BMC Bioinformatics

Research article

Simulation of DNA array hybridization experiments and evaluation of critical parameters during subsequent image and data analysis

Christoph K Wierling*[1], Matthias Steinfath[1], Thorsten Elge[2], Steffen Schulze-Kremer[2], Pia Aanstad[3], Matthew Clark[1], Hans Lehrach[1] and Ralf Herwig[1]

Address: [1]Department of Vertebrate Genomics, Max-Planck-Institut für Molekulare Genetik, Ihnestrasse 73, D-14195 Berlin-Dahlem, Germany, [2]Deutsches Ressourcenzentrum für Genomforschung GmbH, Heubnerweg 6, D-14059 Berlin, Germany and [3]Department of Biochemistry, University of California, San Francisco, 513 Parnassus Ave, San Francisco, CA-94143-0448, USA

E-mail: Christoph K Wierling* - wierling@molgen.mpg.de; Matthias Steinfath - steinfat@molgen.mpg.de; Thorsten Elge - elge@rzpd.de; Steffen Schulze-Kremer - steffen@rzpd.de; Pia Aanstad - aanstad@itsa.ucsf.edu; Matthew Clark - clark@molgen.mpg.de; Hans Lehrach - lehrach@molgen.mpg.de; Ralf Herwig - herwig@molgen.mpg.de

*Corresponding author

Published: 22 October 2002

BMC Bioinformatics 2002, **3**:29

Received: 18 June 2002
Accepted: 22 October 2002

This article is available from: http://www.biomedcentral.com/1471-2105/3/29

Abstract

Background: Gene expression analyses based on complex hybridization measurements have increased rapidly in recent years and have given rise to a huge amount of bioinformatic tools such as image analyses and cluster analyses. However, the amount of work done to integrate and evaluate these tools and the corresponding experimental procedures is not high. Although complex hybridization experiments are based on a data production pipeline that incorporates a significant amount of error parameters, the evaluation of these parameters has not been studied yet in sufficient detail.

Results: In this paper we present simulation studies on several error parameters arising in complex hybridization experiments. A general tool was developed that allows the design of exactly defined hybridization data incorporating, for example, variations of spot shapes, spot positions and local and global background noise. The simulation environment was used to judge the influence of these parameters on subsequent data analysis, for example image analysis and the detection of differentially expressed genes. As a guide for simulating expression data real experimental data were used and model parameters were adapted to these data. Our results show how measurement error can be balanced by the analysis tools.

Conclusions: We describe an implemented model for the simulation of DNA-array experiments. This tool was used to judge the influence of critical parameters on the subsequent image analysis and differential expression analysis. Furthermore the tool can be used to guide future experiments and to improve performance by better experimental design. Series of simulated images varying specific parameters can be downloaded from our web-site: [http://www.molgen.mpg.de/~lh_bioinf/projects/simulation/biotech/]

Background

DNA-array technology is nowadays frequently used for the generation of genome-wide gene expression profiles (see *The chipping forecast,* Nature Genetics Suppl. 21, 1999 for a review). The technology is based on the hybridization of labeled ssDNA to its complementary strand called *probe*. Different probes are fixed as spots on planar surfaces, like glass slides or nylon filters. The arrays are scanned and hybridization signals of the spots are quantified by suitable image analysis software. To gain further biological relevant information complex hybridizations from parallel experiments with different target samples as well as experimental repetitions are carried out. Further data evaluation of these hybridization signals by statistical tests and clustering algorithms yields information about differentially expressed or coregulated genes.

The reliability of data produced by these experiments and their reproducibility are crucial for this research. To ensure both reliability and reproducibility a sophisticated experimental design is necessary. This includes for example the identification of error parameters that affect the hybridization data during the data generation process. Influences of systematic and statistical errors due to biotechnical methods (for example mRNA preparation, PCR, hybridization), as well as due to devices and array-media (for example robots, filters, glass-slides) and their effects on evaluation software and algorithms (image analysis, statistical tests, clustering algorithms) must be estimated. These sources of error are frequently discussed in the context of callibration and normalization of microarray data (e.g. [2,4,6,9]). Here we present a computer simulation, that takes into account several sources of error. It enables scientists to judge which parameters are critical and how the experimental design or data evaluation might be improved.

On the other hand creating simulated data without practical consideration is less helpful because it might lead to artificial data sets that estimate and quantify parameters that are not relevant for the analysis of hybridization data. Thus, data should be adapted and linked to real experiments.

Our tool is designed for that purpose. Hybridization signal intensities taken from experimental data are the input; these data were derived as mean values from six filters each of which spotted with the same set of 14208 zebrafish cDNA clones and hybridized independently with the same complex target of an mRNA pool from zebrafish gastrula stage embryos. The output are series of filter images containing well-defined error parameters. In each series only a single parameter was varied at once in order to measure its effects on data analysis. The range of parameter variation was adapted to real experiments (*experimental reference*).

After creating the simulated data the effect of the error parameters were measured on the subsequent data analysis pipeline. We highlight two modules of this pipeline: Image analysis and statistical analysis of differentially expressed genes, although the simulation tool is not restricted to these applications. We chose image analysis because it is the first module of the data analysis and builds the basis for all further research and statistical analysis of differentially expressed genes because it is one of the most utilized applications of gene arrays.

The images were analyzed with three different image processing programs. Parameters that are judged in this paper are variations of the spot positions caused by different experimental artifacts and different sources of background noise. For gene expression profiling twelve filters with varying local background and experimentally determined signal variations were simulated, six of them correspond to hybridizations with a *treatment* and six of them correspond to hybridizations with a complex *control* target. We analyzed how many experimental repetitions are necessary to detect a given level of differential expression. Here, the significance of the differential expression was judged by P values computed by the Welch t-test (cf. [3]).

Our results show that the simulation tool is a valuable resource for the identification and the rating of sources of error arising in hybridization experiments. The simulated sets can be used as benchmark tests for new data analysis modules such as image analyses coming up in the course of gene expression data analysis.

Methods

Implementation of the simulation tool

The simulation tool is written in the object-oriented scripting language python [http://www.python.org]. Some computation intensive functions are implemented in C and can be used as modules in python. Objects like filters, spots or hybridization-data are stored as persistent objects by the use of Zope [http://www.zope.org]. Figure 1A illustrates the implemented simulation pipeline. It takes as an input a set of expression data (in this paper we used an experimental signal distribution of hybridization data) and their position on the array. During the simulation pipeline several perturbations can be performed. Signal intensities can change due to the up- or down-regulation of gene expression, independent perturbations (that effect signal differences of identical spotted duplicates) or a systematic error during the spotting process due to pin-dependent differences in the amount of transfered PCR-product. Perturbations of systematic or non systematic spot position errors and varying spot shapes are also

