Estimation of fusion protein copy numbers from noisy fluorescence intensity data of mouse embryonic stem cells

Masterarbeit (Biostatistik)

Anton Hilger

Institut für Statistik
Ludwig-Maximilians-Universität München

Computational Modeling in Biology
Institute for Bioinformatics and Systems Biology
Helmholtz Zentrum München

Betreuer: Carsten Marr
Gutachter: Fabian Theis

HelmholtzZentrum münchen
German Research Center for Environmental Health

Abstract

Single cell time-lapse fluorescence microscopy is a well established technique in system biology to investigate protein expression. Our datasets are from genetically modified mouse embryonic stem cells (mESC), which express Nanog and Oct4 proteins with a VENUS fluorophore tag. The concept of fluorescence microscopy is that the more fusion proteins are in the cell, the brighter is the fluorescence signal. However, no absolute fusion protein copy numbers can be derived from the microscopy data without a conversion factor. It has been shown that the conversion factor can be estimated with intensities form mitosis events, which are intensities from mother cell and daughter cells just before and after cell devision. The main assumption is an equal stochastic apportionment of proteins from mother cell to either daughter cell and binomial distribution is used to estimate absolute protein copy numbers.

However, data analysis shows that our fluorescence intensities are observations with notable multiplicative and additive noise from microscopy imaging. In this master thesis two approaches are developed to estimate both noise parameters, additive and multiplicative, in longitudinal time-lapse fluorescence intensities. The variance approach fits a linear model to the noise terms and the likelihood approach maximizes the likelihood. The two methods give similar results, whereas the variance method is striking simple to use. Multiplicative log normal scale parameter is estimated to $\widehat{\sigma}_m = 0.15 \pm 0.04$ and relative additive normal standard deviation to $\widehat{\sigma}_a/\bar{I} = 0.13 \pm 0.05$.

Therefore the existing methods to estimate protein copy numbers from mitosis events have to be extended to incorporate both additive and multiplicative noise. Again two unbiased approaches are derived, a variance approach using a linear model fit and a likelihood approach which needs a numerical 3-dimensional integration for its maximization. For our best datasets we estimate NanogVENUS and Oct4VENUS copy numbers to 218 and 452 respectively. Reported analysis performed with western blot technique indicate copy numbers from 400 000 to 180 millions for Nanog and Oct4 in mESCs. The discrepancy to our derived copy numbers may indicate that the fusion proteins cluster to multimers prior mitosis.

Contents

1 Introduction

Mouse embryonic stem cells (mESC) were first extracted from the inner cell mass of a blastocyst stage of the embryo 1981 [1]. mESC are able to differentiate into all cell types of the organism and stay pluripotent when kept in the right culture conditions [2]. mESC express amongst others the variant homeodomain proteins Oct4 and Nanog, which are regarded important to keep cells pluripotent: "Nanog seems to be a master gene that makes embryonic stem cells grow in the laboratory. In effect this makes stem cells immortal. Being Scottish, I therefore chose the name after the Tir nan Og legend," said Dr Ian Chambers who isolated the Nanog gene 2003 [3]. Together with Oct4 and other factors, Nanog forms a regulatory network to influence several genes which are important for pluripotency [4] [5] [6]. Elevated levels of Nanog activate these genes and may maintain the self-renewal of the mESC [7].

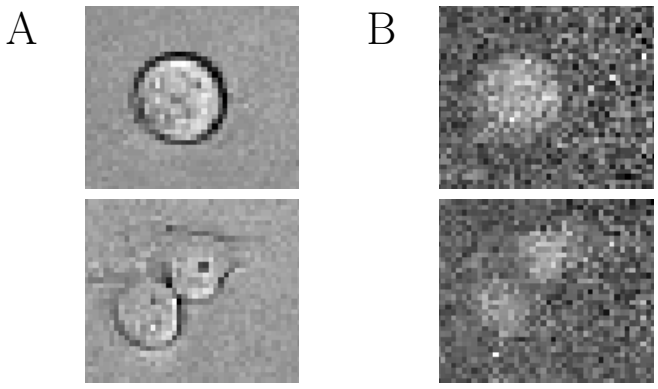

Figure 1: Microscope images of a mESC mother cell and its daughter cells. A: Bright field images. B: NanogVENUS fluorescence images. The images of the mother and of the daughters are taken in 30 minutes intervals.

Recent developments show that cell fate decisions depend crucially on absolute numbers of proteins [8]. The standard in analyzing protein copy numbers in cells is mass spectrometry-based proteomics. Recent developments use labeling of stable isotopes to analyze differential expression levels [9][10]. Another method is to generate a dilution assay of known amounts of fluorophores with western blot technique [11]. When comparing western blots of mESCs fluorophore tagged proteins with these dilution assays their copy numbers can be derived. This technique has already been applied to NanogVENUS and Oct4VENUS fusion proteins in mESCs: Adam Filipczyk from SCD lab of Timm Schroeder at Helmholtz Zentrum München [12] estimated Nanog protein copy numbers to approximately 1.5 millions and Oct4 protein copy numbers to approximately 180 millions per cell. Nick Mullin from the lab of Ian Chambers [13] estimated Nanog protein copy numbers to approximately 400 000 per cell.

In this thesis, we use single mESC fluorescence data generated in the SCD lab of Timm Schroeder post processed in the lab of Fabian Theis, both located at Helmholtz Zentrum München. The mESCs have been genetically manipulated so they express Nanog and Oct4 proteins with VENUS fluorophore tags. The mESCs for the NanogVENUS experiment are preselected with a FACS (=fluorescence activated cell sorting) and only the 5% brightest are taken as initial cells. After illumination with wavelength $\lambda = 515nm$ VENUS tagged proteins re-emit photons with wavelength $\lambda = 528nm$ which can be monitored in vitro with the microscope (see figure 1). One assumption is that on average each fluorophore re-emits the same number of photons (see chapter 3.5.1). Thus the observed intensity in the microscope has a linear relationship with the fusion protein copy number: Double intensity is interpreted as doubled copy number.

However, fluorescence microscopy gives only relative information on the concentration of fluorophore tags. A conversion factor is needed here to derive absolute copy numbers and compare the fluorescence microscopy data to protein numbers derived with other techniques. In the literature statistical methods are described to calculate the conversion factor of proteins in cells from single cell time-lapse microscopy data [14] [15] [16]. In these methods cells are measured before and after mitosis and the variability in the brightness of the two daughter cells is used. Put simply, the higher the number of fusion proteins the lower is the relative variability in the in the daughter cells. A fundamental assumption of these methods is that apportionment of fusion proteins during mitosis to either daughter cell is a stochastic process with equal probability $p = 0.5$. Thus copy number of daughter cells are binomial distributed with size is the copy number of their mother cell.

An example shall illustrate the principle. Petri dish A has 10 mother cells with 10 tagged proteins each, petri dish B has 10 mother cells with 1000 tagged proteins each. After mitosis we have 10 pairs of daughter cells in dish A and dish B with randomly chosen copy numbers (see figure 2). Looking at the relative difference between the copy number in the daughter cells, it is obvious, that the higher the copy number in the mother cell, the lower is the relative variance within the daughter cells. This effect of the binomial distribution can be used to get an estimate for the absolute copy numbers.

Naturally, these methods to estimate the copy numbers are quite sensitive to noise in the intensity measurements. There are various sources of noise in video cameras. Irie et al. [17] distinguish different additive and multiplicative noise sources of a CCD (Charge-coupled Device) camera chip. Possible additive noise sources are read out noise of the chip, dark current shot noise, offset fixed pattern noise, quantization noise,

A	n_{2i}	n_{2i+1}	n_{2i}/n_i	n_{2i+1}/n_i
	6	4	0.6	0.4
	4	6	0.4	0.6
	2	8	0.2	0.8
	4	6	0.4	0.6
	5	5	0.5	0.5

B	n_{2i}	n_{2i+1}	n_{2i}/n_i	n_{2i+1}/n_i
	494	506	0.494	0.506
	504	496	0.504	0.496
	503	497	0.503	0.497
	513	487	0.513	0.487
	498	502	0.498	0.502

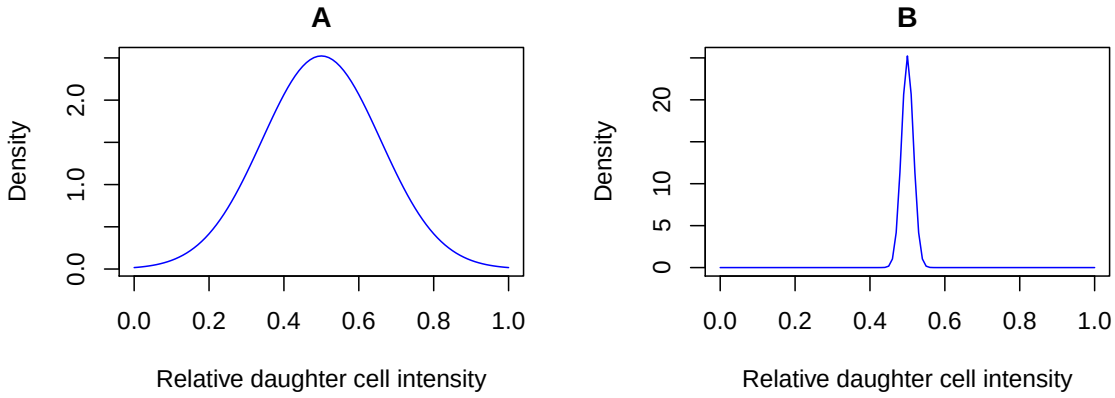

Figure 2: Example of five randomly apportions of fusion proteins from mother to daughter cells, its relative copy numbers and the density lines of its normal approximations. A: Protein copy number mother cell is 10. B: Protein copy number mother cell is 1000.

thermal noise, reset noise, flicker noise or post image capture effects. Illumination depending multiplicative noise sources are photoresponse non-uniformity and photon shot noise. The noise level depends not only on the camera type, but also on the channel (blue, red, green). The authors investigated a commercially available video camera while taking images of a controlled lighting chart and compared this with simulated noisy images of the chart to assign the different noise sources. For the analysed camera they found main contributions from read out noise (additive), photoresponse non-uniformity and photon shot noise (multiplicative). This indicates that typical CCD image data contains multiplicative and additive noise.

Up to now, only additive noise has been studied for the estimation of protein copy numbers [15]. In this thesis we generalize the method to additive plus multiplicative noise. In chapter 2 the used datasets are introduced. We describe fundamental features, apply data cleaning and descriptive data analysis. In chapter 3 the needed methods are derived to estimate noise and signal of the time series to estimate additive and multiplicative noise parameters for intensities and to estimate fusion protein copy numbers from mitosis events. In chapter 4 the new methods are validated with toydata and its capabilities analyzed and finally the methods are applied on the available datasets. The outcome is discussed in chapter 5 and an outlook is given in chapter 6.

2 Descriptive data analysis

2.1 Data overview

Four different fluorescence intensity data sets from NanogVENUS and Oct4VENUS are available (see table 1). In a preselection step only the 5% brightest NanogVENUS mESCs from a bigger group of cells are used as initial cells for our experiment. The preselection was performed with a FACS (=fluorescence activated cell sorting). No such preselection was performed for the Oct4VENUS mESCs. Each data set consists of around 3000 individual cells and includes 1500 mitosis events.

	cells	observations	mitosis events
NanogVENUS raw	3097	32 782	1494
NanogVENUS normalized	3022	32 781	1448
Oct4VENUS raw	2999	61 142	1473
Oct4VENUS normalized	2999	61 142	1473

Table 1: Available fluorescence intensity data. Number of cells, observations and mitosis events before data cleaning.

For raw intensity and normalized intensity data set the same microscope images are used. The raw intensities are the sum over the intensities of the pixels within the segments of the cell. For normalized intensities, additional background correction is performed and the gain of every pixel is considered. For more details on the normalization process please see [18]. The higher number of cells and mitosis events within NanogVENUS raw intensity data set compared to NanogVENUS normalized intensitiy data set is because manual adjustment of the segmentation is done in more cells here. Coincidentally the amount of observations in these two datasets is almost the same.

Time series of mother and its daughter cells give a first overview of the fluorescence intensities (see figure 3 to 6). In this thesis only a randomized subset can be displayed. For better comparison of the effect of normalization process the same cells are chosen from the raw and normalized data set. Fluorescence images are taken every 30 minutes. NanogVENUS fluorescence intensity data are from experiment no. 111115AF6, Oct4VENUS fluorescence intensity data are from experiment no. 110613AF6 with mESCs in culture conditions in one dish. The first label (e.g. p0230) indicates the area within the dish. The second label (001, 002, ...) indicates the number of first mother cell within this area. The third label (AF or MS) indicates the operator who takes care of the experiment. The last label (e.g. 29, 58, 59) indicates the number of mother cell and its daughter cells.

