Biology Direct

REVIEW

CrossMark

Microarray experiments and factors which affect their reliability

Roman Jaksik[1*], Marta Iwanaszko[1,2,3], Joanna Rzeszowska-Wolny[1] and Marek Kimmel[1,3]

Abstract

Oligonucleotide microarrays belong to the basic tools of molecular biology and allow for simultaneous assessment of the expression level of thousands of genes. Analysis of microarray data is however very complex, requiring sophisticated methods to control for various factors that are inherent to the procedures used. In this article we describe the individual steps of a microarray experiment, highlighting important elements and factors that may affect the processes involved and that influence the interpretation of the results. Additionally, we describe methods that can be used to estimate the influence of these factors, and to control the way in which they affect the expression estimates. A comprehensive understanding of the experimental protocol used in a microarray experiment aids the interpretation of the obtained results. By describing known factors which affect expression estimates this article provides guidelines for appropriate quality control and pre-processing of the data, additionally applicable to other transcriptome analysis methods that utilize similar sample handling protocols.

Reviewers: This article was reviewed by Dr. Janet Siefert, Dr. Leonid Hanin, and Dr. I King Jordan.

Keywords: Microarrays, Microarray pre-processing, Quality control, Transcriptome profiling, Measurement bias

Introduction

Oligonucleotide microarrays belong to the most common tools used to describe changes in gene expression levels caused by altering the physical or chemical conditions. Microarrays can be also used to track differential expression patterns among various tissues and thus evaluate variability among individuals [1–3], they are used in SNP (single-nucleotide polymorphism) genotyping [4–7] and identification of transcription factor binding sites using the ChIP-chip (ChIP: chromatin immunoprecipitation) method [8–12]. Microarrays are also used to estimate genomic copy number using Comparative Genomic Hybridization (CGH) arrays [13–16] and in resequencing [17–22].

Microarray analysis offers a variety of methods allowing, among other, identification of genes which might be significant in a specific cellular response mechanism or a particular gene expression pattern that characterizes a particular disease. To obtain significant results, microarray data need to undergo statistical processing to differentiate between signal changes caused by direct experimental factors and arising from the indirect experimental factors such as specific methods used, as well as from inaccuracies of the measurements. This level of processing challenges led to studies of the compatibility of different microarray platforms [23–28] which usually is achieved by standardizing protocols and data analysis pipelines [29, 30]. Selection of an appropriate statistical method for microarray processing is a significant subject of scientific discussion and although microarrays have been in use for more than fifteen years, many issues related to data analysis remain unresolved.

The most discussed issues concern the algorithms used for the data normalization [31, 32], whose goal is to eliminate differences between samples that originate from technical aspects of the microarray handling which may confound the biological differences in a given experimental setup. A similar goal underlies methods used for batch-effect removal, a step which is crucial when comparing datasets that originate from different times and laboratories [33]. Other frequently-discussed issues concern the identification of sample differentiating genes [34, 35] and evaluation of noise level in the sample [36], as well as methods to evaluate contamination or damage

* Correspondence: roman.jaksik@polsl.pl
[1]Systems Biology Group, Faculty of Automatic Control, Electronics and Informatics, Silesian University of Technology, Gliwice, Poland
Full list of author information is available at the end of the article

Jaksik *et al. Biology Direct* (2015) 10:46

Page 2 of 14

on the microarray's surface [37, 38]. The most commonly used microarrays, produced by Affymetrix, are known for additional issues related to their particular design which influence the final results. These include problems resulting from several measurements of expression level for a single gene [39, 40], incorrect assignments of probes to genes [41, 42], incorrect evaluation of the background level and non-specific probe hybridization signals [43], and the effects of distinct probe features on data processing algorithms [44].

The most significant disadvantages of microarrays include the high cost of a single experiment, the large number of probe designs based on sequences of low-specificity, as well as the lack of control over the pool of analyzed transcripts since most of the commonly used microarray platforms utilize only one set of probes designed by the manufacturer. Other weaknesses of microarrays are their relatively low accuracy, precision and specificity [45] as well as the high sensitivity of the experimental setup to variations in hybridization temperature [46], the purity and degradation rate of genetic material [47], and the amplification process [48] which, together with other factors, may impact the estimates of gene expression.

Review
Microarray structure
A typical microarray consists of oligonucleotides which are several dozen nucleotides (nt) long attached to the surface of a glass slide. Using appropriate photolithographic masks, a single nucleotide A, C, T, or G is attached at a time, and therefore it is possible to construct a microarray with hundreds of thousands of different oligonucleotide sequences which are complementary to characteristic fragments of known DNA or RNA sequences. These characteristic fragments are arranged in sets called probes [49]. A sample containing DNA or RNA molecules is spread on the surface of a microarray and its components hybridize specifically with their complementary probes, which are located in multiple copies across the microarray (Fig. 1). The amount of material hybridized to a given probe is determined by a fluorescence-based method and although the relationship is not linear the fluorescence intensity reflects the amount of DNA or RNA of a given gene in the sample [50]. This approach allows quantifying the level of transcripts of thousands of genes in a relatively short time.

The most widespread microarray is the Affymetrix 3′IVT (3′ *in vitro* transcription), i.e. HG-U133A or HG-U133_Plus_2, which is assembled as 11 sets of perfect match (PM) probes consisting of 25 nt sequences, which in most cases were chosen out of 600 nt sequence fragments located near the 3′ end of a specific transcript. For every PM probe on the microarray, a MM (mismatch) probe exists in which all nucleotides but one are identical to those on the corresponding PM probe but the original 13th nucleotide is replaced by a non-complementary one. The rationale behind the MM probes is to gauge the level of nonspecific hybridization [51], although the usefulness of this concept has been doubted (see further on).

The most recent generation of Affymetrix microarrays, such as the HuGene 1.0ST, is constructed using probes similar to the standard PM probes but with affinity not to the noncoding part of the 3′ end but rather to the
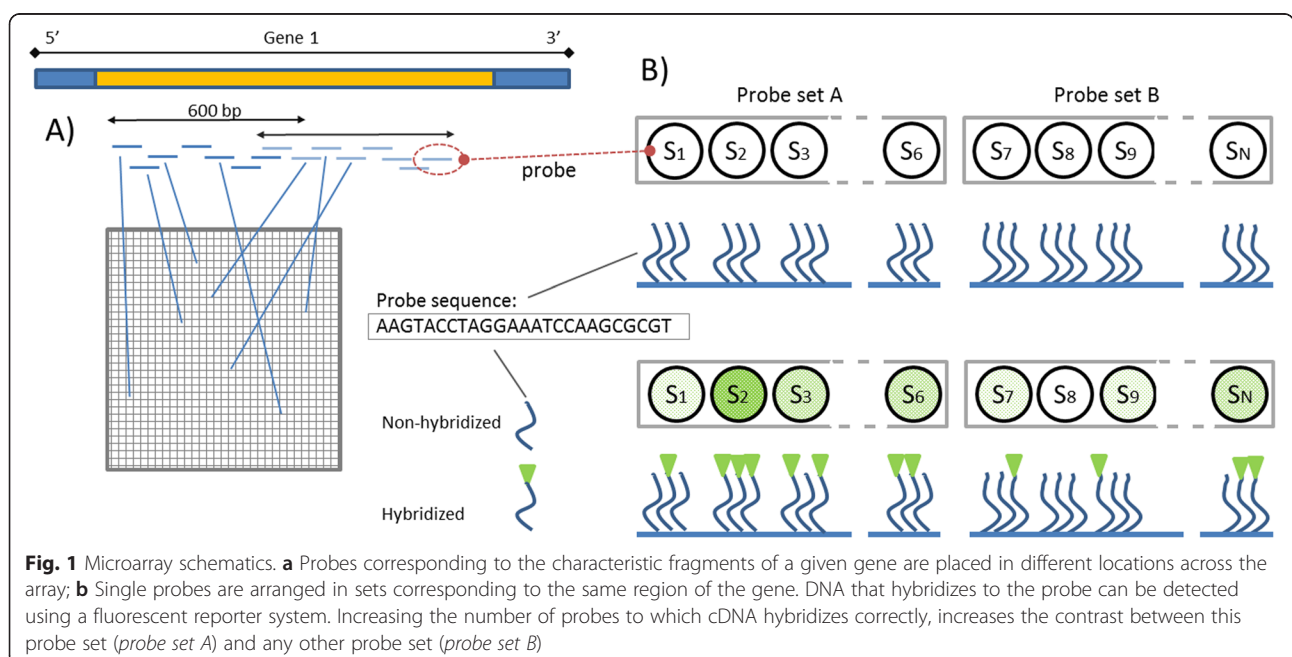

Fig. 1 Microarray schematics. **a** Probes corresponding to the characteristic fragments of a given gene are placed in different locations across the array; **b** Single probes are arranged in sets corresponding to the same region of the gene. DNA that hybridizes to the probe can be detected using a fluorescent reporter system. Increasing the number of probes to which cDNA hybridizes correctly, increases the contrast between this probe set (*probe set A*) and any other probe set (*probe set B*)

Jaksik *et al. Biology Direct* (2015) 10:46

Page 3 of 14

individual exons in a given transcript. In this design the MM probes are replaced by the Background Intensity Probes (BGP), which are designed to evaluate background intensity levels for probes of different sequence characteristics. BGP are a set of about 1000 probes, non-complementary to any human gene sequence, with a variable ratio of GC nucleotides in the sequence. This approach enables a better evaluation of non-specific hybridization across the microarray compared with MM probes, for which the signal often exceeds the PM signal due to probe-specific effects [52]. Additionally, lowering the number of probes which evaluate non-specific hybridization allows inserting of a much higher number of PM probes. The probe set in the new generation of whole transcript microarrays is constructed with two levels, exon and gene level. The exon probe set includes 4 probes on average, which are tailored for individual exons, and then these are clustered, usually in groups of around 25, creating sets for individual genes. Using this approach it is possible to determine levels of individual differently-spliced transcripts.