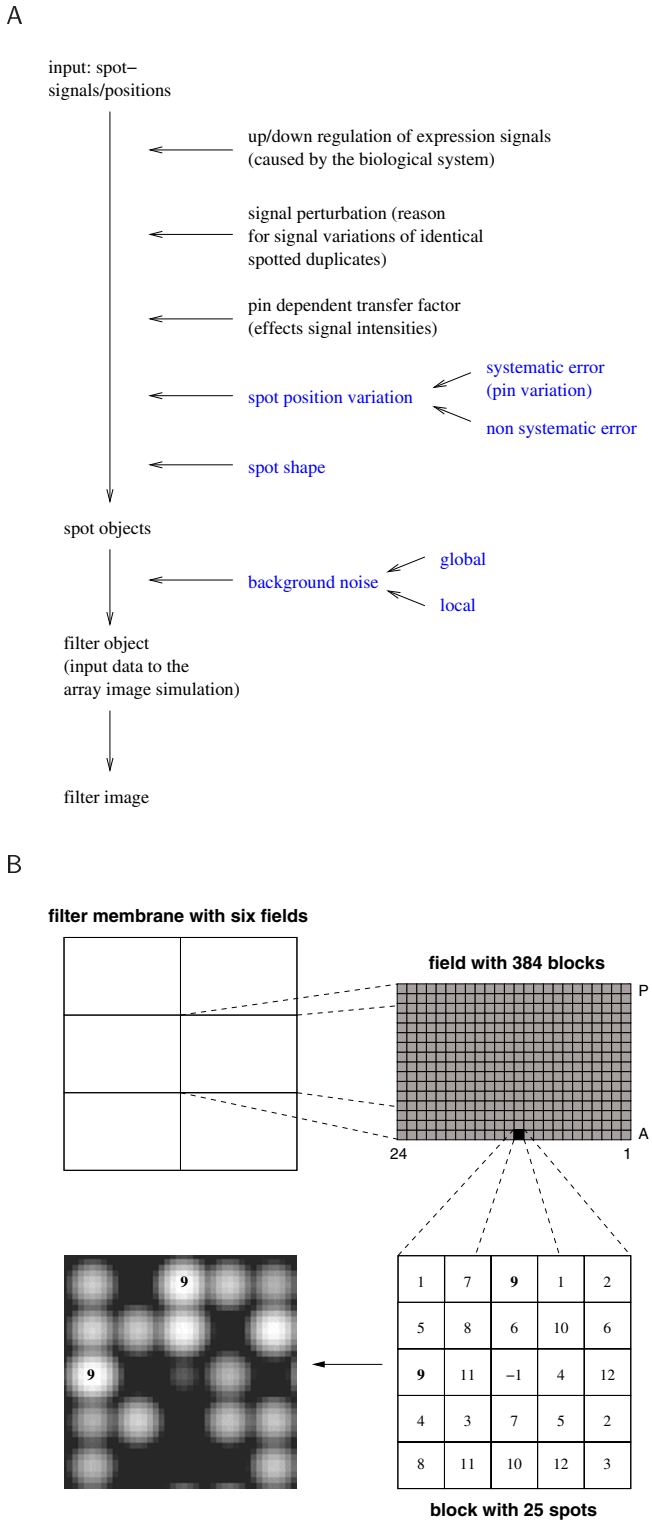


Figure 1
Simulation pipeline and array layout. (**A**) Diagram of the filter simulation pipeline. The parameters highlighted in blue are the parameters that were varied in this paper (cf. Table 1). (**B**) Layout of a filter membrane with 57 600 spot positions. A 5 × 5 spotting pattern is shown; spots with identical position numbers (e.g. No. 9) indicate duplicates. -1 denotes a constant anchor spot which is identical for each block.

Table 1: Definition, modelling and critical effects of simulation parameters.

Parameter	Model	Variation	Critical effect[1]
Spot variation	spot shift (Gaussian distribution)	SD from ideal position	SD > 0.15–0.2 mm ≙ 16.7–22.2%[2]
Pin variation	block shift (Gaussian distribution)	SD from ideal position	SD > 0.12–0.167 mm ≙ 13.3–18.6 %[2,3]
Spot shape	a) two-dimensional Gaussian distribution	a) no variation (fixed SD = 0.1482 mm)	
	b) Crater spot distribution	b) radius of crater	b) radius > 0.1995 mm ≙ 22.2 %[2,4,5]
	c) Plateau spot distribution	c) no variation (fixed radius of cylindric plateau spot = 0.342 mm)	
Global background	additive signal from a Gaussian distribution	fixed mean/SD derived from experimental data	not critical[6]
Local background	additive signal from fractal clouds	signal/background ratio	mean signal/background ratio < 25

[1]Pearson correlation < 0.95. [2] Percent of spot radius relative to the mean spot distance. [3] For VisualGrid and FA; AIDA did not become critical for the parameter range used for the simulations in this paper. [4]Only analysed with FA. [5]For radius ≥ 0.228 mm the automatic gridfind failed. [6]Not critical for global background noise that is comparable to our experimental reference data.

considered. These perturbations result in the input data (filter object, which references its spot objects) used for the array image simulation. Depending on the type of array (filter or glass slide) different levels of global or local background noise can be considered here. The simulation parameters that are under investigation in this paper are listed in Table 1. The output of one array simulation is a parameter file (that contains the values of the variation parameters), a file of the input data for the array image simulation (that contains signal and background intensities and the spot positions) and the image as a 16 bit Tiff-file.

Data sets

The quality of an expression analysis strongly depends on the distribution of the signal intensities and the spot positions on the filter (e.g. outshining effects). To have a realistic situation results of real experiments were used as input data for the construction of the artificial data and the statistical expression analysis.

Experimental macroarray data

A detailed description of the cDNA clone array design, mRNA labeling, hybridization and data capture is given in [3]. PCR products of 14208 zebrafish cDNA clones of a representative library from gastrula stage embryos [1] and 2304 copies of an *Arabidopsis thaliana* cDNA clone were spotted on nylon filter membranes. Clones were spotted in a rectangular grid of blocks with 25 spots (5 × 5) per block by the use of a gadget with 16 × 24 pins corresponding to a 384-well microtiter plate. Figure 1B illustrates the filter design. Due to the experimental procedure a filter is divided into six fields of 384 blocks each. For the 5 × 5 spotting pattern each block contains 25 spots. The zebrafish target derived from mRNA of gastrula stage embryos (6 hours post fertilization) was hybridized to six filter replicates which were spotted with the same set of clones. Each clone was spotted twice in the same block (duplicate) to improve reproducibility.

Design of artificial sample sets

In order to detect differentially expressed genes the cDNA clone array is hybridized in real experiments with two mRNA targets of different origin: one target commonly refers to a reference tissue (control), the second target refers to a certain chemical treatment, a mutant or a disease (treatment).

In our simulation set-up the signals for the control target hybridization were taken from a signal-distribution derived from corresponding experimental data of 14208 clones (see above); the experimental images were analyzed with the in-house developed image analysis FA [10] and medians and the coefficients of variation (CV = standard deviation/mean) were calculated from the replicates of each clone. Figure 2 shows the distributions of these medians and CVs. If reproducibility is perfect the CV is 0, if it is poor the CV tends to higher values. The CVs of the raw data are most frequent in the interval between 0.4 and 0.5 (Fig. 2B). These values are fairly high since a CV of 0.5 for example means that nearly 50 % of the measurement is due to error. However, we want to have a rather upper bound for initial data reproducibility since then error parameter can be identified more clearly. In published stud-

