ICC	CI95% _{ICC}	ICC	CI95% _{ICC}
A498	miRNA	0.996	0.9955 - 0.9965	0.996	0.9955 - 0.9965
	Array	0.000	0.0000 - 0.0000	0.0001	-0.0001 - 0.003
	miRNA*Array	0.004	0.0035 - 0.0045	0.0039	0.0034 - 0.0044
hREF	miRNA	0.9987	0.9985 - 0.9989	0.9986	0.9984 - 0.9988
	Array	0.0000	0.0000 - 0.0000	0.0000	0.0000 - 0.0000
	miRNA*Array	0.0013	0.0012 - 0.0014	0.0014	0.0012 - 0.0016

Table 5.13: *Point estimates and interval estimates (95% confidence level) for the ICC of different sources of variability on normalized data, Illumina platform. Confidence intervals for ICC were estimated using the standard t approach in a parametric bootstrap framework with 1000 simulations [111].*

In figure 5.15 we have plotted, for both samples, the point and interval estimates for the ICC described in tables 5.11-5.14, separately for each factor included in the random effects model. It can be noted the decrease in precision for all the ICC of line A498 when a combination of filtering and normalization is performed, and how much does the normalization step reduce the array effect by almost completely removing it for both samples. Notably, filtering and normalization have a different impact on the two samples when both are used: for A498, this is associated with a substantial loss in precision for the ICC, whereas for hREF it is associated with a substantially unchanged precision, with the exception of some strange effect for miRNA effect. Thus, all considered, these two steps do not seem to

Sample	Source	ICC	CI95% _{ICC}
A498	miRNA	0.9955	0.9925 - 0.9985
	Array	0.0015	-0.0015 - 0.0045
	miRNA*Array	0.0030	0.0026 - 0.0034
hREF	miRNA	0.9845	0.9756 - 0.9934
	Array	0.0042	-0.0045 - 0.0129
	miRNA*Array	0.0114	0.0099 - 0.0129

Table 5.14: Point estimates and interval estimates (95% confidence level) for the ICC of different sources of variability on filtered data (669 probes), Illumina platform. Confidence intervals for ICC were estimated using the standard t approach in a parametric bootstrap framework with 1000 simulations [111].

be really decisive for the purposes of studying platform reliability, at least in the present experiment.

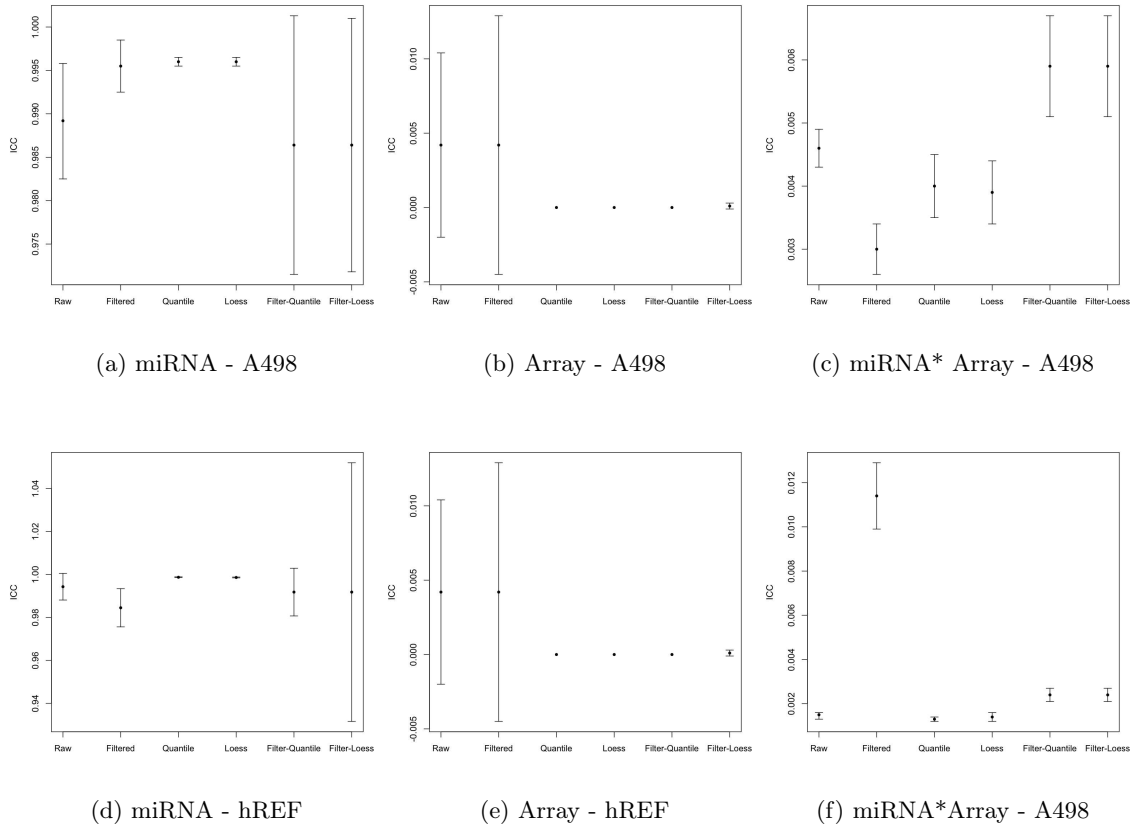


Figure 5.15: ICC and CI95% for all combinations of data filtering and normalization.

To conclude, all the platforms analyzed are characterized by good degree of repeatability, and show a discrete robustness to some pre-processing steps in terms of Intraclass Correlation Coefficients. Illumina, though being the one with the best biological ICC, can not be directly compared to the previous ones because of the different data structure available. Actually, filtering and normalization only slightly affect results in terms of variability explained by each factor, yet have a more evident role in determining estimators precision. In this sense, their impact is seen to be different across samples and platforms, and should thus be carefully evaluated before proceeding to down-stream analysis.

5.2.2 CCC and Overall CCC

5.2.2.1 Affymetrix

The CCC considered for the Affymetrix platform is the RM-CCC described in section 4.1.2.2, equation (4.30). Results for the Affymetrix platform are reported in table 5.15. From these results we can once again note the relevant issue which occurs for line A498, which shows patterns of substantial lack of repeatability: if we consider the threshold 0.96 defined in equation (4.29), section 4.1.2.1, than we would say that the platform can not be considered reliable for line A498. On the other hand, for sample hREF the platform has a good level of repeatability, indicating the actual level of array reliability which in the case of line A498 is substantially underestimated.

Sample	Pair	$\rho_{c,rm}$	CI95% ^a	CI95% ^b	OSCC	CI95% ^b
A498	1 vs 2	0.9021	0.8793 - 0.9207	0.8772 - 0.9242		
	1 vs 3	0.8596	0.8298 - 0.8845	0.8291 - 0.8893	0.8936	0.8703 - 0.9160
	2 vs 3	0.9213	0.9040 - 0.9356	0.9032 - 0.9378		
hREF	1 vs 2	0.9777	0.9733 - 0.9814 [†]	0.973 - 0.981 [†]		
	1 vs 3	0.9829	0.9794 - 0.9858 [†]	0.979 - 0.986 [†]	0.9818	0.977 - 0.985 [†]
	2 vs 3	0.9848	0.9815 - 0.9875 [†]	0.981 - 0.988 [†]		

[†]: the CI95% does not include 0.96.

^a: CI95% estimated using the hyperbolic tangent transformation.

^b: CI95% estimated using 1000 bootstrap samples.

Table 5.15: *Repeated Measures CCC and Overall CCC for Affymetrix technical replicates.*

Additionally, we have also evaluated performance of the platform after averaging across

replicated probes to have a single intensity measure for each microRNA. Results are presented in table 5.16, and appear very different between the two samples considered. In particular, according to the threshold of 0.96 defined in section 4.1.2.1, the line A498 shows patterns of poor repeatability, whereas sample hREF appears highly reliable, with both pairwise CCC and Overall CCC to be around 0.99. Since we are aware of the issues occurred on some technical replicates for A498 (see section 5.1.1), it is likely that the level of repeatability observed for this sample is substantially underestimated, and should not be considered.

Sample	Pair	ρ^a	v^a	u^a	c_b^a	ρ_c	CI95% ^b	CI95% ^c	OCCC	CI95% ^c
A498	1 vs 2	0.973	1.001	0.27	0.965	0.939	0.931 - 0.945	0.921 - 0.952		
	1 vs 3	0.965	1.047	0.400	0.925	0.893	0.881 - 0.904	0.868 - 0.914	0.930	0.911 - 0.944
	2 vs 3	0.972	1.046	0.137	0.99	0.962	0.957 - 0.967	0.952 - 0.971		
hREF	1 vs 2	0.992	0.918	-0.034	0.996	0.988	0.987 - 0.989 [†]	0.985 - 0.99 [†]		
	1 vs 3	0.994	0.957	0.001	0.999	0.993	0.992 - 0.994 [†]	0.992 - 0.994 [†]	0.992	0.99 - 0.993 [†]
	2 vs 3	0.996	1.042	0.036	0.999	0.994	0.993 - 0.995 [†]	0.993 - 0.995 [†]		

[†]: the CI95% does not include 0.96

^a: ρ is the Pearson correlation coefficient, v , u and c_b are the scale shift, the location shift and the accuracy component of ρ_c .

^b: CI95% estimated using the hyperbolic tangent transformation.

^c: CI95% estimated using 1000 bootstrap samples.

Table 5.16: *Lin's CCC and Overall CCC for Affymetrix technical replicates - intensity measures averaged across replicated probes of the same microRNA.*

To assess an eventual effect of normalization on increasing reliability patterns on line A498 replicates, we evaluated both the pairwise CCC and the Overall CCC for both lines using quantile and loess normalized data, without performing any filtering (see table 5.17). As expected, there's a clear effect of normalization, for line A498, since the threshold of 0.96 is now exceeded by almost all pairs, with the exception of the comparison 1 vs 2 for A498. Yet, the CI95% for this CCC estimated through the normal approximation described in section 4.1.2.1 lies above 0.96 (0.962 - 0.971).

In figure 5.16 we have plotted the Overall CCCs, with their Confidence intervals, which refer to data averaged across replicated probes, to have a compact view of what's going on here. It's interesting to note that loess normalization not only returns the highest Overall CCC for A498, but that the related bootstrap confidence interval is the most narrow of the three, whereas on line hREF it is the largest one. Certainly, in this particular case we would want to normalize between arrays to remove the bias which has heavily affected our analysis

Sample	Pair	Quantile				Loess			
		ρ_c	CI95% ^a	OCCC	CI95% ^a	ρ_c	CI95% ^a	OCCC	CI95% ^a
A498	1 vs 2	0.972	0.963 - 0.979 †			0.983	0.978 - 0.988 †		
	1 vs 3	0.967	0.958 - 0.974	0.971	0.963 - 0.977 †	0.978	0.971 - 0.983 †	0.981	0.976 - 0.985 †
	2 vs 3	0.973	0.966 - 0.980 †			0.981	0.977 - 0.986 †		
hREF	1 vs 2	0.992	0.991 - 0.994 †			0.989	0.987 - 0.991 †		
	1 vs 3	0.994	0.993 - 0.995 †	0.994	0.992 - 0.995 †	0.991	0.989 - 0.993 †	0.992	0.989 - 0.993 †
	2 vs 3	0.995	0.993 - 0.996 †			0.994	0.992 - 0.995 †		

†: the CI95% does not include 0.96.

^a: CI95% estimated using 1000 bootstrap samples.

Table 5.17: *Lin's CCC and Overall CCC for Affymetrix normalized technical replicates - intensity measures averaged across replicated probes of the same microRNA.*

so far, yet this would expose us to the risk of over-interpreting the results by ascribing to the platform a performance level which is actually mostly attributable to the normalization method used. Since hREF sample for Affymetrix does not seem to be affected by this bias, we can retrieve the actual performance of the platform by evaluating its behavior for this particular sample, using A498 as a sort of negative control.

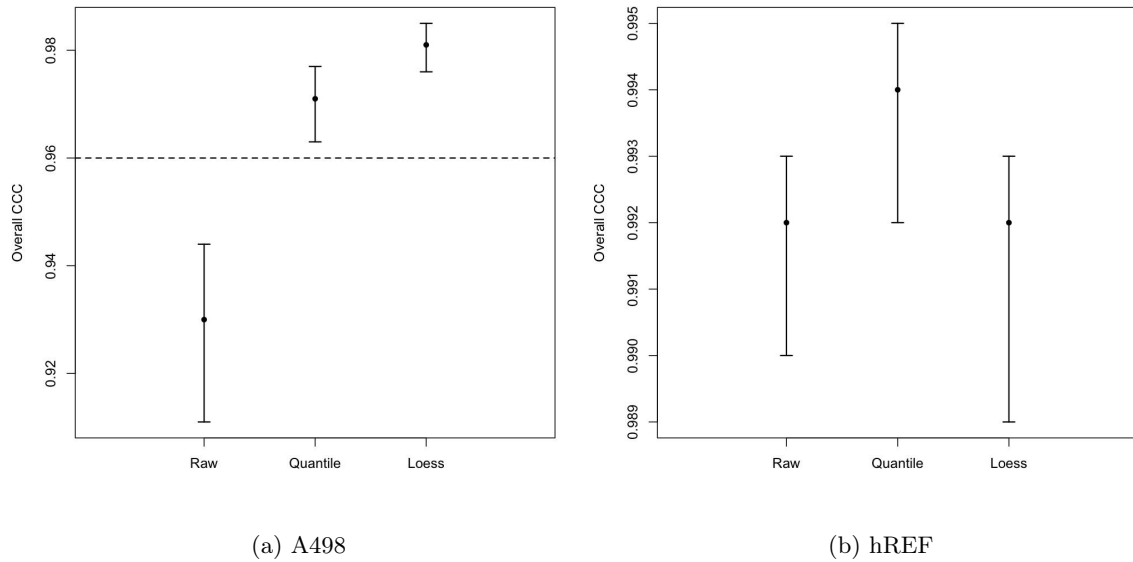


Figure 5.16: *Overall CCC and 95% bootstrap confidence intervals for raw and normalized data using quantile and loess algorithms for Affymetrix data. The dashed line in left panel represents the threshold under which two platforms are no more considered reliable.*

In the light of these considerations, the Affymetrix platform is confirmed to have a good degree of repeatability both at a probe-level (table 5.15) and at a probeset level (table 5.16), suggesting that both replicated probes and replicated arrays have a large degree of consistency.

5.2.2.2 Agilent

Also in this case we have estimated the $\rho_{c,rm}$ to take into account the different probe replicates available for each microRNA. Yet the computational burden was sufficiently heavy and required a powerful hardware to estimate the bootstrap confidence intervals for both the CCC and the overall CCC. Notably, on a standard Macbook (2.4 GHZ Intel Core 2 Duo processor and 4 GB memory) it took more than 3 minutes to estimate 3 pairwise CCC for repeated measures, even because of the extremely high-dimensionality of the data ($851 \text{ miRNA} \times 16 \text{ probes} = 13616 \text{ records}$). Thus estimating 1000 bootstrap samples and related BC_a confidence intervals, which requires also a jackknife estimation to derive the acceleration factor, is computationally extremely intensive, so we relied only on the Normal approximation of the pair-wise CCC to evaluate patterns of repeatability (see table 5.18). Notably, all the pairs of technical replicates appear to have a good repeatability level, which is substantially comparable across samples, yet hREF seems to perform slightly better, as the random effects model analysis had already pointed out. Actually, since no confidence interval has been built around the Overall CCC, we can not exclude that it may contain the threshold value of 0.96, suggestive of poor within-platform repeatability at least for line A498, so that no final conclusion can be drawn at this level.

Sample	Pair	$\rho_{c,rm}$	CI95% ^b	OSCC
A498	1 vs 2	0.9620	0.9520 - 0.9700	0.9693
	1 vs 3	0.9765	0.9697 - 0.9819 [†]	
	2 vs 3	0.9697	0.9617 - 0.9761 [†]	
hREF	1 vs 2	0.9937	0.9926 - 0.9945 [†]	0.9889
	1 vs 3	0.9832	0.9807 - 0.9854 [†]	
	2 vs 3	0.99	0.9884 - 0.9914 [†]	

[†]: the CI95% does not include 0.96

^a: CI95% estimated using the hyperbolic tangent transformation.

Table 5.18: *King's CCC and Overall CCC for Agilent technical replicates.*

Also for the Agilent platforms we have estimated the standard Lin's CCC, yet in this case we have not simply averaged across the different `ProcessedSignal` values but we have extracted the `TotalGeneSignal` estimated by the Feature Extraction software. Results in table 5.19 suggest that at a probeset level the platform has a very good degree of repeatability, which is confirmed for all pairwise comparison of technical replicates and for both samples (all CI95% do not include the threshold value 0.96). Also, it is confirmed that sample hREF shows a better performance with respect to line A498, with a minimum CCC estimate of 0.992 vs a minimum of 0.975 for A498.