Another popular system is the Agilent microarray platform which was built using the SurePrint technology that allows using considerably longer, 60 nt-long probes. While probes are longer than in the Affymetrix system, the number of probes per gene is considerably lower, 8 on average in the most expensive set of exon microarrays (2×400 k) or 2 in the least expensive platform (8×60 k). As the Agilent probes are longer than those in the Affymetrix microarrays, the system tends to be more specific which is an obvious advantage, but on the other hand the lower number of probes per gene makes Agilent microarrays more sensitive to single nucleotide variations. These latter should not affect the signal if they result from amplification errors [53], but they may influence the expression estimates resulting from characteristic features of the sample analyzed. In the case of the Affymetrix microarray system these sources of error will only have a minor impact, as they influence signal only in an individual probe for a transcript or a transcript-specific probe-set. Single nucleotide polymorphisms do not block the hybridization but lower its efficiency, which can be interpreted as a significant decrease of gene expression, a feature which is used to estimate the level of nonspecific hybridization using mismatch probes [54, 55] or to assess allelic frequencies using SNP microarrays [56]. In the Affymetrix systems the signal from one badly designed probe, which may be based on inaccurate data from a sequence database, can be easily eliminated from further analysis [41] without significant decrease in the precision of gene expression estimate, while in the Agilent systems the same design glitch might cause significant difficulties in the evaluation of gene expression levels.

Microarrays provide expression data for thousands of genes, but platform differences contribute to low accuracy of microarrays and for this reason they are only used to identify potentially significant genes in the experimental conditions studied. Precise assessment of the expression level of these presumably significant genes requires additional studies using more accurate methods such as real-time PCR (polymerase chain reaction) which, in turn, are not suitable for large-scale analyses. However, some steps of the microarray protocol are shared by the validation methods, affecting data quality in a similar manner.
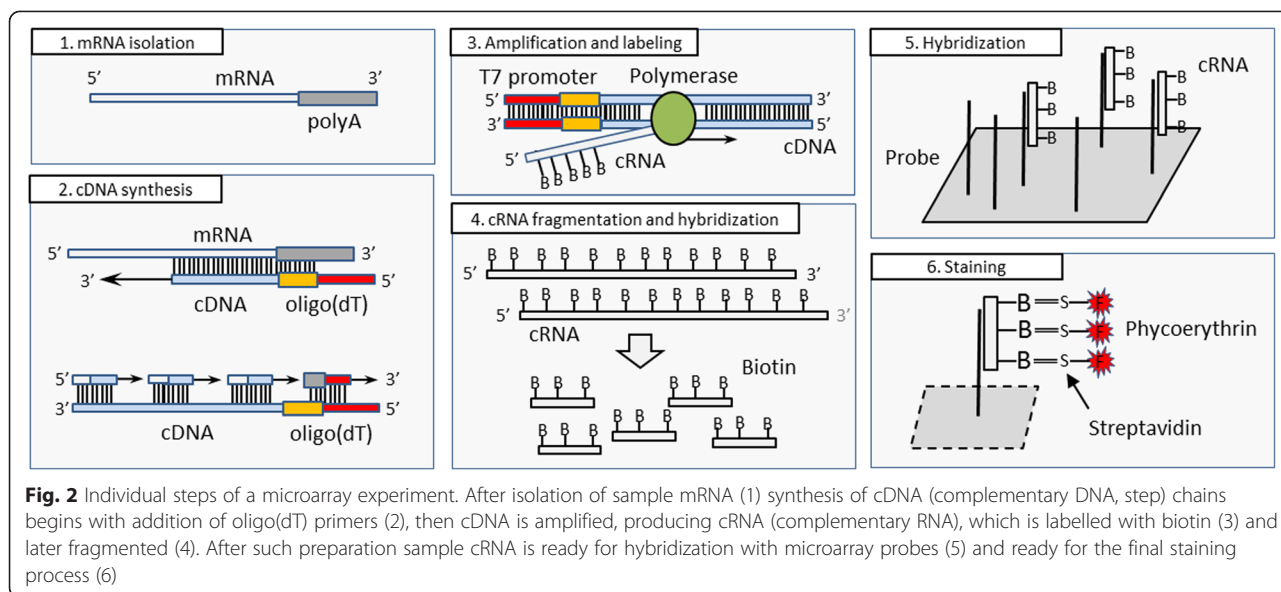
Biological background of microarray experiments

The microarray experiment is a multi-stage process in which the accuracy of each individual step may influence the gene expression estimates. Precise understanding of each step is very important not only for the experimenter but also for the person performing data preprocessing. In order to avoid mistakes that occur during the experiment, its accuracy and the condition of the biological material are controlled in various steps, as shown in Fig. 2. The procedure used in a microarray experiment is very similar across different platforms, and therefore for simplicity the following description is based on the procedure used for the Affymetrix microarrays.
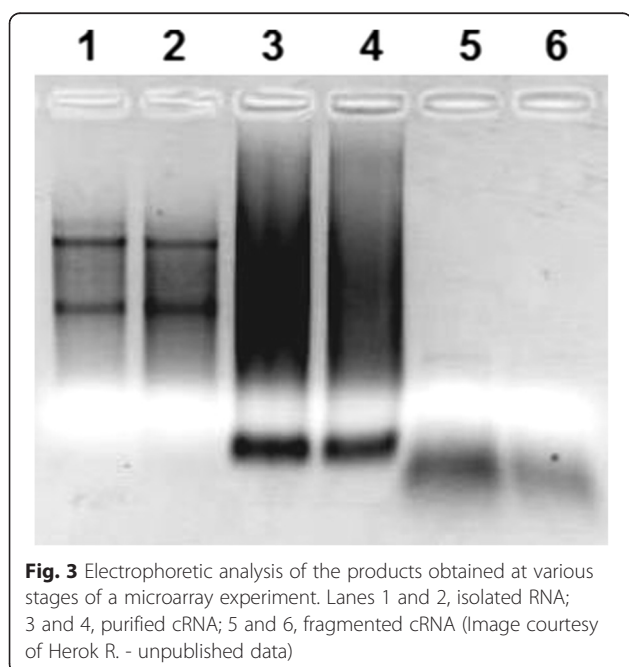
Step I: RNA isolation

In the first step RNA is isolated from the cells and its concentration and extent of degradation is controlled by the use of a spectrophotometer (quantity) and a bioanalyzer (quality). In a high-quality RNA sample ribosomal RNA (rRNA) constitutes over 80 % of the entire RNA, and despite the fact that it is rarely a target of a study in fields other than bacteriology and phylogenetics, its concentration is a good indicator of the overall RNA quality, both before and after the experiment. Prior to hybridization the extent of RNA degradation can be assessed by RNA electrophoresis or a bioanalyzer using the RNA integrity number (RIN) as a benchmark [57]. Fig. 3 shows an example of an image made after electrophoresis in agarose gel. The first two lanes show RNA directly after isolation. The two most distinguishable bands correspond to the 18 and 28S rRNA and their absence would indicate that RNA were highly degraded [58].

RNA quality can also be evaluated after the experiment by analyzing results from a specific group of control probe-sets (as described in Table 1) designed to target certain housekeeping genes (group 1) or rRNA (group 2). As with other probe-sets a single expression intensity is however non-informative since its value, expressed in arbitrary units, depends on the characteristics of the sample [59], the experimental conditions,

Jaksik *et al. Biology Direct* (2015) 10:46

Page 4 of 14

Fig. 2 Individual steps of a microarray experiment. After isolation of sample mRNA (1) synthesis of cDNA (complementary DNA, step) chains begins with addition of oligo(dT) primers (2), then cDNA is amplified, producing cRNA (complementary RNA), which is labelled with biotin (3) and later fragmented (4). After such preparation sample cRNA is ready for hybridization with microarray probes (5) and ready for the final staining process (6)

such as for example the ozone level in the laboratory [60], and the data pre-processing methods used [61]. For this reason arbitrary criteria based solely on single expression intensities are usually ineffective and a comparison between arrays or between probe-sets is required.