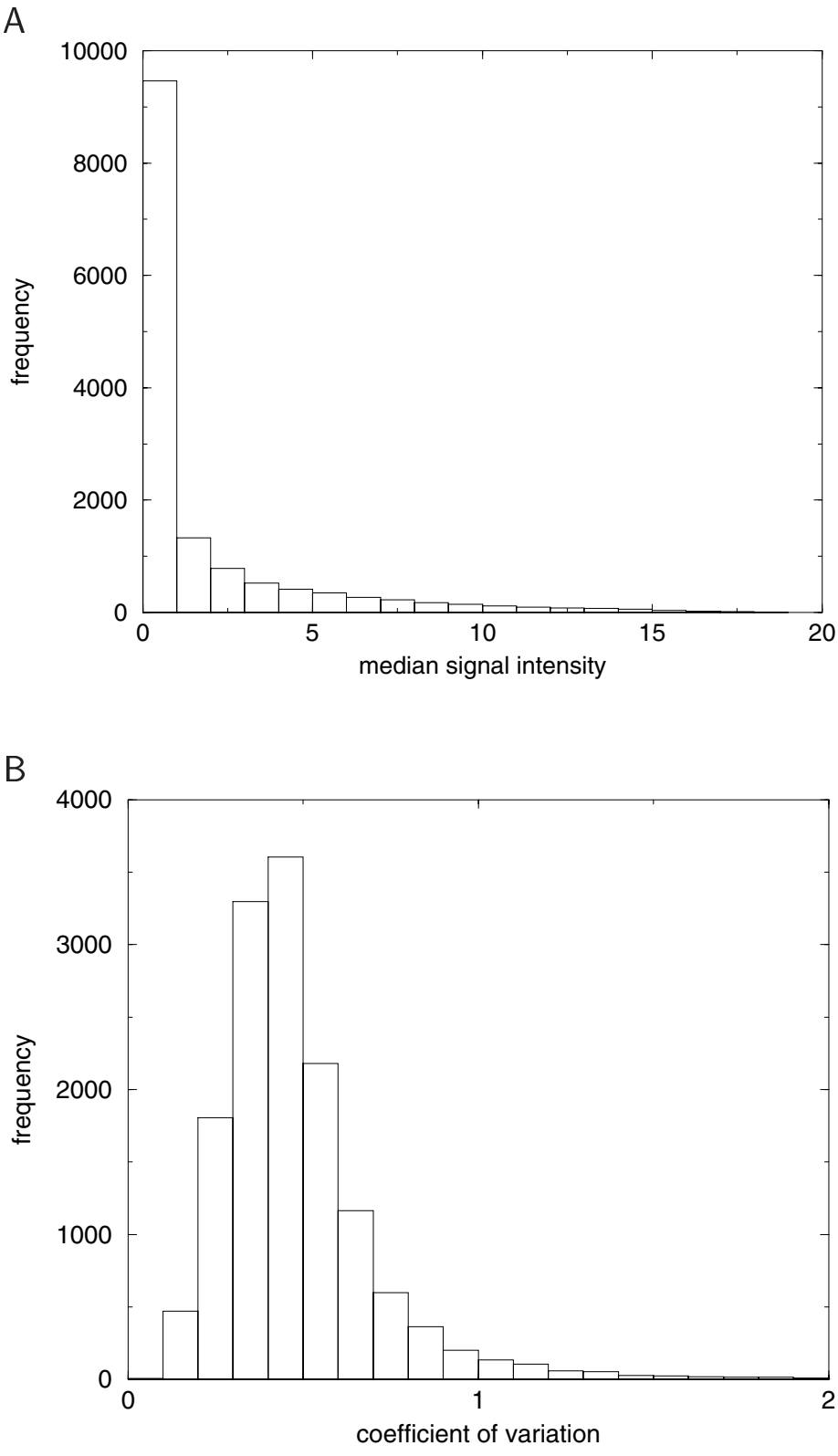


Figure 2
Experimental reference for simulation data. Distribution of the hybridization signals used as experimental reference. (**A**) Histogram ofmedians of 14208 clones from 12 replicates each; (**B**) Histogram of coefficients of variation.

ies the CV is in the area of 10 %–25 % (e.g. [3,7]) since raw data undergoes intensive data normalization and calibration. The signals for the treatment target hybridization were derived from the medians of the experimental reference signals by up-regulating 5000 clones (35.2 % of all clones) randomly. The coefficients of these upregulations – the expression ratios – are uniformly distributed between 1 and 10. The signals of the other 9208 clones remained unchanged. Both signal sets consist of values for the 14208 clones that were screened for differentially expressed genes. The input signal intensity for the spots corresponding to the constant *A. thaliana* cDNA clones of the experimental reference was always the same. For the expression analysis six images were simulated of both signal sets, respectively. Signal intensity variations as described in the following paragraph and local background noise variations (see below) were carried out for each filter. The spotting order was identical with the experimental reference.

Simulation model

Generation of signal intensities

Schuchhardt *et al.*[9] have shown that a strong correlation exists for spot intensities spotted by the same pin. Spots in the same block are spotted by the same pin. Clones that are spotted in different blocks are spotted by different pins. Thus the amount of material that is transfered to the array varies from pin to pin, and this relative pin specific variation can be described for the 384 pins of a gadget by the following pin distribution $P(Y)$:

$$P(Y) = N(1, \sigma_1^2); \sigma_1 = 0.43 \qquad (1)$$

Here $N(1, \sigma_1^2)$ denotes a Gaussian normal distribution with mean 1 and variance σ_1^2. The standard deviation, σ_1, was derived from experimental data. Clones with identical 384-well microtiter plate positions are spotted by the same pin. In the experimental reference *A. thaliana* cDNA of identical amplicons were spotted in each block as a control. Based on this information the mean CV over all pins was calculated and used as σ_1.

On one filter the signal distribution $P(X_{ij})$ of replicates is defined as follows:

$$P(X_{ij}) = N(\gamma_i \cdot z_j, (\gamma_i \cdot z_j \cdot \sigma_2)^2); \sigma_2 = 0.2 \qquad (2)$$

with $i \in \mathbb{N}$; $i = [1, w]$

$j \in \mathbb{N}$; $j = [1, m]$

z_j is the mean signal for clone j taken from the median signal distribution of experimental data (cf. Figure 2), γ_i denotes the pin dependent factor for pin i derived from the distribution, $P(Y)$. For the simulations presented in this paper the number of pins is w = 384 and the number of clones is m = 14208. Using the duplicate correlation (0.8) of the constant experimental *A. thaliana* clone signals and σ_1 one can calculate σ_2 = 0.2, because they are associated with each other (proof is not shown). Thus σ_2 is the CV for identical PCR-products that were spotted by the same pin.

Filter model

The simulated images are generated by an intensity function, which yields for each pixel k an intensity value. The presented model is based on empirical assumptions. It is given by a continuous function of the position $\mathbf{r}$ on the filter, $I(\mathbf{r})$, as follows:

$$I(\mathbf{r}) = \sum_j A_j f(|\mathbf{r} - \mathbf{r}_j|) + g(\mathbf{r}) + \varepsilon \qquad (3)$$

where A_j is the given spot intensity, g is a function that describes the local and global background, $\in$ denotes a stochastic perturbation, and $|\mathbf{r}\text{-}\mathbf{r}_j|$ is the Euclidean distance to the center of spot j. The nine spot centers closest to $\mathbf{r}$ are considered, due to the fact, that the pixelized spot shape is given by a square 19 × 19 pixel matrix and the usual distance between two spot centers is 7.89 pixel for the image resolution used in this paper (0.114 mm/pixel).

Here $f(|\mathbf{r}\text{-}\mathbf{r}_j|)$ is a spot shape distribution which describes the spot shape (see below).

The pixel intensity $\tilde{I}(k)$ is given by:

$$\tilde{I}(k) = \left[\frac{I(\mathbf{r}_k) * 2^N}{max_{\mathbf{r}} I(\mathbf{r})} \right] \qquad (4)$$

with N = 16 for a 16 bit image and $\mathbf{r}_i$ is the center of the pixel k. The square brackets denotes the integer function, that returns the largest integer less than or equal to the value in brackets.

The spot intensities A_j are taken from a real experiment (see above, intensity distribution see Fig. 2).

To determine the location $\mathbf{r}_j$ of the spots we assume that the probes are spotted approximately in an orthogonal grid.

Local distortions

Local distortions of the spots are considered. Due to the experimental procedure two different spot distortions are introduced: spot shifting and pin shifting. Both of them are modeled by randomly Gaussian distributed shifting of the spot-centers relative to their theoretical spot-centers. For spot shifting the distortions are independent for each spot; for pin shifting they are equal for all spots of one block of 5 × 5 spots, because they were spotted by the same pin.

Spot shape

Due to the experimental procedure of the array preparation, the array surface type, and the nature of the fixed DNA material, the spot shapes are different. Here we introduced three distribution models of spot shapes that are based on experimental evidence:

(a) a normalized two-dimensional Gaussian distribution with a given SD (σ):

$$f\left(\left|\mathbf{r}-\mathbf{r}_j\right|\right)=\frac{1}{2\pi\sigma^2}e^{-\frac{(\mathbf{r}-\mathbf{r}_j)^2}{2\sigma^2}} \qquad (5)$$

(b) a normalized two-dimensional Gaussian distribution with a given SD (σ_1) of which another concentric Gaussian-distribution (SD = σ_2) with a scaling-factor S = (0,1) is subtracted. The resulting spot resembles a crater like spot shape.

$$f\left(\left|\mathbf{r}-\mathbf{r}_j\right|\right) = \left(\frac{1}{2\pi\sigma_1^2}e^{-\frac{(\mathbf{r}-\mathbf{r}_j)^2}{2\sigma_1^2}} - S\frac{1}{2\pi\sigma_2^2}e^{-\frac{(\mathbf{r}-\mathbf{r}_j)^2}{2\sigma_2^2}}\right) \times\left(1-S\right)^{-1}$$

(c) a normalized cylindric distributed shape with a given radius *d* that forms a plateau-like spot:

$$f\left(\left|\mathbf{r}-\mathbf{r}_j\right|\right)=\begin{cases}\frac{1}{\pi d^2}, & \text{if} \quad \left|\mathbf{r}-\mathbf{r}_j\right|\le d\\ 0, & \text{if} \quad \left|\mathbf{r}-\mathbf{r}_j\right|> d\end{cases}$$

These spot models were used because they are commonly observable with spotted array data on nylon and glass supports respectively and are frequently assumed as quantification models by image analysis programs. More irregular spot shapes that do not have a common spot distribution can also be observed (e.g. [5]), but are not considered for this paper.