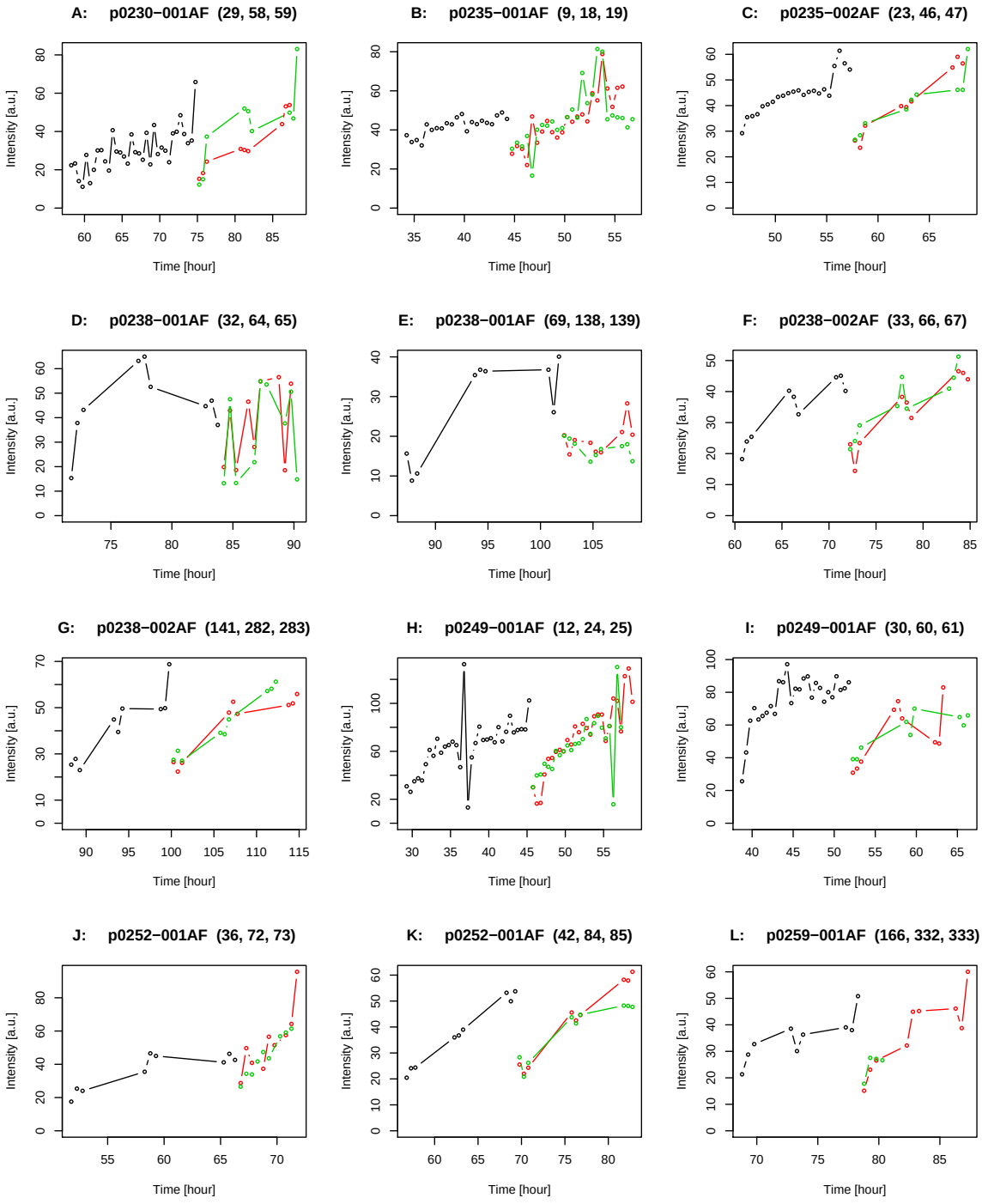

Figure 3: **NanogVENUS raw intensity.** Fluorescence intensity time series before data cleaning for 12 randomly chosen mitosis events. Black points indicate the fluorescence intensity of the mother cell, red and green the fluorescence intensity of its daughter cells.

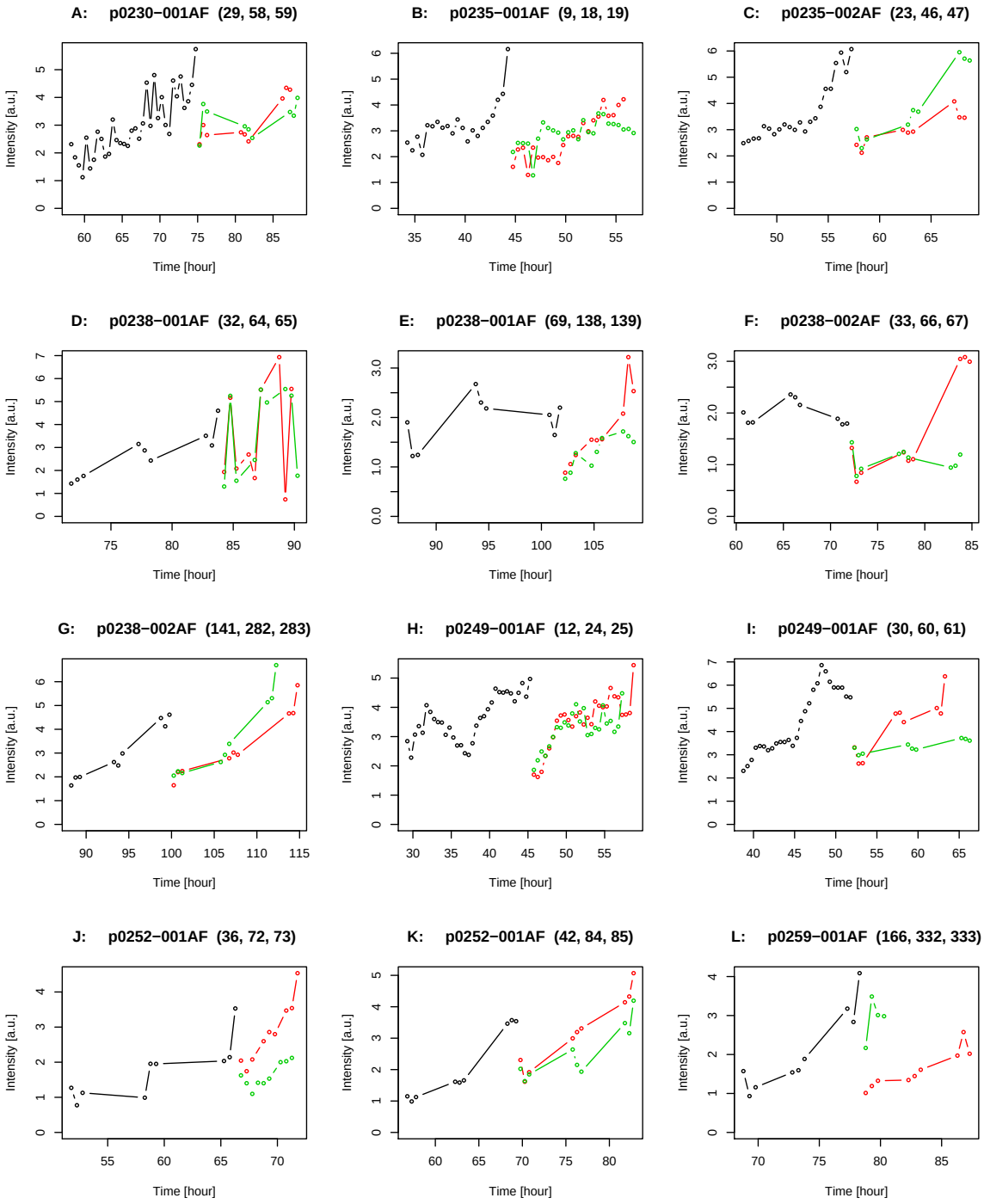

Figure 4: **NanogVENUS normalized intensity.** Fluorescence intensity time series before data cleaning for 12 randomly chosen mitosis events. Black points indicate the fluorescence intensity of the mother cell, red and green the fluorescence intensity of its daughter cells.

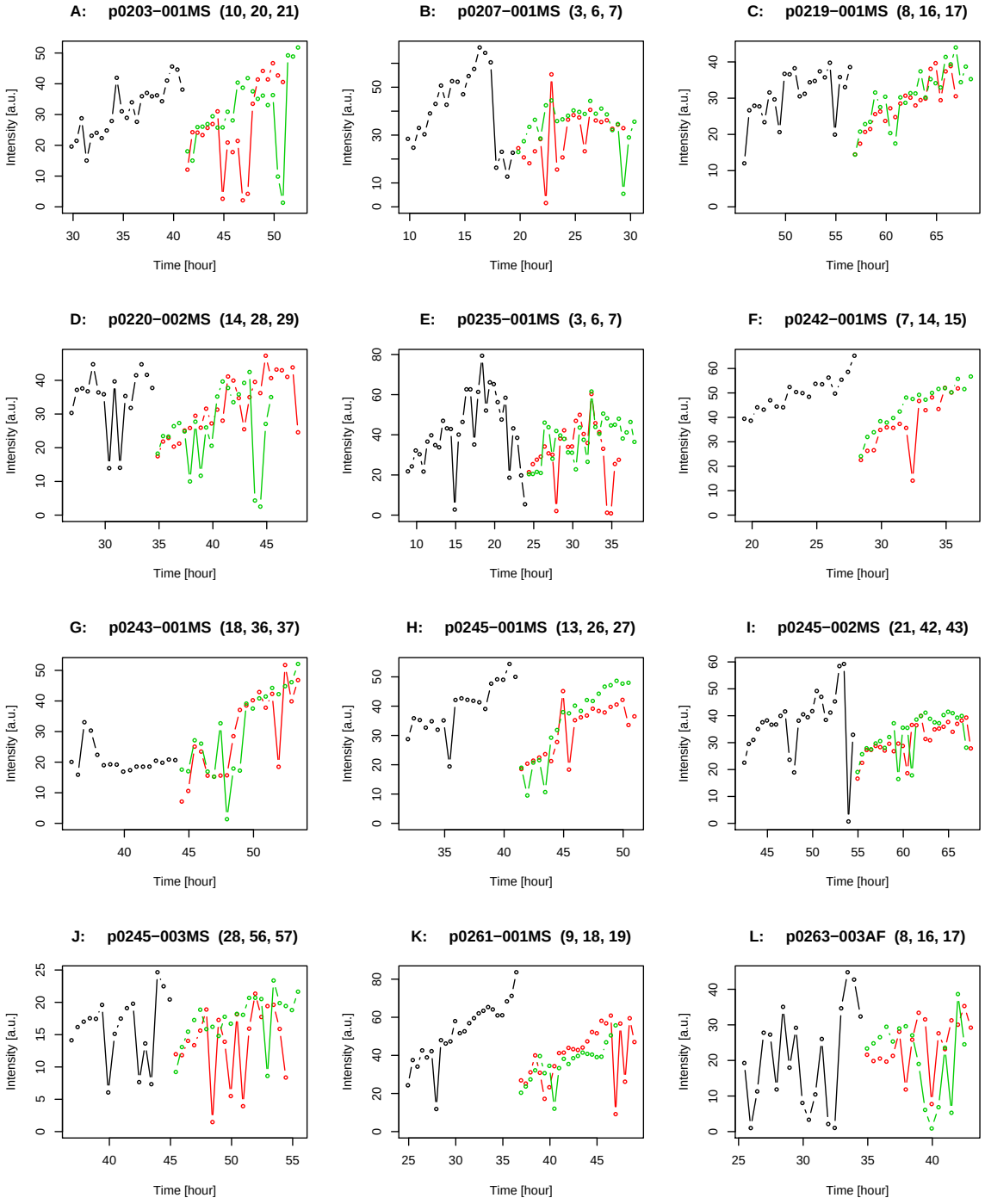

Figure 5: **Oct4VENUS raw intensity.** Fluorescence intensity time series before data cleaning for 12 randomly chosen mitosis events. Black points indicate the fluorescence intensity of the mother cell, red and green the fluorescence intensity of its daughter cells.

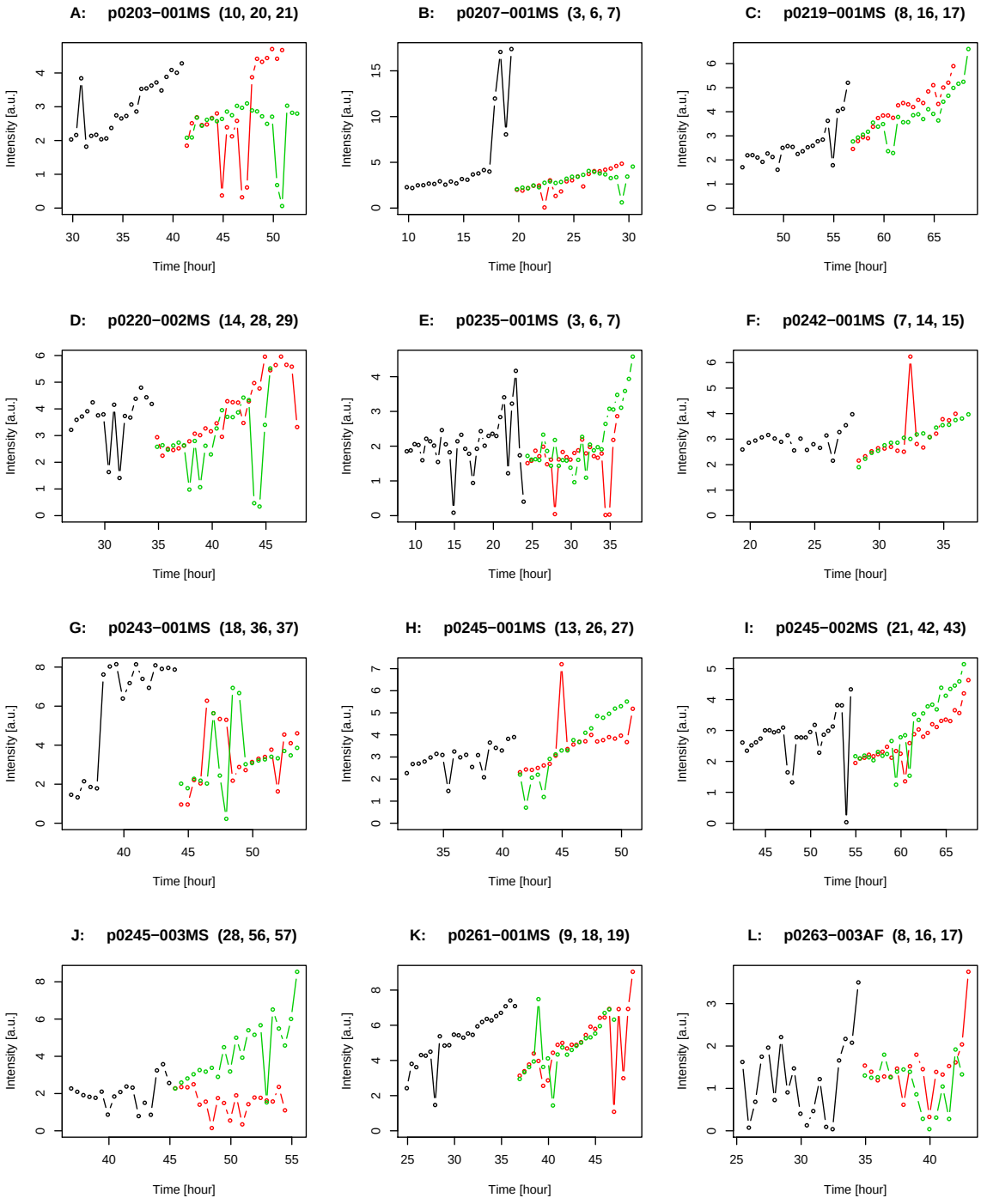

Figure 6: **Oct4VENUS normalized intensity.** Fluorescence intensity time series before data cleaning for 12 randomly chosen mitosis events. Black points indicate the fluorescence intensity of the mother cell, red and green the fluorescence intensity of its daughter cells.