Each of the control probe-sets exists in three variants, each targeting a different region of the selected transcript - its central section and the 3′- and 5′-ends. This allows assessing the degradation rate of individual transcripts by examining the 3′/5′ probe-set signal ratios,

Fig. 3 Electrophoretic analysis of the products obtained at various stages of a microarray experiment. Lanes 1 and 2, isolated RNA; 3 and 4, purified cRNA; 5 and 6, fragmented cRNA (Image courtesy of Herok R. - unpublished data)

which can be compared to the threshold defined by the manufacturer and ratios obtained for other microarrays, in order to assess the homogeneity of degradation level across individual samples. In order to aid the assessment of post-experimental RNA degradation, more complex methods have been developed including RNA degradation plots [62] or mixed effect models based on individual probe and transcript characteristics [63].

Step II: cDNA synthesis

At the beginning of this stage external RNA controls (ERCs) are added which serve as a control of cDNA synthesis independently of the volume and condition of the input material. For this purpose bacterial RNA is used (the so-called poly-A spike) with no homology to known human genes. Following this, the cDNA synthesis process is performed by the use of oligo-dT (a primer with a short sequence of deoxy-thymine nucleotides) or random primers. Oligo-dT binds to the poly-A tails of mRNAs, initiating the synthesis of the complementary strand in a process of reverse transcription (Fig. 4). This process does not work for rRNA molecules since unlike mRNA molecules they do not have a poly-A tail, and for this reason it is not necessary to remove the rRNA prior to this process. However, in some cases rRNA can be polyadenylated in human cells [64], and conversely, not all mRNAs have a poly-A tail, and there are also reports of mRNAs that exist in two forms both polyadenylated and non-polyadenylated [65].

The second strand of the cDNA is then created by using the first strand as a template. Addition of ribonuclease causes RNA cleavage at nonspecific sites, leaving only short fragments attached to the cDNA (Fig. 2). These fragments are then used as primers for the

Table 1 Reference genes found on a typical Affymetrix 3'IVT microarray. Amplification and hybridization control RNAs are added in various proportions and quantities as indicated in the last column. The amplification control transcripts are added using various dilutions which results in an estimated copy numbers ranging from one copy per 6,667 to 100,000 transcripts in the studied RNA sample. The hybridization control consists of biotinylated and fragmented cRNAs added in various amounts that result in a final concentrations ranging from 1.5 to 100 pM

Group	Type	Name	Description
1	Housekeeping genes	AFFX-HSAC07/X00351	ACTB - β-actin gene responsible for the structure of the cell
		AFFX-HUMGAPDH/M33197	GAPDH – enzyme which takes part in glycolysis
		AFFX-HUMISGF3A/M97935	STAT1 – transcription factor
2	Ribosomal RNA	AFFX-HUMRGE/M10098	Gene coding for 18S rRNA subunit
		AFFX-M27830	Gene coding for 28S rRNA subunit
		AFFX-r2-Hs18SrRNA	Gene coding for 18S rRNA subunit - version 2
		AFFX-r2-Hs28SrRNA	Gene coding for 28S rRNA subunit - version 2
3	Amplification control (Poly-A spike)	AFFX-DapX / AFFX-r2-Bs-dap	Dap gene of *B.Subtilis* bacteria - proportions 1:6,667
		AFFX-ThrX / AFFX-r2-Bs-thr	Thr gene of *B.Subtilis* bacteria - proportions 1:25,000
		AFFX-PheX / AFFX-r2-Bs-phe	Phe gene of *B.Subtilis* bacteria - proportions 1:50:000
		AFFX-LysX / AFFX-r2-Bs-lys	Lys gene of *B.Subtilis* bacteria - proportions 1:100,000
4	Hybridization control (Bacterial spike)	AFFX-BioB / AFFX-r2-Ec-bioB	BioB gene of *E.Coli* bacteria – quantity 1.5 pM
		AFFX-BioC / AFFX-r2-Ec-bioC	BioC gene of *E.Coli* bacteria – quantity 5 pM
		AFFX-BioDn / AFFX-r2-Ec-bioD	BioD gene of *E.Coli* bacteria – quantity 25 pM
		AFFX-CreX / AFFX-r2-P1-cre	Cre gene of P1 bacteriophage – quantity 100 pM

polymerase which synthesizes the second strand of the cDNA, removing the remaining mRNA fragments found on its way. Measurement of cDNA concentration, which allows standardizing it across various samples, is not a part of the standard experimental procedure for eukaryotic cells, due to the presence of other nucleic acid species that affect the spectrophotometric measurement, whose removal requires additional cDNA purification. This step is strongly influenced by any previous RNA degradation, which leads to the creation of truncated mRNAs (from the 5′-end) [66]. When oligo-dT primers are used during cDNA synthesis these truncated mRNAs are read from the 3′-end only to the position of truncation, and the remaining part is lost due to the lack of poly-A. In such a situation probes located further from the 3′-end usually show lower signal intensity, a phenomenon which is the basis of RNA degradation plots used to assess the mRNA quality [62]. In order to reduce this effect, on 3′IVT microarrays probes from a single set are selected based on a very small region of 600 bp located close to the 3′-end of the mRNA. To

further reduce this bias sophisticated methods have been developed that take into account the location of regions targeted by probes in order to correct the signal intensities [67, 68].

The 3′-end bias does not occur when random primers are used for the cDNA synthesis. Random primers do not require a poly-A tail since they can attach to any region of the mRNA and not only to its 3′-end, promoting synthesis in a 3′ ➔ 5′ direction, and a very strong 5′-end bias can be observed as shown in ref. [69].

Although many of the available cDNA synthesis kits include a combination of oligo-dT and random primers, kits based solely on oligo-dT are commonly used especially for the 3′-IVT platform where 3′-UTR sequences are of the highest importance since they are targeted by oligonucleotide probes.

Oligo-dT-based cRNA synthesis introduces an additional bias that may affect the results of a microarray experiment. First of all, because of the mRNA degradation problem, oligo-dT primers are a good choice only if the region of interest is located in the vicinity of the

Fig. 4 cDNA synthesis based on a commercial T7-oligo (dT) primer. The arrow indicates the direction of synthesis, red font indicates the promoter sequence used in the amplification. The green region is the sequence spacer that separates the primer from the $(T)_{24}$ motif

Jaksik *et al. Biology Direct* (2015) 10:46

Page 6 of 14

3´-UTR, since large distances between the region targeted by probes and the poly-A can decrease the precision of expression level estimates [70]. If the analysis requires the entire transcript as in the case of WT (whole-transcript) microarray platforms where individual exons are analyzed, random primers are required. Additionally, oligo-dT is assumed to bind only to the poly-A tail of the transcript, requiring a long continuous strand of A nucleotides, as shown in Fig. 2. However, partial primer complementarity (i.e. complementarity of only 8 adenine nucleotides in the primer's sequence) is sufficient for the reaction initiation, and due to the random nature of the attachment it can also bind to the A-strands found commonly in the UTRs [71]. Further, with increasing concentration of oligo-dT the chance of attaching multiple oligonucleotides to a single mRNA are increased. In such situation the synthesis may start from two distinct regions but the reaction located closer to the 3´-end might be blocked by the second reaction, again producing truncated cDNA products [71]. This phenomenon can therefore affect the entire probe-set signal intensity of the targeted transcript if its sequence includes simple repeats built predominantly of A nucleotides.

Step III: Amplification and labeling

In this step the newly-synthesized cDNA is replicated (amplified) in a process of *in vitro* transcription. The goal of this step is to obtain a large quantity of cRNA containing biotinylated C and U nucleotides that will be required in the subsequent steps [58]. For this purpose another fragment of oligo-dT is used, marked in red in Fig. 2, which serves as a promoter for the T7 bacteriophage polymerase.

The efficiency of this reaction and its consistency between samples has a decisive impact on the final experimental outcomes [72]. There are many factors which influence the efficiency of this reaction including the structural properties of the cDNA itself which, depending on the GC content, can affect the efficiency of the polymerase [73] and form secondary structures [74]. This step is completed with a cleanup and quantification of the cRNA which allows for control of the total reaction yield and purity of the sample. The product of the amplification reaction can be observed in lanes three and four of the electrophoresis gel (Fig. 3). rRNA is no longer visible, and due to the variability in length of the cRNAs there are no easily distinguishable bands visible on the gel.

Post-experimental control of cRNA level variations, utilizes the signals of probes targeting a reference RNA (poly-A spike) added prior to cDNA synthesis and signals of housekeeping genes which should be on a similar level across all samples. The poly-A spike contains transcripts of five *B. subtilis* genes (Dap, Lys, Phe, Thr, and Trp) which are added in various proportions to the isolated RNA. Since they all include a poly-A tail they undergo the same procedure as the RNA analyzed, independently of its condition. Lys gene RNA is added at the lowest concentration (1:100,000 of the total RNA) which is close to the sensitivity level of the microarray. Its detection in at least half of the microarrays of a given experiment is a good indicator of a properly conducted procedure. The remaining reference RNAs are added in increasing concentrations Lys < Phe < Thr < Dap with Dap being the highest and close to the probe signal intensity saturation level.