Background noise

Two different sources of background noise can be distinguished: a global background due to the scanner noise or filter surface and a local background due to inhomogeneous hybridization to the filter that looks like smear.

Global background noise

The global background is described by a randomly Gaussian distributed noise that is equal for the whole filter. It can be varied by its mean and SD.

Local background noise

As a model for the local background fractal clouds as described in [8] are used. They are generated with the *midpoint displacement method* with a fractal dimension of 0.4 and then scaled to a given minimum/maximum-range, which defines the intensity level of this background. The model was chosen for local background, because the intensity level of a given pixel depends on its neighbors. This results in images that look quite the same as the background of experimental images. By the use of a pseudo random number generator reproducible fractals were created.

Data evaluation and quality measurement

Image analysis

To illustrate the power of using simulated data for the judgment of image analysis software, we used the following programs: (1.) FA: which was developed at the Max-Planck-Institute for Molecular Genetics [10]. It is fully automated – no manual effort for the positioning of the grid is necessary, (2.) AIDA: Raytest, Germany [www.raytest.de], which needs some manual interaction for the positioning of the grid, (3.) Visual Grid: GPC Biotech, Germany [www.gpc-ag.com], for which the whole grid has to be adapted manually. These programs have been chosen, because they are frequently used at our institute and have already been utilized for massive image analysis (FA [10]; Visual Grid [3]). Furthermore, they are representative for the different levels of automation of image analyses.

Evaluation of gridfind and quantification quality

The following two steps are essential for the analysis of hybridization images: *gridfind* and *quantification*. First the gridfind has to locate the exact positions of the spots and then the signal intensities are assigned to each spot by the quantification. For instance the image analysis FA does a Gaussian spot shape fit for quantification [10]. The per-

formance of the different image analysis programs are tested by the following quality parameters:

• The mean distance between simulated and calculated spot centers.

• The Pearson correlation between simulated and calculated intensities.

The first parameter measures the quality of the gridfind. The second is a measure for the quality of the whole image processing.

Statistical evaluation of differential expression

For testing statistical significance of differential expression we calculated P values according to the Welch test [11]. This test is an unpaired t-test. It assumes that the two samples ("treatment" and "control") are distributed according to Gaussian distributions with means, $\mu_{treatment}$ and $\mu_{control}$ respectively, and judges the hypothesis if $\mu_{treatment} = \mu_{control}$. Here, in contrast to Student's t-test, it is not assumed that both sample distributions have the same variance. The test statistic, T, has the form

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \quad (8)$$

Here, $\bar{x}$ and $\bar{y}$ denote the sample means, S_x^2 and S_y^2 denote the sample variances and n and m are the respective sizes of the treatment and the control sample. High and low values of the test statistic then indicate significantly different sample means. This test has been applied in several studies on differential expression of array data, for example [3] and [2].

Results

The quality of an expression profile analysis based on array data is highly dependent on the number of repeated sample measurements, and of the array preparation, hybridization and signal quantification procedure. The latter can be increased either by improved array preparation and hybridization or better algorithms of the image analysis software that can handle preparation errors. The improvement of both methods is limited. Major critical parameters are local distortions of the spots, variations of the spot shape and outshining effects due to neighbor spots or massive background noise. These parameters have been analyzed in this paper (see Table 1). In the following we simulated series of images by changing only one parameter at once.

Local distortion

In the following the spots have constant Gaussian shape without background noise. Thus only the effects of local distortions are tested. Figure 3 and 4 show the influence of spot-shifts on the gridfind and quantification.

Spot shifting

Spot shifting was simulated with SDs between 0 and 0.342 mm from its ideal positions (Fig. 3). The mean distance between adjacent spot centers was 0.9 mm. For the three image analysis programs which were under investigation this parameter became critical (correlation < 0.95) for the quantification, which is also influenced by the quality of the gridfind, for SDs in the range of 0.15–0.2 mm (Fig. 3B). This is about a fifth of the distance of adjacent spot centers.

In figure 3C we focused only on the quality of the gridfind. The error given by the mean distance of the calculated spot center after image analysis from its simulated center is relatively linear to its perturbation for all tested image analysis programs. The low quality for Aida for small perturbations is due to a missing sub-pixel precision. This means, that if e.g. the simulated spot center is not identical with the center of a pixel, the output-result from Aida lacks this sub-pixel precision.

Pin shifting

The error due to pin variations is a systematic error for all spots in the same block, because they were spotted by the same pin (Fig. 4A). Perturbations with SDs between 0 and 0.2 mm were simulated. This error became critical (correlation < 0.95) for SDs of the pin shifting greater than 0.12 mm for Visual Grid and greater than 0.167 mm for FA. The error of the gridfind was linear to its perturbation (Fig. 4C). Here again the low quality for Aida for small perturbations is due to the missing sub-pixel precision.

Figure 5 shows the distribution of block center shifts measured for experimental data (the block centers were manually determined with Visual Grid). For the results mentioned above this means that the error due to pin shifting is for the majority of blocks never in the critical area. But in general this can become a critical parameter strongly depending on the used devices (e.g. spotting robots).

Spot shape

The spot shape, that depends on several spotting procedure specific properties like the spotting method, the carrier surface or the probe viscosity, was modeled as a two-dimensional Gaussian distributed shape, a crater-like shape (Figure 6A-6J) and a plateau shape (Figure 6K). A mean SD of 0.1482 mm for a two-dimensional Gaussian distributed spot shape was handeled by all three image

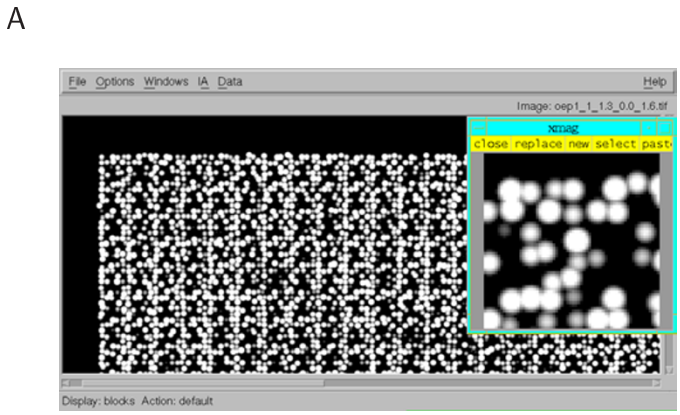


Figure 3
Spot shifting. Every spot was shifted randomly relative to the ideal grid position by a Gaussian distributed distance with a given standard deviation σ. (**A**) simulated image, σ = 0.1824 mm, (**B**) Pearson correlation of simulated and calculated intensities dependent on the standard deviation of the spot centers from their ideal grid nodes, (**C**) mean distance between the calculated and the simulated spot centers dependent on the standard deviation of the spot centers from their ideal grid nodes. The vertical lines in (B) and (C) correspond to the image in (A). In (B) and (C) each point in the plot is determined by a single analysis of a simulated image, respectively.