Sample	Pair	ρ^a	v^a	u^a	c_b^a	ρ_c	CI95% ^b	CI95% ^c	OSCC	CI95% ^c
A498	1 vs 2	0.988	1.166	0.054	0.987	0.975	0.973 - 0.978 †	0.971 - 0.980 †		
	1 vs 3	0.989	1.158	-0.137	0.980	0.970	0.966 - 0.972 †	0.963 - 0.977 †	0.975	0.969 - 0.979 †
	2 vs 3	0.994	0.993	-0.178	0.984	0.978	0.975 - 0.980 †	0.973 - 0.982 †		
hREF	1 vs 2	0.999	1.045	0.030	0.999	0.997	0.997 - 0.998 †	0.997 - 0.998 †		
	1 vs 3	0.997	1.087	0.097	0.992	0.989	0.988 - 0.990 †	0.988 - 0.991 †	0.994	0.993 - 0.995 †
	2 vs 3	0.998	1.040	0.066	0.997	0.995	0.995 - 0.996 †	0.995 - 0.996 †		

†:the CI95% does not include 0.96.

^a: ρ is the Pearson correlation coefficient, v , u and c_b are the scale shift, the location shift and the accuracy component of ρ_c .

^b: CI95% estimated using the hyperbolic tangent transformation.

^c: CI95% estimated using 1000 bootstrap samples.

Table 5.19: *Lin's CCC and Overall CCC for Agilent technical replicates - intensity measures averaged across replicated probes of the same microRNA.*

The eventual effect of normalization was evaluated also in this case using quantile and loess normalized data (results in table 5.20). There seems to be a slight improvement in the performance of Agilent platform, in particular for line A498, whereas sample hREF good level of repeatability is substantially confirmed, yet all the CI95% OSCC exceed the fixed threshold of 0.96, so that the platform can be considered reliable, irrespective of the normalization method considered. The plots of figure 5.17 highlight the increases in performance one can obtain by normalizing data, which however are not much relevant in terms of the absolute values of the OSCC, but provide useful information on the gain in the precision of confidence intervals when these are estimated on normalized data.

In comparison with Affymetrix, Agilent appears slightly more reliable on both samples. Additionally, the two normalization methods here considered appear to be much more consistent with one another on Agilent, rather than for Affymetrix, with a higher precision (more narrow confidence intervals) on both samples. This could be possibly due

Sample	Pair	Quantile				Loess			
		ρ_c	CI95% ^a	OCCC	CI95% ^a	ρ_c	CI95% ^a	OCCC	CI95% ^a
A498	1 vs 2	0.9911	0.9891 - 0.9926 †			0.9902	0.9874 - 0.9922 †		
	1 vs 3	0.9932	0.9916 - 0.9944 †	0.9929	0.9913 - 0.9941 †	0.9923	0.9902 - 0.9940 †	0.9922	0.9904 - 0.9937 †
	2 vs 3	0.9944	0.9930 - 0.9955 †			0.9941	0.9926 - 0.9954 †		
hREF	1 vs 2	0.9982	0.9977 - 0.9985 †			0.9982	0.9979 - 0.9985 †		
	1 vs 3	0.9976	0.9971 - 0.9980 †	0.9978	0.9973 - 0.9982 †	0.9968	0.9963 - 0.9973 †	0.9976	0.9972 - 0.9979 †
	2 vs 3	0.9976	0.9970 - 0.9981 †			0.9977	0.9973 - 0.9981 †		

†: the CI95% does not include 0.96.

^a: CI95% estimated using 1000 bootstrap samples.

Table 5.20: *Lin's CCC and Overall CCC for Agilent normalized technical replicates - intensity measures averaged across replicated probes of the same microRNA.*

that Affymetrix replicated probe intensities were simply averaged within each probeset to obtain a single intensity measure, whereas the algorithm which transform the replicated `ProcessedSignal` into the final `TotalGeneSignal` is rather complex and provides some adjustments for outliers and possible confounding factors which make the TGS a very robust estimate of the levels of expression.

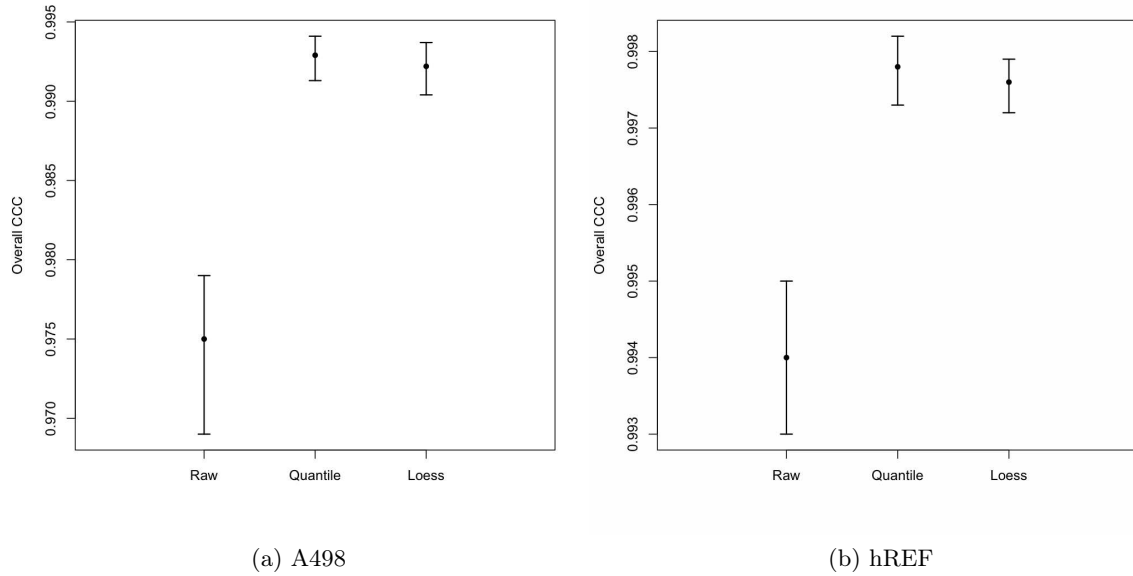


Figure 5.17: *Overall CCC and 95% bootstrap confidence intervals for raw and normalized data using quantile and loess algorithms for Agilent data.*

5.2.2.3 Illumina

Because of the different data structure, i.e. no data on probe replicates available, for Illumina arrays we could only estimate Lin's CCC and the Overall CCC from Barnhart *et al.*. All the pairwise CCCs and overall CCC confidence intervals do not include the threshold value 0.96, indicating that the platform has good patterns of repeatability, for both samples considered. Additionally, it is confirmed that sample hREF has a slightly better performance with respect to A498.

Sample	Pair	ρ^a	v^a	u^a	c_b^a	ρ_c	CI95% ^b	CI95% ^c	OCCC	CI95% ^c
A498	1 vs 2	0.998	0.993	0.055	0.998	0.996	0.996 - 0.997 [†]	0.9954 - 0.9970 [†]		
	1 vs 3	0.994	0.995	-0.156	0.988	0.982	0.980 - 0.984 [†]	0.9787 - 0.9843 [†]	0.989	0.9873 - 0.9907 [†]
	2 vs 3	0.995	1.002	0.102	0.995	0.990	0.988 - 0.991 [†]	0.9871 - 0.9912 [†]		
hREF	1 vs 2	0.999	0.987	0.124	0.992	0.991	0.990 - 0.992 [†]	0.9900 - 0.9918 [†]		
	1 vs 3	0.998	0.977	0.096	0.995	0.993	0.992 - 0.994 [†]	0.9921 - 0.9939 [†]	0.994	0.9936 - 0.9949 [†]
	2 vs 3	0.999	0.990	-0.028	0.999	0.9988	0.9987 - 0.999 [†]	0.9987 - 0.9990 [†]		

[†]:the CI95% does not include 0.96.

^a: ρ is the Pearson correlation coefficient, v , u and c_b are the scale shift, the location shift and the accuracy component of ρ_c .

^b: CI95% estimated using the hyperbolic tangent transformation.

^c: CI95% estimated using 1000 bootstrap samples.

Table 5.21: *Lin's CCC and Overall CCC for Illumina technical replicates.*

Potential effects of normalization were assessed also in this case by evaluating pairwise CCC and Overall CCC on data normalized using quantile and loess normalization (see table 5.22). Also in this case we see a mild increase in the OCCC, to which corresponds a moderate decrease in the precision of the confidence intervals, which are slightly larger than what occurs for raw data.

Sample	Pair	Quantile				Loess			
		ρ_c	CI95% ^a	OCCC	CI95% ^a	ρ_c	CI95% ^a	OCCC	CI95% ^a
A498	1 vs 2	0.9978	0.9973 - 0.9982 [†]			0.998	0.9976 - 0.9984 [†]		
	1 vs 3	0.9950	0.9941 - 0.9958 [†]	0.996	0.9951 - 0.9965 [†]	0.995	0.9942 - 0.9959 [†]	0.996	0.9955 - 0.9968 [†]
	2 vs 3	0.9950	0.9940 - 0.9959 [†]			0.995	0.9944 - 0.9962 [†]		
hREF	1 vs 2	0.9988	0.9985 - 0.9990 [†]			0.9984	0.9981 - 0.9987 [†]		
	1 vs 3	0.9982	0.9978 - 0.9985 [†]	0.9987	0.9985 - 0.9989 [†]	0.9976	0.9972 - 0.9980 [†]	0.9984	0.9981 - 0.9986 [†]
	2 vs 3	0.9993	0.9991 - 0.9994 [†]			0.9992	0.9991 - 0.9993 [†]		

[†]: the CI95% does not include 0.96.

^a: CI95% estimated using 1000 bootstrap samples.

Table 5.22: *Lin's CCC and Overall CCC for Illumina normalized technical replicates.*

In figure 5.18 we represent the Overall CCCs computed on raw and normalized data for both samples, alongside their bootstrap confidence intervals. Even in this case it is clear that the main effect of normalization lies not in the point estimate but rather in the interval estimates: both quantile and loess method reduce the width of the interval from 0.076 for raw data in line A498 to 0.014 and 0.013 for quantile and loess normalized data respectively, and from 0.0013 for raw data in sample hREF to 0.0004 and 0.0005 for quantile and loess normalized data.

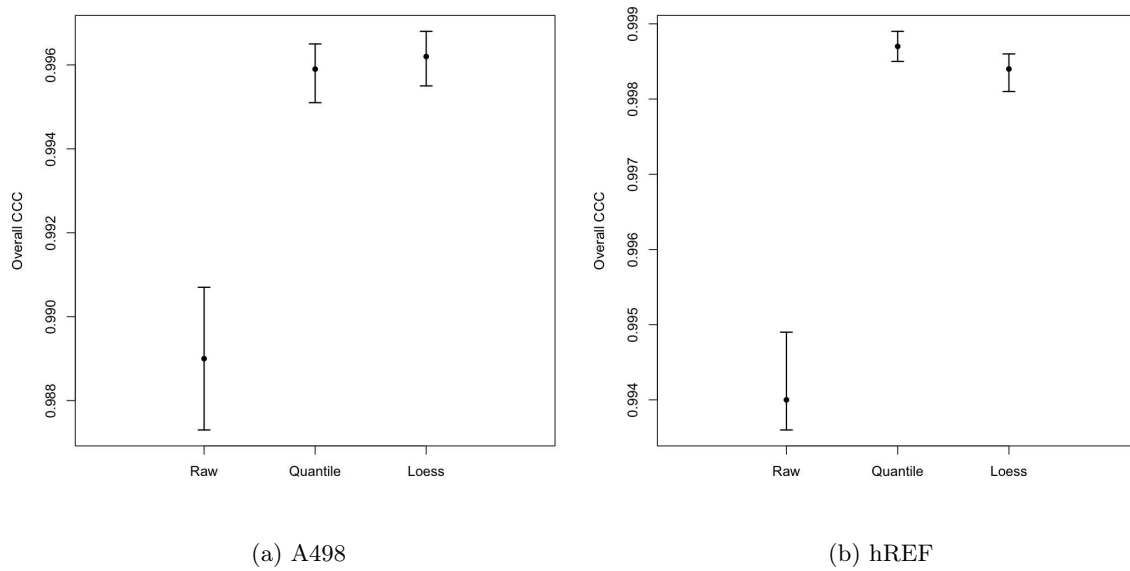
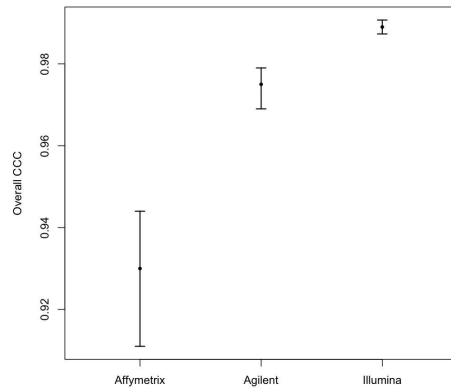
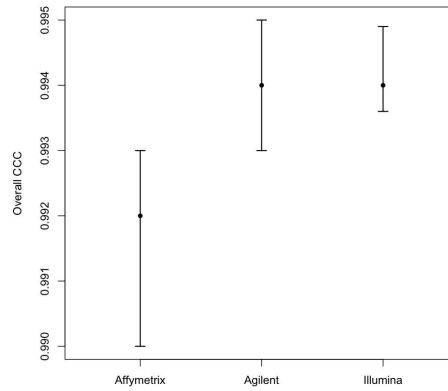


Figure 5.18: Overall CCC and 95% bootstrap confidence intervals for raw and normalized data using quantile and loess algorithms for Illumina data.

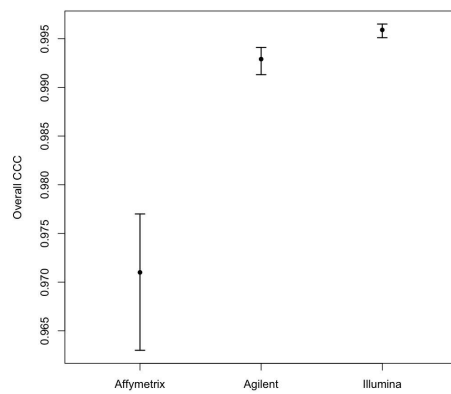
To directly compare all the three platforms we have plotted the Overall CCC estimated on probeset-level data, considering both samples and both raw and normalized data (see figure 5.19). With the exception of raw data for hREF sample, the hierarchy appears substantially conserved, with Illumina being the best performing platform and Affymetrix the worst. Notably, Agilent and Illumina show similar patterns in all the situations, and this could possibly be related to a lab-effect, since they were processed at a different site with respect to Affymetrix.



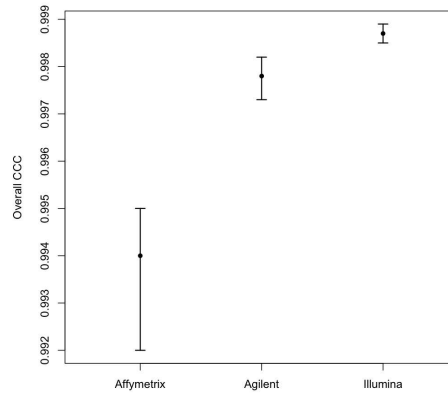
(a) A498 - Raw data



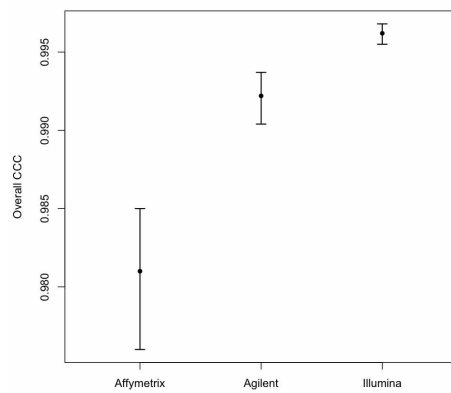
(b) hREF - Raw data



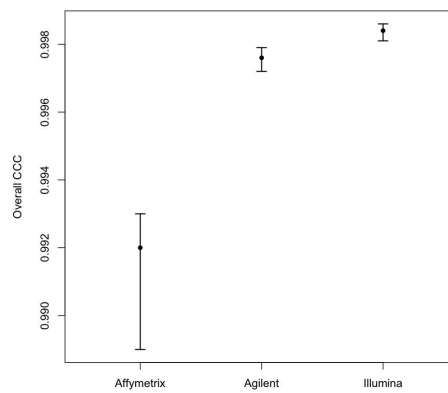
(c) A498 - Quantile Normalization



(d) hREF - Quantile Normalization



(e) A498 - Loess Normalization



(f) hREF - Loess Normalization

Figure 5.19: Overall CCC and 95% bootstrap confidence intervals for raw and normalized data using quantile and loess algorithms for all the platforms at a probeset level.