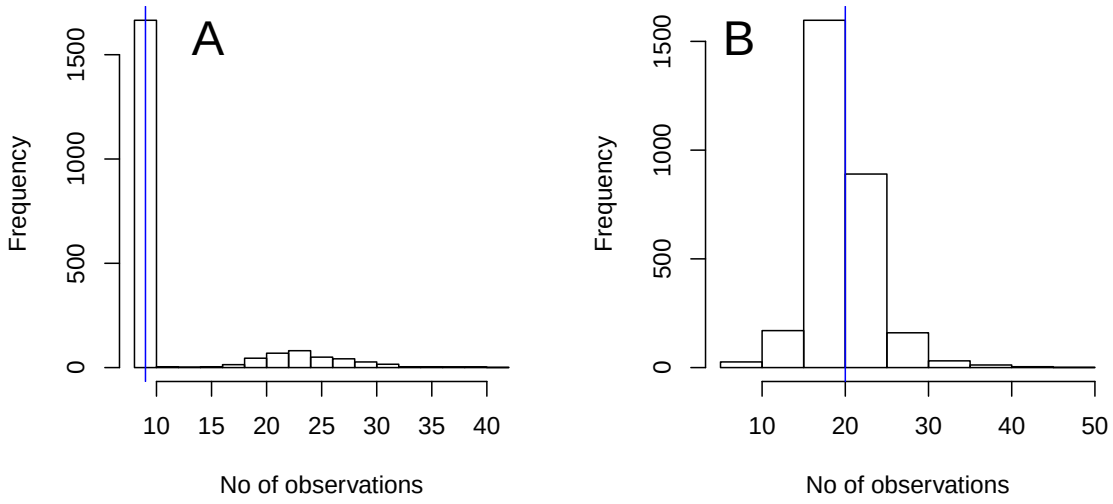

Figure 7: Number of observations per cell. A: NanogVENUS und B: Oct4VENUS. Raw intensity data sets. Blue line indicates the median.

For NanogVENUS, the majority of cells has nine intensity observations only (see figure 7). This is because the segmentation of the cells has been manually adjusted within this data set to get more reliable NanogVENUS fluorescence intensities for the first three, last three and middle three images of every cell. Oct4VENUS data set is not reworked like this and is in lower quality therefore. The median number of cell observations for Oct4VENUS is 20. Raw intensity and normalized intensity data set show the same distribution in number of observations.

One cell cylce last for around 10 hours. The fluorescence intensities in raw intensity data set are about 10 times as high as the fluoresence intensities in normalized intensity data sets. The fluorescence intensity levels of NanogVENUS and Oct4VENUS are of the same magnitude. The fluorescence intensity seems to increase within a cellcycle in all cells. The variation from observation to observation is quite high in some cells (see figure 4 A, figure 5 L). This indicates a fair amount of measurement noise. It seems that fluorescence intensities below 3.0 in raw intensities and 0.3 in normalized intensities have a higher tendency to be incorrect measurements (see figure 5 E, figure 6 I). Also fluorescence intensities above 100 (raw) respectively 10 (normalized) seems to be incorrect more likely (see figure 3 H, figure 6 B).

In some cells time series of the fluorescence intensity changes considerably from raw to normalized intensity (see figure 3 B to figure 4 B; figure 5 B to figure 6 B). The normalized intensities are often more regular as raw intensities (see figure 3 H to figure 4 H; figure 5 C to figure 6 C). As expected, for some cells the intensity of the mother

cell (black) is around the sum of the intensities of its daughters (red, green) at the moment of mitosis (see figure 4 G, figure 6 A). However for some cells the intensity of the daughters does not make the intensity of its mother (see figure 3 J, figure 5 E).

2.2 Data cleaning

To minimize invalid measurements data cleaning is performed. Following rules are used in this order.

1. Fluorescence intensities must be within $[0.3, 10.0]$ for raw intensity and within $[3.0, 100.0]$ for normalized intensity data sets.

2. Minimum number of observations per cell is 9.

Additional data cleaning rules are applied for fluorescence intensities of mitosis events.

3 Fluorescence intensities of mother and both daughter cells must exist.

4 Last three fluorescence intensities of mother cell and first three fluorescence intensities of both daughter cells must be measured within 6 timepoints (2.5 hours).

5 The ratio of median and mean of the last three mother cell intensities respectively of the first three daughter cell intensities must be within $[0.8, 1.2]$.

With data cleaning the number of cells for NanogVENUS data sets are reduced by ca. 35% and for Oct4VENUS data sets by 4% (see tables 1 and 2). The number of observations are reduced for NanogVENUS data sets by ca. 30% and for Oct4VENUS data sets by 5%, the number of mitosis events are reduced for NanogVENUS data sets by ca. 18% and for Oct4VENUS data sets by 16% (see tables 1 and 2). Mitosis fluorescence intensities consists of the median of the last three fluorescence intensities of one mother cell and the median of the first three fluorescence intensities of its daughter cells.

	cells	observations	mitosis events
NanogVENUS raw	2034	23 887	1232
NanogVENUS normalized	1850	22 317	1191
Oct4VENUS raw	2891	57 857	1235
Oct4VENUS normalized	2888	57 185	1233

Table 2: Number of cells, observations and mitosis events after data cleaning. Mitosis fluorescence intensities consists of the median of the last three fluorescence intensities of one mother cell and the median of the first three fluorescence intensities of its daughter cells.

For the following chapters only fluorescence intensities after data cleaning are used.

2.3 Data analysis

The median of fluorescence intensities in NanogVENUS raw intensity data set is 42.5. The median in the normalized intensity data set is 2.82. This means that the normalization process reduces fluorescence intensity here by a factor 15. The median of fluorescence intensities in Oct4VENUS raw intensity data set is 31.9 and after normalization 2.85. NanogVENUS shows 1/3 higher intensities than Oct4VENUS in raw intensity data sets, but after normalization the difference is almost vanished (see table 3).

	Min.	1st Qu.	Median	3rd Qu.	Max.	Mean	SD
NanogVENUS raw	4.05	31.12	42.47	52.75	99.87	43.42	16.57
NanogVENUS normalized	0.31	1.90	2.82	3.88	9.85	3.00	1.48
Oct4VENUS raw	3.00	23.31	31.86	40.54	99.73	32.57	13.55
Oct4VENUS normalized	0.30	2.05	2.85	3.73	10.00	3.01	1.42

Table 3: 5 point summary plus mean and standard deviation of fluorescence intensities.

The fluorescence intensities of all four data sets are right-skewed (see figures 8 A to 11 A). Their right tail is longer and the bulk of the values and the median lie to the left of the mean. In contrary, the logarithmized fluorescence intensities of all four data sets are left-skewed (see figures 8 B to 11 B). A Kolmogorov-Smirnov test with the null hypothesis that the fluorescence intensities follows a normal distribution can be refused for all four data sets with alpha error $p < 2.2 \cdot 10^{-16}$. The same test performed with null hypothesis that the fluorescence intensities follows log-normal distribution can also be refused for all four data sets with the same alpha error. Therefore it can be concluded that NanogVenus and Oct4VENUS fluorescence intensities are neither normally nor log-normally distributed. No bimodality can be seen in the histograms.

Mother cell intensities as well as daughter cell intensities are right-skewed as well (see figures 8C to 11C and figures 8D to 11D). Median of mother cell intensity is 52.3 (NanogVENUS raw), 4.0 (NanogVENUS normalized), 43.5 (Oct4VENUS raw) and 4.0 (Oct4VENUS normalized). Median of daughter cell intensity is 26.7 (NanogVENUS raw), 2.08 (NanogVENUS normalized), 23.1 (Oct4VENUS raw) and 2.27 (OctVENUS normalized). This indicates an fluorescence intensity gain during mitosis, because the median of the daughter cells is higher than the half of the median of the mother cells. The mean fluorescence intensity gain is from 0.24 for NanogVENUS normalized intensity to 4.79 for OCT4VENUS raw intensity data set (see figures 8E to 11E and table 4).

Fluorescence intensity gain is not independent from mother cell fluorescence intensity.

	Mean	SD
NanogVENUS raw	0.71	11.60
NanogVENUS normalized	0.24	0.88
Oct4VENUS raw	4.79	12.14
Oct4VENUS normalized	0.58	0.93

Table 4: Fluorescence intensity gain during mitosis.

Scatter plots show the dependency of the daughter cells intensity from their mother cell intensity and its regression line (red) (see figures 8 F to 11 F). With constant fluorescence intensity gain and a assuming linear relation between the copy number and the fluorescence intensity, we expect a regression line with slope 0.5. However the estimated slope values of the regression lines between daughter and mother cell fluorescence intensities range from 0.23 to 0.48. The slope seems to be higher for NanogVENUS and for normalized intensity data sets. Deviation of the slope from 0.5 might be explained by nonlinear proportion between the copy number of protein and fluorescence intensity. The smaller slope values in raw intensity and for Oct4VENUS data sets indicate that fluorescence intensity increases here more than linear with copy number. Normalization process and manual adjustment of the segmentation seems to help to support the assumption of linear relation between the copy number and fluorescence intensity.

Another assumption is homogeneous partitioning of the fusion protein from mother cell to its daughter cells. Inhomogenous partitioning would mean that for example the probability for one cell is $p_1 = 0.7$ and for its sister it is $p_2 = 0.3$. Histograms of the ratio between daughter cell intensity devided by the sum of both daughter cell intensities show only a central unimodal distribution in all four data sets (see figure 8 G to 11 G), so the assumption of homogeneous partitioning can not be rejected.

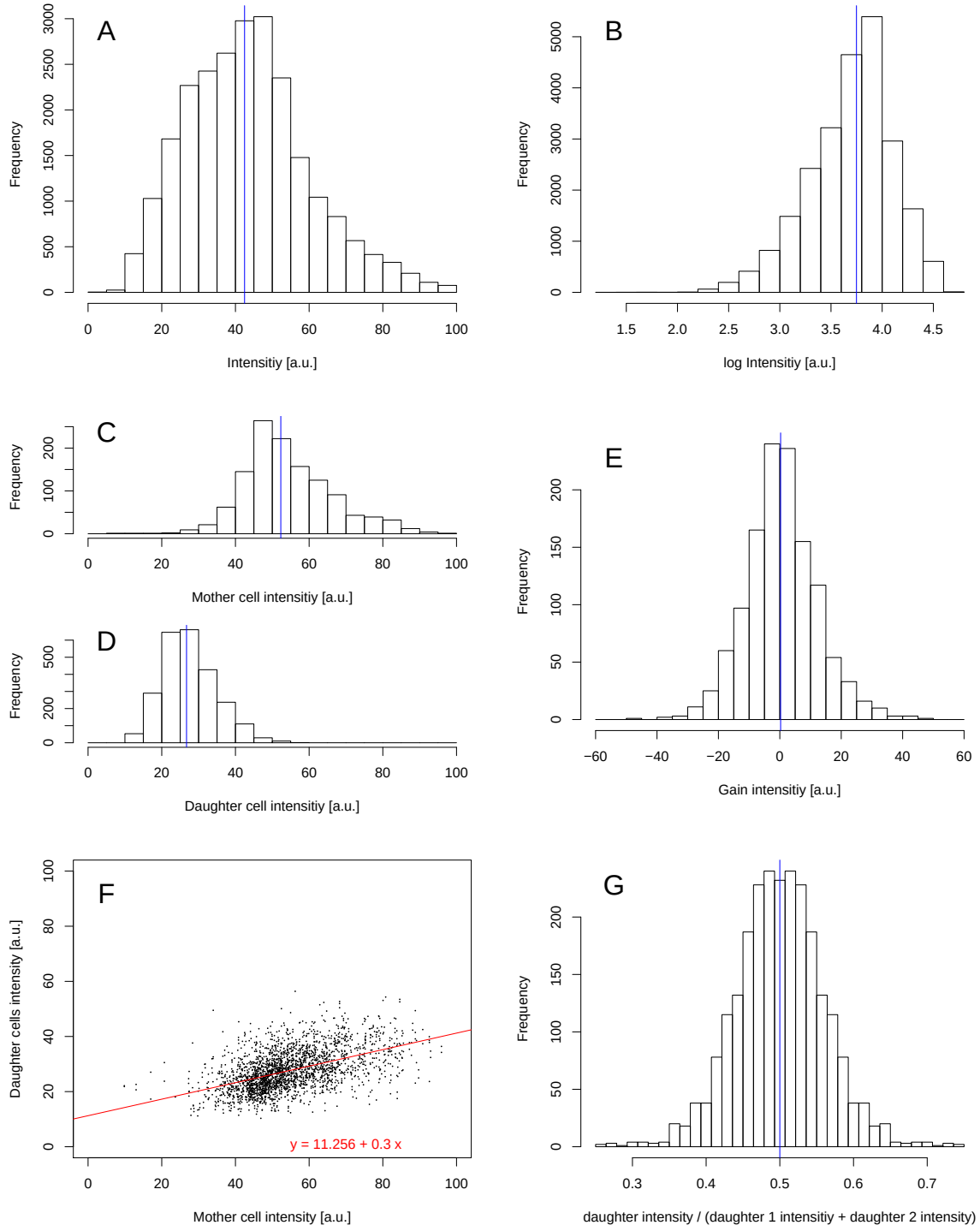

Figure 8: **NanogVENUS raw intensity.** Graphical data analysis. A: Histogram all intensity observations (n=23 887). B: Histogram intensities in logistic scale (n=23 887). C. Histogram mother cell intensities (n=1232). D: Histogram daughter cell intensities (n=2464). E: Histogram intensity gain during mitosis (n=1232). F: Scatterplot daughter cell intensity versus mother cell intensity (n=2464) with regression line. G: Histogram ratio of one daughter cell intensity devided by sum both daughter cells intensity both daughter cells (n=2464). The blue lines indicate the median values.