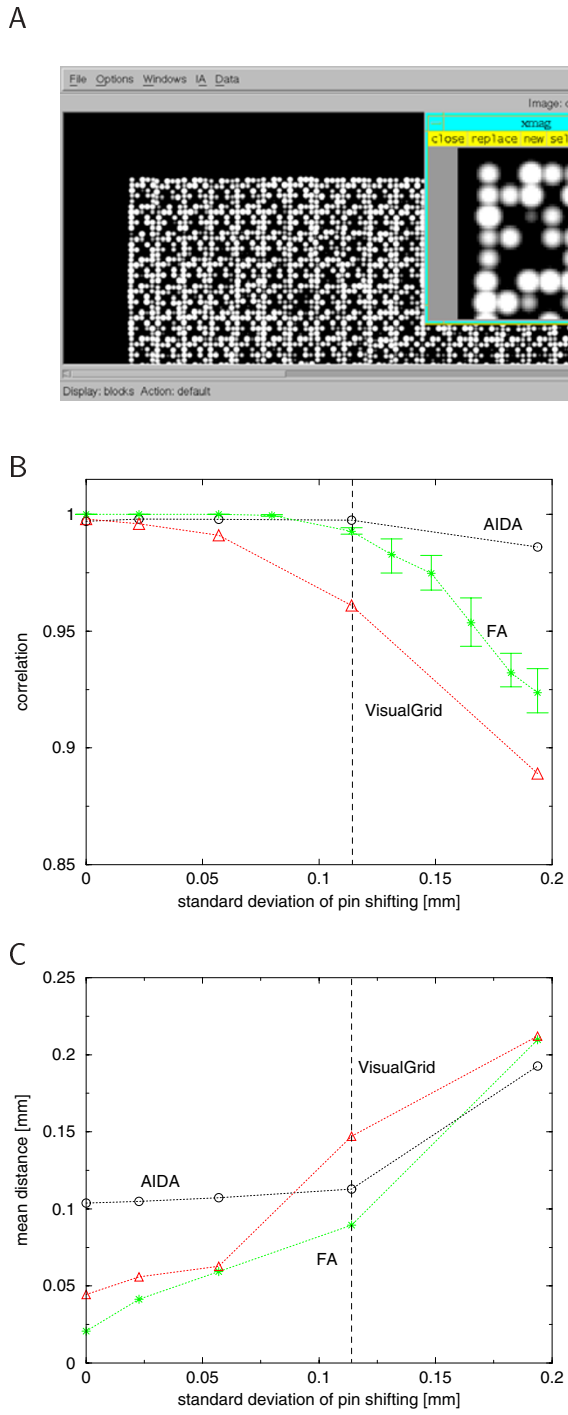


Figure 4

Pin shifting. Every block was shifted randomly relative to its ideal position by a Gaussian distributed distance with a given standard deviation σ. (**A**) simulated image, σ = 0.114 mm, (**B**) Pearson correlation of simulated and calculated intensities dependent on the standard deviation of the block centers from their ideal positions (for AIDA and Visual Grid each data point is determined by a single analysis of a simulated image and for FA three different images has been analyzed for each σ, the asterisk is the mean and the error bar shows the minimum and maximum value of the three repetitions), (**C**) mean distance between the calculated and simulated spot centers dependent on the standard deviation of the block centers from their ideal positions (each data point is determined by a single analysis of a simulated image). The vertical lines in (B) and (C) correspond to the simulated image in (A).

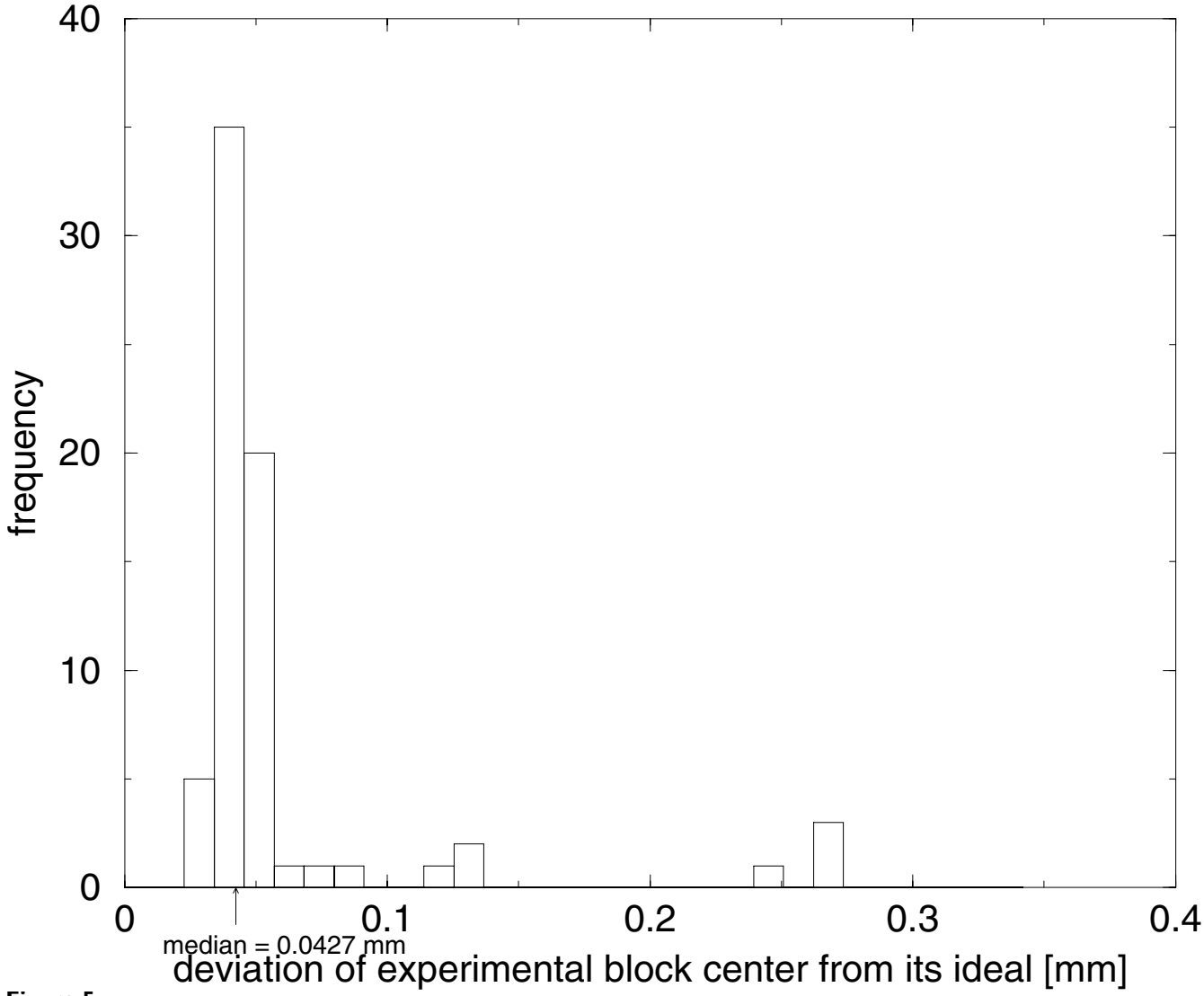


Figure 5
Experimental block center deviation. Histogram of the distance of experimental block centers from their ideal block centers (computed from 12 experimental filter-images each containing 48 × 48 blocks with 5 × 5 spots respectively). Block positions were manually tagged by the use of Visual Grid and distances to the ideal grid – given by field corners – were calculated.

analyses (correlation always > 0.99). Crater-like spot shapes were simulated with crater-radii ranging from 0.0285 mm to 0.285 mm (in 0.0285 mm steps; σ_1 = 0.1482 mm). To judge the influence of this parameter the images were analysed by FA: Up to a crater-radius of 0.1995 mm FA analysed them without any problems (correlation always > 0.99). For crater-radii of 0.228 mm and above (Figure 6H-6J) FA failed due to problems during the gridfind. A third very idealized spot shape – a plateau-like spot shape – was also simulated, to see if this can be handeled by FA. Therefore a filter with plateau spots with a radius of 0.342 mm (not overlapping with neighbor-spots; half distance between two neighbor-spots is 0.44973 mm) was simulated and has been analyzed by FA without any problems (correlation > 0.99).

Background noise

In the following all images have constant Gaussian spot-shapes and all spot centers are located at the ideal grid nodes. Thus the gridfind has only to cope with the background noise.