All the platforms can thus be considered reliable, net of batch effects occurred in the hybridization phase of the experiment. From this comparison we can devise a possible confounding effect related to the laboratories where the experiment was independently conducted, that should be carefully considered when it comes to interpretation of results, in particular when performing cross-platform comparison. In this perspective, array normalization will be a useful tool to actually understand patterns of between-platform concordance when technical issues that are not strictly connected to the array are removed.

5.3 Agreement Assessment

5.3.1 Raw data

To perform agreement evaluation, we averaged miRNAs intensities across technical replicates for each array and for each sample. Then, we evaluated pair-wise arrays agreement in terms of miRNA lying within the modified agreement interval described in section 4.2.2. Estimates of the measurement error model for error-variance ratio λ equal to 1 presented in table 5.23 show that the relationship between Agilent and Illumina is the one which is closest to the agreement line with intercept 0 and slope 1 for both samples, whereas models which include the Affymetrix platform for line A498 show a very negatively large intercept (-12,4128 and -17.4064) which possibly reflects the technical bias already highlighted for this samples. However, if we consider line hREF we can note that also in this case Affymetrix is the array which deviates most from the line of perfect agreement with both Illumina and Agilent, whereas these two show patterns very close to concordance (slope = 1.0925, CI95%: 1.0371 - 1.1479).

	<i>Pair</i>	a_0		b_0	
		<i>Estimate</i>	CI95%	<i>Estimate</i>	CI95%
<i>A498</i>	Agilent vs Affymetrix	-12.4128	(-16.9575 , -11.8681)	2.8265	(2.7461 , 2.9068)
	Illumina vs Affymetrix	-17.4064	(-18.6406 , -16.1722)	4.2889	(4.1679 , 4.4098)
	Illumina vs Agilent	2.1916	(1.7773 , 2.6058)	1.3254	(1.2407 , 1.4100)
<i>hREF</i>	Agilent vs Affymetrix	-6.1037	(-6.4471 , -5.7603)	1.9610	(1.8964 , 2.0255)
	Illumina vs Affymetrix	-4.7630	(-5.5006 , -4.0254)	2.4418	(2.3472 , 2.5363)
	Illumina vs Agilent	3.6033	(3.3791 , 3.8274)	1.0925	(1.0371 , 1.1479)

Table 5.23: *Estimates of the linear measurement error model, $\lambda = 1$*

Since the confidence intervals suggest an intercept different from 0 and a slope different from 1 for all comparisons in both samples, we build the agreement intervals following formula 4.49, and visualize results for lines A498 and hREF in figure 5.20. At the top of each graph we have reported the proportion of microRNA which are found to be in agreement after bias correction, i.e. miRNA that lie within the agreement interval.

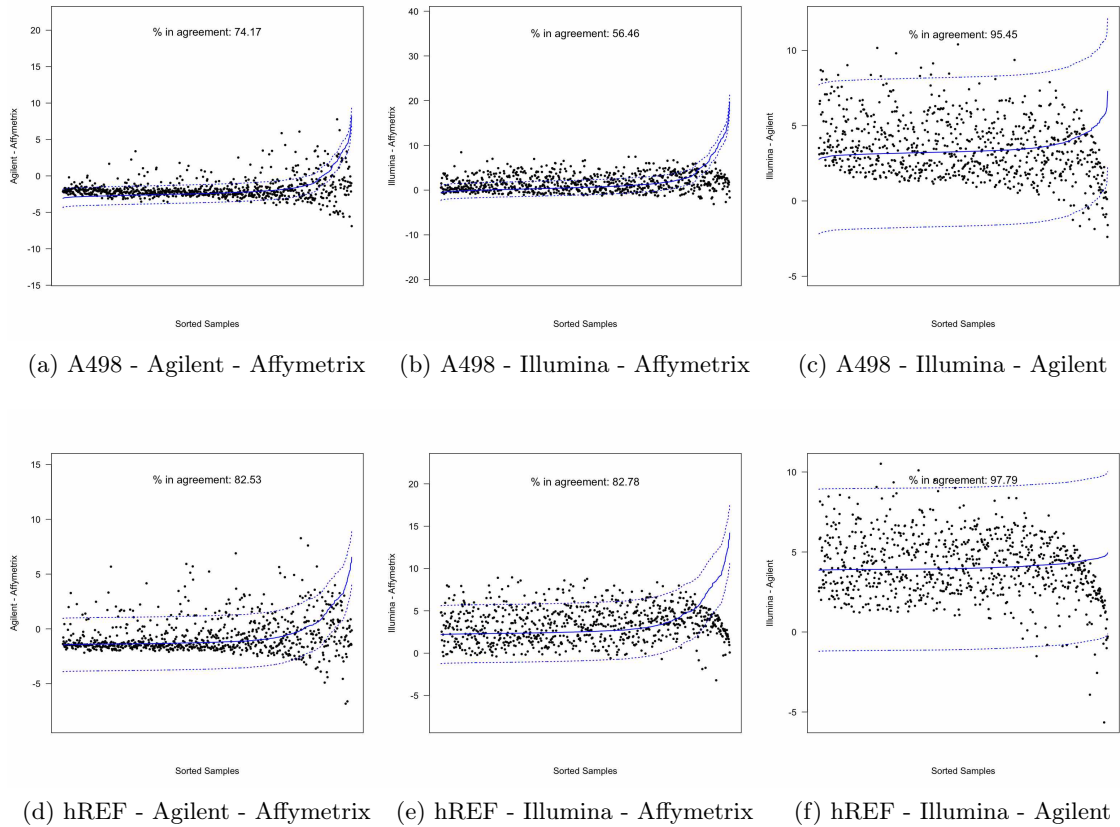


Figure 5.20: *Agreement intervals for the modified Bland-Altman method, $\lambda = 1$.*

If we consider A498 comparisons (upper panels) results are substantially worse when Affymetrix platform is considered: 74.17% (CI95%: 71.02 - 77.15) of microRNA, i.e. 603, lie within the agreement interval for the comparison Agilent - Affymetrix, 56.46% (CI95%: 52.97 - 59.90), i.e. 459, for the comparison Illumina - Affymetrix. Comparing Illumina and Agilent results in 95.45% (CI95%: 93.78 - 96.78), i.e. 776, microRNA in the agreement interval. hREF comparisons, on the other hand, show better patterns of agreement after

bias correction, resulting in 82.53% (CI95%: 79.75 - 85.08) of microRNA, i.e. 671, within the agreement interval for the comparison Agilent - Affymetrix, 82.78% (CI95%: 80.01 - 85.31) for the comparison Illumina - Affymetrix, i.e. 673, and 97.79% (CI95%: 96.52 - 98.68) for the comparison of Illumina and Agilent, i.e. 795 microRNA out of 813. Confidence intervals for the proportions were computed using the Clopper-Pearson exact method [130].

For a comparison with the classical Bland-Altman method, we need to estimate the average and difference between paired observations, and then to plot them on a graph with the corresponding LoA described in equation (4.34) to assess eventual agreement between methods. In table 5.24 we report estimates of the average difference, its standard deviation and the related Limits of Agreement.

	<i>Pair</i>	$\bar{D}$	σ_D	<i>LoA</i>
A498	Agilent - Affymetrix	-1.824	1.463	(-4.750, 1.102)
	Illumina - Affymetrix	1.660	1.895	(-2.130, 5.450)
	Illumina - Agilent	3.484	1.934	(-0.384, 7.353)
hREF	Agilent - Affymetrix	0.654	1.679	(-4.013, 2.704)
	Illumina - Affymetrix	3.413	2.031	(-0.649, 7.475)
	Illumina - Agilent	4.067	1.920	(0.228, 7.907)

Table 5.24: *Estimates for the Bland-Altman method*

Graphical results for the classical Bland-Altman method are reported in figure 5.21. To interpret results from these plots one should evaluate whether the computed LoA are sufficiently small or not: if they are, then the two methods can be considered to agree, otherwise they can not. One first limitation is that to say if the interval is small enough, one should have some reference value to compare it with, and in miRNA microarray this is a non trivial issue. Additionally, assuming that some previous knowledge exists which allows to draw conclusions on the actual concordance between methods, there are two key assumptions which are rarely met in common practice (and the plots of figure 5.21 make no exception): 1) the variance of the differences is constant over the range of measurements, and 2) the difference between the methods is at least approximately constant over the range of measurements [113, 64]. Because of these reasons, the standard Bland-Altman approach is not a suitable approach in our experiment since both quantities show a heavy variability,

thus posing relevant problems in terms of interpretation. Moreover, the modified method proposed allows to assess agreement with less *a priori* information than required with the standard method, since the choice of the threshold k previously discussed, though not trivial, can be regarded as easier than that of a reference agreement interval.

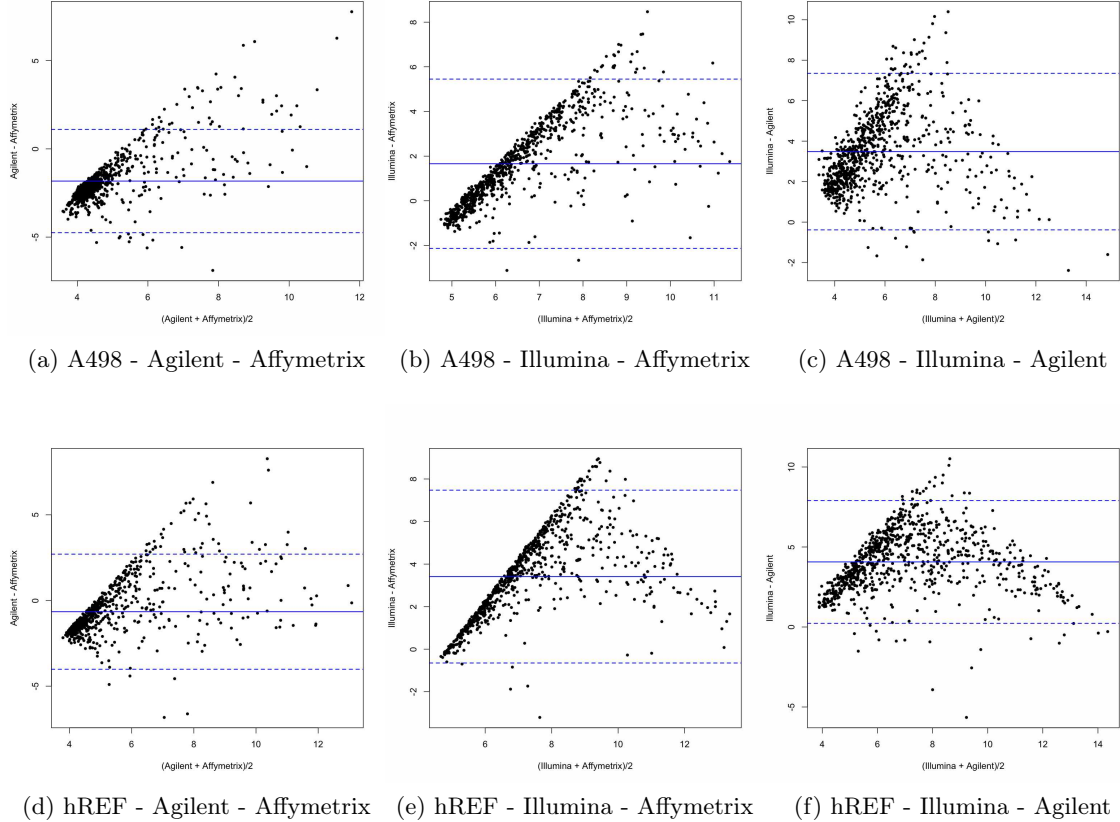


Figure 5.21: *Classical Bland-Altman plots*

Despite these advantages, the results of this modified version of the Bland-Altman plot showed so far rely on the validity of the assumption $\lambda = 1$ which, though reasonable in many practical situations, could be misleading in the context of microarrays. That is, assuming an error variance which is the same for very different platforms may be not the best choice, as the exploratory analysis of section 5.1 showed. For this reason, we have estimated the error variance for each platform within each sample by means of the random effects model described in (4.40), then the estimates of λ were obtained by computing ratios

of the error variances, and CI95% were estimated using formula (4.41). From table 5.25 we can see that the estimates for the comparison Illumina - Agilent are quite close to 1 for both samples (1.125 (CI95%: 1.039 - 1.218) for hREF and to 1.352 (CI95%: 1.248 - 1.463)), whereas the estimates for the other comparison are quite different from 1, in particular for Affymetrix-related comparisons in A498 line. Given these estimates, we expect the results for the concordance analysis to be substantially similar for the Illumina - Agilent comparison, whereas relevant differences are expected for the remaining pairs, in particular for samples A498.

		λ	CI95%
A498	<i>Agilent - Affymetrix</i>	4.125	3.810 - 4.466
	<i>Illumina - Affymetrix</i>	5.576	5.150 - 6.037
	<i>Illumina - Agilent</i>	1.352	1.248 - 1.463
hREF	<i>Agilent - Affymetrix</i>	2.608	2.409 - 2.824
	<i>Illumina - Affymetrix</i>	2.935	2.711 - 3.178
	<i>Illumina Agilent</i>	1.125	1.039 - 1.218

Table 5.25: *Estimates of λ and CI95%. Values obtained as ratio of σ_ε^2 (error variance of Y) and σ_δ^2 (error variance of X)*

Estimates of model parameters using these values for λ produces results presented in table 5.26. By comparing these estimates with those presented in table 5.23, we can see that for all comparisons and in both samples the intercepts decrease and the slopes increase, though these changes are much steeper for the Affymetrix-related comparisons, whereas for Illumina vs Agilent the differences do not appear to be that relevant.

		a_0		b_0	
	<i>Pair</i>	<i>Estimate</i>	CI95%	<i>Estimate</i>	CI95%
<i>A498</i>	Agilent vs Affymetrix	-7.902	(-7.965, -7.838)	2.048	(2.021, 2.076)
	Illumina vs Affymetrix	-6.497	(-6.556, -6.437)	2.407	(2.380, 2.434)
	Illumina vs Agilent	2.804	(2.545, 3.063)	1.171	(1.104, 1.238)
<i>hREF</i>	Agilent vs Affymetrix	-4.160	(-4.250, -4.071)	1.618	(1.585, 1.651)
	Illumina vs Affymetrix	-0.668	(-0.779, -0.558)	1.720	(1.683, 1.756)
	Illumina vs Agilent	3.757	(3.555, 3.958)	1.062	(1.009, 1.114)

Table 5.26: *Estimates of the linear measurement error model, λ estimated.*

Graphical results are reported in figure 5.22. Line A498 (upper panels) shows better results with respect to those obtained when $\lambda = 1$, in particular when Affymetrix platform is involved. In fact, an increase of almost 5% is seen when Agilent and Affymetrix are compared: from 74.17% (CI95%: 71.02 - 77.15) with $\lambda = 1$ to 78.84% (75.87 - 81.6) with λ estimated. Additionally, if we consider the Illumina-Affymetrix comparison, there's an increase in the proportion of concordant miRNA from 56.46% (CI95%: 52.97 - 59.90) to 73.43% (CI95%: 70.25 - 76.44). For hREF, Illumina and Agilent show similar patterns of concordance with respect to the value of λ : 97.79% (CI95%: 96.52 - 98.68), i.e. 795 miRNA, for $\lambda = 1$ and 97.54% (CI95%: 96.22 - 98.49), i.e. 793 miRNA, for $\lambda = 1.125$ (estimated value). The comparison Affymetrix - Agilent shows a slight increase in the percentage of concordant miRNA, with only 84.26% (CI95%: 81.57 - 86.70), i.e. 685, miRNA in the agreement interval, whereas Illumina and Affymetrix show the largest increase in proportion of concordant miRNA: 89.91% (CI95%: 87.64 - 91.9).