The amplification products no longer have the T7 promoter, although the spacer sequence between the promoter and the $(T)_{24}$ primer (green in Fig. 3) is also amplified [75]. Since this fragment is copied with each cRNA its quantity is very large, and since it can bind to probes having a similar sequence it might affect their signal intensity [69]. It is believed that the process of amplification might be the source of inconsistent signals among samples, as it depends highly on the experiment conditions and the transcript structure [74, 76], becoming the main motivation for the development of microarray protocols that do not require RNA amplification [77].

Step IV and V: cRNA fragmentation and hybridization

cRNAs obtained in the previous step are cut into 50–100 nt fragments shown in lanes five and six of the electrophoresis gel (Fig. 3). After this, another set of external RNA controls (ERCs) that originates from P1 bacteriophage and *E. coli* bacteria (termed bacterial spikes) is added to the RNA pool. Similarly to the poly-A spike, bacterial RNA is added in various concentrations with the following relations satisfied: bioB < bioC < bioD < Cre (group 4 in Table 1). BioB, bioC and bioD originate from the *E. coli* genes used in the synthesis of biotin, while Cre is isolated from P1 bacteriophage where its gene product serves as a recombinase [78]. This bacterial spike is already converted to cRNA and fragmented allowing to control the hybridization process, independently of the efficiency of labeling and amplification used in the previous steps to obtain cRNA [58]. After this the mixture of various cRNAs is transferred on to the microarray chip, initiating the hybridization process.

Hybridization is the most time-consuming step of the entire microarray procedure. During approximately 16 h, in which microarrays are incubated in a hybridization oven set to 45 °C, the cRNA binds to the specific probes attached to the glass surface of the microarray chip. The dynamics of the hybridization process depends on many factors which, as in the amplification step, depend on both the reaction conditions and structural properties of

Jaksik *et al. Biology Direct* (2015) 10:46

Page 7 of 14

the individual cRNA molecules which may significantly affect the experimental outcomes [79, 80]. Prolonged hybridization can cause sample drying and uneven distribution of the material on the surface of the chip. Additionally, evaporation of some of the water can change the salt concentration in the buffers and significantly affect the efficiency of the process [81].

The main purpose of the bacterial spikes added before the hybridization step is to control the consistency of hybridization conditions across all samples, assessing the overall microarray performance [82]. Flaws in the experimental procedure cause either variations in expression intensity range or in the relations among individual bioB, bioC, bioD and Cre transcripts, although one has to remember that flaws in the hybridization process affect other transcripts as well. For this reason, hybridization inconsistencies should be also visible in probe-sets targeting other cRNAs, including the poly-A spike controls. If variations are only present in the bacterial spikes, the problem most likely originates from inaccuracies in their preparation or their concentration in the pre-hybridized cRNA. All of the possible scenarios for housekeeping genes, poly-A, and bacterial spike controls are summarized in Table 2.

Bacterial spike controls are a good indicator of problems that may occur during the hybridization procedure, although they fail to detect uneven hybridization, since the probe-set intensity is obtained after summarizing signals of over 20 individual probes, spread over the entire surface of the microarray (3′IVT arrays) or located in a small region at the middle of the array (WT arrays). For this purpose the quality control of each sample should include the analysis of an image of the microarray surface, which is either a complete scan saved in a DAT file, or more commonly a recreated image based on the individual probe intensities stored in a CEL file [83, 84].

The main assumption made in design of a microarray is that probes targeting a single transcript are placed

Table 2 Problems detected by different control probe-sets and their possible reasons[a]

Housekeeping genes	Poly-A spike	Bacterial spike	Possible reason
error	ok	ok	Poor quality of the mRNA analyzed
error	error	ok	Problems during amplification/ labeling
error	error	error	Problems during hybridization/ washing
ok	ok	error	Inaccurate preparation of bacterial spike
ok	error	ok	Inaccurate preparation of bacterial poly-A spike

[a]Other possible combinations of errors rarely occur in practice

randomly on its surface. For this reason, variations in the signal intensity of specific regions suggest reasons other than the biological variation between the analyzed mRNAs. Such differences among regions, termed image artifacts, are mostly caused by bubbles of air or small levels of impurities, which were added into the microarray cartridge with the experimental solutions [85]. Such artifacts appear very commonly, although they usually have a very small size and are handled efficiently by summarization methods, which are insensitive to a small number of outlying values. The main problem occurs when the artifact covers a significant percentage of the array surface or its intensity is extremely high and close to the saturation level of the probes. Such artifacts are mainly caused by uneven hybridization and affect not only the expression estimates from probes located in its region, but also the remaining probe signals. This latter effect is due to data processing, which utilizes expression levels of all or of a significant fraction of the probes on the microarray [38].

Microarray surface artifacts can be visualized by either creating an image, based on single probe expression intensities in a convenient (usually logarithmic) scale, or by analyzing differential images created by subtracting the signal of each probe on a single microarray from that on another reference array created by, for example, calculating the median intensity level of each probe across all microarrays in a single experiment [37]. If a defective array is found, probes affected by an aberration may be separated and removed from the subsequent data analysis or even recreated using imputation techniques [38, 37, 85, 86]. Microarrays affected by a very large aberration should be removed from the study, as they no longer serve as a reliable source of information.

Step VI: Washing and staining
Washing follows the cRNA hybridization and is used to remove cRNA non-specifically bound to the microarray surface. Again, in this step small variations in the reaction conditions may affect the expression estimates [87]. Depending on the conditions of the washing process (temperature, salt concentration, calcium and magnesium ion levels in the buffer) non-specifically bound cRNA is removed with varying efficiency, affecting the sensitivity and background level of the entire microarray. The binding strength of cRNAs depends not only on their complementarity level but also on the temperature of the hybridization and their sequence characteristics, mainly the GC content [88] and specific base positions inside the sequence [44]. Separation between the binding strength of non-specifically bound GC-rich cRNA and GC-poor cRNA with perfect complementarity is not very sharp, affecting the final intensity level of cRNAs depending on their sequence characteristics [50], which

Jaksik *et al. Biology Direct* (2015) 10:46

Page 8 of 14

can be only reduced using sequence-based normalization approaches during the data pre-processing step [89, 90].

The washing process is followed by staining of the hybridized cRNA using a streptavidin-phycoerythrin complex (Fig. 2). Streptavidin is a protein with high binding affinity to the biotinylated nucleotides used in the cRNA preparation, while phycoerythrin is a fluorescent dye used for quantitation of the hybridized cRNA. The quality of the fluorophore used significantly affects the fluorescence intensity of the microarray, decreasing its sensitivity if it is exposed for too long to daylight [91].

Step VII: Scanning

In this step the microarray cartridge is placed in the microarray scanner where the fluorescence of the phycoerythrin bound to the cRNA is excited using a laser. The level of fluorescence is measured by the scanner's detector and is assumed to be proportional to the amount of cRNA bound to the corresponding probe. The length of this process depends on the size of the microarray and in most cases lasts around 10 min for a single array. During the scanning process all arrays are placed inside the scanner's chamber so that the fluorescence intensity is not affected by differences in the length of exposure to daylight, which could increase the differences among microarrays in both the scale of the measurements and the sensitivity level. It is advised to scan each microarray only once, since each subsequent scan decreases the fluorescence intensity by 10–20 %, due to decay of the fluorophore [92]. The fluorescence intensity of cyanine-based dyes also used in microarray experiments, such as Cy5, can be further affected by the ozone concentration in the laboratory, a factor which is both time- and location-dependent, and can become a major source of among-experiment inconsistencies [60, 93].

Step VIII: Data pre-processing

The last stage involves data pre-processing which starts by analyzing the microarray image stored in the DAT file, whose goal is to obtain single fluorescence intensity for each probe based on the 16 pixels of the original microarray image. This step is performed by the Affymetrix software and returns a CEL file as an output, in which each probe, at a specific position on the microarray, has a signal intensity assigned to it. These individual probe intensities are used in the subsequent preprocessing steps, during which each array is standardized by first estimating and then subtracting the background signal in order to reduce the effect of non-specific hybridization [44]. Following step is to perform normalization procedure which reduces the differences in probe intensities that originate from differences in experimental conditions and cRNA concentration [31, 94]. The final step of pre-processing is the summarization, in

which a single expression estimate is calculated for each probe-set based on the intensity of the individual probe signals [95]. Summarization step is highly dependent on the quality of the probe and probeset definitions which are in many cases low due to inaccurate transcriptome data at the time of microarray design. This can result in probesets targeting transcripts of multiple genes due to low probe specificity, probes that do not map any of the known transcripts [41, 42] or multiple probesets that map the same gene [39, 40], requiring the development of methods used for the validation of existing probes and for probeset redefinition [41, 42, 96].