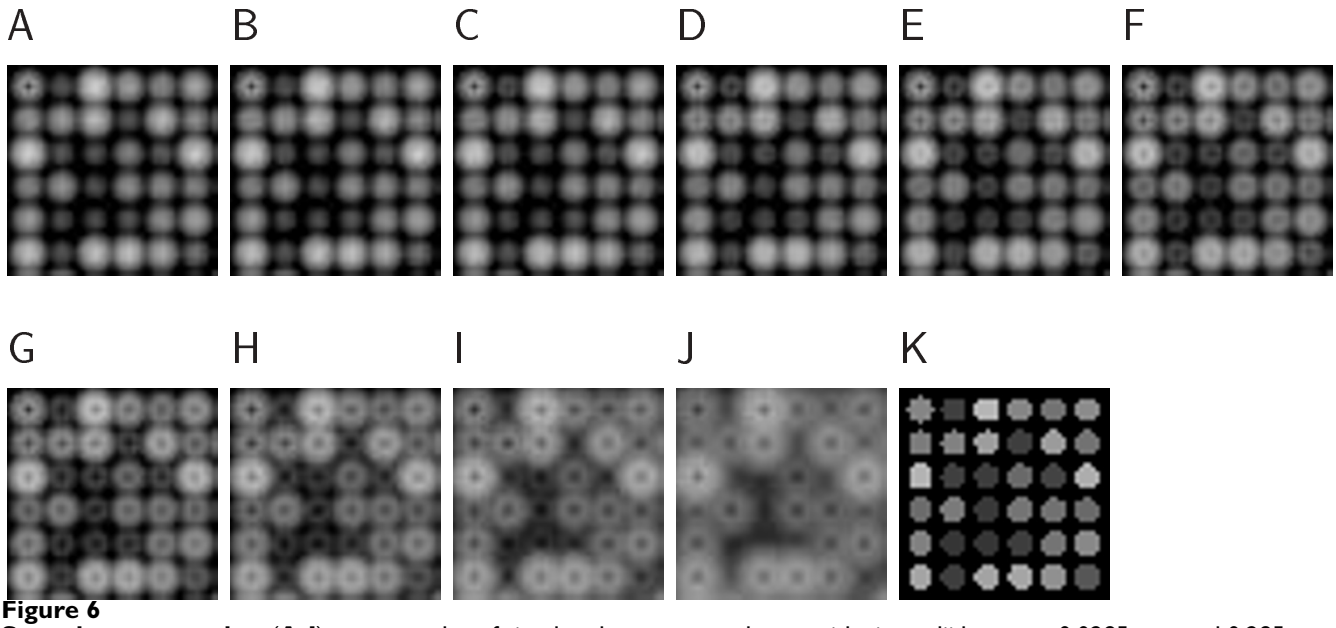


Figure 6
Spot shape examples. (**A-J**) are examples of simulated crater spot shapes with rim radii between 0.0285 mm and 0.285 mm in 0.0285 mm steps. (**K**) is an example of a plateau spot shape (radius = 0.342 mm).

Global background noise
From the border (non-spotted) area of an experimental filter image with a 16 bit depth the noise level was found to be about 16000 with a standard deviation of about 4000; the distribution is similar to Gaussian (data not shown).

The simulated image shown in Figure 7A has Gaussian background noise with μ = 16000 and σ = 4000. The grid was nearly perfectly detected by all image analysis programs for this image. The correlation between input and output intensities were always higher than 0.99; so a realistic global background noise as give by the experimental reference does not influence the quantification of the programs.

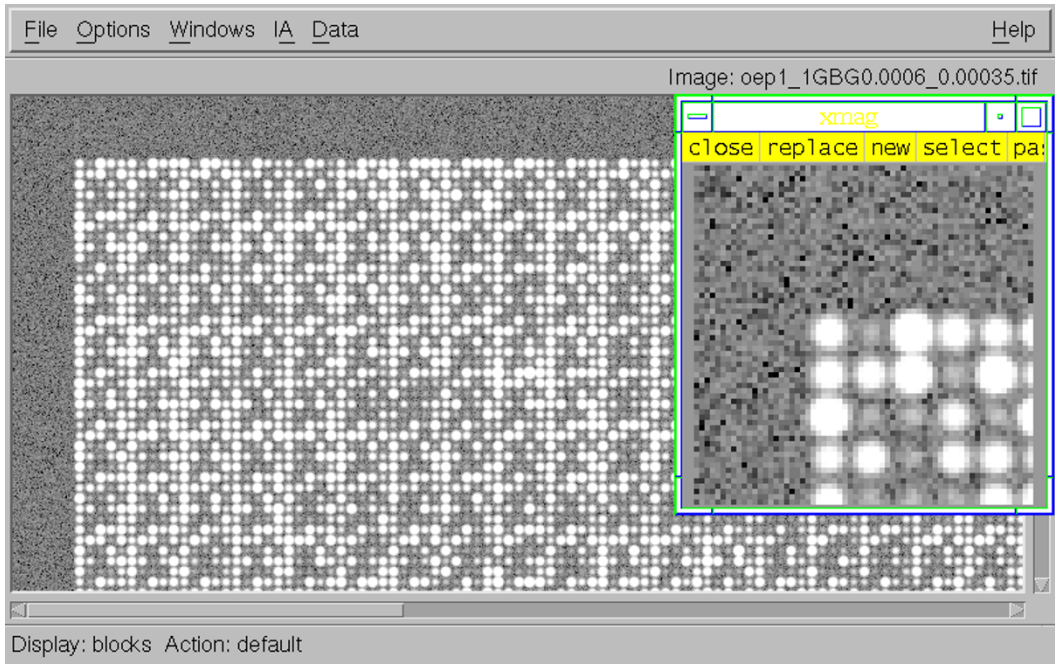
Local background noise
As a model for the local background fractal clouds as described in [8] were used (Fig. 7B).

Figure 8 shows the effect of local background-noise on the image analysis. For mean signal/background ratios above 25 this error did not become critical for any of the three programs. Below a ratio of 20 correlation is decreasing rapidly, especially for AIDA. Correlations for Visual Grid and FA are decreasing significantly for mean signal/background ratios below 13. At this point the signal/background ratio becomes critical for all programs, and thus it was chosen for a further statistical test series (see below).

Influence of background noise on the expression test
We tested the influence of local background noise on the quality of the expression analysis in dependence of the number of repetitions. The significance of differentially expressed genes was judged by the use of the Welch test as described in [3].

A series of six images with variations of signal intensities due to replicated spotting of duplicates and a varying transfer quality for different pins as described in the methods was simulated. Furthermore different local backgrounds with intensities scaled in the same way as given for the mean signal/background ratio of 13 as described above were added. This was done for a control set with 14208 different test clones and for a test set, for which signal intensities of 5 000 clones are up-regulated with factors between 1 and 10. Images were analyzed by the three image processing programs. The source signal sets used for the individual image simulations as well as the analyzed data were used for the statistical significance test. This was done for two, four and six images of the control and test series, respectively. This corresponds to samples of four, eight and twelve signals per clone and series. Figure 9 shows the results. The rate of false positive clones is always low (false positive rate < 0.02). For the input data (Fig. 9A) with expression ratios below 1.45 only 42 % of the regulated clones (sample size 12) could be identified as regulated with a P value < 0.01 as significance level. For expression ratios above 1.45 and sample size 12 almost all regulated clones could be identified – for ratios above 1.9

A



B

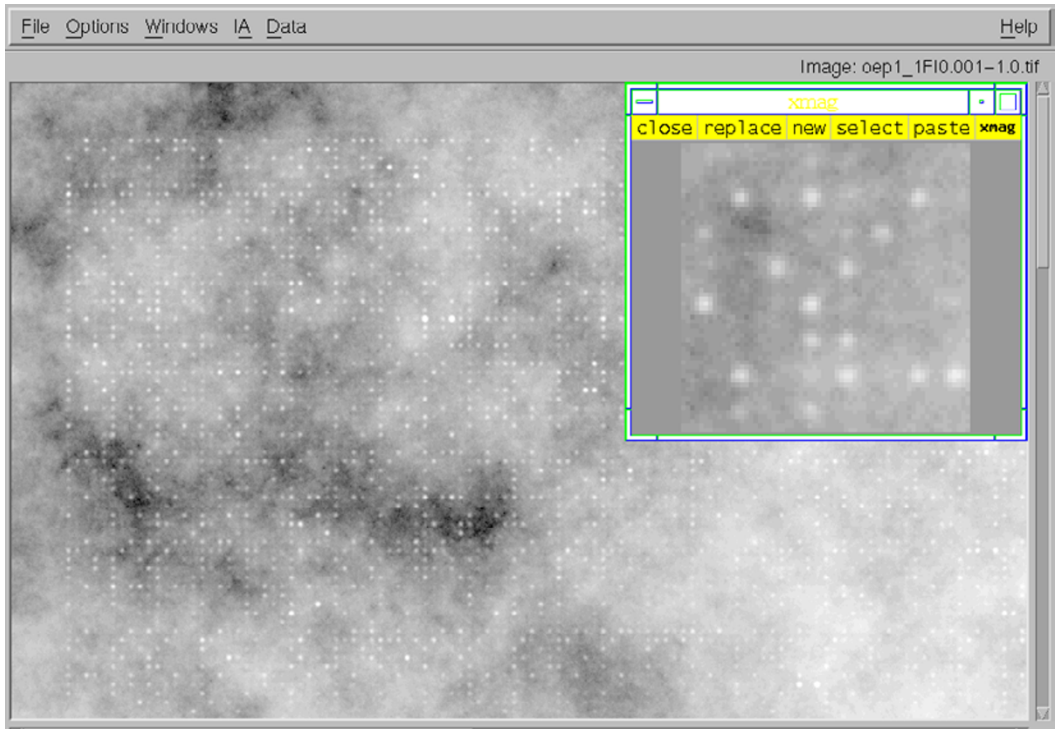


Figure 7
Background noise examples. Examples for filter images with simulated global (**A**) and local (**B**) background noise.