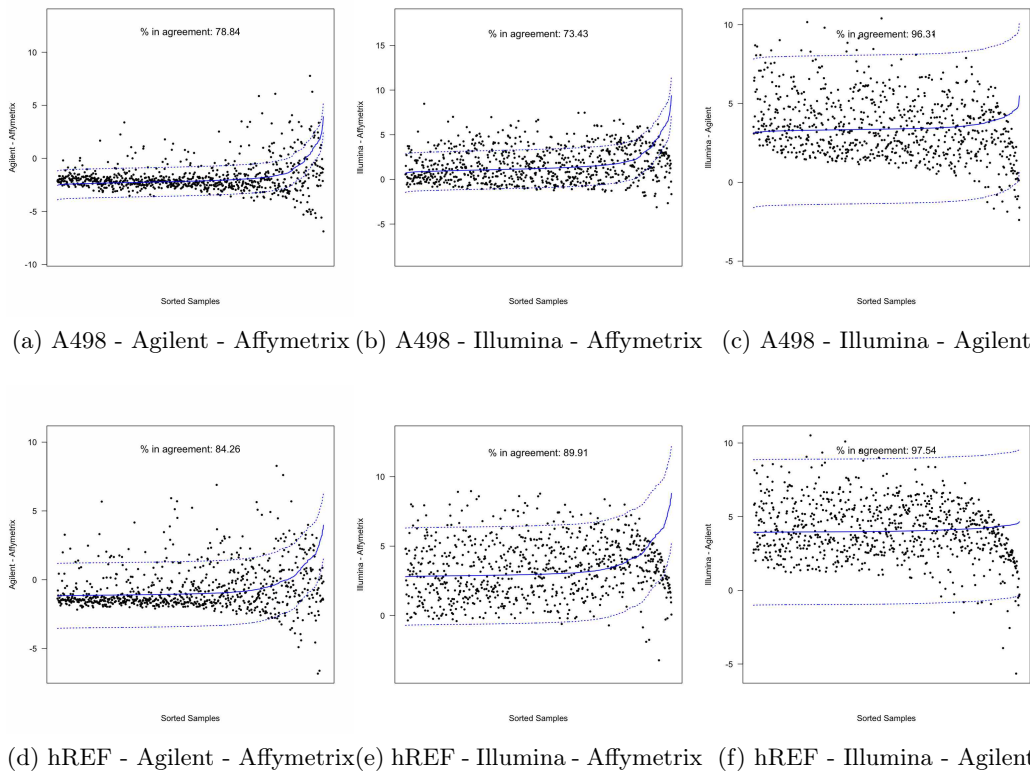


Figure 5.22: Agreement intervals for the modified Bland-Altman method, λ estimated.

In table 5.27 detailed results on percentages of microRNA lying within the agreement intervals visualized in figures 5.20 and 5.22 are reported. By comparing results not only according to the parameter λ but also with respect to the different samples, we can note that Affymetrix platform shows moderate-to-low levels of concordance with both Agilent and Illumina irrespectively of the sample considered. In general, the worst comparison appears to be the one between Affymetrix and Illumina, which shows a relevantly lower proportion of concordant miRNAs for all λ and samples.

	<i>Comparison</i>	$\lambda = 1$		λ estimated	
		% (CI95%)	<i>n</i>	% (CI95%)	<i>n</i>
A498	<i>Agilent - Affymetrix</i>	74.17 (71.02 - 77.15)	603	78.84 (75.87 - 81.60)	641
	<i>Illumina - Affymetrix</i>	56.46 (52.97 - 59.90)	459	73.43 (70.25 - 76.44)	597
	<i>Illumina - Agilent</i>	95.45 (93.78 - 96.78)	776	96.31 (94.77 - 97.50)	783
hREF	<i>Agilent - Affymetrix</i>	82.53 (79.75 - 85.08)	671	84.26 (81.57 - 86.70)	685
	<i>Illumina - Affymetrix</i>	82.78 (80.01 - 85.31)	673	89.91 (87.63 - 91.90)	731
	<i>Illumina - Agilent</i>	97.79 (96.52 - 98.68) [†]	795	97.54 (96.23 - 98.49) [†]	793

[†]: the platform pair is in agreement

Table 5.27: *Number and proportion of miRNA lying within the agreement intervals, estimated according to the measurement error model parameters estimated by setting $\lambda = 1$ and by estimating it via random effects models. Confidence intervals for the proportions were computed using the Clopper-Pearson exact method [130].*

To have a better view of these results, we have plotted the point and interval estimates separately for each sample in figure 5.23, but keeping the same range for the Y-axis, to better highlight differences. Let us focus on hREF results, since the technical issues on A498 lines may lead to biased conclusions: Illumina and Agilent show similar patterns of moderate-to-poor agreement with Affymetrix platform when $\lambda = 1$ is considered, but when we estimate it the number of concordant miRNAs seems to increase slightly more for the Illumina-Affymetrix than for the Agilent-Affymetrix comparison. Notably, Illumina and Agilent show very satisfactory levels of agreement in both samples for both choices for the parameter λ .

To see which are actually the microRNAs which do not agree, in the scatterplots of figure 5.24 we have plotted the intensity values for all platform pairs and the fitted linear measurement error model (for $\lambda = 1$), highlighting which are the non concordant microRNAs. The main problems seem to occur for poorly expressed microRNAs on Affymetrix, which

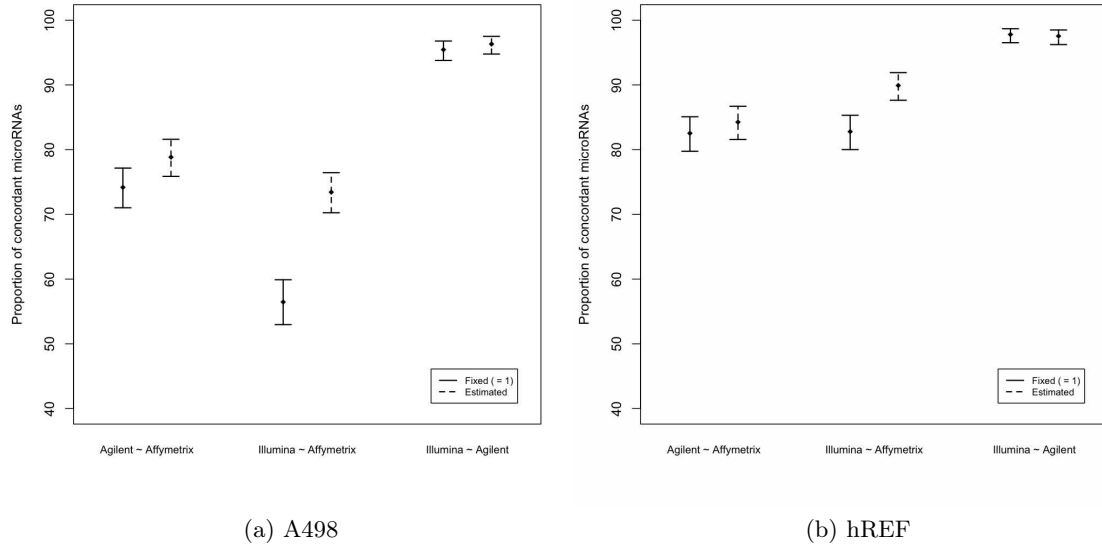


Figure 5.23: *Confidence intervals for the proportion of concordant miRNAs according to different choices/estimates for the parameter λ .*

are highly discordant from both Illumina and Agilent, with qualitatively similar patterns between samples. Additionally, we can see that the fitted line for the comparison Illumina - Agilent (solid red) is almost perfectly parallel to the line of perfect agreement (dashed blue), thus suggesting that the proportional bias (defined by the slope of the model) is neglectable and a different agreement interval could be computed. If we consider the agreement interval described in (4.47) for the comparison Illumina - Agilent, than the proportion of miRNA in agreement for line hREF is equal to 98.65% (CI95%: 97.60 - 99.32) and for line A498 to 95.69% (CI95%: 94.06 - 96.98), always considering $\lambda = 1$. Notably, these values are slightly better than those seen for the "full model" agreement interval which corrected for both a fixed and proportional bias.

Finally, to assess agreement we need to choose a value for the threshold k , the maximum number of subjects (i.e. microRNA) that we are willing to accept to lie outside the estimated agreement interval. By choosing a value of $k = 41$ ($\simeq 0.05 * 813$), we can say that only Illumina and Agilent can be considered to be in agreement, whereas Affymetrix shows moderate-to-poor concordance with both Agilent and Illumina.

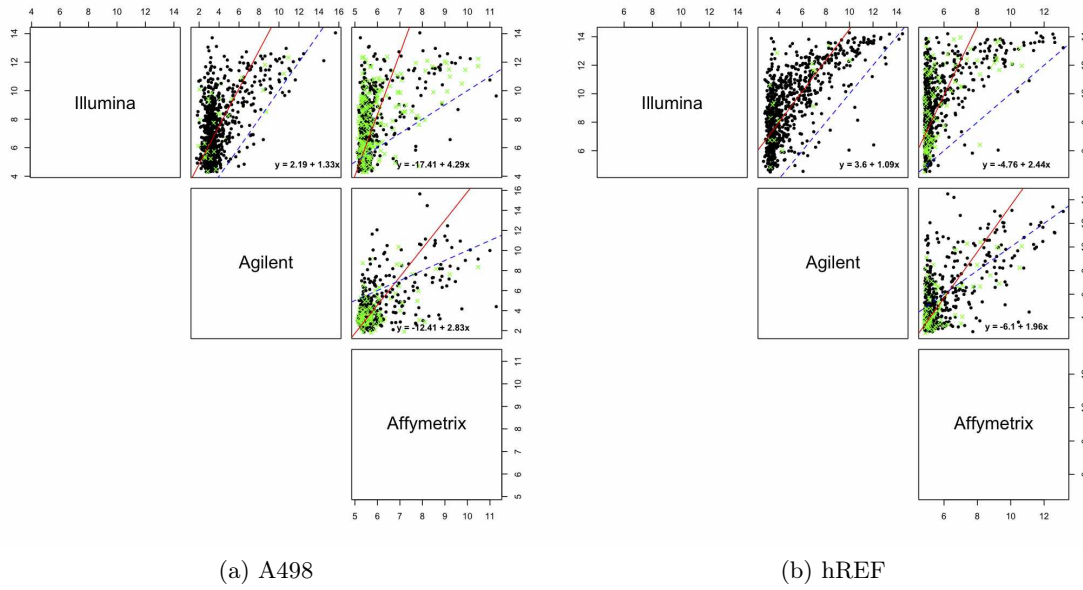


Figure 5.24: *Pair-plots of intensity values for all platform comparisons for both samples. The solid red line represents the linear measurement error model estimated with $\lambda = 1$, the dashed blue line is the line of perfect agreement (intercept = 0 and slope = 1) and the green labelled crosses are the non concordant microRNA.*

5.3.2 Normalized data

As for repeatability analysis, we evaluated possible effects of normalization on results of cross-platform evaluation. That is, we performed quantile and loess normalization and the measurement error model was estimated for normalized data, by considering both $\lambda = 1$ and estimating it via random effects model. For the sake of brevity, only proportion and number of microRNAs lying within the modified agreement intervals are reported in table 5.28 and 5.29 for quantile and loess normalization, respectively. Also in this case the agreement intervals were computed according to formula (4.49), since all confidence intervals for the intercepts did not include 0 and those for the slope did not include 1. With respect to results presented in table 5.27, results are opposite for the two samples: in fact, whereas A498 benefits from both normalization algorithms and the proportion of concordant microRNA increases both when λ is set to 1 and when is estimated, sample hREF shows decreases in these percentages, which are sometimes substantially lower when normalization (both of them) is performed. For instance, if we consider the Illumina -

Affymetrix comparison on non normalized data we have that, for $\lambda = 1$ there are 82.78% of microRNAs within the agreement interval (CI95%: 80.01 - 85.31), but if we normalize there are only 71.34% (CI95%: 68.10 - 74.43) and 71.83% (CI95%: 68.60 - 74.90) for quantile and loess normalization, respectively.

		$\lambda = 1$		λ estimated	
	<i>Comparison</i>	% (CI95%)	<i>n</i>	% (CI95%)	<i>n</i>
A498	<i>Agilent - Affymetrix</i>	78.84 (75.87, 81.6)	641	82.41 (79.62, 84.97)	670
	<i>Illumina - Affymetrix</i>	70.48 (67.21, 73.6)	573	83.89 (81.18, 86.35)	682
	<i>Illumina - Agilent</i>	98.52 (97.44, 99.24) [†]	801	98.77 (97.75, 99.41) [†]	803
hREF	<i>Agilent - Affymetrix</i>	79.46 (76.52, 82.19)	646	81.55 (78.71, 84.16)	663
	<i>Illumina - Affymetrix</i>	71.34 (68.1, 74.43)	580	80.93 (78.06, 83.58)	658
	<i>Illumina - Agilent</i>	96.43 (94.92, 97.6)	784	96.19 (94.63, 97.39)	782

[†]: the platform pair is in agreement

Table 5.28: Number and proportion of miRNA lying within the agreement intervals, estimated according to the measurement error model parameters estimated by setting $\lambda = 1$ and by estimating it via random effects models. Data were normalized according to quantile normalization algorithm. Confidence intervals for the proportions were computed using the Clopper-Pearson exact method [130].

		$\lambda = 1$		λ estimated	
	<i>Comparison</i>	% (CI95%)	<i>n</i>	% (CI95%)	<i>n</i>
A498	<i>Agilent - Affymetrix</i>	79.95 (77.03, 82.65)	650	83.27 (80.53, 85.77)	677
	<i>Illumina - Affymetrix</i>	68.76 (65.45, 71.93)	559	83.03 (80.27, 85.54)	675
	<i>Illumina - Agilent</i>	98.4 (97.28, 99.15) [†]	800	98.52 (97.44, 99.24) [†]	801
hREF	<i>Agilent - Affymetrix</i>	79.21 (76.26, 81.95)	644	81.18 (78.32, 83.81)	660
	<i>Illumina - Affymetrix</i>	71.83 (68.60, 74.90)	584	82.90 (80.14, 85.43)	674
	<i>Illumina - Agilent</i>	96.06 (94.49, 97.29)	781	95.82 (94.2, 97.09)	779

[†]: the platform pair is in agreement

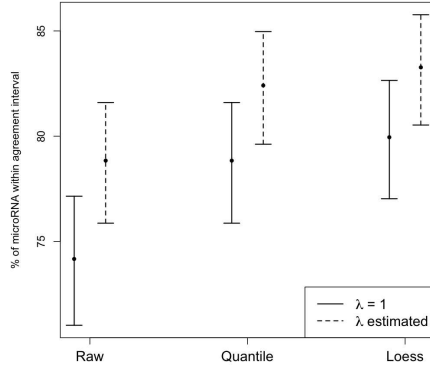
Table 5.29: Number and proportion of miRNA lying within the agreement intervals, estimated according to the measurement error model parameters estimated by setting $\lambda = 1$ and by estimating it via random effects models. Data were normalized according to loess normalization algorithm. Confidence intervals for the proportions were computed using the Clopper-Pearson exact method [130].

To compare results between normalized and non-normalized data, as well as between different normalization methods and strategies for the choice of λ , we have plotted the proportion of microRNA within the agreement interval alongside its confidence intervals in

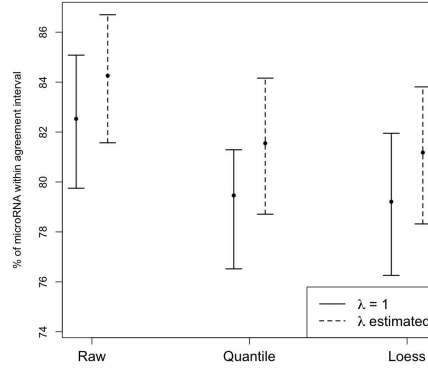
figure , separated for sample and platform-pair comparison. In the left panels results refer to line A498 and in the right panels to sample hREF, whereas within each panel solid line refers to $\lambda = 1$ and dashed line to an estimate via random effects model. As previously discussed, the effect of normalization is very different on the two samples. That is, if we consider line A498 we can see that all the comparisons show increased pattern of agreement when data are normalized, irrespective of the method used, whereas for sample hREF the opposite situation occurs, i.e. normalizing data results in a loss of concordance for all comparison, in particular for the Illumina-Affymetrix, the one that shows the worst results also for line A498. It is relevant to note that using an estimated value for λ leads generally to a higher proportion of microRNA within the agreement interval, on both samples and for both type of data (normalized and non normalized), suggesting that a proper estimation of λ in this context is fundamental to be able to correctly assess patterns of agreement.

This last point deserves a particular attention, because it is crucial for the correct assessment of between-platforms agreement. Illumina and Agilent do not seem to suffer much from the choice of the value of λ , and in fact this is very close to 1 irrespective of the normalization method considered, so that results are almost the same between normalization methods for this comparison (lower panels of figure 5.25). Actually, our method is just one of all possible choices to estimate measurement error variability, which has the advantage to be relatively simple and intuitive, but relies on some distributional assumption to be fulfilled, such as residuals homoscedasticity and normality, that we have seen may present some problem in microarray data (see section 5.2.1).