Selection of the pre-processing strategy can have a very large impact on the experimental outcomes [94] and often requires a few assumptions which are not always acceptable. The main assumption made by pre-processing methods is that the total level of mRNA in the cell does not vary significantly among samples, regardless of the experimental conditions and cell lines used. This assumption is required for the standardization approaches based on mean and median scaling or more complex approaches, such as quantile normalization [31], and its natural consequence is that the amount of differentially-expressed features with increased or decreased levels will be always similar. For example, in the case of global transcript level changes in cells with inhibited transcription, one might expect to detect predominantly transcript down-regulation, whereas after applying quantile normalization it is very probable that a significant number of up-regulated transcripts will be observed, due to intensity distribution transformations.

Another important assumption is forced by the massively parallel experimentation of the microarray technique which allows for assessing expression level of thousands of genes simultaneously. We have to assume that the reaction conditions for each individual gene were similar while knowing that due to various molecular properties of the analyzed RNA/DNA fragments it is impossible to properly optimize each of the individual reactions. Most of the data processing methods make this assumption although some standardization methods also exist that utilize probe and RNA/DNA sequence information in order to reduce the signal differences resulting from sub-optimal amplification and hybridization conditions that affect gene expression estimates to a varying degree [89, 90].

Conclusions

Despite successful studies of reproducibility [27] and specificity [97], microarrays have been often subject of criticism as a method which fails to identify relevant information that can be transferred directly into clinical applications [98]. The main reason is that statistical significance often differs from biological relevance due to a

Jaksik *et al. Biology Direct* (2015) 10:46

Page 9 of 14

very limited number of samples or to the influence of other factors, such as cellular heterogeneity or variability of the morphological features, which are difficult to separate from the studied features. This highlights the importance of experimental design which utilizes an adequate number of samples and biological replicates to answer questions defined in the project.

In this article we describe the experimental protocol used for Affymetrix microarrays and important factors that may influence its outcomes, summarized in Fig. 5. The entire procedure has been subject of many changes since the first Affymetrix microarrays were released, mainly involving different sets of reagents which allow obtaining a higher yield of reactions with shorter incubation times. The most important changes include the transition from one-step to two-step cDNA synthesis in 2004, and the addition of whole transcript (WT) microarrays in 2009, which utilize different sets of reagents produced by Ambion. Apart from the differences in cDNA synthesis, which as described in step two of the experimental protocol, is based on random primers instead of oligo-dT, significant modification was also made to the labeling process of WT microarrays., The labeling takes place after cDNA fragmentation and uses terminal deoxynucleotidyl transferase (TdT) that adds labeled nucleotides only at the 3′-end of the cRNA. Despite similar methods of oligonucleotide probe design, changes in the experimental procedure might explain the differences in transcript level estimates obtained using various platforms, and their modifications with time might be a source of the inconsistencies observed among experiments

conducted by different laboratories and even in the same laboratory, when separated by long periods of time.

The capabilities of microarray studies are limited, since the measurement of transcript levels provides only a rough estimate of the intracellular conditions at a specific time point, and is affected by a plethora of experiment-specific factors. The process of discovery of new drugs, using expression or genotyping microarrays, is therefore uneven in pace and in some cases might be even misleading. However, microarrays can be successfully used to validate the effects of existing drugs by helping to identify their targets and off-target effects [99]. Microarrays are becoming less popular due to the decreasing costs of the RNA-seq methods [100], although one has to remember that some of the steps used in the microarray procedure, with their drawbacks and limitations, are also utilized in other techniques including the RNA-seq approaches [101–103]. Despite the evolution of experimental procedures, the fundamental principles behind microarray experiments remain similar and their understanding is also an essential step towards appropriate interpretation of the data provided by more advanced but related methods.

Reviewers' comments
Reviewer's report 1: Dr. Janet Siefert
Jaksik et al. have written a review article on the use of microarrays and the cautions and complications of using the results of them to evaluate research data. The Kimmel lab has considerable experience, over several years, with microarray data so the expertise of his team
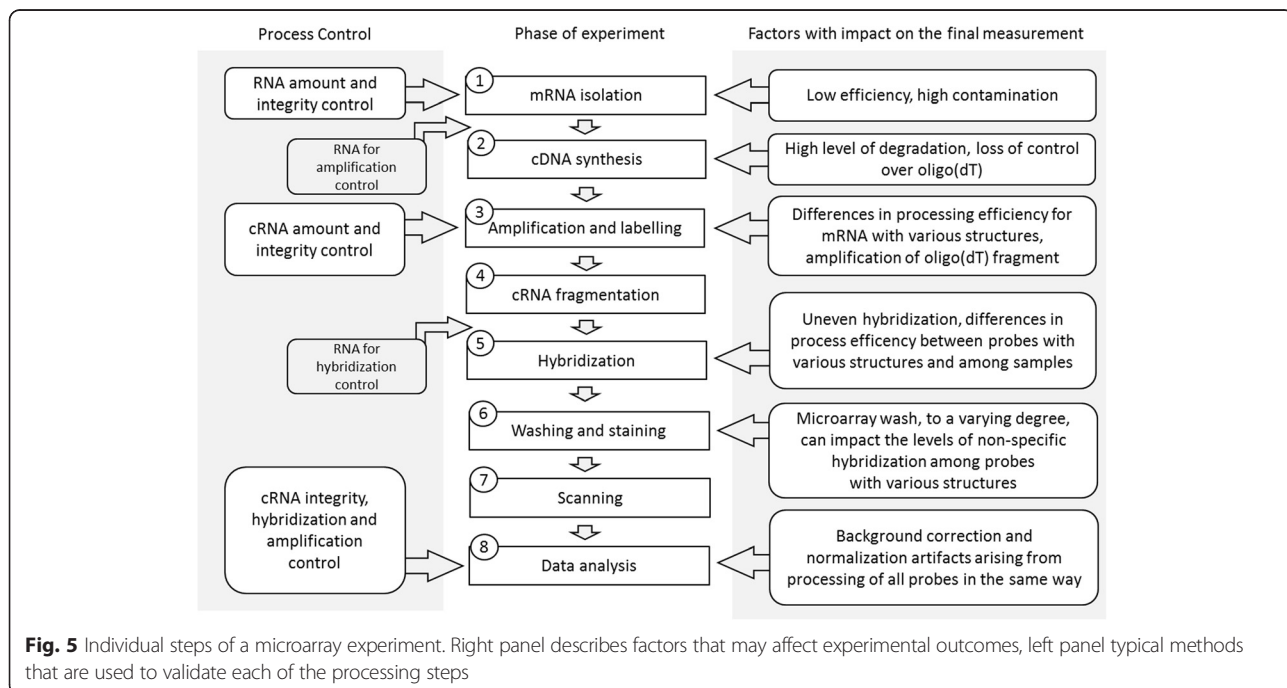

Fig. 5 Individual steps of a microarray experiment. Right panel describes factors that may affect experimental outcomes, left panel typical methods that are used to validate each of the processing steps

Jaksik *et al. Biology Direct* (2015) 10:46

Page 10 of 14

to evaluate and write such an article is well placed. I find their review to be comprehensive and thorough. It will be of considerable use to anyone considering employing microarray experiments as well as those who need to troubleshoot previous use of microarrays and accompanying statistical evaluations. Although the published literature offers a number of articles reviewing microarray use, it is the expertise of this team, as statisticians actively working with numerous microarray data sets, that makes this article valuable to the researching community.

Reviewer's report 2: Dr. Leonid Hanin

This is an interesting article that lists and analyzes in detail various sources of errors and inconsistencies associated with microarray technology. It represents an important step towards answering the following fundamental question: Is microarray technology a reliable tool for furthering our understanding of biological systems at the genomic level or is it bound to produce a lot of biological/technological artefacts and largely generate false knowledge? I believe the main deficiency of the article is that it is entirely qualitative in that no quantitative estimates of the impact of various factors identified in this work or of their relative importance were given. The basic question that a researcher utilizing microarrays would ask is whether impact of these factors is minor or major. The article does not provide any information or opinion in this regard. Given that the processes collectively forming microarray technology are either biochemical or physical in nature, estimating the effects of various factors on gene expression signals quantitatively seems to be in principle possible. For example, here are two relevant publications, just from the top of my head, about the physics of DNA/RNA hybridization, see also references therein:

1. E. Carlon, T. Heim (2006), Thermodynamics of RNA/DNA hybridization in high-density oligonucleotide microarrays, Physica A 362: 433–449.

2. A. Ferrantini, E. Carlon (2008), On the relationship between perfect matches and mismatches in Affymetrix Genechips, Gene 422: 1–6.

On a more technical level, I have the following questions and comments.

1. One of the sources of uncertainty in determination of gene expression levels that was not mentioned in the article is errors in gene finding. While for well-annotated genomes of model organisms they are probably insignificant, for many other organisms type I and II errors in gene finding may be as high as 10–15 %.