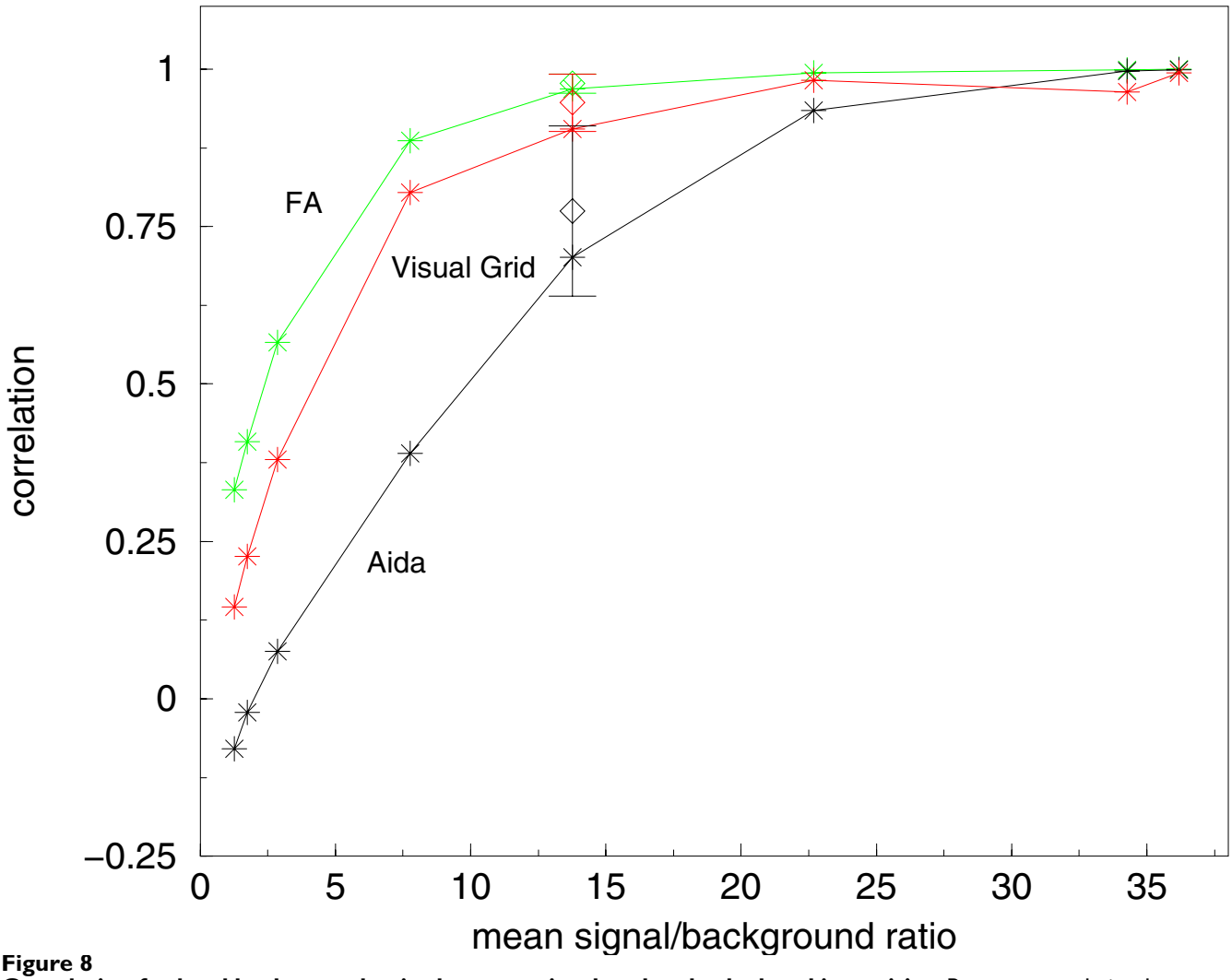


Figure 8
Correlation for local background noise between simulated and calculated intensities. Pearson correlation between simulated and calculated intensities depending on the intensity-level of the fractal background given by the mean of all signal/background ratios over all spots. Each data point (asterisk) corresponds to the results of one image analysis. The used fractal background image was always identical except for the signal/background ratio of 13. For this ratio 7 different fractal background images were simulated; correlation mean μ (diamond) and standard deviations (error bars representing the interval $\mu \pm \sigma$) were calculated.

already with a sample size of 8 nearly all regulated clones could be identified significantly. For a sample size of 4 not even for ratios between 9.55 and 10.0 more than 93 % of the regulated clones could be identified, while for sample size 8 and 12 98.5 % were found. After image analysis the number of identified regulated clones decreased significantly. With the image analysis FA more than 90 % significant clones with sample size 12 could be found for ratios above 1.9 (Fig. 9B), for AIDA (Fig. 9C) and Visual Grid (Fig. 9D) not until ratios above 3.7. Especially for expression ratios between 1.45 and 1.9 with FA (sample size 12) 89 % of the regulated clones could be identified, while AIDA identified only 67 % and Visual Grid 61 %. However, the area below expression ratios of 2 seems to be critical for this kind of expression analysis. For expression ratios above 2 the differences between sample size 8 and 12 are relatively small in comparison to sample size 4.

Figure 9E shows a comparison of the CVs for sample size 12 of the input data signals and the signals quantified by the three different image processing programs. The medians of the CVs are increasing in the following order: input data (0.19), FA (0.21), AIDA (0.29), Visual Grid (0.34).