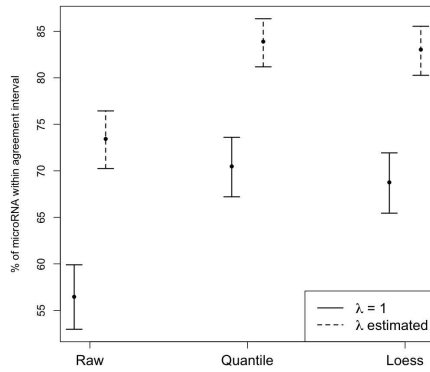
Our results suggest that Affymetrix platform shows moderate-to-poor patterns of concordance with both Illumina and Agilent, since all the confidence intervals include the threshold defined for the agreement of 95% of microRNA lying within the agreement intervals. Notably, only Illumina and Agilent are characterized by good level of agreement after bias correction, but only for sample hREF if we consider raw data, and only for line A498 if we perform normalization. It has however to be pointed out that these considerations strictly depend on our choice for the value of k , the "agreement" threshold.



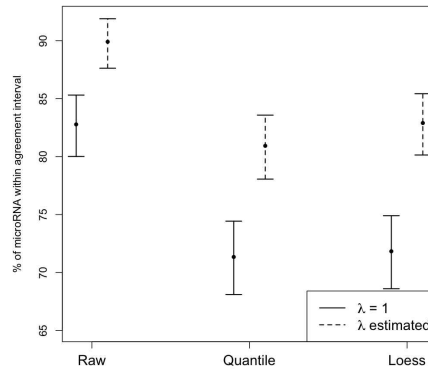
(a) A498 - Agilent vs Affymetrix



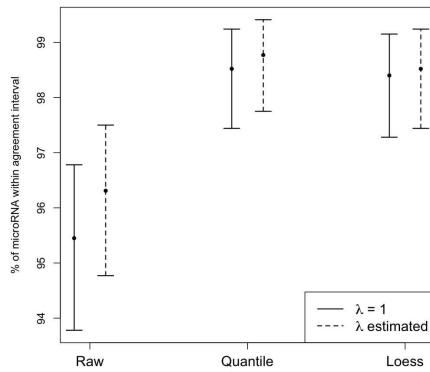
(b) hREF - Agilent vs Affymetrix



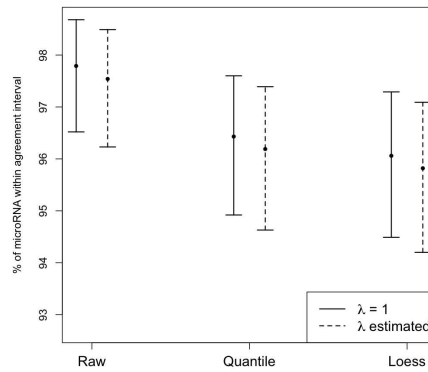
(c) A498 - Illumina vs Affymetrix



(d) hREF - Illumina vs Affymetrix



(e) A498 - Agilent vs Illumina



(f) hREF - Agilent vs Illumina

Figure 5.25: Proportion of miRNA within the agreement intervals, according to different data (normalized and non normalized), different choices for λ (set to 1 and estimated) and platform-pair.

In this sense, the choice of using $k = 41$ is quite conservative, but could be not the more appropriate in a complex setting like that of microarrays, where measurements from a single miRNA are influenced by several uncontrollable factors and could thus "differ" from those obtained on another array merely by chance. Thus, if we defined a less stringent threshold at $k = 81$ (90% of microRNA required to lie within the agreement intervals), we would find that Illumina and Agilent are declared concordant throughout all scenarios here considered, net of the correction by both a fixed and a proportional bias, estimated by the parameters of a linear measurement error model. Actually, the proportional bias to correct for between these two platforms is almost irrelevant, so that we could almost say that Agilent and Illumina mainly differ by a fixed shift.

Discussion

Contents

6.1 Within-Platform Repeatability	116
6.2 Between-Platforms Agreement	120

6.1 Within-Platform Repeatability

Technical evaluation of microarray reliability is an issue that researchers have been dealing with since the dawn of genomics era, and miRNA microarray are not different in this sense. Given the complex and noisy nature of microarray data, it is fundamental to address such an issue with proper statistical methods, to ensure that biological discoveries are not affected by some bias. Nonetheless, many published works make use of correlation coefficients (Pearson and/or Spearman) or compare list of differentially expressed miRNAs to evaluate these technical aspects (see literature in section 2.4 for a detailed review). In this project we have chosen to adopt a completely different approach, trying to quantify all sources of variability involved in a microarray experiment, and have derived estimates of the Intraclass Correlation Coefficient for all the factors included in the model. One advantage of the proposed approach is that we are not only able to assess whether microarray data are repeatable or not, but also, and most important, we can say which factors influence data repeatability.

For instance, a well known problem in gene expression analysis is the summarization of probes within a probe set. Several algorithms exist which incorporate this step in the pre-processing step of microarray data (RMA from Irizarry *et al.* [12, 131], for instance, is well known for gene expression and has been modified to be used also for microRNAs [88]), but it is not so unusual to run into a "broken" probe with inconsistent expression/intensity levels

when compared to the other probes in the probeset. Reason for such a difference are quite different, and vary from error in the spotted sequence of the probe to cross-hybridization and so on. Since it is important to understand if these probes are really consistent, considering data at a probe-level is extremely important and may shed more light on some technical aspects of the platforms.

Our method provided evidence that the different probes within a probe set convey similar informations, with low variability associated to the "probe" factor, even if the concept of probe is actually quite different from platform to platform. For instance, on Agilent we see that $\approx 10\%$ of total experimental variability is related to the probes, whereas on Affymetrix this percentage is much lower, at least for sample hREF. Yet, despite this difference in magnitude, we must consider that probes within a probeset for Agilent array are not exact replicates, i.e. there are differences in the nucleotide sequence, so that what we are looking is not only a technical component, but rather a mixture of technical (the differences in intensities between probes with identical sequences) and biological (the differences in intensities between probes with different sequences) effects.

Another relevant focus, which is what is mainly studied throughout literature, is what we have called "array effect", i.e. the variability related to technical array replicates. We found it to be substantially very low, suggesting that all the considered platforms provide very stable measurements of intensities, at least when a single-lab experiment is performed. As already discussed however, repeatability is just a component of platform reliability, the reproducibility being the other one. This last property, however, could not be "tested" because of the single-lab nature of our study. Yet, the method here proposed could be implemented also in a multicentre project, by adding an additional effect to the model. This, though feasible, is certainly not trivial, since it would add a further degree of complexity to an already largely parameterized model, resulting in a 4-ways nested-factorial model, for which specification within software packages may become somehow difficult. As a matter of fact, we chose to analyze separately the two samples rather than including another variable in the model to keep the computational demands from increasing too much. This approach, though not the more effective, could be applied also in presence of data from multiple laboratories by comparing results between the same model applied to data coming from different labs rather than directly evaluating variability connected to a "lab effect".

The use of random effects models is somehow new in its application to microarray data reliability, though it has been a landmark in many areas for many years. Its main advantage is the possibility to derive measures allowing to assess how much does each factor contribute to overall data variability, i.e. the ICC, simply by starting from the estimates of the variance components produced by the model. Here we estimated confidence intervals for the ICC by simulating 1000 parametric bootstrap samples and using the standard t approach, which led in some situations to inconsistent results, with CI95% taking values below 0 and above 1. Actually, some methods for an approximate estimation of confidence intervals for the ICC exist in the 2-way ANOVA framework, yet literature is still lacking as long as higher-dimensional studies are concerned. Here, these interval estimates allowed us to directly compare results between samples and between platforms, a task that could not be addressed by considering σ^2 estimates of variance components. Despite the fact that ICC have been criticized because of their dependence on the measurements range [56], they can be extremely useful in getting a picture of different sources of variability even in complex situations.

This method here was applied to a very specific context, but it can be equally applied in a large variety of settings, because of its flexibility. Depending on the experimental setting one can specify a model accordingly, eventually deciding to keep some factors as fixed rather than random, and then estimating variance components and Intraclass Correlation Coefficients for the desired effects to gain insights on the overall variability of the assay.

Jointly with this model-based analysis, the evaluation of concordance between replicates provided also useful information on the repeatability patterns of the different platforms. By evaluating the CCC estimates in a non-inferiority setting using a very conservative threshold of 0.96, we were able to determine if a platform was repeatable, i.e. if the different technical replicates were concordant. Here we found that Agilent and Illumina performed very well on both samples, whereas Affymetrix had some technical bias on line A498, that was seen to be solved when normalizing data. The main challenge here was the computational demand for estimating bootstrap confidence intervals for the repeated measures CCC and the related overall CCC, which resulted in some days of calculations to derive an appropriate number of bootstrap estimates to compute the BC_a bootstrap intervals. Notably, these estimates were not possible for Agilent because of the high dimensionality of the

data, so that only point estimates were computed, alongside a normal approximation for the pairwise CCC confidence intervals. Our results have highlighted that Agilent and Illumina are the platforms which show the best pattern of both reliability and agreement, whereas Affymetrix appears to have some technical problem. If we consider data at a probeset level, the OCCC for line A498 is relevantly lower than that for line hREF for all platforms with non-overlapping confidence intervals and for all set of miRNAs considered (common human miRNAs, human miRNAs array-wise, all miRNAs). Overall, Illumina (OCCC from 0.989 to 0.994) shows better performances with respect to Agilent (from 0.975 to 0.994) and, in particular, Affymetrix (from 0.927 to 0.993), and this is better seen for line A498. Since this line suffered from technical issues on Affymetrix and, partially, also on Agilent, the related OCCC possibly underestimates the "real" degree of reproducibility of the platforms, which can be better appreciated on line hREF where it ranges from 0.992 (CI95%: 0.990 - 0.993) to 0.993 (CI95%: 0.993 - 0.994), with values almost identical to those of Agilent and Illumina, thus suggesting also in this case a good performance of the array.

It should be noted that the repeatability threshold was substantially subjective, but was chosen with the goal of being reasonably conservative. The original work [82] suggested a threshold of 0.9, yet this value seemed too little stringent, possibly meaning that also weakly repeatable platforms may have passed it. Because of these considerations, we modified the minimum requirements to have a sufficiently high threshold, so that platforms passing it could be declared reliable with a good level of confidence.

As discussed in section 3.4.2, we evaluated effect of normalization and filtering by applying a standard filter and two commonly used normalization methods to our data, and compared results "before" and "after" for both random effects model and CCC estimation. With respect to the latter, normalization was found not to relevantly affect results, with the exception of line A498 where it led to an increased Overall CCC above the threshold of 0.96 for repeatability. In this sense, though it can offer a powerful solution to highlight possible biases in the data, normalization should be handled very carefully when there is no evidence of particular technical artifacts. Moreover, it should be remarked that normalization is not suitable for all purposes: if we want to know if there is some gene which is differentially expressed between two experimental groups it's important that we remove as much noise as possible to avoid spurious or misleading findings, and normalization/filtering are simple

and useful tools to accomplish this task. However, if our study wants to assess a platform as a whole and to evaluate its global characteristics, rather than simply keep the "tip of the iceberg", than the noise is likely to contain useful information and discarding it or reducing it may not be the best strategy to answer our questions.

6.2 Between-Platforms Agreement

Assessing agreement between different methods of measurement is a task which has rarely been addressed in microarray literature, where the focus has always been more on the evaluation of a linear relationship between platforms or with "gold-standard" assays such as qPCR, mainly via computation of correlation coefficients. Such an approach has often been biased by the selection of miRNAs used to do the computations, in that often only miRNAs concordantly detectable between platforms [81, 80] were chosen, possibly leading to an overestimate of the real level of correlation between arrays. To avoid this issue we have considered all common human microRNA that were common (i.e. matched by name) on all the platforms considered for the experiment. Additionally, correlation coefficients are computed assuming that intensity/expression values do not suffer from some measurement error, thus leading to possible under-estimates of the real level of correlation between platforms [114], whereas the proposed method takes into account the presence of an additive measurement error in the measurements from both platforms being compared and incorporates it in the linear model to estimate relationship between arrays.

The method is a modified version of the well-known Bland-Altman plot, a graphical technique commonly used to assess agreement between clinical measurements, and we have applied it in the field of microRNA array platforms by using it in a slightly different way. In a nutshell, commonly there are n subjects on which we measure some biological quantities using k measurement methods ($k \geq 2$) and our goal is to evaluate whether measurements from the k methods agree using information on n samples. Had we followed this procedure, each miRNA should have been evaluated separately (since the microRNA is the biological quantity of interest) and the intensities in the n samples (in our case the two cell lines) would have been compared between the k platforms (in our case 3) jointly for both cell lines. Actually, we have considered the microRNAs to be "subjects" (so that $n = 813$)

and the whole profile of intensity on a platform to be the vector of measurements to be compared between different platforms (that is, $k = 3$), separately for each cell line.

By setting a "strong" threshold at 95%, equivalent to 772 microRNAs, we conclude that Agilent and Illumina arrays are concordant according to agreement interval derived from the estimation of the measurement error model, irrespective of the method for choosing the value of λ . The choice of this threshold is subjective, and should depend on the issue at hand; our choice was due to the fact that it is likely that a few miRNAs exist which do not agree because of unwanted technical issues not attributable to the platform itself, but also to the need to reduce the number of false positives (i.e. falsely concordant platforms). This last point is crucial in the field of microarrays, since the comparability of results from different platforms as a tool for validating a lab's own results has gained much relevance in omic research and relies on the assumption that the two platforms are "linked" in that what they say about the profile of intensity/expression for a sample is similar, net of the different measurement and analytic scale of the platform itself. Here we found that only Illumina and Agilent could be considered to be in agreement, and only if we considered raw data on sample hREF. By reducing the stringency of the threshold and setting it to $k = 81$ (i.e. 90% of miRNA within the agreement interval), we find that Agilent and Illumina are called concordant for both lines, using both normalized and non-normalized data, whereas Affymetrix-related comparison show substantial poor patterns of concordance for both lines, irrespective of the normalization method considered.

Actually, however, there's a possible confounding effect in our results that was not possible to control, i.e. the site where the array were processed. As introduced in Chapter 3, Affymetrix arrays were processed at a lab, Agilent and Illumina at another one, so that when we do comparison between Affymetrix and the other platforms we are not simply comparing microarray technologies, but also different laboratories, in a completely confounded way. In this perspective, we can not be sure to which extent does the laboratory effect encompasses the platform effect, i.e. how much the differences between Affymetrix and the other platforms are due to different sites of processing and how much to actual assay differences. Similarly, we can not exclude that the great degree of concordance between Illumina and Agilent could be sharpened by this confounding effect. One alternative could have been to have all the arrays processed at a single site, so that the shared laboratory effect would

cancel out, yet this was not possible for the facilities involved in the experiment did not have all the instrumentation needed for performing the experiment with all the arrays. This problem is also present in the paper from Yauk *et al.* [81], which compares four different microarray technologies processed at three different sites, yet the authors state that *"our primary interpretation of the data is that the platforms show a high level of similarity to each other despite being hybridized in different labs, at different times, with different protocols and amounts of expertise"*. Whereas these conclusion can be appropriate for repeatability analysis, though being confident in the validity of our results we have no way to discriminate between "biological" agreement, i.e. concordance due to real similarity between platforms, and "technical" agreement, i.e. concordance or discordance related to the lab where the platform was processed.

Another relevant point concerns the choice/estimation of the λ parameter. We have already discussed that setting it equal to 1 is a strong assumption which, though reasonable in many practical situations, is likely to be violated in microarrays. To circumvent this issue, we estimated the error variance by means of replication, fitting a random effects linear model to each platform and sample, but also this strategy presents some drawbacks, since results for the nested-factorial random effects model clearly shows that the assumptions underlying the model were questionable. In the context of random effects models, a Bayesian framework exists that might result in better estimates, yet its performance need to be carefully studied [132].

Notably, the measurement error model considered here is just one of the possible models that could be fitted. From scatterplots of figure 5.24 it is easy to note that the linearity of the relationship between discordant platforms can be questioned, so that the differences we have found could be due to a non-linearity of the relationship or to a lack-of-fit of the regression line, which can be accounted for by considering different functional forms for the x variable, both in a linear and in a non-linear context such as, for instance, non linear measurement error models [115] and regression splines [133].

Conclusions and Future Developments

MiRNA microarray technology is a recent technique which allows the profiling of more than thousand human microRNA, small non-coding molecules of RNA which are known to play a fundamental role in the onset and development of several pathologies. Despite its wide usage however, microarray platforms for the profiling of microRNA expression have received only little attention, when compared to gene expression microarrays, in terms of both their reliability and concordance between technologies. Additionally, the few studies that addressed this issues mainly adopted poor statistical methods that led to misleading and sometimes partial answer to these questions. Correlation coefficients, overlap of significantly modulated genes between groups and other metrics have been commonly used to evaluate miRNA microarray performance, yet they have been substantially misused and over-interpreted.