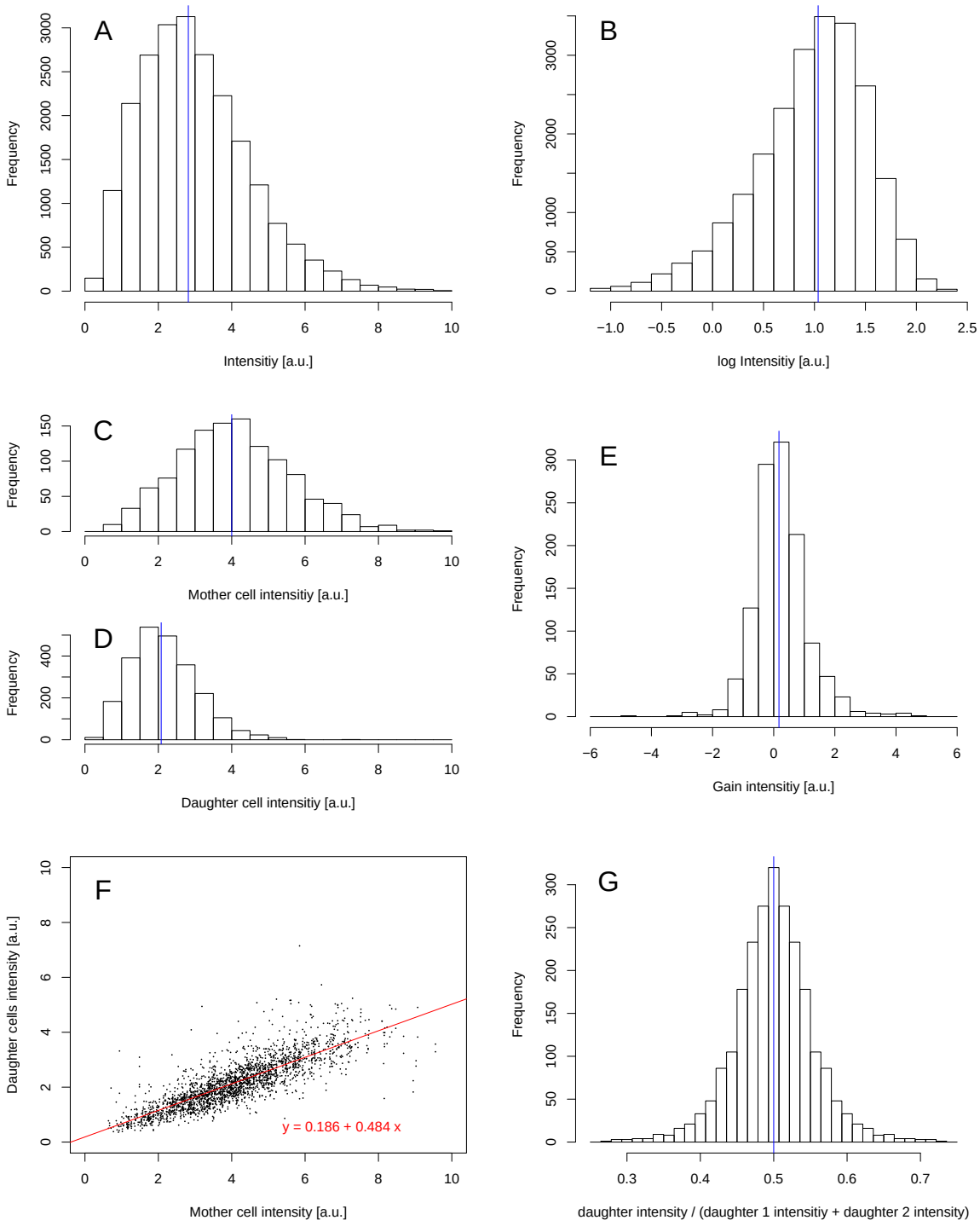

Figure 9: **NanogVENUS normalized intensity.** Graphical data analysis. A: Histogram all intensity observations (n=22 317). B: Histogram intensities in logistic scale (n=22 317). C. Histogram mother cell intensities (n=1191). D: Histogram daughter cell intensities (n=2382). E: Histogram intensity gain during mitosis (n=1191). F: Scatterplot daughter cell intensity versus mother cell intensity (n=2382) with regression line. G: Histogram ratio of one daughter cell intensity devided by sum both daughter cells intensity both daughter cells (n=2382). The blue lines indicate the median values.

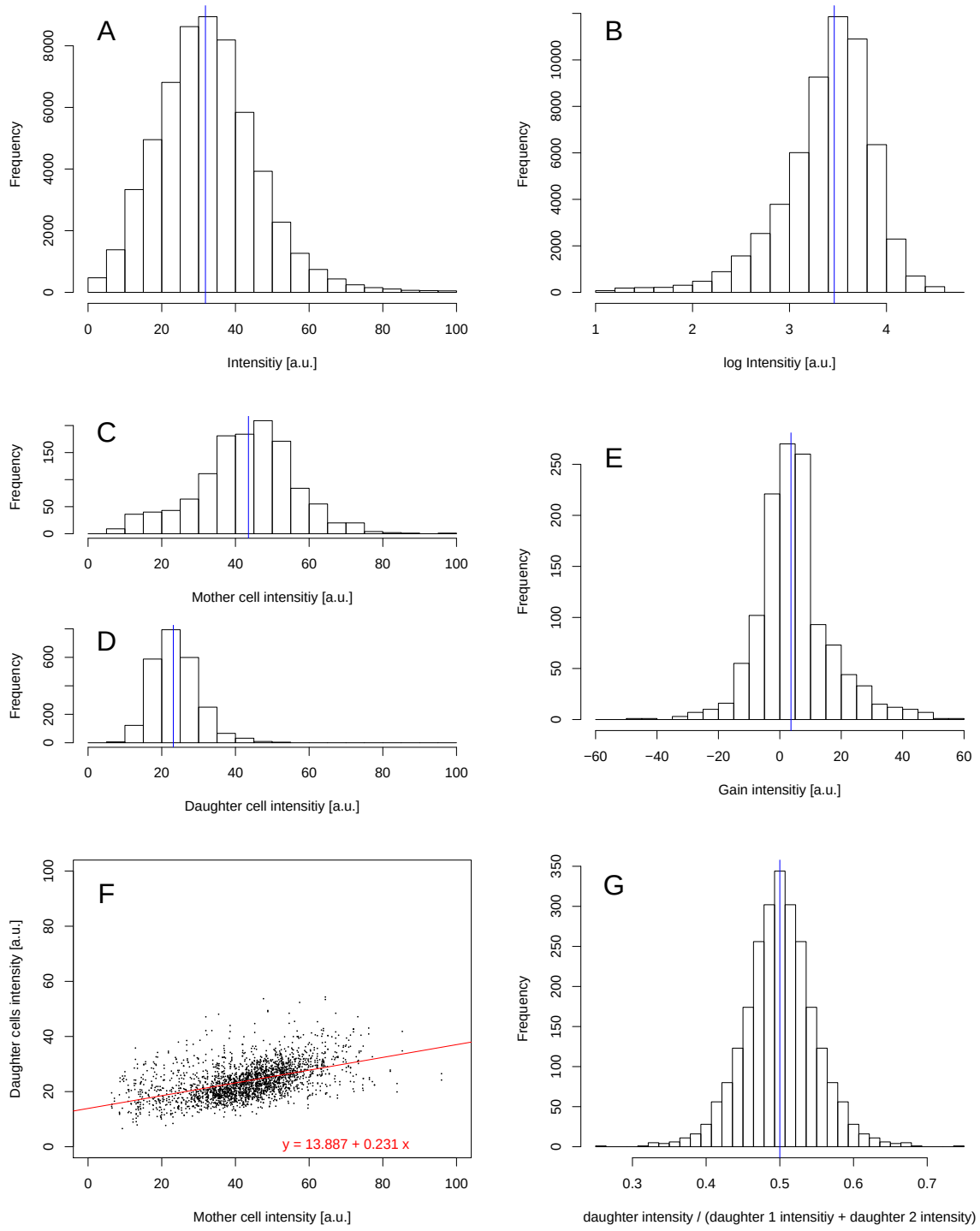

Figure 10: **Oct4VENUS raw intensity.** Graphical data analysis. A: Histogram all intensity observations (n=57 857). B: Histogram intensities in logistic scale (n=57 857). C. Histogram mother cell intensities (n=1235). D: Histogram daughter cell intensities (n=2470). E: Histogram intensity gain during mitosis (n=1235). F: Scatterplot daughter cell intensity versus mother cell intensity (n=2470) with regression line. G: Histogram ratio of one daughter cell intensity devided by sum both daughter cells intensity both daughter cells (n=2470). The blue lines indicate the median values.

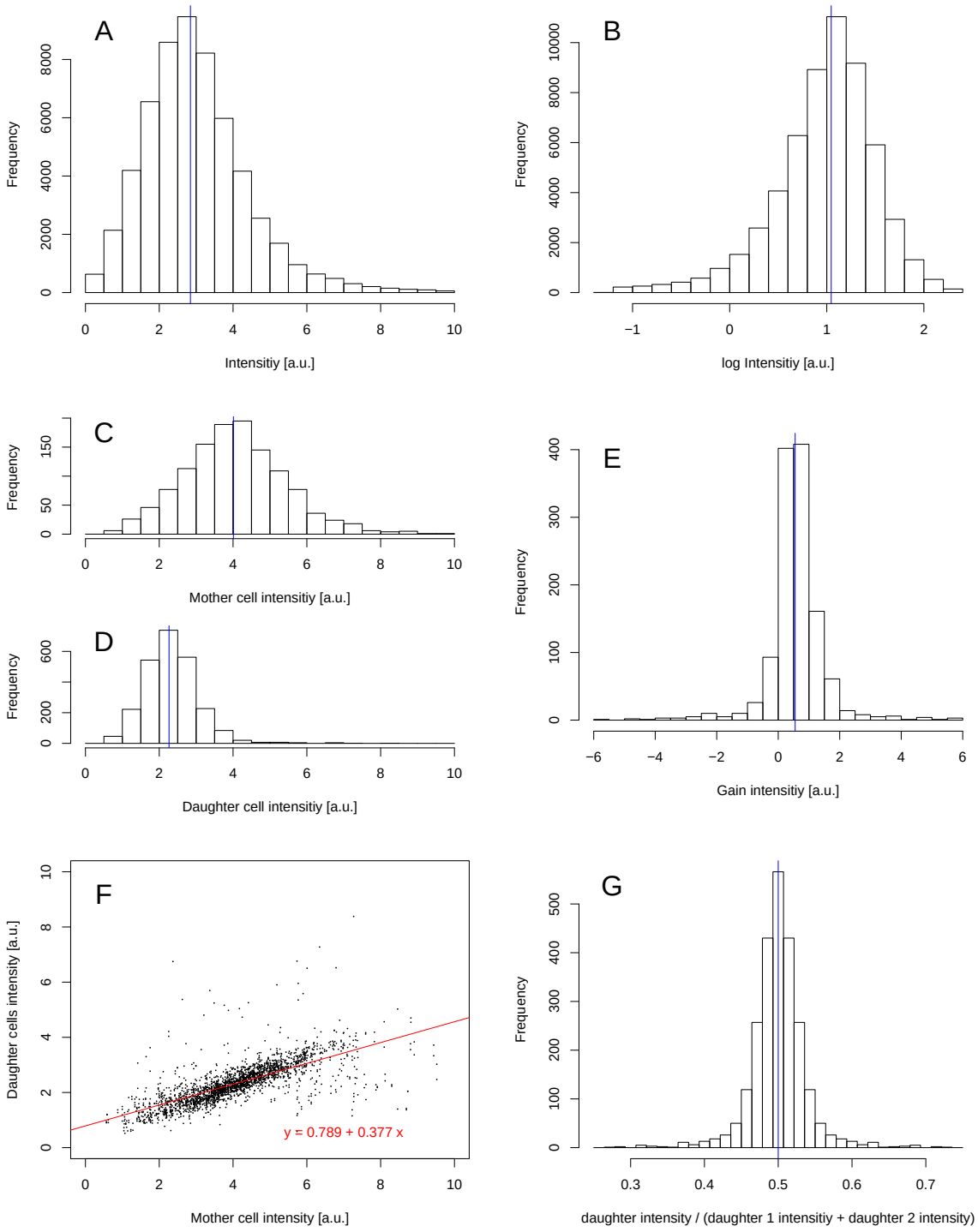

Figure 11: **Oct4VENUS normalized intensity.** Graphical data analysis. A: Histogram all intensity observations (n=57 185). B: Histogram intensities in logistic scale (n=57 185). C. Histogram mother cell intensities (n=1233). D: Histogram daughter cell intensities (n=2466). E: Histogram intensity gain during mitosis (n=1233). F: Scatterplot daughter cell intensity versus mother cell intensity (n=2466) with regression line. G: Histogram ratio of one daughter cell intensity devided by sum both daughter cells intensity both daughter cells (n=2466). The blue lines indicate the median values.

3 Method

3.1 Nomenclature

Following nomenclature and abbreviations are used in this thesis.

Nomenclature		
I	Intensity	Fluorescence intensity with noise of one cell
$I^{'}$	Signal	Fluorescence intensity without noise of one cell
n	Copy number	Number of fusion proteins in one cell
ν	Conversion factor	Parameter of linear relation between copy number n and signal $I^{'}$
N	Number of cells or mitosis events	
i	Index for mother cell	
$2i$	Index for first daughter cell	
$2i+1$	Index for second daughter cell	
ϵ_m	Variable multiplicative noise	
ϵ_a	Variable additive noise	
ϵ_0	Variable total noise	
ϵ_i	Residuum of observation i	
ϵ_{ij}	Residuum of observation j of unit i	
σ_m	Parameter multiplicative noise	
σ_a	Parameter additive noise	
$\sigma_a/\bar{I}$	Parameter relative additive noise	
$\bar{I}$	Mean intensity within a dataset	
age	Lifetime of cell starting from mitosis of its mother cell	
$gain$	Fusion protein gain during mitosis	

Table 5: Nomenclature.

Please note that the daughter cells are not ordered, that is to say the "first" daughter cell is not bigger, brighter or rounder than the other one.

Abbreviation	
mESC	Mouse embryonic stem cell
LMM	Linear Mixed Model
ACF	Auto Correlation Function

Table 6: Used abbreviations.

3.2 Additive and multiplicative noise

We consider two types of noise:

- Multiplicative measurement noise. The random variable ϵ_m is log-normally distributed.
- Additive measurement noise. The random variable ϵ_a is normally distributed.

This means that the intensity I is a function of the signal $I^{'}$ and two independent noise terms, an additive and a multiplicative, $\epsilon_a,\ \epsilon_m$ contribution:

$$I = I^{'} \cdot \epsilon_m + \epsilon_a \qquad with: \ \ \epsilon_a \sim N\left(0, \sigma_a^2\right) \qquad \epsilon_m \sim LN\left(-\frac{\sigma_m^2}{2}, \sigma_m^2\right) \tag{1}$$

Note that $E(\epsilon_m) = 1$ and $var(\epsilon_m) = \exp\left(\sigma_m^2\right) - 1$.

The densities of the independent multiplicative and additive variables are given by

$$f_{\epsilon_a}(\epsilon_a) = \frac{1}{\sqrt{2\pi}\sigma_a}\ \exp\left(-\frac{\epsilon_a^2}{2\sigma_a^2}\right) \tag{2}$$

$$f_{\epsilon_m}(\epsilon_m) = \frac{1}{\sqrt{2\pi}\sigma_m\epsilon_m}\ \exp\left(-\frac{\left(\ln(\epsilon_m) + \frac{\sigma_m^2}{2}\right)^2}{2\sigma_m^2}\right). \tag{3}$$

The total noise ϵ_0 is defined as:

$$\epsilon_0 = I - I^{'} = I^{'}(\epsilon_m - 1) + \epsilon_a$$

Three noise scenarios are considered in the following chapters:

- Additive measurement noise only: $\sigma_a > 0$, $\sigma_m = 0$.
- Multiplicative measurement noise only: $\sigma_a = 0$, $\sigma_m > 0$.
- Multiplicative plus additive measurement noise: $\sigma_a > 0$, $\sigma_m > 0$.