Authors' response: Description of probe design flaws resulting from inaccurate transcriptome data were added to the description of the data pre-processing step.

"Summarization step is highly dependent on the quality of the probe and probeset definitions which are in many cases low due to inaccurate transcriptome data at the time of microarray design. This can result in probesets targeting transcripts of multiple genes due to low probe specificity, probes that do not map any of the known transcripts [41, 42] or multiple probesets that map the same gene [39, 40], requiring the development of methods used for the validation of existing probes and for probeset redefinition [41, 42, 96]. "

2. Another factor that was mentioned only very briefly in the text but probably deserves more discussion is heterogeneity of biological material from which mRNA is extracted. It may include different types of cells, cells in various phases of their life cycle, quiescent and proliferating cells, etc.

Authors' response: Cellular heterogeneity is a major problem in many biological studies and can indeed significantly affect the results of a microarray study. However because this problem is unrelated to the technical aspects of the microarray protocol, we find it to be outside of the scope of this article.

3. It was mentioned, again very briefly, in the Conclusions section that utilizing several replications may improve design of microarray experiments. I think this is a very important point. I find it pretty appalling that a lot of microarray experiments with so many sources of variation and error were based on a single run of the process! What minimum replication number would the authors of the article recommend?

Authors' response: Based on our experience we suggest using at least 3 replicates for studies based on cell lines. From a statistical point of view a minimum of 3 samples are required for the estimate of standard deviation to be valid. Higher number of replicates might be highly beneficial for experiments dealing with poor quality material or studies aiming to detect small differences in gene expression. For experiments based on samples extracted from multiple patients, replicates are usually not necessary since the confidence level increases with the number of patients studied.

4. It was stated in the Microarray Structure section that the "amount of material hybridized to a given probe… is related to the amount of DNA or RNA of a given gene in the sample. "What is this functional relationship? Is the magnitude of the optical signal produced by microarray chip proportional to the copy number of a gene's RNA transcript in the sample? What happens to

Jaksik *et al. Biology Direct* (2015) 10:46

Page 11 of 14

this relationship when microarray output data are pre-processed?

> *Authors' response: The fluorescence intensity of probe is proportional to the RNA level of corresponding gene although, as shown by Held et al. in 2006 the relationship is not linear whether or not the data are pre-processed.*

Finally, here are few minor comments.

> *Authors' response: Minor comments of the reviewer have been taken into account. Corresponding changes in the manuscript are highlighted in grey.*

1. Introduction, line 2. "physical or chemical conditions". Perhaps biological conditions too?

2. Introduction, line 4. It seems like "their" should be inserted between "evaluate" and "variability".

3. Introduction, paragraph 2, sentence 2. Aren't "specific methods used" and "inaccuracies of the measurements" themselves "experimental factors"?

4. Introduction, last paragraph. What is the difference between "accuracy" and "precision"? Also, how is specificity defined in the case of microarrays?

> *Author's response: Definitions of accuracy and precision can be found in ref. 45.*
>
> *In the case of microarrays specificity refers to the ability of a probe to bind a unique target sequence. A specific probe provides signal proportional only to the amount of the target sequence, while non-specific probe signal will be a result of interaction with more than one target sequence. The specificity of a probe can be diminished by cross-hybridization, also called non-specific hybridization, a phenomenon in which sequences that are not strictly complementary according to the Watson–Crick rules bind to each other.*

5. Introduction, last sentence. Delete "have".

6. Microarray Structure, paragraph 2. "…600 nt sequence fragments located near the 3′ end of a specific transcript (Fig. 1). "This is not clear from Fig. 1.

7. Microarray Structure, last paragraph. Platform differences contribute to low accuracy of microarrays, so "despite" seems to be out of place.

8. Step I: RNA isolation, paragraph 1. "…ribosomal RNA… is rarely studied". I think rRNA is studied quite extensively in phylogeny and pharmacology.

9. Step I: cDNA synthesis, paragraph 1. Shouldn't the figure referred to here be 4 rather than 3?

10. Conclusions, paragraph 1. What is the significance of "morphological features" in this context?

11. Caption to Fig. 1. What is the purpose of probe sets A and B? Also, does "corresponds to" mean "proportional to", see technical comments 4?

12. Figure 2. I think it would be better if the steps of microarray experiment shown in the figure correspond to the steps described under Biological background of microarray experiments.

13. Table 2. What about the other two combinations of control probe-set outcomes involving an error?

14. The authors are encouraged to proofread their submission. There are a few places with missing or extra commas, instances where article "the" can (or perhaps should) be removed, etc.

Reviewer's report 3: Dr. I King Jordan
This reviewer provided no comments for publication.

Author details
[1]Systems Biology Group, Faculty of Automatic Control, Electronics and Informatics, Silesian University of Technology, Gliwice, Poland. [2]Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA. [3]Department of Statistics, Rice University, Houston, TX, USA.

References
1. Ramsay G. DNA chips: state-of-the art. Nat Biotechnol. 1998;16(1):40–4. doi:10.1038/nbt0198-40.
2. Stoughton RB. Applications of DNA microarrays in biology. Annu Rev Biochem. 2005;74:53–82. doi:10.1146/annurev.biochem.74.082803.133212.
3. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol. 1996;14(13):1675–80. doi:10.1038/nbt1296-1675.
4. Erickson S, MacLeod SL, Hobbs CA. Cheek swabs, SNP chips, and CNVs: assessing the quality of copy number variant calls generated with subject-collected mail-in buccal brush DNA samples on a high-density genotyping microarray. BMC Med Genet. 2012;13:51. doi:10.1186/1471-2350-13-51.
5. Gardner S, Thissen JB, McLoughlin KS, Slezak T, Jaing CJ. Optimizing SNP microarray probe design for high accuracy microbial genotyping. J Microbiol Methods. 2013;94(3):303–10.
6. Clarke W, Parkin IA, Gajardo HA, Gerhardt DJ, Higgins E, et al. Genomic DNA enrichment using sequence capture microarrays: a novel approach to

Jaksik *et al. Biology Direct* (2015) 10:46

Page 12 of 14

discover Sequence Nucleotide Polymorphisms (SNP) in Brassica napus L. PLoS One. 2013;8(12):e81992.

7. Masimba P, Gare J, Klimkait T, Tanner M, Felger I. Development of a simple microarray for genotyping HIV-1 drug resistance mutations in the reverse transcriptase gene in rural Tanzania. Trop Med Int Health. 2014;19(6):664–71.

8. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, et al. Identification and analysis of functional elements in 1 % of the human genome by the ENCODE pilot project. Nature. 2007;447:799–816.

9. Kaufmann K, Muiño JM, Østerås M, Farinelli L, Krajewski P, Angenent GC. Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP). Nat Protoc. 2010;5(3):457–72.

10. Cauchy P, Benoukraf T, Ferrier P. Processing ChIP-chip data: from the scanner to the browser. Methods Mol Biol. 2011;719:251–68.

11. Dowell N, Sperling AS, Mason MJ, Johnson RC. Chromatin-dependent binding of the S. cerevisiae HMGB protein Nhp6A affects nucleosome dynamics and transcription. Genes Dev. 2010;24(18):2031–42.

12. Makeyev A, Bayarsaihan D. ChIP-chip identifies SEC23A, CFDP1, and NSD1 as TFII-I target genes in human neural crest progenitor cells. Cleft Palate Craniofac J. 2013;50(3):347–50.

13. Hegde M, Chin EL, Mulle JG, Okou DT, Warren ST, et al. Microarray-based mutation detection in the dystrophin gene. Hum Mutat. 2008;29:1091–9.

14. Rouleau E, Lefol C, Tozlu S, Andrieu C, Guy C, et al. High-resolution oligonucleotide array-CGH applied to the detection and characterization of large rearrangements in the hereditary breast cancer gene BRCA1. Clin Genet. 2007;72:199–207.

15. Aston E, Whitby H, Maxwell T, Glaus N, Cowley B, et al. Comparison of targeted and whole genome analysis of postnatal specimens using a commercially available array based comparative genomic hybridisation (aCGH) microarray platform. J Med Genet. 2008;45(5):268–74.

16. Ahn J, Mann K, Walsh S, Shehab M, Hoang S, et al. Validation and implementation of array comparative genomic hybridisation as a first line test in place of postnatal karyotyping for genome imbalance. Mol Cytogenet. 2010;3(9). doi:10.1186/1755-8166-3-9.

17. Hartmann A, Thieme M, Nanduri LK, Stempfl T, Moehle C, et al. Validation of microarray-based resequencing of 93 worldwide mitochondrial genomes. Hum Mutat. 2009;30(1):115–22.

18. Zwick M, Kiley MP, Stewart AC, Mateczun A, Read TD. Genotyping of bacillus cereus strains by microarray-based resequencing. PLoS One. 2008;3(7):e2513.

19. Berthet N, Deletoile A, Passet V, Kennedy GC, Manuguerra JC, et al. Reconstructed ancestral sequences improve pathogen identification using resequencing DNA microarrays. PLoS One. 2010;5(12):e15243.