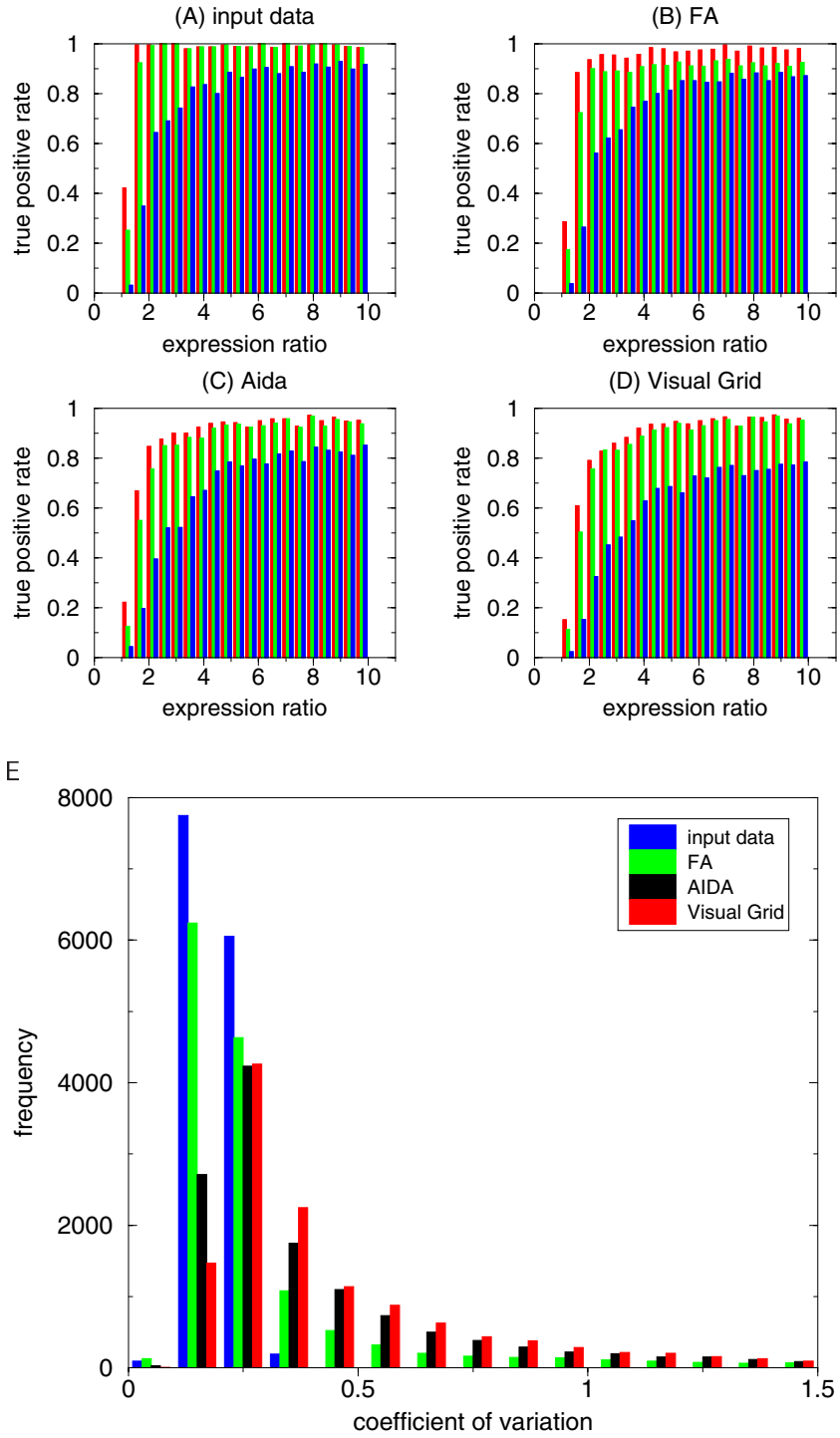


Figure 9

Results of statistical tests for simulated fold-changes. True positive rates of detected simulated fold-changes (P value < 0.01) as given by the Welch test. For all test results the false positive rate is below 0.02. (Histogram intervals have a width of 0.45. The absolute number of regulated clones per interval ranges between 217 and 289.) (**A**) Simulated signals without image analysis (input for the image simulation); and after image analysis of the simulated images with FA (**B**), Aida (**C**), Visual Grid (**D**). For all expression ratio intervals results for 12 (red), 8 (green) and 4 (blue) repetitions are given. (**E**) Histogram of the distribution of the coefficients of variation for sample size 12; The medians of the coefficients of variation are the following: input data: 0.19, FA: 0.21, AIDA: 0.29, Visual Grid: 0.34.

This result shows that data reproducibility increases with the level of automation of the image analysis programs.

Discussion and Conclusions

In this paper we presented a simulation for complex hybridization experiments. This was used to judge critical experimental parameters in the light of the following data analysis. We studied critical parameter of the image analysis by the use of three different image analysis programs representing different levels of automation of the grid-finding and signal quantification. We showed that local distortions of the spot centers like non systematic spot shifting as well as systematic errors resulting in block shifting due to pin errors did not become critical for the reference experiments with the image analysis programs. Also global background noise did not become critical for the experiments studied in this paper. Local background noise might become critical for filter experiments in some cases. Here we showed by the use of fractal clouds as background – which looks very similar to the smear in real experiments – that a mean signal/background ratio below 13 might become critical for some image analysis. However, for the automation of complex hybridizations it might be very helpful to check these parameters during the following data analysis pipeline. This can help to identifiy bad experiments more efficiently. Furthermore it might help to detect sources of error during the experimental procedure or improvements that were made. Although it is possible to get a higher quality of the results by an improvement of the experimental procedure and data analysis algorithms, it is always limited (not at least by the available resources). Furthermore variations of biological material can be expected. To cope with this limitations repetitions of the experiments are indispensable. Not at least due to the fact that array experiments are still very expensive one wants to know how many repetitions are necessary to ensure a certain quality for your expression analysis. For this purpose we did statistical analysis with 4, 8 and 12 repetitions using a Gaussian distributed noise of the input data with $\sigma_2 = 0.2$. Here the image analyses had to cope with changing local backgrounds with the same intensity level. The results of the statistical analysis indicate that for the different image analyses expression ratios below 2 become critical. The relatively poor performance for Visual Grid indicated by the distribution of the CVs is probably due to the fact that this program does no local alignment of the spot position. Since here ideal spot positions were simulated this can explain the relatively good correlation found in Fig. 8 for this program. But due to the manual positioning of the global grid this might become a significant source of error. AIDA and FA do local alignments for the spot positions whereby this source of noise due to manual interaction does not occur.

Automated expression analysis by chip technology will become more and more important in the future, e.g. in biology for comprehensive studies of any kind of developmental processes or in medicine for the study of genetically reasoned diseases. Therefore it is essential to have a well characterized chip technology and subsequent data analysis. This can be supported significantly by well defined models and a whole process simulation. By using well characterized radioactively labeled filter cDNA-arrays, we showed in this paper, that the simulation of this biotechnological method reveals for several parameters the level when they become critical for the follow up data analysis and how this can be improved. Furthermore the simulation environment can also be easily used for the study of cDNA arrays based on glass slides, where e.g. background noise seems to be less critical, but distortions of spot positions and less well characterized spot shapes are more critical.

Authors' Contributions

CKW has written the simulation tool, performed the simulations, designed parts of the model, participated in the image and expression analysis and drafted the manuscript. MS designed parts of the model, participated in expression analysis and assisted in writing the manuscript. TE carried out image analysis. SSK participated in study coordination. PA carried out the mRNA and macroarray preparation and hybridization. MC prepared the used zebrafish cDNA library. HL participated in the study design and coordination. RH conceived of the study, perfomed the statistical analysis, and participated in its design, coordination and manuscript preparation. All authors read and approved the final manuscript.

Abbreviations

CV – coefficient of variation; SD – standard deviation.

Acknowledgments

This work has been financed by the Max-Planck-Society and the BMBF grant No. 01GR0105. We thank M. Albrecht for performing image analysis of experimental images.

References

1. Clark MD, Henning S, Herwig R, Clifton SW, Marra MA, Lehrach H, Johnson SL, the WU-GSC EST group: **An oligonucleotide fingerprint normalized and EST characterized zebrafish cDNA library.** *Genome Research* 2001, **11:**1594-1602
2. Dudoit S, Yang YH, Speed TP, Callow MJ: **Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.** *Statistica Sinica* 2002, **12(1):**111-139
3. Herwig R, Aanstad P, Clark M, Lehrach H: **Statistical evaluation of differential expression on cDNA nylon arrays with replicated experiments.** *Nucleic Acids Research* 2001, **29(23):**e117
4. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression.** *Bioinformatics* 2002, **18(Suppl. 1):**S96-S104
5. Jain AN, Tokuyasu TA, Snijders AM, Segraves R, Albertson DG, Pinkel D: **Fully automatic quantification of microarray image data.** *Genome Research* 2002, **12:**325-332

6. Kepler TB, Crosby L, Morgan KT: **Normalization and analysis of DNA microarray data by self-consistency and local regression.** *Genome Biology* 2002, **3(7):**1-0037
7. Salin H, Vujasinovic T, Mazurie A, Maitrejean S, Menini C, Mallet J, Dumas S: **A novel sensitive microarray approach for differential screening using probes labelled with two different radioelements.** *Nucleic Acids Research* 2002, **30(4):**e17
8. Saupe D: **Algorithms for random fractals** *In The Science of Fractal Images. (Edited by: HO Peitgen, D Saupe) Springer-Verlag, New York, Berlin, Heidelberg* 1988
9. Schuchhardt J, Beule D, Malik A, Wolski E, Eickhoff H, Lehrach H, Herzel H: **Normalization strategies for cDNA microarrays.** *Nucleic Acids Research* 2000, **28(10):**e47
10. Steinfath M, Wruck W, Seidel H, Radelof U, Lehrach H, O'Brien J: **Automated image analysis for array hybridization experiments:** *Bioinformatics* 2001, **17:**634-641
11. Welch BL: **The generalization of Student's problem when several different population variances are involved.** *Biometrika* 1947, **34:**28-35