Figure 12 illustrates additive (A), multiplicative (B) and multiplicative plus additive noise (C) in simulated data. Total noise versus signal scattergrams show the homogeneity of variance of the noise term. For additive noise (see figure 12 A), the variance of total noise ϵ_0 is independent from the signal $I^{'}$ and the random variable total noise is so called homoscedastic. For multiplicative noise (see figure 12 B) and multiplicative plus additive noise (see figure 12 C) the spread of the total noise ϵ_0 increases with increasing signal $I^{'}$ and the random variable total noise is so called heteroscedastic.

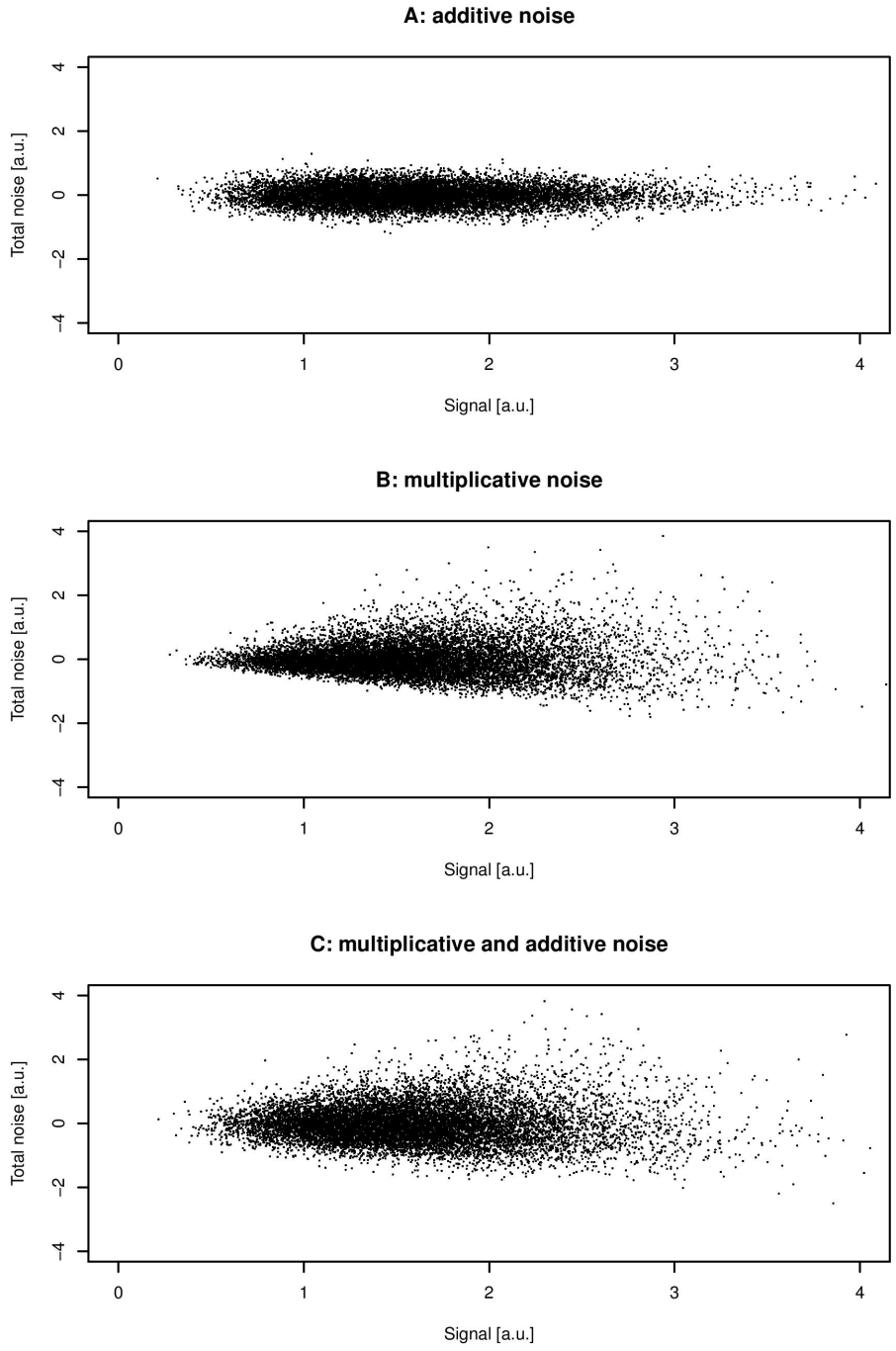

Figure 12: **Simulated data.** Illustration of additive noise (A), multiplicative noise (B) and multiplicative plus additive noise (C).

3.3 Estimation of signal $I^{'}$ and total noise ϵ_0

To estimate noise parameters σ_m and σ_a from fluorescence intensity, estimates for the signal $I^{'}$ and the total noise ϵ_0 are needed. Since single cell time-lapse fluorescence intensities consists of repeated measurements of the same unit (cell), we can apply a Linear Mixed Model (LMM) for longitudinal data to get an estimate for $I^{'}$ and ϵ_0. The assumption of uncorrelated error terms within LMM is analysed with the estimated autocorrelation function (ACF).

3.3.1 Linear Mixed Model

A LMM can be written in general form as [19]:

$$y_{ij} = \beta_0 + \beta_1 x_{i1j} + ... + \beta_n x_{inj} + z_{i0} + z_{i1} x_{i1j} + ... + z_{im} x_{imj} + \epsilon_{ij}$$

$$z_{ik} \sim N\left(0, d_k^2\right) \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$

y_{ij} is the observation j of unit i, $\beta_0, ..., \beta_n$ are the so called fixed effect parameters, $z_{i0}, ..., z_{im}$ are the random effect parameters of unit i with $m \leq n$, ϵ_{ij} is the error term of observation j of unit i.

In this context, observations are intensities, units are cells and fixed and random effects are intensity intercept, age and age^2. age is the time in hour from mitosis of mother cell, where cell is born, until its own mitosis or end of observation.

Six different LMMs are used:

A LMM with fixed linear effect and random intercept:
$I_{ij} = \beta_0 + \beta_1 \cdot age_{ij} + z_{i0} + \epsilon_{ij}$
with fixed intercept β_0, fixed linear β_1 in age and random intercept z_{i0}.

B LMM with fixed linear effect and random intercept and slope:
$I_{ij} = \beta_0 + \beta_1 \cdot age_{ij} + z_{i0} + z_{i1} \cdot age_{ij} + \epsilon_{ij}$
with fixed intercept β_0, fixed linear β_1 in age and random intercept z_{i0} and random linear z_{i1} effect in age.

C LMM with fixed parabola and random intercept:
$I_{ij} = \beta_0 + \beta_1 \cdot age_{ij} + \beta_2 \cdot age_{ij}^2 + z_{i0} + \epsilon_{ij}$
with fixed intercept β_0, fixed linear β_1 and fixed quadratic effect β_2 in age and random intercept z_{i0}.

D LMM with fixed parabola and random intercept and slope:
$I_{ij} = \beta_0 + \beta_1 \cdot age_{ij} + \beta_2 \cdot age_{ij}^2 + z_{i0} + \epsilon_{ij}$
with fixed intercept β_0, fixed linear β_1 and fixed quadratic effect β_2 in age and random intercept z_{i0} and random linear z_{i1} effect in age.

E LMM with fixed parabola and random parabola:
$I_{ij} = \beta_0 + \beta_1 \cdot age_{ij} + \beta_2 \cdot age_{ij}^2 + z_{i0} + z_{i1} \cdot age_{ij} + z_{i2} \cdot age_{ij}^2 + \epsilon_{ij}$
with fixed intercept β_0, fixed linear β_1 and fixed quadratic effect β_2 in age and random intercept z_{i0}, random linear z_{i1} and random quadratic effect z_{i2} in age.

F LMM with fixed exponential and random exponential:
$\ln(I_{ij}) = \beta_0 + \beta_1 \cdot age_{ij} + z_{i0} + z_{i1} \cdot age_{ij} + \epsilon_{ij}$
with fixed intercept β_0 and fixed linear effect β_1 in age and random intercept z_{i0} and random linear effect z_{i1} in age for the logarithmised observation.

In model A to E the estimates for the signal $\widehat{I'}$ of cell i for observation j are the predicted values and the estimates for total noise ϵ_0 are the residuals of the models. In model F the estimates for signal $\widehat{I'}$ are the exponential of the predicted values from the model and the total noise $\widehat{\epsilon_0}$ is the difference between intensity I and estimated signal $\widehat{I'}$.

A linear random intercept:
$\widehat{I'}_{ij} = \beta_0 + \beta_1 \cdot age_{ij} + z_{i0} \quad \text{and} \quad \widehat{\epsilon_0}_{ij} = \epsilon_{ij}$

B linear random slope:
$\widehat{I'}_{ij} = \beta_0 + \beta_1 \cdot age_{ij} + z_{i0} + z_{i1} \cdot age_{ij} + \epsilon_{ij} \quad \text{and} \quad \widehat{\epsilon_0}_{ij} = \epsilon_{ij}$

C parabola random intercept
$\widehat{I'}_{ij} = \beta_0 + \beta_1 \cdot age_{ij} + \beta_2 \cdot age_{ij}^2 + z_{i0} \quad \text{and} \quad \widehat{\epsilon_0}_{ij} = \epsilon_{ij}$

D parabola random slope:
$\widehat{I'}_{ij} = \beta_0 + \beta_1 \cdot age_{ij} + \beta_2 \cdot age_{ij}^2 + z_{i0} + z_{i1} \cdot age_{ij} \quad \text{and} \quad \widehat{\epsilon_0}_{ij} = \epsilon_{ij}$

E parabola random parabola:
$\widehat{I'}_{ij} = \beta_0 + \beta_1 \cdot age_{ij} + \beta_2 \cdot age_{ij}^2 + z_{i0} + z_{i1} \cdot age_{ij} + z_{i2} \cdot age_{ij}^2 + \epsilon_{ij} \quad \text{and} \quad \widehat{\epsilon_0}_{ij} = \epsilon_{ij}$

F exponential random exponential:
$\widehat{I'}_{ij} = \exp\left(\beta_0 + \beta_1 \cdot age_{ij} + z_{i0} + z_{i1} \cdot age_{ij}\right) \quad \text{and} \quad \widehat{\epsilon_0}_{ij} = I_{ij} - \widehat{I'}_{ij}$

3.3.2 Autocorrelation

For a linear model, as well as for a LMM, we assume that the error terms are uncorrelated. This assumption can be analysed with the estimation of an autocorrelation

function (ACF) of the residuals of the LMM. Autocorrelation is the correlation of a variable with itself, but shifted with 1, 2, and more lags.

Within a LMM, the ACF might be estimated for the residuals of every cell individually. This leads to n_0 estimated coefficients for every lag. These coefficients are displayed with boxplots.

3.3.3 Model comparison

There are several approaches to compare different linear models. Most common are AIC, BIC and $\bar{R}^2$. AIC and BIC are based on the maximized log-likelihood with penalty for the complexity of the model. Within LMMs the complexity is rated as the number of fixed and random effects [19]. The number of random effects are rated here with the same complexity as the number of fixed effects although that for the random effects n_0 individual parameters (for every cell) must be estimated whereas for the fixed effect only one overall parameter is estimated. This penalization is appropriate if the LMM is used for prediction only, however we estimate the signal and noise with our models and rate this proposed penalization as not advisable. Moreover AIC and BIC use the log-likelihood, which is not straightforward for our model F with logarithmized response variable.

Because of these reason we use adjusted R^2 to compare different LMMs:

$$\overline{R}^2 = 1 - \frac{n-1}{n-p}(1-R^2) \tag{4}$$

with

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

and number of observations N and number of regressors p. For every random effect we take the number n_0 of estimated individual parameters for penalization,
$p =$ (number of fixed effects) $+ n_0 \cdot$ (number of random effects).

3.4 Estimation of noise parameters from longitudinal data

In this chapter methods to estimate the noise parameters σ_a and σ_m from fluorescence intensity data are introduced. These methods are not limited to time lapse fluorescence data but also applicable for all longitudinal observations with additive and/or multiplicative noise, if an estimate for the signal $I^{'}$ and the noise ϵ_0 can be derived.

The estimation of noise parameters is quite simple for additive or multiplicative noise only. For the combined situation, additive plus multiplicative noise, two different approaches are shown. The first approach uses a likelihood technique and needs numeric integration, the second one uses a variance approach and fits a linear model to the noise values.

3.4.1 Additive noise

For a model with normally distributed additive noise only, $\sigma_a > 0$, $\sigma_m = 0$ ($I = I^{'} + \epsilon_a$), the parameter σ_a can be estimated with the maximum likelihood estimator for the standard error of the normal distribution.

$$\widehat{\sigma_a} = \sqrt{\frac{1}{1-N}\sum_{i=1}^{N}(\epsilon_{a_i} - \bar{\epsilon}_a)^2} = \sqrt{\frac{1}{1-N}\sum_{i=1}^{N}\epsilon_{a_i}^2}$$

3.4.2 Multiplicative noise

For a model with log-normally distributed multiplicative noise only,$\sigma_a = 0$, $\sigma_m > 0$ ($I = I^{'} \cdot \epsilon_m$), the parameter σ_m can be estimated with the maximum likelihood estimator for the parameter of the log-normal distribution.

$$\widehat{\sigma_m} = \sqrt{\ln\left(\frac{var(\epsilon_m)}{E^2(\epsilon_m)} + 1\right)} = \sqrt{\ln\left(var(\epsilon_m) + 1\right)}$$

with $\widehat{var(\epsilon_m)} = \frac{1}{1-n}\sum_{i=1}^{n}(\epsilon_{m_i} - \bar{\epsilon}_m)^2$.

3.4.3 Multiplicative plus additive noise

There is no closed-form solution for the maximum likelihood estimator of the parameters σ_m and σ_a for a model with multiplicative plus additive noise. Two different approaches are discussed here: (i) Likelihood technique using numerical integration and (2) a variance approach which fits a linear model to the noise values ϵ_0.