In this project we have proposed to resort to a random effects modeling framework, which allows us evaluate how much does each source of variability contribute to overall experimental variation. By means of an estimate of the Intraclass Correlation Coefficient we were then able to directly compare results between platforms, and to assess to which extent technical effects do actually affect intensity measurements. Jointly, the evaluation of Concordance Correlation Coefficients at both probe and probeset level in a non-inferiority setting provides useful and integrative information to the assessment of platform repeatability.

The use of a modified version of the Bland-Altman plot is a powerful tool to determine which are the reasons for possible disagreement between platforms, and allows to directly take into account the error in the measurements of intensities, an issue which is not always

given the proper importance in microarray investigations, which often rely on simple correlation coefficients to assess the existence of some linear relationship, without any further characterization. The proposed approach could be used also for repeatability assessment, in conjunction with what has been already discussed.

We suggest that this kind of analysis should be performed on raw data, prior to any filtering and/or normalization, opposite to what happens in the majority of published works, not to alter the information provided by the array itself. Normalization, in particular, should only be considered if some technical effect has occurred that may mask true information, such as batch effects. Even in this case, however, comparison of results before and after normalizing should be made, to better understand the impact of this procedure. Additionally, since no gold standard exists for miRNA microarray, if data have to be normalized to retrieve the information they contain, then more than just one algorithm should be considered and deep evaluation of results has to be performed.

Since online databases of genomic data have now become a large resource from which to draw, the described approaches will be applied to publicly available data and the deriving results compared with the original ones. In this perspective, studies assessing both repeatability and reproducibility will be considered for selection. Moreover, given the large degree of flexibility of the proposed methods, other and more recent genomic platforms, such as Next Generation Sequencing, will be considered for analysis. In particular, attention will be devoted to RNASeq, a technology allowing to directly count how many copies of a given transcript are present in a portion of biological material, for which the urge of platform performance evaluation is very urgent.

R functions for Repeatability Analysis

A.1 ANOVA table for nested-factorial random effects models

```
# Function to estimate a three-way ANOVA table.

# data = a 4-column dataframe, containing three categorical variables (the
# main effects) and one continuous outcome (the response variable). The
# continuous outcome needs to be in the last column, and if some hierarchy
# (i.e. nesting) exists in the data this must be respected in the columns
# ordering.

# source = optional effects names for the ANOVA output table. Defaults to
# c('P', 'B', 'W', 'P*B', 'P*W', 'B*W', 'P*B*W', 'Error') for full-factorial
# designs and to c('P', 'B', 'P*B', 'WinP', '(WinP)*B', 'Error') for
# nested-factorial designs

# subset = index to select subset of the data. Defaults to all data.

# type = type of design. Options are 'crossed' and 'nested.factorial'. This
# must be specified or the function will return an error.

ANOVA.crossed.3ways <- function(data, source, subset = 1:nrow(data),
  type = c("crossed", "nested.factorial")) {

  # categories for each factor
  data <- data[subset, ]
  data <- data[order(data[, 2], data[, 1]), ]
  lev.P <- nlevels(data[, 1])
  lev.B <- nlevels(data[, 2])
  lev.W <- nlevels(data[, 3])
  lev.PB <- nlevels(data[, 1]) * nlevels(data[, 2])
}
```

```

lev.PW <- nlevels(data[, 1]) * nlevels(data[, 3])
lev.BW <- nlevels(data[, 2]) * nlevels(data[, 3])
lev.PBW <- nlevels(data[, 1]) * nlevels(data[,
  2]) * nlevels(data[, 3])
n <- nrow(data)/lev.PBW

# degrees of freedom estimates
df.P <- lev.P - 1
df.B <- lev.B - 1
df.W <- lev.W - 1
df.PB <- df.P * df.B
df.PW <- df.P * df.W
df.BW <- df.B * df.W
df.PBW <- df.P * df.B * df.W
df.err <- lev.P * lev.B * lev.W * (n - 1)

# strata means for all factors
mu <- mean(data[, 4], na.rm = TRUE)
mu.P <- tapply(data[, 4], data[, 1], mean, na.rm = TRUE)
mu.B <- tapply(data[, 4], data[, 2], mean, na.rm = TRUE)
mu.W <- tapply(data[, 4], data[, 3], mean, na.rm = TRUE)
mu.PB <- tapply(data[, 4], list(data[, 1], data[,
  2]), mean, na.rm = TRUE)
mu.PW <- tapply(data[, 4], list(data[, 3], data[,
  1]), mean, na.rm = TRUE)
mu.BW <- tapply(data[, 4], list(data[, 3], data[,
  2]), mean, na.rm = TRUE)
mu.PBW <- tapply(data[, 4], list(data[, 3], data[,
  1], data[, 2]), mean, na.rm = TRUE)

if (type == "crossed") {

  # Main Effect 1: P
  SS.P <- (sum((mu.P - mu)^2)) * (lev.B * lev.W * n)
  MS.P <- SS.P/df.P

  # Main Effect 2: B
  SS.B <- (sum((mu.B - mu)^2)) * (lev.P * lev.W * n)
  MS.B <- SS.B/df.B

  # Main Effect 3: W
  SS.W <- (sum((mu.W - mu)^2)) * (lev.P * lev.B * n)
  MS.W <- SS.W/df.W

  # 2-way Interaction 1: P*B
  v.muP <- rep(mu.P, lev.B)

```

```

v.muB <- rep.int(mu.B, rep(lev.P, length(mu.B)))
v.muPB <- as.vector(mu.PB)
SS.PB <- n * lev.W * sum((v.muPB - v.muP - v.muB + mu)^2)
MS.PB <- SS.PB/df.PB

# 2-way Interaction 2: P*W
v.muP2 <- rep.int(mu.P, rep(lev.W, lev.P))
v.muW <- rep(mu.W, lev.P)
v.muPW <- as.vector(mu.PW)
SS.PW <- n * lev.B * sum((v.muPW - v.muP2 - v.muW + mu)^2)
MS.PW <- SS.PW/df.PW

# 2-way interaction 3: B*W
v.muW2 <- rep(mu.W, lev.B)
v.muB2 <- rep.int(mu.B, rep(lev.W, lev.B))
v.muBW <- as.vector(mu.BW)
SS.BW <- n * lev.P * sum((v.muBW - v.muB2 - v.muW2 + mu)^2)
MS.BW <- SS.BW/df.BW

# 3-way interaction: P*B*W
v.muP3 <- rep(rep.int(mu.P, rep(lev.W, length(mu.P))), lev.B)
v.muB3 <- rep.int(mu.B, rep(lev.P * lev.W, lev.B))
v.muW3 <- rep(mu.W, lev.P * lev.B)
v.muPB2 <- as.vector(mu.PB[rep.int(1:nrow(mu.PB),
  rep(lev.W, nrow(mu.PB))), ])
v.muPW2 <- rep(as.vector(mu.PW), lev.B)
v.muBW2 <- as.vector(mu.BW[, rep.int(1:ncol(mu.BW),
  rep(lev.P, ncol(mu.BW)))])
v.muPBW <- NULL
for (i in 1:lev.B) v.muPBW <- c(v.muPBW, as.vector(mu.PBW[, i]))
SS.PBW <- n * sum((v.muPBW - v.muPB2 - v.muPW2 -
  v.muBW2 + v.muP3 + v.muB3 + v.muW3 - mu)^2)
MS.PBW <- SS.PBW/df.PBW

# Error component
SS.err <- sum((data[, 4] - v.muPBW)^2)
MS.err <- SS.err/df.err

if (is.nan(MS.err))
  MS.err <- 0

# Variance components estimates
sigma.P <- (MS.P - MS.PW - MS.PB + MS.PBW)/(lev.B *lev.W)
sigma.B <- (MS.B - MS.PB - MS.BW + MS.PBW)/(lev.P *lev.W)
sigma.W <- (MS.W - MS.PW - MS.BW + MS.PBW)/(lev.P *lev.B)
sigma.PB <- (MS.PB - MS.PBW)/lev.W

```

```

sigma.PW <- (MS.PW - MS.PBW)/lev.B
sigma.BW <- (MS.BW - MS.PBW)/lev.P
sigma.PBW <- (MS.PBW - MS.err)/n
sigma.err <- MS.err

sigma <- c(sigma.P, sigma.B, sigma.W, sigma.PB,
          sigma.PW, sigma.BW, sigma.PBW, sigma.err)
sigma.corr <- ifelse(sigma >= 0, sigma, 0)
stdev <- sqrt(sigma.corr)

# building up the ANOVA table
if (is.null(source)) {
  source <- c("P", "B", "W", "P*B", "P*W",
             "B*W", "P*B*W", "Error")
}

varcomp.table <- data.frame(Fonte = source,
  SS = c(SS.P, SS.B, SS.W, SS.PB, SS.PW,
        SS.BW, SS.PBW, SS.err), df = c(df.P,
        df.B, df.W, df.PB, df.PW, df.BW, df.PBW,
        df.err), MS = c(MS.P, MS.B, MS.W, MS.PB,
        MS.PW, MS.BW, MS.PBW, MS.err), VarComp = sigma.corr,
  StDev = stdev, Quota = sigma.corr/sum(sigma.corr))
varcomp.table
}

if (type == "nested.factorial") {
  df.WinP <- df.W + df.PW
  df.WinP.B <- df.PBW + df.BW

  # Main Effect 1: P
  SS.P <- (sum((mu.P - mu)^2)) * (lev.B * lev.W * n)
  MS.P <- SS.P/df.P

  # Main Effect 2: B
  SS.B <- (sum((mu.B - mu)^2)) * (lev.P * lev.W * n)
  MS.B <- SS.B/df.B

  # Nested Effect: WinP
  SS.W <- (sum((mu.W - mu)^2)) * (lev.P * lev.B * n)
  v.muP2 <- rep.int(mu.P, rep(lev.W, lev.P))
  v.muW <- rep(mu.W, lev.P)
  v.muPW <- as.vector(mu.PW)
  SS.PW <- n * lev.B * sum((v.muPW - v.muP2 - v.muW + mu)^2)
  SS.WinP <- SS.W + SS.PW
  MS.WinP <- SS.WinP/df.WinP

```

```

# 2-way interaction1: P*B
v.muP <- rep(mu.P, lev.B)
v.muB <- rep.int(mu.B, rep(lev.P, length(mu.B)))
v.muPB <- as.vector(mu.PB)
SS.PB <- n * lev.W * sum((v.muPB - v.muP - v.muB + mu)^2)
MS.PB <- SS.PB/df.PB

# Nested Interaction Factor: B*W
v.muW2 <- rep(mu.W, lev.B)
v.muB2 <- rep.int(mu.B, rep(lev.W, lev.B))
v.muBW <- as.vector(mu.BW)
SS.BW <- n * lev.P * sum((v.muBW - v.muB2 - v.muW2 + mu)^2)
v.muP3 <- rep(rep.int(mu.P, rep(lev.W, length(mu.P))), lev.B)
v.muB3 <- rep.int(mu.B, rep(lev.P * lev.W, lev.B))
v.muW3 <- rep(mu.W, lev.P * lev.B)
v.muPB2 <- as.vector(mu.PB[rep.int(1:nrow(mu.PB),
  rep(lev.W, nrow(mu.PB))), ])
v.muPW2 <- rep(as.vector(mu.PW), lev.B)
v.muBW2 <- as.vector(mu.BW[, rep.int(1:ncol(mu.BW),
  rep(lev.P, ncol(mu.BW)))])
v.muPBW <- NULL
for (i in 1:lev.B) v.muPBW <- c(v.muPBW, as.vector(mu.PBW[, i]))
SS.PBW <- n * sum((v.muPBW - v.muPB2 - v.muPW2 -
  v.muBW2 + v.muP3 + v.muB3 + v.muW3 - mu)^2)
SS.WinP.B <- SS.PBW + SS.BW
MS.WinP.B <- SS.WinP.B/(df.WinP.B)

# Error
SS.err <- sum((data[, 4] - v.muPBW)^2)
MS.err <- SS.err/df.err

if (is.nan(MS.err))
  MS.err <- 0

# Variance Components Estimates
sigma.P <- (MS.P - MS.WinP - MS.PB + MS.WinP.B)/(lev.B * lev.W)
sigma.B <- (MS.B - MS.PB)/(lev.P * lev.W)
sigma.PB <- (MS.PB - MS.WinP.B)/(lev.W)
sigma.WinP <- (MS.WinP - MS.WinP.B)/lev.B
sigma.WinP.B <- MS.WinP.B
sigma.err <- MS.err
sigma <- c(sigma.P, sigma.B, sigma.PB, sigma.WinP,
  sigma.WinP.B, sigma.err)
sigma.corr <- ifelse(sigma >= 0, sigma, 0)
stdev <- sqrt(sigma.corr)

```

```

# building up the ANOVA table
if (is.null(source)) {
  source <- c("P", "B", "P*B", "WinP", "(WinP)*B", "Error")
}

varcomp.table <- data.frame(Source = source,
  SS = c(SS.P, SS.B, SS.PB, SS.WinP, SS.WinP.B,
    SS.err), df = c(df.P, df.B, df.PB,
    df.WinP, df.WinP.B, df.err), MS = c(MS.P,
    MS.B, MS.PB, MS.WinP, MS.WinP.B, MS.err),
  VarComp = sigma.corr, StDev = stdev,
  Quota = sigma.corr/sum(sigma.corr))
varcomp.table
}
}

```

A.2 Overall CCC - point estimate

```

# Function to estimate the Overall Concordance Correlation Coefficient
# from Barnhart et al. (2002)

# data: a n x p data frame containing measurements on n subjects
# by p observers/raters

occc <- function(data) {
  # define all pair-wise combinations of raters
  comb <- combn(1:ncol(data), 2)

  # initialize vectors for pair-wise CCC
  accuracy <- NULL
  precision <- NULL
  ccc <- NULL
  weights.ccc <- NULL

  # for every pair-wise combination estimate the
  # needed quantities (variances, mean, accuracy,
  # precision, CCC, weight for the OCCC)
  for (i in 1:ncol(comb)) {
    data.ccc <- data[, c(comb[1, i], comb[2, i])]
    var.1 <- var(data.ccc[, 1])
    var.2 <- var(data.ccc[, 2])
  }
}

```

```
mu1 <- mean(data.ccc[, 1], na.rm = TRUE)
mu2 <- mean(data.ccc[, 2], na.rm = TRUE)
prec <- cor(data.ccc[, 1], data.ccc[, 2])
acc <- (2 * sqrt(var.1) * sqrt(var.2))/(var.1 +
  var.2 + (mu1 - mu2)^2)
CCC <- prec * acc
weight <- var.1 + var.2 + (mu1 - mu2)^2
accuracy <- c(accuracy, acc)
precision <- c(precision, prec)
ccc <- c(ccc, CCC)
weights.ccc <- c(weights.ccc, weight)
}

# estimate the OCCC and store results in a list
OCCC <- sum(weights.ccc * ccc)/sum(weights.ccc)
results.pairs <- data.frame(Sample.1 = comb[1,
  ], Sample.2 = comb[2, ], precision = precision,
  accuracy = accuracy, ccc = ccc, weights = weights.ccc)
results.all <- list(ccc.comp = results.pairs, OCCC = OCCC)
return(results.all)
}
```