20. Kathiravel U, Keyser B, Hoffjan S, Kötting J, Müller M, et al. High-density oligonucleotide-based resequencing assay for mutations causing syndromic and non-syndromic forms of thoracic aortic aneurysms and dissections. Mol Cell Probes. 2013;27(2):103–8.

21. Vanhomwegen J, Berthet N, Mazuet C, Guigon G, Vallaeys T, et al. Application of high-density DNA resequencing microarray for detection and characterization of botulinum neurotoxin-producing clostridia. PLoS One. 2013;8(6):e67510.

22. Hadiwikarta W, Van Dorst B, Hollanders K, Stuyver L, Carlon E, Hooyberghs J. Targeted resequencing of HIV variants by microarray thermodynamics. Nucleic Acids Res. 2013;41(18):e173.

23. Barnes M, Freudenberg J, Thompson S, Aronow B, Pavlidis P. Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. Nucleic Acids Res. 2005;33:5914–23.

24. Beekman J, Boess F, Hildebrand H, Kalkuhl A, Suter L. Gene expression analysis of the hepatotoxicant methapyrilene in primary rat hepatocytes: an interlaboratory study. Environ Health Perspect. 2006;114:92–9.

25. Dobbin K, Beer DG, Meyerson M, Yeatman TJ, Gerald WL, et al. Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. Clin Cancer Res. 2005;11:565–72.

26. Saitoh T, Yamamoto M, Miyagashi M, Taira K, Nakanishi M, et al. A20 is a negative regulator of IFN regulatory factor 3 signaling. J Immunol. 2005;174:1507–12.

27. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, et al. The microarray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol. 2006;24:1151–61.

28. Guo L, Lobenhofer EK, Wang C, Shippy R, Harris SC, et al. Rat toxicogenomic study reveals analytical consistency. Nat Biotechnol. 2006;24:1162–9.

29. Irizarry R, Warren D, Spencer F, Kim IF, Biswal S, et al. Multiple-laboratory comparison of microarray platforms. Nat Methods. 2005;2:345–50.

30. Hockley S, Mathijs K, Staal YC, Brewer D, Giddings I, et al. Interlaboratory and interplatform comparison of microarray gene expression analysis of HepG2 cells exposed to benzo(a)pyrene. OMICS. 2009;12(2):115–25.

31. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19(2):185–93.

32. Li C, Hung Wong W. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol. 2001;2(8). RESEARCH0032. http://www.genomebiology.com/2001/2/8/research/0032.

33. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics. 2007;8(1):118–27. doi:10.1093/biostatistics/kxj037.

34. Vardhanabhuti S, Blakemore SJ, Clark SM, Ghosh S, Stephens RJ, Rajagopalan D. A comparison of statistical tests for detecting differential expression using affymetrix oligonucleotide microarrays. OMICS. 2006;10(4):555–66. doi:10.1089/omi.2006.10.555.

35. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology. 2004;3:Article3. doi:10.2202/1544-6115.1027.

36. Calza S, Raffelsberger W, Ploner A, Sahel J, Leveillard T, Pawitan Y. Filtering genes to improve sensitivity in oligonucleotide microarray data analysis. Nucleic Acids Res. 2007;35(16):e102. doi:10.1093/nar/gkm537.

37. Suarez-Farinas M, Pellegrino M, Wittkowski KM, Magnasco MO. Harshlight: a "corrective make-up" program for microarray chips. BMC Bioinformatics. 2005;6:294. doi:10.1186/1471-2105-6-294.

38. Moffitt RA, Yin-Goen Q, Stokes TH, Parry RM, Torrance JH, Phan JH, et al. caCORRECT2: Improving the accuracy and reliability of microarray data in the presence of artifacts. BMC Bioinformatics. 2011;12:383. doi:10.1186/1471-2105-12-383.

39. Jaksik R, Polanska J, Herok R, Rzeszowska-Wolny J. Calculation of reliable transcript levels of annotated genes on the basis of multiple probe-sets in affymetrix microarrays. Acta Biochim Pol. 2009;56(2):271–7. doi:20091781.

40. Schneider S, Smith T, Hansen U. SCOREM: statistical consolidation of redundant expression measures. Nucleic Acids Res. 2011;40(6):e46. doi:10.1093/nar/gkr1270.

41. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res. 2005;33(20):e175. doi:10.1093/nar/gni179.

42. Ferrari F, Bortoluzzi S, Coppe A, Sirota A, Safran M, Shmoish M, et al. Novel definition files for human GeneChips based on GeneAnnot. BMC Bioinformatics. 2007;8:446. doi:10.1186/1471-2105-8-446.

43. Kroll KM, Barkema GT, Carlon E. Modeling background intensity in DNA microarrays. Phys Rev E Stat Nonlinear Soft Matter Phys. 2008;77(6 Pt 1):061915.

44. Wu ZJ, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. A model-based background adjustment for oligonucleotide expression arrays. J Am Stat Assoc. 2004;99(468):909–17. doi:10.1198/016214504000000683.

45. Draghici S, Khatri P, Eklund AC, Szallasi Z. Reliability and reproducibility issues in DNA microarray measurements. Trends Genet. 2006;22(2):101–9. doi:10.1016/j.tig.2005.12.005.

46. Blair S, Williams L, Bishop J, Chagovetz A. Microarray temperature optimization using hybridization kinetics. Methods Mol Biol. 2009;529:171–96. doi:10.1007/978-1-59745-538-1_12.

47. Opitz L, Salinas-Riester G, Grade M, Jung K, Jo P, Emons G, et al. Impact of RNA degradation on gene expression profiling. BMC Med Genet. 2010;3:36. doi:10.1186/1755-8794-3-36.

48. Croner RS, Lausen B, Schellerer V, Zeittraeger I, Wein A, Schildberg C, et al. Comparability of microarray data between amplified and non amplified RNA in colorectal carcinoma. J Biomed Biotechnol. 2009;2009:837170. doi:10.1155/2009/837170.

49. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc Natl Acad Sci U S A. 1994;91(11):5022–6.

50. Held GA, Grinstein G, Tu Y. Relationship between gene expression and observed intensities in DNA microarrays - a modeling study. Nucleic Acids Res. 2006;34(9):e70. doi:10.1093/nar/gkl122.

51. Affymetrix. GeneChip Expression Analysis - Technical Manual. 2004:185.

52. Wang Y, Miao ZH, Pommier Y, Kawasaki ES, Player A. Characterization of mismatch and high-signal intensity probes associated with

Jaksik *et al. Biology Direct* (2015) 10:46

Page 13 of 14

affymetrix genechips. Bioinformatics. 2007;23(16):2088–95. doi:10.1093/bioinformatics/btm306.

53. Schneider J, Buness A, Huber W, Volz J, Kioschis P, Hafner M, et al. Systematic analysis of T7 RNA polymerase based in vitro linear RNA amplification for use in microarray experiments. BMC Genomics. 2004;5:29. doi:10.1186/1471-2164-5-29.

54. Urakawa H, El Fantroussi S, Smidt H, Smoot JC, Tribou EH, Kelly JJ, et al. Optimization of single-base-pair mismatch discrimination in oligonucleotide microarrays. Appl Environ Microbiol. 2003;69(5):2848–56.

55. Deng Y, He Z, Van Nostrand JD, Zhou J. Design and analysis of mismatch probes for long oligonucleotide microarrays. BMC Genomics. 2008;9:491. doi:10.1186/1471-2164-9-491.

56. LaFramboise T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic Acids Res. 2009;37(13):4181–93. doi:10.1093/nar/gkp552.

57. Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, et al. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol. 2006;7:3. doi:10.1186/1471-2199-7-3.

58. Affymetrix. 3′ IVT Express Kit User Manual. 2012. http://www.affymetrix.com.

59. Grillo G, Turi A, Licciulli F, Mignone F, Liuni S, Banfi S, et al. UTRdb and UTRsite (RELEASE 2010): A collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Res. 2010;38(Database issue):D75–80. doi:10.1093/nar/gkp902.

60. Fare TL, Coffey EM, Dai H, He YD, Kessler DA, Kilian KA, et al. Effects of atmospheric ozone on microarray data quality. Anal Chem. 2003;75(17):4672–5.

61. Mignone F, Grillo G, Licciulli F, Iacono M, Liuni S, Kersey PJ, et al. UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Res. 2005;33(Database issue):D141–6. doi:10.1093/nar/gki021.

62. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy–analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3):307–15. doi:10.1093/bioinformatics/btg405.

63. Archer KJ, Guennel T. An application for assessing quality of RNA hybridized to Affymetrix GeneChips. Bioinformatics. 2006;22(21):2699–701. doi:10.1093/bioinformatics/btl459.

64. Slomovic S, Laufer D, Geiger D, Schuster G. Polyadenylation of ribosomal RNA in human cells. Nucleic Acids Res. 2006;34(10):2966–75. doi:10.1093/nar/gkl357.