3.4.3.1 Likelihood approach

The joint density function for ϵ_m and ϵ_a is the product of the single density functions $f_{\epsilon_m}(\epsilon_m)$ (see equation 2) and $f_{\epsilon_a}(\epsilon_a)$ (see equation 3) since the multiplicative and additve noise variables are independent.

$$\begin{aligned} f_{\epsilon_m,\epsilon_a}(\epsilon_m,\epsilon_a) &= f_{\epsilon_m}(\epsilon_m)\cdot f_{\epsilon_a}(\epsilon_a) \\ &= \frac{1}{2\pi\sigma_m\sigma_a\epsilon_m}\ \exp\left(-\frac{\left(\ln(\epsilon_m)+\frac{\sigma_m^2}{2}\right)^2}{2\sigma_m^2}-\frac{\epsilon_a^2}{2\sigma_a^2}\right) \end{aligned} \tag{5}$$

With transformation for bivariate density functions (appendix 7.1.2) this joint density function can be transformed to get the joint density function for total noise ϵ_0 and additive noise ϵ_a, $f_{\epsilon_0,\epsilon_a}(\epsilon_0,\epsilon_a)$. Total noise is $\epsilon_0 = I^{'}(\epsilon_m-1)+\epsilon_a$ and the transformation is $T(\epsilon_0,\epsilon_a) = (\epsilon_0 = I^{'}(\epsilon_m-1)+\epsilon_a, \epsilon_a = \epsilon_a)$. The inverse transformation function is $T^{-1}(\epsilon_0,\epsilon_a) = \left(\frac{\epsilon_0-\epsilon_a}{I'}+1, \epsilon_a\right)$. The density after transformation is

$$f_{\epsilon_0,\epsilon_a}(\epsilon_0,\epsilon_a) = \frac{1}{2\pi\sigma_m\sigma_a\left(\frac{\epsilon_0-\epsilon_a}{I'}+1\right)}\ \exp\left(-\frac{\left(\ln\left(\frac{\epsilon_0-\epsilon_a}{I'}+1\right)+\frac{\sigma_m^2}{2}\right)^2}{2\sigma_m^2}-\frac{\epsilon_a^2}{2\sigma_a^2}\right)\cdot\frac{1}{I'}\ ,$$

while the Jacobi determinant is $\frac{1}{I'}$:

$$J(y_1,y_2) = \begin{vmatrix} \frac{\partial\frac{\epsilon_0-\epsilon_a}{I'}+1}{\partial\epsilon_0} & \frac{\partial\epsilon_a}{\partial\epsilon_0} \\ \frac{\partial\frac{\epsilon_0-\epsilon_a}{I'}+1}{\partial\epsilon_a} & \frac{\partial\epsilon_a}{\partial\epsilon_a} \end{vmatrix} = \begin{vmatrix} \frac{1}{I'} & 0 \\ -\frac{1}{I'} & 1 \end{vmatrix} = \frac{1}{I'}\ .$$

The desired univariate density function for ϵ_0 is obtained after integration of the joint density function,

$$\begin{aligned} f_{\epsilon_0}(\epsilon_0) &= \int f_{\epsilon_0,\epsilon_a}(\epsilon_0,\epsilon_a)\ d\epsilon_a \\ &= \int_{-\infty}^{I^{'}+\epsilon_0} \frac{1}{2\pi\sigma_m\sigma_a\left(\epsilon_0-\epsilon_a+I'\right)}\ \exp\left(-\frac{\left(\ln\left(\frac{\epsilon_0-\epsilon_a}{I'}+1\right)+\frac{\sigma_m^2}{2}\right)^2}{2\sigma_m^2}-\frac{\epsilon_a^2}{2\sigma_a^2}\right)\ d\epsilon_a\ . \end{aligned}$$

The integration limits are because the multiplicative noise variable is log-normally distributed, $\epsilon_m \sim LN\left(\frac{\sigma_m^2}{2}, \sigma_m^2\right)$. With co-domain $\epsilon_m \in]0,\infty[$ follows $\epsilon_a \in]-\infty, I^{'}+\epsilon_0[$

with $\epsilon_a = \epsilon_0 - I^{'}(\epsilon_m - 1)$.

Likelihood L of the parameters σ_a^2 and σ_m^2 given some data is the product of the density function for each observation ϵ_0.

$$
\begin{aligned}
L(\sigma_m^2, \sigma_a^2|\boldsymbol{\epsilon_0}) &= \prod_{i=1}^{n} f_{\epsilon_0}(\epsilon_{0_i}|\sigma_m^2, \sigma_a^2) \\
&= \prod_{i=1}^{n} \int_{-\infty}^{I^{'}+\epsilon_0} \frac{1}{2\pi\sigma_m\sigma_a\left(\epsilon_{0_i} - \epsilon_a + I^{'}\right)} \exp\left(-\frac{\left(\ln\left(\frac{\epsilon_{0_i}-\epsilon_a}{I^{'}}+1\right)+\frac{\sigma_m^2}{2}\right)^2}{2\sigma_m^2} - \frac{\epsilon_a^2}{2\sigma_a^2}\right) d\epsilon_a
\end{aligned}
$$

The log-likelihood l is the logarithm of the likelihood L.

$$
\begin{aligned}
&l(\sigma_m^2, \sigma_a^2|\boldsymbol{\epsilon_0}) = \\
&= \sum_{i=1}^{n} \ln\left(\int_{-\infty}^{I^{'}+\epsilon_0} \frac{1}{2\pi\sigma_m\sigma_a\left(\epsilon_{0_i} - \epsilon_a + I^{'}\right)} \exp\left(-\frac{\left(\ln\left(\frac{\epsilon_{0_i}-\epsilon_a}{I^{'}}+1\right)+\frac{\sigma_m^2}{2}\right)^2}{2\sigma_m^2} - \frac{\epsilon_a^2}{2\sigma_a^2}\right) d\epsilon_a\right)
\end{aligned} \tag{6}
$$

The log-likelihood l may be evaluated and maximized numerically for σ_m^2 and σ_a^2 to get the maximum likelihood estimator for the noise parameters.

Alternatively to the use of the density transformation we can get the density function $f_{\epsilon_a}(\epsilon_0)$ through integration of the joint density $f_{\epsilon_m,\epsilon_a}(\epsilon_m, \epsilon_a)$ from equation 5.

$$
\begin{aligned}
f_{\epsilon_a}(\epsilon_0) &= \int_0^{\infty} f_{\epsilon_m}(\epsilon_m) \cdot f_{\epsilon_a}(\epsilon_a = \epsilon_0 - I^{'}(\epsilon_m - 1))\, d\epsilon_m \\
&= \int_0^{\infty} \frac{1}{2\pi\sigma_m\sigma_a\epsilon_m} \exp\left(-\frac{\left(\ln(\epsilon_m)+\frac{\sigma_m^2}{2}\right)^2}{2\sigma_m^2} - \frac{\left(\epsilon_0 - I^{'}(\epsilon_m - 1)\right)^2}{2\sigma_a^2}\right) d\epsilon_m
\end{aligned} \tag{7}
$$

Maximization of the log-likelihood using this alternative density function in equation 7 leads to the same maximum likelihood estimator for σ_m^2 and σ_a^2.

Confidence intervals for ML-estimators

Let the parameter set θ be one dimensional. Under certain regularity constraints (which are independent identically distributed observations or at least the information in the data increases with sample size, estimates of parameters lie not on boundary and/or the boundary is not dependent from the parameter itself and the number of nuisance parameters must not increase with number of observations [19]) the maximum likelihood estimator $\widehat{\theta}$ is asymptotically normally distributed and the log-likelihood $l(\widehat{\theta}, x)$ has the form of a parabola at the maximum likelihood estimator $\widehat{\theta}$ in large samples. With the second derivative of the log-likelihood, l'', an asymptotic 95% confidence interval can be written as

$$CI_{95\%} = \widehat{\theta} \pm 1.96 \cdot \frac{1}{-l''(\widehat{\theta}, x)} \ .$$

Our parameter set $\theta = (\sigma_m^2, \sigma_a^2)$ is two dimensional. This implies that the confidence interval is a confidence region. For easier interpretation the confidence interval are calculated for each parameter individually with the second parameter is set to the value of its maximum likelihood estimator.

$$CI_{\sigma_m 95\%} = \widehat{\sigma_m} \pm 1.96 \cdot \frac{1}{-\frac{\partial^2}{\partial^2 \sigma_m^2} l(\widehat{\sigma_m^2}, \widehat{\sigma_a^2} | \epsilon_{\mathbf{0}})} \tag{8}$$

$$CI_{\sigma_a 95\%} = \widehat{\sigma_a} \pm 1.96 \cdot \frac{1}{-\frac{\partial^2}{\partial^2 \sigma_a^2} l(\widehat{\sigma_m^2}, \widehat{\sigma_a^2} | \epsilon_{\mathbf{0}})} \tag{9}$$

These are the diagonal elements of the negative Hessian matrix or the so called observed information.

3.4.3.2 Variance approach

In this paragraph a novel method to estimate the noise parameter from fluorescence data with additive and multiplicative noise is described. It compares the variances and fits a linear model to the noise values ϵ_0.

The total noise of one observation is $\epsilon_0 = I - I' = I'(\epsilon_m - 1) + \epsilon_a$. To derive an estimate for the noise parameters σ_m and σ_a the variance of the noise of one observation ϵ_0 shall be considered.

$$
\begin{aligned}
var(\epsilon_0 = I^{'}(\epsilon_m - 1) + \epsilon_a) &= var(I^{'}\epsilon_m + \epsilon_a) \\
&= I^{'2} var(\epsilon_m) + var(\epsilon_a) + 2I^{'} cov(\epsilon_m, \epsilon_a) \\
&= I^{'2}\left(e^{\sigma_m^2} - 1\right) + \sigma_a^2 + 0 \\
&\stackrel{!}{=} E(\epsilon_0^2) - E^2(\epsilon_0) \\
&= E(\epsilon_0^2)
\end{aligned} \tag{10}
$$

The last step holds true, because of $E(\epsilon_0) = E(I^{'}(\epsilon_m - 1) + \epsilon_a) = I^{'}(E(\epsilon_m) - 1) + E(\epsilon_a) = 0 + 0$.

Thus a linear model might be fitted, $\epsilon_{0_i}^2 = \beta_0 + \beta_1 I_i^{'2} + \epsilon_i$ to estimate $E(\epsilon_0^2)$. An estimator for the noise parameters σ_m and σ_a can be derived with comparing the coefficients from the linear model with equation 10

$$
\begin{aligned}
\widehat{\sigma_a}^2 &= \widehat{\beta_0} \\
I^{'2}\left(e^{\widehat{\sigma_m}^2} - 1\right) &= \widehat{\beta_1}
\end{aligned}
$$

and the estimator can be written as

$$
\widehat{\sigma_a} = \sqrt{\widehat{\beta_0}} \tag{11}
$$

$$
\widehat{\sigma_m} = \sqrt{\ln\left(\widehat{\beta_1} + 1\right)}\,. \tag{12}
$$

3.5 Estimation of copy numbers from mitosis events

3.5.1 Assumptions for copy numbers estimation

For all described methods to estimate copy numbers of proteins following three assumptions hold:

Assumption 1: $I^{'} = \nu \cdot n$

Assumption 2: $n_{2i} \sim Bi(n_i, p = 0.5)$

Assumption 3: $n_i = n_{2i} + n_{2i+1}$

About assumption 1: The fluorescence intensity I is regarded linear in the copy number of fusion proteins n with conversion factor ν.

About assumption 2: Each fusion protein from the mother cell has the same chance to be allocated in one of the two daughter cells, so copy number n_{2i} of daughter cell follows a binomial distribution with probability 0.5. A close approximation for $n_i \cdot p > 5$ of this binomial distribution is a normal distribution with mean $n_i p = n_i/2$ and variance $n_i \cdot p(1-p) = n_i/4$.

About assumption 3: Conversion of fusion protein holds during mitosis. Adjustments are done in the methods which consider multiplicative plus additive noise to include the observed fusion protein gain (see table 4).

Following variances can be derived when using these assumptions:

$$\begin{aligned} var(n_{2i}) &= n_i\, p(1-p) = \frac{1}{4} n_i \\ var(I'_{2i}) &= var(\nu\, n_{2i}) = \nu^2 \frac{1}{4} n_i = \frac{1}{4} I'_i \nu \end{aligned} \tag{13}$$

$$\begin{aligned} var(n_{2i} - n_{2i+1}) &= var(n_{2i} - (n_i - n_{2i})) = var(2n_{2i}) = n_i \\ var(I'_{2i} - I'_{2i+1}) &= \nu^2 \cdot var(n_{2i} - n_{2i+1}) = I'_i \nu \end{aligned}$$

$$var\left(\frac{I'_{2i} - I'_{2i+1}}{I'_i}\right) = \frac{1}{n_i} \tag{14}$$

Please note that the signal I'_i and the copy number n_i of the mother cell are not random variables.

3.5.2 No noise

3.5.2.1 Likelihood approach

This method is an "approximate solution" [15], because it does not take any measurement error into account. Rosenfeld and colleagues [15] observed a fusion protein in *Escherichia coli* which is not expressed nor degraded during microcolony growth. By washing out the inducer they are able to stop the fluorescent gene expression. Therefore the colony contains a fixed amount of fluorophores which keeps constant within a lineage tree. In contrast, our data is based on cells, which express NanogVENUS and Oct4VENUS fusion proteins during their cell cycle. Conversion holds approximately only for one mitosis event. Therefore we adjust the original approach to fit it to our data.

Without measurement error, $I = I^{'}$, constant number of proteins means that the intensity is preserved during one cell devision. Mother fluorescence intensity is then equal the sum of its daughters, $I_i = I_{2i} + I_{2i+1}$.

The density of the approximated normal distribution for n_{2i} is

$$f_{n_{2i}}(n_{2i}|n_i) = \sqrt{\frac{2}{\pi}} \frac{1}{\sqrt{n_i}} \exp\left(-\frac{(2n_{2i} - n_i)^2}{2n_i}\right) . \tag{15}$$

The density $f_{n_{2i}}(n_{2i}|n_i)$ can be transformed with theorem of transformation of univariate probability density functions (see chapter 7.1.1) to get $f_{I_{2i}}(I_{2i}|I_i, \nu)$.