R functions for Concordance Assessment

B.1 Liao-Capen modified Bland-Altman approach

```
# Function to estimate model parameters from a linear measurement error
# model and to plot agreement intervals according to Liao and Capen
# method

# data: a n x k matrix of n subjects on which k methods were used to
# measure a given quantity

# alpha: the 1 - confidence level for confidence intervals of model
# parameters

# meas.err: a vector of length k giving the measurement error variance
# for the k methods (defaults to NULL)

# variance.ratio: a vector of length k*(k-1)/2 giving the values of
# parameter lambda for different paired comparisons of methods
# (defaults to NULL)

# plot: logical. If TRUE, than agreement intervals are plotted with
# proportion of miRNA lying within (defaults to TRUE)

# mult: logical. If TRUE, the 4 different agreement intervals are plotted,
# else a single plot for each interval is returned (defaults to FALSE)

# NOTE: if meas.err and variance.ratio are not specified, than
# lambda is automatically set to 1 for all pair-wise comparisons

liao.capen = function(data,alpha = 0.05,meas.err = NULL,
                      variance.ratio = NULL,plot=TRUE,mult=FALSE){

  # create index of pairwise comparisons and checks values for lambda
  index = combn(1:ncol(data),2)
  if(is.null(meas.err) & is.null(variance.ratio)){
    meas.err = unlist(rep(1,ncol(data)))
```

```

}
  if(!is.null(meas.err)){
    meas.err = unlist(meas.err)
    if(length(meas.err) != ncol(index)){
      stop("Vector of measurement errors is too short")
    }

    }
  if(is.null(meas.err) & !is.null(variance.ratio)){
    meas.err = NULL
    if(length(variance.ratio) != ncol(index)){
      stop("Vector of measurement errors is too short")
    }
  }
}

list.res = vector(mode="list",length=ncol(index))
n = nrow(data)
t.alpha = qt(1-(alpha/2),df= n-1)
z.alpha = qnorm(1 - (alpha/2))
for (i in 1:ncol(index)){
  data2 = data[,index[,i]]
  if(!is.null(meas.err)){
    vr = meas.err[index[,i]][1]/meas.err[index[,i]][2]
    names(vr) = NULL
  }
  else{
    vr = variance.ratio[i]
  }

  # estimates the parameters, the related CI95% and the agreement intervals
  x = data2[,1]
  y = data2[,2]
  diff = y - x
  mean = apply(data2,1,mean,na.rm = TRUE)
  mu.x = mean(x,na.rm = TRUE)
  mu.y = mean(y,na.rm = TRUE)
  Sxx = sum((x - mu.x)^2)
  Syy = sum((y - mu.y)^2)
  Sxy = sum((x - mu.x)*(y-mu.y))
  b0 = (- (vr*Sxx - Syy) + sqrt((vr*Sxx - Syy)^2 + 4*vr*Sxy^2))/(2*Sxy)
  a0 = mu.y - b0*mu.x
  sigma.X = (1/n)*(Sxx - (Sxy/b0))
  sigma.Y = vr*sigma.X
  delta.nb = c(- t.alpha*sqrt(1 + vr)*sigma.X,t.alpha
               *sqrt(1 + vr)*sigma.X)
  sigma.b = ((1 + vr*b0^2)*(Sxx*Syy - Sxy^2))/((Sxx - vr*Syy)^2

```

```

      + 4*vr*Sxy^2)
sigma.a = sigma.X^2/n + mu.x^2*sigma.b^2
CI.fb = c(a0 - (z.alpha*sqrt(sigma.a/n)),
          a0 + (z.alpha*sqrt(sigma.a/n)))
b0.low = b0 - (z.alpha*sqrt(sigma.b/n))
b0.upp = b0 + (z.alpha*sqrt(sigma.b/n))
CI.pb = data.frame(CI.low = (b0.low - 1)*x, CI.upp = (b0.upp - 1)*x)
delta.fb = c(a0 - t.alpha*sqrt(1 + vr)*sigma.X,
             a0 + t.alpha*sqrt(1 + vr)*sigma.X)
delta.pb = data.frame(pblow = (b0 - 1)*x - t.alpha*sqrt(1 + vr)*sigma.X,
diff = diff, pbupp = (b0 - 1)*x + t.alpha*sqrt(1 + vr)*sigma.X)
delta.pb = delta.pb[order(x),]
b0X = (b0 - 1)*x
b0X = b0X[order(x)]
delta.fb.pb = data.frame(pbfb.low = a0 + (b0 - 1)*x -
                        t.alpha*sqrt(1 + vr)*sigma.X,
                        diff = diff, pbfb.upp = a0 + (b0 - 1)*x +
                        t.alpha*sqrt(1+vr)*sigma.X)
delta.fb.pb = delta.fb.pb[order(x),]
a0.b0X = a0 + (b0 - 1)*x
a0.b0X = a0.b0X[order(x)]
label.x = names(data2)[1]
label.y = names(data2)[2]
min = min(delta.nb,delta.fb,min(delta.pb),min(delta.fb.pb))
max = max(delta.nb,delta.fb,max(delta.pb),max(delta.fb.pb))
nb.agr = sum((diff > delta.nb[1]) & (diff < delta.nb[2]))/n
fb.agr = sum((diff > delta.fb[1]) & (diff < delta.fb[2]))/n
pb.agr = sum((delta.pb$diff > delta.pb$pblow) &
             (delta.pb$diff < delta.pb$pbupp))/n
fbpb.agr = sum((delta.fb.pb$diff > delta.fb.pb$pbfb.low) &
               (delta.fb.pb$diff < delta.fb.pb$pbfb.upp))/n
names(list.res)[i] = paste(names(data2)[2],names(data2)[1],sep=" - ")
list.res[[i]] = list(a0 = a0, b0 = b0, vara0 = sigma.a,
                    varb0 = sigma.b, CI95.a0 = CI.fb,
                    CI95.b0 = c(b0.low,b0.upp),varX = sigma.X,
                    varY = sigma.Y, variance.ratio = vr,
                    delta = delta.nb, agreement = nb.agr,
                    delta.fixed = delta.fb,
                    agreement.fixed = fb.agr, delta.prop = delta.pb,
                    agreement.prop = pb.agr,delta.fixed.prop = delta.fb.pb,
                    agreement.prop.fixed = fbpb.agr)

# plotting facilities for results visualization. On Windows running
# systems quartz() is to be turned into win.graph() for compatibility
if(plot){
quartz()

```

```

if(mult){
par(mfrow=c(2,2))}
plot(1:n,diff,main="No bias",xlab="Samples",
      ylab=paste(label.y," - ",label.x,sep=""),
      ylim = range(min,max), pch="", xaxt="n", las=2, cex=0.45)
points(1:n,diff,pch=21,bg = "blue",col="blue",cex=0.45)
#axis(side=1,at = 1:n,labels=NULL,cex.axis=0.6,las = 2)
abline(h = c(delta.nb,0),lty= c(2,2,1),lwd=1.5)
text(n/2,max*0.9,labels = paste("% in agreement: ",
      round(nb.agr*100,digits=2),sep=""),
      cex=0.8,adj=0.5)

plot(1:n,diff,main="Fixed bias",xlab="Samples",ylab=
      paste(label.y," - ", label.x,sep=""),
      ylim = range(min,max), pch="", xaxt="n", las=2, cex=0.45)
points(1:n,diff,pch=21,bg = "blue",col="blue",cex=0.45)
#axis(side=1,at = 1:n,labels=NULL,cex.axis=0.6,las = 2)
abline(h = c(delta.fb,a0),lty= c(2,2,1),lwd=1.5)
text(n/2,max*0.9,labels = paste("% in agreement: ",
      round(fb.agr*100,digits=2),sep=""),
      cex=0.8,adj=0.5)

plot(1:n,delta.pb$diff,main="Proportional bias",xlab="Sorted Samples",
      ylab=paste(label.y," - ",label.x,sep=""),
      ylim = range(min,max), pch="", xaxt="n", las=2, cex=0.45)
points(1:n,delta.pb$diff,pch=21,bg = "blue",col="blue",cex=0.45)
axis(side=1,at = 1:n,labels=NULL,cex.axis=0.6,las = 2)
lines(1:n,delta.pb$pb_low,lty= 3, lwd=1.5)
lines(1:n,delta.pb$pb_upp,lty= 3, lwd=1.5)
lines(1:n,b0X,lty=1,lwd=1.5)
text(n/2,max*0.9,labels = paste("% in agreement: ",
      round(pb.agr*100,digits=2),sep=""),cex=0.8,adj=0.5)

plot(1:n,delta.fb.pb$diff,main="Proportional and Fixed bias",
      xlab="Sorted Samples",
      ylab=paste(label.y," - ",label.x,sep=""),ylim = range(min,max), pch="",
      xaxt="n", las=2, cex=0.45)
points(1:n,delta.fb.pb$diff,pch=21,bg = "blue",col="blue",cex=0.45)
axis(side=1,at = 1:n,labels=NULL,cex.axis=0.6,las = 2)
lines(1:n,delta.fb.pb$pbfb_low,lty= 3, lwd=1.5)
lines(1:n,delta.fb.pb$pbfb_upp,lty= 3, lwd=1.5)
lines(1:n,a0.b0X,lty=1,lwd=1.5)
text(n/2,max*0.9,labels = paste("% in agreement: ",
      round(fbpb.agr*100,digits=2),sep=""),
      cex=0.8,adj=0.5)
}

```

```
}  
return(list.res)  
}
```

Bibliography

- [1] Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, and Arnheim N. Enzymatic amplification of β -globin genomic sequences and restriction site analysis. *Science*, 230(4732):1350–1354, 1985.
- [2] Schena M, Shalon D, Davis RW, and Brown PO. Monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.
- [3] Morin RD, Bainbridge M, Fejes A, Hirts M, Krzywinski M, Pugh TJ, McDonald H, Varhol R, Jones SJM, and Marra MA. Profiling the hela s3 transcriptome using randomly primed cdna and massively parallel short-read sequencing. *Biotechniques*, 45:81–94, 2008.
- [4] Miska EA, Alvarez-Saavedra E, Townsend M, Yoshii A, Sestan N, Rakic P, Constantin-Paton M, and Horvitz HR. Microarray analysis of microrna expression in the developing mammalian brain. *Genome Biol*, 5:R68, 2004.
- [5] Mei R, Galipeau PC, Prass C, Berno A, Ghandour G, Patil N, Wolff RK, Chee MS, Reid BJ, and Lockhart DJ. Genome-wide detection of allelic imbalance using human snps and high-density dna arrays. *Genome Res*, 10:1126 – 1137, 2000.
- [6] Gitan RS, Shi H, Chen CM, Yan PS, and Huang THM. Methylation-specific oligonucleotide microarray: A new potential for high-throughput methylation analysis. *Genome Res*, 12:158–164, 2002.
- [7] Farber CR and Lusis AJ. Integrating global gene expression analysis and genetics. In Rao DC and Gu CC, editors, *Genetic Dissection of Complex Traits*, Advances in Genetics, pages 571 – 601. Elsevier, 2nd edition, 2008.
- [8] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, and Lander ES. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531 – 537, 1999.
- [9] Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker mG, Watson D, Park T, Hiller W, Fisher ER, Wickerham DL, Bryant J, and Wolmark N. A multi-gene assay to predict recurrence of tamoxifen-treaten, node-negative breast cancer. *N Engl J Med*, 351:2817 – 2826, 2004.
- [10] MicroArray Quality Control Consortium. Making the most out of microarrays. *Nat Biotechnol*, 24:1039, 2006.
- [11] Bolstad BM, Irizarry RA, Astrand M, and Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19:185 – 193, 2003.

- [12] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, and Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249 – 264, 2003.
- [13] Johnson WE, Li C, and Rabinovic A. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8:118 – 127, 2007.
- [14] Chen C, Grennan k, Badner J, Zhang D, Gershon E, Jin L, and Liu C. Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods. *PLoS One*, 6:e17238, 2011.
- [15] Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:3, 2004.
- [16] Kooperberg C, Aragaki A, Strand AD, and Olson JM. Significance testing for small microarray experiments. *Statist Med*, 24:2281 – 2298, 2005.
- [17] Dudoit S, Shaffer JP, and Boldrick JC. Multiple hypothesis testing in microarray experiments. *Stat Sci*, 18:71 – 103, 2003.
- [18] International Organization for Standardization, Geneva, Switzerland. *Accuracy (True-ness and Precision) of Measurement Methods and Results - Parts 1 - 6, ISO9725*, 1994.
- [19] Yauk CL and Berndt ML. Review of the literature examining the correlation among dna microarray technologies. *Environ Mol Mutagen*, 48:380–394, 2007.
- [20] Miron M, Woody OZ, Marcil A, Murie C, Sladek R, and Nadon R. A methodology for global validation of microarray experiments. *BMC Bioinformatics*, 7:333, 2006.
- [21] Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, and Golub TR. MicroRNA expression profiles classify human cancers. *Nature*, 435:843 – 838, 2005.
- [22] Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, Prabhu AL, Zhao Y, McDonald H, Zeng T, Hirst M, Eaves CJ, and Marra M. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res*, 18:610 – 621, 2008.
- [23] L I Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45:255–268, 1989.
- [24] H X Barnhart, M Haber, and J Song. Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics*, 58:1020–1027, 2002.
- [25] J J Z Liao and Capen R. An improved bland-altman method for concordance assessment. *Int J Biostat*, 7:9, 2011.
- [26] Ambros V. MicroRNA pathways in flies and worms: Growth, death, fat, stress, and timing. *Cell*, 113:673–676, 2003.

- [27] Lai E C. micrnas: Runts of the genome assert themselves. *Curr Biol*, 13:R925–R936, 2003.
- [28] Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, Barzilai A, Einat P, Einav U, Meiri E, Sharon E, Spector Y, and Bentwich Z. Identification of hundreds of conserved and nonconserved human micrnas. *Nat Genet*, 37:766–770, 2005.
- [29] Esquela-Kerscher A and Slack F J. Oncomirs - micrnas with a role in cancer. *Nat Rev Cancer*, 6:259–269, 2006.
- [30] Hebert S S, Horre S, Nicolai S, Bergmans B, Papadopoulou A S, Delacourte A, and De Strooper B. Micrna regulation of alzheimer’s amyloid precursor protein expression. *Neurobiol Dis*, 33:422–428, 2009.
- [31] Dehwah M A and Huang Q. Micrnas and type 2 diabetes/obesity. *J Genet Genomics*, 39:11–18, 2012.
- [32] Paraboschi E M, Solda’ G, Gemmati D, Orioli E, Zeri G, Benedetti M D, Salviati A, Barizzone N, Leone M, Duga S, and Asselta R. Genetic association and altered gene expression of mir-155 in multiple sclerosis patients. *Int J Mol Sci*, 12:8695–8712, 2011.
- [33] Fasanaro P, Greco S, Ivan M, Capogrossi M C, and Martelli F. micrna: Emerging therapeutic targets in acute ischemic diseases. *Pharmacol Ther*, 125:92–104, 2010.
- [34] Chen C Z. Micrnas as oncogenes and tumor suppressors. *N Engl J Med*, 353:1768–1771, 2005.
- [35] Lee R C, Feinbaum R L, and Ambros V. The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. *Cell*, 75:843–854, 1993.
- [36] Chen C, Ridzon D A, Broomer A J, Zhou Z, Lee D H, Nguyen J T, Barbisin M, Xu N L, Mahuvakar V R, Andersen M R, Lao K Q, Livak K J, and Guegler K J. Real-time quantification of micrnas by stem-loop rt-pcr. *Nucleic Acids Res*, 33:e179, 2005.
- [37] Gentleman R, Carey V J, Huber W, Irizarry R A, and Dudoit S. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, 2005.
- [38] Lee M L T. *Analysis of Microarray Gene Expression Data*. Kluwer Academic Publisher, 2004.
- [39] Huang J C, Morris Q D, Hughes T R, and Frey B J. Genxhc: a probabilistic generative model for cross-hybridization compensation in high-density genome-wide microarray data. *Bioinformatics*, 21:i222–i231, 2005.
- [40] Okoniewski M J and Miller C J. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics*, 7:276, 2006.
- [41] Klebanov L, Chen L, and Yakovlev A. Revisiting adverse effect of cross-hybridization in affymetrix gene expression data: do they matter for correlation analysis? *Biol Direct*, 2:28, 2007.

- [42] Affymetrix Corporation. *Data sheet - GeneChip Human Genome Arrays*.
- [43] <http://genecards.weizmann.ac.il/geneannot/>.
- [44] Bassani N, Ambrogi F, Coradini D, Boracchi P, and Biganzoli E. Validation of gene expression profile in genomic data through complementary use of cluster analysis and pca-related biplots. *Int J Stats Med Res*, 1:162–173, 2012.
- [45] Scherer S. *Guide to the Human Genome*. Cold Spring Harbor Laboratory Press, 2010.
- [46] Yin J Q, Zhao R C, and Morris K V. Profiling microRNA expression with microarrays. *Trends Biotechnol*, 26:70–76, 2008.
- [47] Zhao B, Ding S, Li W, and Jin Y. Hybridization kinetics analysis of an oligonucleotide microarray for microRNA detection. *Acta Biochim Biophys Sin*, 43:551–555, 2011.
- [48] Duan D, Zheng K, Shen Y, Cao R, Jiang L, Lu Z, Yan X, and Li J. Label-free high-throughput microRNA expression profiling from total rna. *Nucleic Acids Res*, 39:e154, 2011.
- [49] Dunn G. Review papers: Design and analysis of reliability studies. *Stat Methods Med Res*, 1:123–157, 1992.
- [50] Altman D. *Practical Statistics for Medical Research*. Chapman & Hall, 1991.
- [51] Mullins E. *Statistics for the Quality Control Chemistry Laboratory*. The Royal Society of Chemistry, 2003.
- [52] The power of replicates, 2010.
- [53] Cronbach L J, Gleser G C, Nanda H, and Rajaratnam N. *The Dependability of Behavioural Measurements: Theory of Generalizability for Scores and Profiles*. Wiley, New York, 1972.
- [54] Rao C R and Sinharay S, editors. *Handbook of Statistics, Volume 26: Psychometrics*. Elsevier, 2007.
- [55] Fisher R A. *Statistical Methods for Research Workers*. Oliver and Boy Ltd, Edinburgh, 11th edition, 1950.
- [56] Müller R and Büttner P. A critical discussion of intraclass correlation coefficients. *Stat Med*, 13:2465–2476, 1994.
- [57] Armitage P and Colton T, editors. *Encyclopedia of Biostatistics*. Wiley, 2nd edition, 2005.
- [58] Cesana B M. Letter: Comparison of measurement methods requires appropriate statistical methodology. *Am J Ophthalmol*, 137:787–788, 2004.
- [59] Atkinson G and Nevill A. Comment on the use of concordance correlation to assess the agreement between two variables. *Biometrics*, 53:775–777, 1997.