65. Yang L, Duff MO, Graveley BR, Carmichael GG, Chen LL. Genomewide characterization of non-polyadenylated RNAs. Genome Biol. 2011;12(2):R16. doi:10.1186/gb-2011-12-2-r16.

66. Bemmo A, Benovoy D, Kwan T, Gaffney DJ, Jensen RV, Majewski J. Gene expression and isoform variation analysis using Affymetrix Exon Arrays. BMC Genomics. 2008;9:529. doi:10.1186/1471-2164-9-529.

67. Fasold M, Binder H. AffyRNADegradation: control and correction of RNA quality effects in GeneChip expression data. Bioinformatics. 2013;29(1):129–31. doi:10.1093/bioinformatics/bts629.

68. Fasold M, Binder H. Estimating RNA-quality using GeneChip microarrays. BMC Genomics. 2012;13:186. doi:10.1186/1471-2164-13-186.

69. Jaksik R, Marczyk M, Polanska J, Rzeszowska-Wolny J. Sources of high variance between probe signals in affymetrix short oligonucleotide microarrays. Sensors. 2014;14(1):532–48. doi:10.3390/S140100532.

70. Boelens MC, te Meerman GJ, Gibcus JH, Blokzijl T, Boezen HM, Timens W, et al. Microarray amplification bias: loss of 30 % differentially expressed genes due to long probe - poly(A)-tail distances. BMC Genomics. 2007;8:277. doi:10.1186/1471-2164-8-277.

71. Nam DK, Lee S, Zhou G, Cao X, Wang C, Clark T, et al. Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. Proc Natl Acad Sci U S A. 2002;99(9):6152–6. doi:10.1073/pnas.092140899.

72. Wilson CL, Pepper SD, Hey Y, Miller CJ. Amplification protocols introduce systematic but reproducible errors into gene expression studies. BioTechniques. 2004;36(3):498–506.

73. Arezi B, Xing W, Sorge JA, Hogrefe HH. Amplification efficiency of thermostable DNA polymerases. Anal Biochem. 2003;321(2):226–35.

74. Degrelle SA, Hennequet-Antier C, Chiapello H, Piot-Kaminski K, Piumi F, Robin S, et al. Amplification biases: possible differences among deviating gene expressions. BMC Genomics. 2008;9:46. doi:10.1186/1471-2164-9-46.

75. Kerkhoven RM, Sie D, Nieuwland M, Heimerikx M, De Ronde J, Brugman W, et al. The T7-primer is a source of experimental bias and introduces

76. Duftner N, Larkins-Ford J, Legendre M, Hofmann HA. Efficacy of RNA amplification is dependent on sequence characteristics: Implications for gene expression profiling using a cDNA microarray. Genomics. 2008;91(1):108–17. doi:10.1016/j.ygeno.2007.09.004.

77. Sudo H, Mizoguchi A, Kawauchi J, Akiyama H, Takizawa S. Use of non-amplified RNA samples for microarray analysis of gene expression. PLoS One. 2012;7(2):e31397. doi:10.1371/journal.pone.0031397.

78. Sauer B, Henderson N. Site-specific DNA recombination in mammalian cells by the Cre recombinase of bacteriophage P1. Proc Natl Acad Sci U S A. 1988;85(14):5166–70.

79. Sykacek P, Kreil DP, Meadows LA, Auburn RP, Fischer B, Russell S, et al. The impact of quantitative optimization of hybridization conditions on gene expression analysis. BMC Bioinformatics. 2011;12:73. doi:10.1186/1471-2105-12-73.

80. Koltai H, Weingarten-Baror C. Specificity of DNA microarray hybridization: characterization, effectors and approaches for data correction. Nucleic Acids Res. 2008;36(7):2395–405. doi:10.1093/Nar/Gkn087.

81. Affymetrix. Gene Expression Assay and Data Analysis - Hybridization time. 2012. http://www.affymetrix.com/support/help/faqs/ge_assays/faq_15.jsp.

82. Tong W, Lucas AB, Shippy R, Fan X, Fang H, Hong H, et al. Evaluation of external RNA controls for the assessment of microarray performance. Nat Biotechnol. 2006;24(9):1132–9. doi:10.1038/nbt1237.

83. Reimers M, Weinstein JN. Quality assessment of microarrays: Visualization of spatial artifacts and quantitation of regional biases. BMC Bioinformatics. 2005;6:166. doi:10.1186/1471-2105-6-166.

84. Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A. 2001;98(1):31–6. doi:10.1073/pnas.011404098011404098.

85. Song JS, Maghsoudi K, Li W, Fox E, Quackenbush J, Shirley LX. Microarray blob-defect removal improves array analysis. Bioinformatics. 2007;23(8):966–71. doi:10.1093/bioinformatics/btm043.

86. Petri T, Berchtold E, Zimmer R, Friedel CC. Detection and correction of probe-level artefacts on microarrays. BMC bioinformatics. 2012;13:114. doi:10.1186/1471-2105-13-114.

87. Binder H, Krohn K, Burden CJ. Washing scaling of GeneChip microarray expression. BMC bioinformatics. 2010;11:291. doi:10.1186/1471-2105-11-291.

88. Skvortsov D, Abdueva D, Curtis C, Schaub B, Tavare S. Explaining differences in saturation levels for Affymetrix GeneChip arrays. Nucleic Acids Res. 2007;35(12):4154–63. doi:10.1093/nar/gkm348.

89. Hulsman M, Mentink A, van Someren EP, Dechering KJ, de Boer J, Reinders MJ. Delineation of amplification, hybridization and location effects in microarray data yields better-quality normalization. BMC bioinformatics. 2010;11:156. doi:10.1186/1471-2105-11-156.

90. Royce TE, Rozowsky JS, Gerstein MB. Assessing the need for sequence-based normalization in tiling microarray experiments. Bioinformatics. 2007;23(8):988–97. doi:10.1093/bioinformatics/btm052.

91. Munier M, Jubeau S, Wijaya A, Morancais M, Dumay J, Marchal L, et al. Physicochemical factors affecting the stability of two pigments: R-phycoerythrin of Grateloupia turuturu and B-phycoerythrin of Porphyridium cruentum. Food Chem. 2014;150:400–7. doi:10.1016/j.foodchem.2013.10.113.

92. Affymetrix. Gene Expression Assay and Data Analysis - Microarray scanning. 2012. http://www.affymetrix.com/estore/support/help/faqs/ge_assays/faq_8.jsp.

93. Branham WS, Melvin CD, Han T, Desai VG, Moland CL, Scully AT, et al. Elimination of laboratory ozone leads to a dramatic improvement in the reproducibility of microarray gene expression measurements. BMC Biotechnol. 2007;7:8. doi:10.1186/1472-6750-7-8.

94. Park T, Yi SG, Kang SH, Lee S, Lee YS, Simon R. Evaluation of normalization methods for microarray data. BMC bioinformatics. 2003;4:33. doi:10.1186/1471-2105-4-33.

95. Hochreiter S, Clevert DA, Obermayer K. A new summarization method for Affymetrix probe level data. Bioinformatics. 2006;22(8):943–9. doi:10.1093/bioinformatics/btl033.

96. Marczyk M, Jaksik R, Polanski A, Polanska J. Affymetrix chip definition files construction based on custom probe set annotation database. Stud Comput Intell. 2011;381:135–44. doi:10.1007/978-3-642-23418-7.

97. Canales RD, Luo Y, Willey JC, Austermiller B, Barbacioru CC, Boysen C, et al. Evaluation of DNA microarray results with quantitative gene expression platforms. Nat Biotechnol. 2006;24(9):1115–22. doi:10.1038/nbt1236.

variability between microarray platforms. PLoS One. 2008;3(4):e1980. doi:10.1371/journal.pone.0001980.

Jaksik *et al. Biology Direct* (2015) 10:46

Page 14 of 14

98. Webb PM, Merritt MA, Boyle GM, Green AC. Microarrays and epidemiology: not the beginning of the end but the end of the beginning. Cancer Epidemiol Biomark Prev. 2007;16(4):637–8. doi:10.1158/1055-9965.EPI-07-0156.

99. Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, et al. Drug target validation and identification of secondary drug target effects using DNA microarrays. Nat Med. 1998;4(11):1293–301. doi:10.1038/3282.

100. Shendure J. The beginning of the end for microarrays? Nat Methods. 2008;5(7):585–7. doi:10.1038/nmeth0708-585.

101. Zheng W, Chung LM, Zhao H. Bias detection and correction in RNA-Sequencing data. BMC bioinformatics. 2011;12:290. doi:10.1186/1471-2105-12-290.

102. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40(10):e72. doi:10.1093/nar/gks001.

103. Lahens NF, Kavakli IH, Zhang R, Hayer K, Black MB, Dueck H, et al. IVT-seq reveals extreme bias in RNA sequencing. Genome Biol. 2014;15(6):R86. doi:10.1186/gb-2014-15-6-r86.