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left|\frac{\partial g^{-1}(y)}{\partial y}\right|$$

Here, $y = I_{2i}$, $x = n_{2i}$ and $g(n_{2i}) : I_{2i} = \nu\, n_{2i}$. It follows that

$$f_{I_{2i}}(I_{2i}|I_i, \nu) = \sqrt{\frac{2}{\pi}} \frac{1}{\sqrt{\nu I_i}} \exp\left(-\frac{(2I_{2i} - I_i)^2}{2I_i\nu}\right) , \tag{16}$$

where n_i can not be observed and is replaced with $n_i = I_i/\nu$. The likelihood $L(\nu|I_i, I_{2i})$ for the conversion factor ν given fluorescence intensities I_i and I_{2i} is the product over all N independent mitosis events:

$$L(\nu|I_i, I_{2i}) = \left(\frac{2}{\pi}\right)^{N/2} \nu^{-N/2} \left(\prod_i^N I_i^{-0.5}\right) \cdot \exp\left(-\frac{1}{\nu}\sum_i^N \frac{(2I_{2i} - I_i)^2}{2I_i}\right)$$

The maximum likelihood estimator $\hat{\nu}$ is found by differentiation of the log likelihood $l(\nu|I_i, I_{2i})$,

$$l(\nu|I_i, I_{2i}) = \frac{N}{2}\ln\left(\frac{2}{\pi}\right) - \frac{N}{2}\ln(\nu) + \ln\left(\prod_i^N I_i^{-0.5}\right) - \frac{1}{\nu}\sum_j^N \frac{(2I_{2i} - I_i)^2}{2I_i}$$

and setting it to zero:

$$\frac{\partial l(\nu|I_i, I_{2i})}{\partial \nu} = 0 - \frac{N}{2\nu} + 0 + \frac{1}{\nu^2}\sum_j^N \frac{(2I_{2i} - I_i)^2}{2I_i} = 0$$

With assumption 3, conversion of fusion proteins, it follows $I_i = I_{2i} + I_{2i+1}$ and $\hat{\nu}$ can be written as

$$\hat{\nu} = \frac{1}{N}\sum_j^N \frac{(I_{2i} - I_{2i+1})^2}{I_i} \,. \tag{17}$$

In the original paper the density transformation is performed via Bayes theorem and integration over copy numbers, which gives the same result, but is more complex.

3.5.2.2 Variance approach

This method uses a variance approach, which is important because its describtion in literature [15] motivated us to apply this principle to more complex situations (see chapters 3.4.3.2 and 3.5.4.2). The variance of the intensity I_{2i} of the daughter cell is (see equation 13)

$$var(I_{2i}) = \frac{1}{4}\nu I_i \,. \tag{18}$$

The same variance can also be derived with expectation values:

$$\begin{aligned} var(I_{2i}) &= E\left(I_{2i} - E(I_{2i})\right)^2 \\ &= E\left(I_{2i} - \frac{1}{2}I_i\right)^2 \\ &= E\left(I_{2i} - \frac{I_{2i} + I_{2i+1}}{2}\right)^2 \\ &= E\left(\frac{I_{2i} - I_{2i+1}}{2}\right)^2 \end{aligned} \tag{19}$$

Please see the appendix for variance and covariance rules (see chapter 7.2). The ex-

pectation values from equation 19 can be estimated with a linear model with residuum ϵ_i:

$$\left(\frac{I_{2i} - I_{2i+1}}{2}\right)^2 = \beta_1 I_i + \epsilon_i \tag{20}$$

Thus an estimate for the conversion factor $\widehat{\nu}$ can be derived from coefficient comparison of the estimates of the linear model (see equation 20), which estimates the variance of equation 19, with the variance from equation 18.

$$\widehat{\nu} = 4\widehat{\beta_1}$$

3.5.3 Additive noise

In this method, Rosenfeld and colleagues [15] assume additive normally distributed measurement error.

$$I_i = I_i^{'} + \epsilon = \nu\, n_i + \epsilon\,, \qquad with \qquad \epsilon \sim N(0, \sigma_a^2)\,.$$

In [15], the authors considered the whole lineage tree of one mother cell in cells which does not express fusion proteins. We consider every mitosis event separately without using lineage tree information because our cells express NanogVENUS and Oct4VENUS fusion proteins continuously during their cell cycle.

Using probability product rule and assuming that σ_a and ν are independent, the probability of one cell devision satisfies

$$P(\underline{I}, \underline{I^{'}}, \sigma_a^2, \nu) = P(\underline{I}|\underline{I^{'}}, \sigma_a^2, \nu)\, P(\underline{I^{'}}|\sigma_a^2, \nu)\, P(\sigma_a^2)\, P(\nu)\,.$$

where $\underline{I} = (I_i, I_{2i}, I_{2i+1})$ and $\underline{I^{'}} = \nu(n_i, n_{2i}, n_{2i+1})$ are intensities and signals of one mitosis event. $P(\sigma_a^2)$ and $P(\nu)$ are taken as uniform bounded distributions and can be ignored for the maximum likelihood calculation. $P(\underline{I}|\underline{I^{'}}, \sigma_a^2, \nu)$ is the product of the normal distribution for the noise $\epsilon_{0_i} = I_i - I_i^{'}$ for one mother and two daughter cells.

$$P(\underline{I}|\underline{I^{'}}, \sigma_a^2, \nu) = \left(\frac{1}{\sqrt{2\pi\sigma_a^2}}\right)^3 \exp\left[-\frac{1}{2\sigma_a^2}\left((I_i - I_i^{'})^2 + (I_{2i} - I_{2i}^{'})^2 + (I_{2i+1} - I_{2i+1}^{'})^2\right)\right]$$

$P(\underline{I^{'}}|\sigma_a^2, \nu)$ is the probability for the signal and therefore independent of σ_a^2: $P(\underline{I^{'}}|\sigma_a^2, \nu) = P(\underline{I^{'}}|\nu)$. Conversen of proteins during mitosis requires $I_{2i+1}^{'} = I_i^{'} - I_{2i}^{'}$ and so $P(I_{2i+1}^{'}|I_i^{'}, I_{2i}^{'}, \nu) = 1$. With this follows $P(\underline{I^{'}}|\sigma_a^2, \nu) = P(I_{2i}^{'}|I_i^{'}, \nu)P(I_i^{'})$.

$P(I'_{2i}|I'_i, \nu)$ is an even binomial distribution, which can be approximated with a normal distribution (see equation 15). $P(I'_i)$ is the priori for the protein number of the mother cell. This priori is taken to be uniformly bounded and can be omitted for the maximum likelihood derivation, ending to

$$P(\underline{I'}|\sigma_a^2, \nu) \sim \sqrt{\frac{2}{\pi}} \frac{1}{\sqrt{\nu I'_i}} \exp\left(-\frac{(2I'_{2i} - I'_i)^2}{2I'_i \nu}\right). \tag{21}$$

For σ_a^2 the empirical variance of the difference from the intensities can be used. The authors derive these formulas for the whole lineage trees and use different numeric approximation for the empirical variance. Here, a simplier estimator can be used:

$$\begin{aligned} var(I_i - I_{2i} - I_{2i+1}) &= var(I'_i - I'_{2i} - I'_{2i+1} + \epsilon_i - \epsilon_{2i} - \epsilon_{2i+1}) \\ &= var(0) + 3\sigma_a^2 \end{aligned}$$

Using the maximum likelihood estimator for the variance of a normal distribution, this leads to

$$\widehat{\sigma_a}^2 = \frac{1}{3} \frac{1}{N-1} \sum_j^N (I_i - I_{2i} - I_{2i+1})^2 . \tag{22}$$

Using $I'_{2i+1} = I'_i - I'_{2i}$ (assumption 3), the probability for the intensities of one mitosis event is:

$$\begin{aligned} P(\underline{I}, \underline{I'}, \nu) \sim & \frac{1}{\widehat{\sigma_a}^3} \exp\left(-\frac{1}{2\widehat{\sigma_a}^2}\left((I_i - I'_i)^2 + (I_{2i} - I'_{2i})^2 + (I_{2i+1} - I'_i + I'_{2i+1})^2\right)\right) \\ & \cdot \frac{1}{\sqrt{\nu I'_i}} \exp\left(-\frac{(2I'_{2i} - I'_i)^2}{2I'_i \nu}\right) \end{aligned}$$

This is also the likelihood for the conversion factor and the unobserved signal values

$$\begin{aligned} L(I'_i, I'_{2i}, \nu|\underline{I}) \sim & \widehat{\sigma_a}^{-3} \exp\left[-\frac{1}{2\widehat{\sigma_a}^2}\left((I_i - I'_i)^2 + (I_{2i} - I'_{2i})^2 + (I_{2i+1} - I'_i + I'_{2i+1})^2\right)\right] \\ & \cdot \frac{1}{\sqrt{\nu I'_i}} \exp\left[-\frac{(2I'_{2i} - I'_i)^2}{2I'_i \nu}\right] \end{aligned}$$

For maximization of the likelihood $L(I_i^{'}, I_{2i}^{'}, \nu|\underline{I})$, it is helpful to take derivates of the log-likelihood $l(I_i^{'}, I_{2i}^{'}, \nu|\underline{f})$ with respect to variables $(I_i^{'}, I_{2i}^{'}, \nu)$. With

$$\begin{aligned} l(I_i^{'}, I_{2i}^{'}, \nu|\underline{I}) \sim & -3\,\ln(\widehat{\sigma_a}) - \frac{1}{2\widehat{\sigma_a}^{\,2}}\left((I_i - I_i^{'})^2 + (I_{2i} - I_{2i}^{'})^2 + (I_{2i+1} - I_i^{'} + I_{2i+1}^{'})^2\right) \\ & - \frac{1}{2}\ln(\nu) - \frac{1}{2}\ln(I_i^{'}) - \frac{(2I_{2i}^{'} - I_i^{'})^2}{2I_i^{'}\nu}\,, \end{aligned}$$

this is

$$\begin{aligned} \frac{\partial l(I_i^{'}, I_{2i}^{'}, \nu|\underline{I})}{\partial I_i^{'}} &= \frac{1}{2\widehat{\sigma_a}^{\,2}}\left((I_i + I_{2i+1} - 2I_i^{'} + I_{2i}^{'}\right) - \frac{1}{2I_i^{'}} - \frac{4I_{2i}^{'2} - I_i^{'2}}{2I_i^{'}\nu} \\ \frac{\partial l(I_i^{'}, I_{2i}^{'}, \nu|\underline{I})}{\partial I_{2i}^{'}} &= \frac{1}{2}I_i^{'}\left(1 - \frac{\nu}{6}(I_{2i} + I_{2i+1} - I_i)\right) \\ \frac{\partial l(I_i^{'}, I_{2i}^{'}, \nu|\underline{I})}{\partial \nu} &= \frac{1}{2\nu} + \frac{(2I_{2i} - I_i)^2}{\nu^2 I_i^{'}}\,. \end{aligned}$$

Setting the partial derivative equal zero, solve and elliminate $I_i^{'}$ and $I_{2i}^{'}$, the maximum likelihood estimator for the conversion factor $\widehat{\nu}$ for mitosis event i can be written as

$$\widehat{\nu} = \frac{(I_{2i} - I_{2i+1})^2 - 2\widehat{\sigma_a}^{\,2}}{\frac{1}{3}(2I_i + I_{2i} + I_{2i+1})}. \tag{23}$$

The estimator for the conversion factor $\widehat{\nu_0}$ of all observed mitosis events is the mean of $\widehat{\nu}$.

$$\widehat{\nu_0} = \frac{1}{N}\sum_{j}^{N}\frac{(I_{2i} - I_{2i+1})^2 - 2\widehat{\sigma_a}^{\,2}}{\frac{1}{3}(2I_i + I_{2i} + I_{2i+1})} \tag{24}$$

with the empirical estimator for the variance

$$\widehat{\sigma_a}^{\,2} = \frac{1}{3}\frac{1}{N-1}\sum_{j}^{N}(I_i - I_{2i} - I_{2i+1})^2\,.$$

With $\sigma_a = 0$ in equation 24 the estimator for the model without noise of equation 17 is received.

3.5.4 Multiplicative plus additive noise

We assume that the fluorescence intensities has additive normally distributed noise plus independent multiplicative log-normally distributed noise.

3.5.4.1 Likelihood approach

Several different proposals for random variables might be considered: absolute difference $I_{2i} - I_{2i+1}$ or relative difference $\frac{I_{2i}-I_{2i+1}}{I_i}$ of daughter cell intensities or daughter cell intensities only, I_{2i}. Here, the density function of the difference between the daughter cell intensities is used:

$$\begin{aligned}\Delta_i(I'_{2i}, \epsilon_{m_1}, \epsilon_{m_2}, \epsilon_{a_0}) &:= I_{2i} - I_{2i+1} \\ &= I'_{2i}\epsilon_{m_1} + \epsilon_{a_1} - I'_{2i+1}\epsilon_{m_2} - \epsilon_{a_2} \\ &= I'_{2i}\epsilon_{m_1} - (I'_i - I'_{2i})\epsilon_{m_2} + \epsilon_{a_1} - \epsilon_{a_2} \\ &= I'_{2i}(\epsilon_{m_1} + \epsilon_{m_2}) - I'_i\epsilon_{m_2} + \epsilon_{a_0}\end{aligned}$$

while $\epsilon_{a_0} = \epsilon_{a_1} - \epsilon_{a_2} \sim N(0, 2\sigma_a^2)$. Mother cell signal I'_i is not a random variable.