- [60] I-Kuei L and Chinchilli V. Rejoinder to the letter to the editor from atkinson and nevell. *Biometrics*, 53:777–778, 1997.
- [61] Altman D G and Bland J M. Measurement in medicine: the analysis of method comparison studies. *Statistician*, 32:307–317, 1983.
- [62] Bland J M and Altman D G. Statistical method for assessing agreement between two methods of clinical measurement. *Lancet*, 1:307–310, 1986.
- [63] Bland J M and Altman D G. Measuring agreement in method comparison studies. *Stat Methods Med Res*, 8:136–160, 1999.
- [64] Carstensen B. Comparing methods of measurement: Extending the loa by regression. *Stat Med*, 29:401–410, 2010.
- [65] Patterson T A, Lobenhofer E K, Fulmer-Smentek S B, Collins P J, Chu T M, Bao W, Fang H, Kawasaki E S, Hager J, Tikhonova I R, Walker S J, Zhang L, Hurban P, de Longueville F, Fuscoe J C, Tong W, Shi L, and Wolfinger R D. Performance comparison of one-color and two-color platforms within the microarray quality control (maq) project. *Nat Biotechnol*, 24:1140–1150, 2006.
- [66] Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*, 57: 289–300, 1995.
- [67] Tan P K, Downey T J, Spitznagel Jr E L, Xu P, Fu D, Dimitrov D S, Lempicki R A, Raaka B M, and Cam M C. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res*, 31:5676–5684, 2003.
- [68] Wang H, He X, Band M, Wilson C, and Liu L. A study of inter-lab and inter-platform agreement of dna microarray data. *BMC Genomics*, 6:71, 2005.
- [69] Li C and Wong W H. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc Natl Acad Sci USA*, 98:31–36, 2001.
- [70] de Reynies A, Geromin D, Cayuela J M, Petel F, Dessen P, Sigaux F, and Rickman D S. Comparison of the latest commercial short and long oligonucleotide microarray technologies. *BMC Genomics*, 7:51, 2006.
- [71] Kuo W P, Jenssen T K, Butte A J, Ohno-Machado L, and Kohane I S. Analysis of matched mrna measurements from two different microarray technologies. *Bioinformatics*, 18:405–412, 2002.
- [72] Yuen T, Wurmbach E, Pfeiffer R L, Ebersole B J, and Sealfon S C. Accuracy and calibration of commercial oligonucleotide and custom cdna microarrays. *Nucleic Acids Res*, 30:e48, 2002.
- [73] Lee J K, Bussey K J, Gwadry F G, Reinhold W, Riddick G, Pelletier S L, Nishizuka S, Szakacs G, Annereaus J P, Shankavaram U, Lababidi S, Smith L H, Gottesman M M, and Weinstein J N. Comparing cdna and oligonucleotide array data: concordance of gene expression across platform for the nci-60 cancer cells. *Genome Biol*, 4:R82, 2003.

- [74] Järvinen A K, Hautaniemi S, Edgren H, Auvinen P, Saarela J, Kallioniemi O P, and Monni O. Are data from different gene expression microarray platforms comparable? *Genomics*, 83:1164–1168, 2004.
- [75] Mah N, Thelin A, Lu T, Nikolaus S, Kuhbacher T, Gurbuz Y, Eickhoff H, Kloppel G, Lehrach H, Mellgard B, Costello C M, and Schreiber S. A comparison of oligonucleotide and cdna-based microarray systems. *Physiol Genomics*, 16:361–370, 2004.
- [76] Park P J, Cao Y A, Lee S Y, Kim J W, Chang M S, Hart R, and Choi S. Current issues for dna microarrays: platform comparison, double linear amplification and universal rna reference. *J Biotechnol*, 112:225–245, 2004.
- [77] Larkin J K, Frank B C, Gavras H, Sultana R, and Quackenbush J. Independence and reproducibility across microarray platforms. *Nat Methods*, 2:337–343, 2005.
- [78] Draghici S, Khatri P, Eklund AC, and Szallasi Z. Reliability and reproducibility issues in dna microarray measurements. *Trends Genet*, 22:101 – 109, 2006.
- [79] Chen J J, Hsueh H M, Delongchamp R R, Lin C J, and Tsai C A. Reproducibility of microarray data: a further analysis of microarray quality control (maq) data. *BMC Bioinformatics*, 8:412, 2007.
- [80] Sato F, Tsuchiya S, Terasawa K, and Tsujimoto G. Intra-platform repeatability and inter-platform comparability of microrna microarray technology. *PLoS ONE*, 4:e5540, 2009.
- [81] Yauk C L, Rowan-Carroll A, Stead J D H, and Williams A. Cross-platform analysis of global microrna expression technologies. *BMC Genomics*, 11:330, 2010.
- [82] Liao C T, Lin C Y, and Liu J P. Noninferiority tests based on concordance correlation coefficient for assessment of the agreement for gene expression data from microarray experiments. *J Biopharm Stat*, 17:309–327, 2007.
- [83] Giard D J, Aaronson S A, Todaro G J, Kersey J H, Dosik H, and Parks W P. In vitro cultivation of human tumors: establishment of cell lines derived from a series of solid tumors. *J Natl Cancer Inst*, 51:1417–1423, 1973.
- [84] <http://www.mirbase.org/>.
- [85] Affymetrix, Santa Clara, California. *Affymetrix[©] miRNA QC Tool Guide*, 2008.
- [86] Holder D, Raubertas R F, Pikounis V B, Svetnik V, and Soper K. Statistical analysis of high density oligonucleotide arrays: a safer approach. In *Proceedings of the ASA Annual Meeting*. American Statistical Association, 2001.
- [87] Agilent Technologies. *Agilent Feature Extraction Reference Guide*, 2007.
- [88] P López-Romero, M A González, S Callejas, A Dopazo, and R A Irizarry. Processing of agilent microrna array data. *BMC Research Note*, 3:18, 2010.
- [89] R Foundation for Statistical Computing, Vienna, Austria. *R: A Language and Environment for Statistical Computing*, 2011.

- [90] P López-Romero. Pre-processing and differential expression analysis of agilent mi-crona arrays using the *agimicrona* bioconductor library. *BMC Genomics*, 12:64, 2011.
- [91] R C Gentleman, V J Carey, D M Bates, B Bolstad, M Dettling, S Dudoit, B Ellis, L Gautier, Y Ge, J Gentry, K Hornik, T Hothorn, W Huber, S Iacus, R Irizarry, F Leisch, C Li, M Maechler, A J Rossini, G Sawitzki, C Smith, G Smyth, L Tierney, J Y Yang, and J Zhang. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol*, 5:R80, 2004.
- [92] Dunning M. *Analysis of Bead-summary data using beadarray*, October 2012.
- [93] Ritchie M E, Silver J, Oshlack A, Holmes m, Diyagama D, Holloway A, and Smyth G K. A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, 23:2700–2707, 2007.
- [94] Hackstadt A J and Hess A M. Filtering for increased power for microarray data analysis. *BMC Bioinformatics*, 10:11, 2009.
- [95] Archer K J and Reese S E. Detection call algorithms for high-throughput gene expression microarray data. *Brief Bioinform*, II:244–252, 2009.
- [96] McClintick J N and Edenberg H J. Effects of filtering by present call on analysis of microarray experiments. *BMC Bioinformatics*, 7:49, 2006.
- [97] Bourgon R, Gentleman R, and Huber W. Independent filtering increases detection power for high-throughput experiments. *Proc Natl Acad Sci USA*, 107:9546–9551, 2010.
- [98] Calza S, Raffelsberger W, Ploner A, Sahel J, Leveillard T, and Pawitan Y. Filtering genes to improve sensitivity in oligonucleotide microarray data analysis. *Nucleic Acids Res*, 35:e102, 2007.
- [99] Y J Hua, K Tu, Z Y Tang, Y X Li, and H S Xiao. Comparison of normalization methods with microrna microarray. *Genomics*, 92:122–128, 2008.
- [100] Y Rao, Y Lee, and D Jarjoura. A comparison of normalization techniques for microrna microarray data. *Stat Appl Mol Genet Biol*, 7:22, 2008.
- [101] S Pradervand, J Weber, J Thomas, M Bueno, P A Wirapati, K Lefort, G P Dotto, and K Harshman. Impact of normalization on mirna microarray expression profiling. *RNA*, 15:493–501, 2009.
- [102] B Wang, X F Wang, P Howell, X Qian, K Huang, A I Riker, J Ju, and Y Xi. A personalized microrna microarray normalization method using a logistic regression model. *Bioinformatics*, 26:228–234, 2010.
- [103] C Suo, A Salim, K S Chia, Y Pawitan, and S Calza. Modified least-variant set normalization for mirna microarray. *RNA*, 16:2293–2303, 2010.

- [104] Huber W, von Heydebreck A, Sültmann H, Poutska A, and Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18:S96–S104, 2002.
- [105] Dudoit S, Yang Y H, Callow M J, and Speed T P. Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. *Stat Sin*, 12:111–139, 2002.
- [106] Cleveland W S and Devlin S J. Locally-weighted regression: an approach to regression analysis by local fitting. *J Am Stat Assoc*, 83:596–610, 1988.
- [107] Laird N M and Ware J H. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.
- [108] Satterthwaite F E. An approximate distribution of estimates of variance components. *Biometr Bull*, 2:110–114, 1946.
- [109] Montgomery D C. *Design and Analysis of Experiments*. Wiley, New York, 7th edition, 2009.
- [110] Burch B D. Nonparametric bootstrap confidence intervals for variance components applied to interlaboratory comparisons. *J Agric Biol Environ Stat*, 17:228–245, 2012.
- [111] Efron B and Tibshirani R J. *An introduction to the Bootstrap*. New York: Chapman & Hall/CRC, 1993.
- [112] King T S, Chinchill V M, and Carrasco J L. A repeated measures concordance correlation coefficient. *Stat Med*, 26:3095–3113, 2007.
- [113] Carstensen B, Simpson J, and Gurrin L C. Statistical models for assessing agreement in method comparison studies with replicate measurements. *Int J Biostat*, 4:16, 2008.
- [114] Bland J M and Altman D G. Measurement error and correlation coefficients. *B M J*, 313:41–42, 1986.
- [115] Carroll R J, Ruppert D, Stefanski L A, and Crainiceanu C M. *Measurement Error in Nonlinear Models: A Modern Perspective*. New York: Chapman & Hall/CRC, 2006.
- [116] Fuller W A. *Measurement Error Models*. New York: John Wiley & Sons, 1987.
- [117] Casella G and Berger R L. *Statistical Inference*. Belmont, California: Duxbury Press, 1990.
- [118] Liao J J Z, Capen R C, and Schofield T L. Assessing the reproducibility of an analytical method. *J Chromatogr Sci*, 44:119–122, 2006.
- [119] Lopez-Romero P. *AgiMicroRna: Processing and Differential Expression Analysis of Agilent microRNA chips*. R package version 2.8.0.
- [120] Du P, Kibbe W A, and Lin S M. lumi: a pipeline for processing illumina microarray. *Bioinformatics*, 24:1547–1548, 2008.

- [121] Benjamin Milo Bolstad. *preprocessCore: A collection of pre-processing functions*. R package version 1.20.0.
- [122] Bates D, Maechler M, and Bolker B. *lme4: Linear mixed-effects models using Eigen and Eigen*, 2012. URL <http://CRAN.R-project.org/package=lme4>. R package version 0.999999-0.
- [123] Mark Stevenson with contributions from Telmo Nunes, Javier Sanchez, Ron Thornton, Jenö Reiczigel, Jim Robison-Cox, and Paola Sebastiani. *epiR: An R package for the analysis of epidemiological data*, 2012. URL <http://CRAN.R-project.org/package=epiR>. R package version 0.9-45.
- [124] Josep Lluís Carrasco, Josep Puig Martínez, and Lluís Jover. *cccrm: CCC-RM*, 2012. URL <http://CRAN.R-project.org/package=cccrm>. R package version 1.1.
- [125] Canty A and Ripley B. *boot: Bootstrap R (S-Plus) Functions*, 2012. R package version 1.3-7.
- [126] Nobre J S and Singer J dM. Residual analysis for linear mixed models. *Biom J*, 49: 863–875, 2007.
- [127] Rocke D M and Durbin B. Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics*, 19:966–972, 2003.
- [128] Durbin B and Rocke D M. Estimation of transformation parameters for microarray data. *Bioinformatics*, 19:1360–1367, 2003.
- [129] Scherer A, editor. *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*. Wiley, 2009.
- [130] Clopper C and Pearson E S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26:404–413, 1934.
- [131] Irizarry R A, Bolstad B M, Collin F, Cope L M, Hobbs B, and Speed T P. Summaries of affymetrix genechip probe level data. *Nucleic Acids Res*, 31:e15, 2003.
- [132] Chung Y, Rabe-Hesketh S, Gelman A, Liu J, and Dorie V. Avoiding boundary estimates in linear mixed models through weakly informative priors. Technical Report 284, U.C. Berkeley Division of Biostatistics, 2012.
- [133] Berry S M, Carroll R J, and Ruppert D. Bayesian smoothing and regression splines for measurement error problems. *J Am Stat Assoc*, 97:160–168, 2002.

Ringraziamenti

Il primo e piu' grande GRAZIE e' sicuramente per Elisa, mia moglie, il mio piu' grande "successo" di questi anni, che mi e' sempre stata vicina in questi anni e mi ha sostenuto e sopportato anche quando probabilmente sarebbe stato piu' facile (e piu' giusto) mandarmi a quel paese.

Un enorme ringraziamento va poi a tutti i miei parenti, vicini e lontani, naturali e acquisiti, perche' non e' facile cercare di seguirmi quando provo a rispondere alla domanda "Ma di cosa ti stai occupando adesso al lavoro?", e perche' sicuramente se sono arrivato fin qui e' merito anche loro.

Ci sono poi gli amici di una vita, quelli che conosci da quando hai memoria e con cui sei cresciuto (almeno in altezza), e quelli che conosci solo (!!) da una decina d'anni ma ti sembra che sia molto di piu'. Non posso citarli tutti, ma li voglio ricordare qui, perche' anche loro hanno almeno una volta nella vita dovuto sorbirsi i miei discorsi "sull'importanza della statistica" e sul senso del "capire quello che stai facendo".

Come non ringraziare Anna, la mia testimone di nozze, che mi si e' sorbita nel corso della scrittura di questa tesi. A buon rendere, ma chere amie!

Ringrazio enormemente tutti i miei colleghi, passati e presenti, per il sostegno, sia scientifico che umano, e per avere diligentemente ascoltato i miei dubbi ed esserci concessi qualche risata assieme. In particolare, vorrei ringraziare il mio tutor, il Prof. Elia Biganzoli, che fra un impegno e l'altro e' riuscito anche a trovare il tempo di seguirmi e di darmi continue opportunita' di crescita, sia professionale che personale. E grazie anche al dott. Federico Ambrogi, che fra un corso ai podologi e uno ai fisioterapisti riusciva a dare un occhio ai miei infiniti report tecnici.

Un ringraziamento un po' particolare lo voglio fare anche alle mie fedeli scarpe da corsa e ai miei scarponi da montagna, che in questi anni di Dottorato mi hanno portato ovunque glielo chiedessi. E alle mie bici, che mi hanno scarrozzato per tutte le strade dove le ho portato senza mai lamentarsi (troppo). Anche loro hanno avuto la loro parte, in fin dei conti.

Un ultimo GRAZIE lo rivolgo ad una cara amica che non c'e' piu'. Mi ha insegnato la fiducia in me stesso, il donarsi senza riserve e il mettere sempre tutto te stesso in cio' che fai, al 100%. Questa tesi e' anche per lei. Ciao Elena, e grazie di cuore. Per tutto.