Because the four random variables $I'_{2i}, \epsilon_{m_1}, \epsilon_{m_2}, \epsilon_{a_0}$ are independent, the joint density function is the product of the single density functions. With the approximation of the binomial distribution with the normal distribution, $I'_{2i} = \nu \cdot n_{2i} \sim N\left(\frac{1}{2}I'_i, \frac{1}{4}\nu I_i\right)$

$$\begin{aligned}f_{I'_{2i},\epsilon_{m_1},\epsilon_{m_2},\epsilon_{a_0}}(I'_{2i}, \epsilon_{m_1}, \epsilon_{m_2}, \epsilon_{a_0}) &= f_{I'_{2i}}(I'_{2i}) \cdot f_{\epsilon_{m_1}}(\epsilon_{m_1}) \cdot f_{\epsilon_{m_2}}(\epsilon_{m_2}) \cdot f_{\epsilon_{a_0}}(\epsilon_{a_0}) \\ &= \frac{1}{\sqrt{2\pi\frac{1}{4}\nu I'_i}} \exp\left(-\frac{\left(I'_{2i} - \frac{1}{2}I'_1\right)^2}{2\frac{1}{4}\nu I'_i}\right) \cdot \frac{1}{\sqrt{2\pi}\sigma_m\epsilon_{m_1}} \exp\left(-\frac{\left(\ln(\epsilon_{m_1}) + \frac{\sigma_m^2}{2}\right)^2}{2\sigma_m^2}\right) \\ &\quad \cdot \frac{1}{\sqrt{2\pi}\sigma_m\epsilon_{m_2}} \exp\left(-\frac{\left(\ln(\epsilon_{m_2}) + \frac{\sigma_m^2}{2}\right)^2}{2\sigma_m^2}\right) \cdot \frac{1}{\sqrt{2\pi}\sqrt{2}\sigma_a} \exp\left(-\frac{\epsilon_{a_0}^2}{2 \cdot 2\sigma_a^2}\right).\end{aligned}$$

The density of the difference between the two daughter cells intensities $\Delta_i = I_{2i} - I_{2i+1} = I'_{2i}(\epsilon_{m_1} - \epsilon_{m_2}) - I'_i\epsilon_{m_2} + \epsilon_{a_0}$ is derived via integration, similar to equation 7.

$$
\begin{aligned}
&f_{\Delta}(\Delta_i) = \\
&\int_0^{I_i'} \int_0^{\infty} \int_0^{\infty} f_{I_{2i}'}(I_{2i}') f_{\epsilon_{m_1}}(\epsilon_{m_1}) f_{\epsilon_{m_2}}(\epsilon_{m_2}) f_{\epsilon_{a_0}}(\epsilon_{a_0} = \Delta_i - I_{2i}'(\epsilon_{m_1} - \epsilon_{m_2}) + I_i'\epsilon_{m_2})\, d\epsilon_{m_1}\, d\epsilon_{m_2}\, dI_{2i}' \\
&= \int_0^{I_i'} \int_0^{\infty} \int_0^{\infty} \frac{1}{2\pi^2\sqrt{2}\sigma_m^2\sigma_a\epsilon_{m_1}\epsilon_{m_2}\sqrt{\nu I_i'}} \cdot \exp\left(-\frac{2\left(I_{2i}' - \frac{1}{2}I_1'\right)^2}{\nu I_i'} - \frac{\left(\ln(\epsilon_{m_1}) + \frac{\sigma_m^2}{2}\right)^2}{2\sigma_m^2} \right) \\
&\qquad \cdot \exp\left(-\frac{\left(\ln(\epsilon_{m_2}) + \frac{\sigma_m^2}{2}\right)^2}{2\sigma_m^2} - \frac{(\Delta_i - I_{2i}'(\epsilon_{m_1} + \epsilon_{m_2}) + I_i'\epsilon_{m_2})^2}{4\sigma_a^2} \right) d\epsilon_{m_1}\, d\epsilon_{m_2}\, dI_{2i}'
\end{aligned}
\tag{25}
$$

We need to estimate the signal from the mother cell, I_i', which is not a random variable. This can be done with the intensities of the mother and both daughter cells and the fusion protein gain during mitosis (see table 4):

$$\widehat{I_i'} = \frac{1}{2}(I_i + I_{2i} + I_{2i+1} + \widehat{gain}) \tag{26}$$

Fusion protein gain is assumed to be constant for one data set. It is the mean difference between the sum of the daughter cells intensities minus the mother cell intensity.

$$\widehat{gain} = \bar{I}_{2i} + \bar{I}_{2i+1} - \bar{I}_i$$

The log-likelihood l of the parameters σ_m^2, σ_a^2 and ν is the logarithm over the product of the densitiy function $f_{\Delta}(\Delta_i)$ for all observed differences between the daughter cell intensities Δ_i.

$$l(\sigma_m^2, \sigma_a^2, \nu|\mathbf{\Delta}) = \sum_{i=1}^{N} \ln f_{\Delta}(\Delta_i) \tag{27}$$

The log-likelihood l can evaluated and maximized numerically for σ_m^2, σ_a^2 and ν to get the maximum likelihood estimator for the noise parameter and the conversion factor.

3.5.4.2 Variance approach

A second, numerically easier method to estimate copy numbers from mitosis fluorescence data with multiplicative plus additive noise is based on the comparison of

variances, similar to the approaches used in chapter 3.4.3.2 and 3.5.2.2. To derive an estimate for the conversion factor, the variance of the difference of the observed fluorescence intensities of the daughter cells $var(I_{2i} - I_{2i+1})$ is analysed. All used equations for variance and covariance are listed in chapter 7.2.

For mother and daughter intensities, the following expectation values can be derived:

$$E(I'_{2i}) = E(I'_{2i+1}) = \frac{1}{2} I'_i$$
$$E(I'_{2i} - I'_{2i+1}) = 0$$

The variance of the difference of the observed daughter cell intensity can be split into

$$\begin{aligned} var(I_{2i} - I_{2i+1}) &= var\left(I'_{2i} - I'_{2i+1} + \epsilon_0(I'_{2i}) - \epsilon_0(I'_{2i+1})\right) \\ &= var\left(I'_{2i} - I'_{2i+1}\right) + var\left(\epsilon_0(I'_{2i})\right) + var\left(\epsilon_0(I'_{2i+1})\right) \\ &\quad + 2cov\left(I'_{2i} - I'_{2i+1},\, \epsilon_0(I'_{2i}) - \epsilon_0(I'_{2i+1})\right) \\ &\quad - 2cov\left(\epsilon_0(I'_{2i}),\, \epsilon_0(I'_{2i+1})\right) . \end{aligned} \tag{28}$$

Using the variance of a product $var(XY) = E^2(X)var(Y) + E^2(Y)var(X) + var(X)var(Y)$, the single terms of equation 28 can be written as

$$\begin{aligned} var(\epsilon_0(I'_{2i})) &= var(I'_{2i}(\epsilon_m - 1) + \epsilon_a) \\ &= var(I'_{2i}(\epsilon_m - 1)) + \sigma_a^2 \\ &= E^2(I'_{2i})\, var(\epsilon_m - 1) + E^2(\epsilon_m - 1)\, var(I'_{2i}) + var(I'_{2i})\, var(\epsilon_m - 1) + \sigma_a^2 \\ &= \left(\frac{I'_i}{2}\right)^2 \left(e^{\sigma_m^2} - 1\right) + 0\frac{1}{4} I_i \nu + \frac{1}{4} I'_i \nu \left(e^{\sigma_m^2} - 1\right) + \sigma_a^2 \\ &= I_i'^2 \frac{1}{4} \left(e^{\sigma_m^2} - 1\right) + I'_i \frac{1}{4} \nu \left(e^{\sigma_m^2} - 1\right) + \sigma_a^2 \end{aligned}$$

and

$$
\begin{aligned}
&cov\left(I'_{2i} - I'_{2i+1},\, \epsilon_{0_1}(I'_{2i}) - \epsilon_{0_2}(I'_{2i+1})\right) \\
&\quad = cov\left(I'_{2i} - I'_{2i+1},\, I'_{2i}(\epsilon_{m_1} - 1) + \epsilon_{a_1} - \left(I'_{2i+1}(\epsilon_{m_2} - 1) + \epsilon_{a_2}\right)\right) \\
&\quad = cov\left(I'_{2i} - I'_{2i+1},\, I'_{2i}(\epsilon_{m_1} - 1) - I'_{2i+1}(\epsilon_{m_2} - 1)\right) + cov\left(I'_{2i} - I'_{2i+1},\, \epsilon_{a_1} + \epsilon_{a_2}\right) \\
&\quad = E\left(\left[I'_{2i} - I'_{2i+1} - E\left(I'_{2i} - I'_{2i+1}\right)\right] \times \right. \\
&\qquad \left. \left[I'_{2i}(\epsilon_{m_1} - 1) - I'_{2i+1}(\epsilon_{m_2} - 1) - E\left(I'_{2i}(\epsilon_{m_1} - 1) - I'_{2i+1}(\epsilon_{m_2} - 1)\right)\right]\right) + 0 \\
&\quad = E\left(\left[I'_{2i} - I'_{2i+1}\right]\left[I'_{2i}(\epsilon_{m_1} - 1) - I'_{2i+1}(\epsilon_{m_2} - 1)\right]\right) \\
&\quad = E\left[I'^2_{2i}(\epsilon_{m_1} - 1) + I'^2_{2i+1}(\epsilon_{m_2} - 1) - I'_{2i}I'_{2i+1}(\epsilon_{m_1} - 1) - I'_{2i}I'_{2i+1}(\epsilon_{m_2} - 1)\right] \\
&\quad = 0
\end{aligned}
$$

The last step holds because of $E(X,Y) = cov(X,Y) + E(X)E(Y)$ and
$E[I'^2_{2i}(\epsilon_m - 1)] = cov(I'^2_{2i},\, (\epsilon_m - 1)) + E(I'^2_{2i})\, E(\epsilon_m - 1) = 0 + 0.$
The last term in equation 28 can be evaluated to

$$
\begin{aligned}
&cov\left(\epsilon_0(I'_{2i}),\, \epsilon_0(I'_{2i+1})\right) \\
&\quad = cov\left(I'_{2i}(\epsilon_{m_1} - 1) + \epsilon_{a_1},\, I'_{2i+1}(\epsilon_{m_2} - 1) + \epsilon_{a_2}\right) \\
&\quad = E\left(\left[I'_{2i}(\epsilon_{m_1} - 1) + \epsilon_{a_1} - E\left(I'_{2i}(\epsilon_{m_1} - 1) + \epsilon_{a_1}\right)\right] \cdot \right. \\
&\qquad \left. \left[I'_{2i+1}(\epsilon_{m_2} - 1) + \epsilon_{a_2} - E\left(I'_{2i+1}(\epsilon_{m_2} - 1) + \epsilon_{a_2}\right)\right]\right) \\
&\quad = E\left(\left[I'_{2i}(\epsilon_{m_1} - 1) + \epsilon_{a_1}\right]\left[I'_{2i+1}(\epsilon_{m_2} - 1) + \epsilon_{a_2}\right]\right) \\
&\quad = E\left[I'_{2i}I'_{2i+1}(\epsilon_{m_1} - 1)(\epsilon_{m_2} - 1) + \epsilon_{a_1}\epsilon_{a_2} + \epsilon_{a_1}I'_{2i+1}(\epsilon_{m_2} - 1) + \epsilon_{a_2}I'_{2i}(\epsilon_{m_1} - 1)\right] \\
&\quad = 0\,.
\end{aligned}
$$

Using this, the variance of the difference within the intensities of the daughter cells can be expressed as

$$
\begin{aligned}
var(I_{2i} - I_{2i+1}) &= I'_i\nu + 2\left(I'^2_i\frac{1}{4}\left(e^{\sigma^2_m} - 1\right) + I'_i\frac{1}{4}\nu e^{\sigma^2_m} + \sigma^2_a\right) \\
&= 2\sigma^2_a + I'_i\nu\frac{1}{2}\left(1 + e^{\sigma^2_m}\right) + I'^2_i\frac{1}{2}\left(e^{\sigma^2_m} - 1\right) \\
&\approx 2\sigma^2_a + \widehat{I'_i}\nu\frac{1}{2}\left(1 + e^{\sigma^2_m}\right) + \widehat{I'_i}^2\frac{1}{2}\left(e^{\sigma^2_m} - 1\right). \qquad (29)
\end{aligned}
$$

For the last step I'_i is replaced with an estimate for it, similar to equation 26

$$\widehat{I'_i} = \frac{1}{2}(I_i + I_{2i} + I_{2i+1} + \widehat{gain}).$$

A second way to derive the variance of the difference in daughter cell intensities is to use the expectation values is needed:

$$\begin{aligned} var(I_{2i} - I_{2i+1}) &= E(I_{2i} - I_{2i+1})^2 - E^2(I_{2i} - I_{2i+1}) \\ &= E(I_{2i} - I_{2i+1})^2 \end{aligned} \tag{30}$$

A linear model can be fitted to estimate the variance from equation 30

$$(I_{2i} - I_{2i+1})_j^2 = \beta_0 + \beta_1 \widehat{I_i'} + \beta_2 \widehat{I_{i\,i_j}'}^2 + \epsilon_j \,. \tag{31}$$

With coefficient comparison from equation 29 and 30),

$$\begin{aligned} \widehat{\beta_0} =& 2\widehat{\sigma_a^2} \\ \widehat{\beta_1} =& \widehat{\nu}\frac{1}{2}\left(1 + e^{\widehat{\sigma_m}^2}\right) \\ \widehat{\beta_2} =& \frac{1}{2}\left(e^{\widehat{\sigma_m^2}} - 1\right) \end{aligned}$$

an estimator for the conversion factor can be derived

$$\widehat{\nu} = \frac{\widehat{\beta_1}}{1 + \widehat{\beta_2}} \;. \tag{32}$$

Confidence intervals for $\widehat{\nu}$ may be derived with the standard errors of $\widehat{\beta_1}$ and $\widehat{\beta_2}$ or with bootstraping.
The estimated number $\widehat{n_i}$ of proteins of cell i is the product of the conversion factor with its fluorescence intensitiy.

$$\widehat{n_i} = \widehat{\nu} I_i$$

4 Results

4.1 Estimation of signal and total noise

In chapter 2 four different mESC time lapse fluorescence datasets are discussed (see table 1) for which six different LMMs are estimated (see chapter 3.3.1).

A linear random intercept: $I_{ij} = \beta_0 + \beta_1 \cdot age_{ij} + z_{i0} + \epsilon_{ij}$

B linear random slope: $I_{ij} = \beta_0 + \beta_1 \cdot age_{ij} + z_{i0} + z_{i1} \cdot age_{ij} + \epsilon_{ij}$

C parabola random intercept: $I_{ij} = \beta_0 + \beta_1 \cdot age_{ij} + \beta_2 \cdot age_{ij}^2 + z_{i0} + \epsilon_{ij}$

D parabola random slope: $I_{ij} = \beta_0 + \beta_1 \cdot age_{ij} + \beta_2 \cdot age_{ij}^2 + z_{i0} + z_{i1} \cdot age_{ij} + \epsilon_{ij}$

E parabola random parabola:
$I_{ij} = \beta_0 + \beta_1 \cdot age_{ij} + \beta_2 \cdot age_{ij}^2 + z_{i0} + z_{i1} \cdot age_{ij} + z_{i2} \cdot age_{ij}^2 + \epsilon_{ij}$

F exponential random exponential: $\ln(I_{ij}) = \beta_0 + \beta_1 \cdot age_{ij} + z_{i0} + z_{i1} \cdot age_{ij} + \epsilon_{ij}$

All models are estimated with *R*-software and the package *nlme* [20], [21] which fits a linear as well as a nonlinear mixed-effects model in the formulation described by Lindstrom and Bates [22]. For every unit (cell) an individual expectation line is estimated in the LMM. For model A with linear fixed effect and random intercept the expectation lines of the different cells are shifted straight lines with the same gradient (see figure 13 A). For model B with linear fixed effect and random slope the expectation lines of the different cells are straight lines with individual gradients (see figure 13 B). For model C with fixed quadratic effect and random intercept the expectation lines of the different cells are parabolas with individual intercepts (see figure 13 C) and so on. For model F with fixed and random exponential effect the expectation lines are exponential lines (see figure 13 F). In the diagrams only a subset of 6 randomly choosen cells can be displayed. The corresponding diagrams for NanogVENUS raw intensity dataset, Oct4VENUS raw and normalized intesnity datasets are in the appendix, figures 30 to 32.

The expectation lines of the different LMMs for a cell can be quite different. The expectation lines of the model A are straight lines whereas the curvatures for the same cells for model E are very bended (e.g. blue lines in figures 13 A and E).

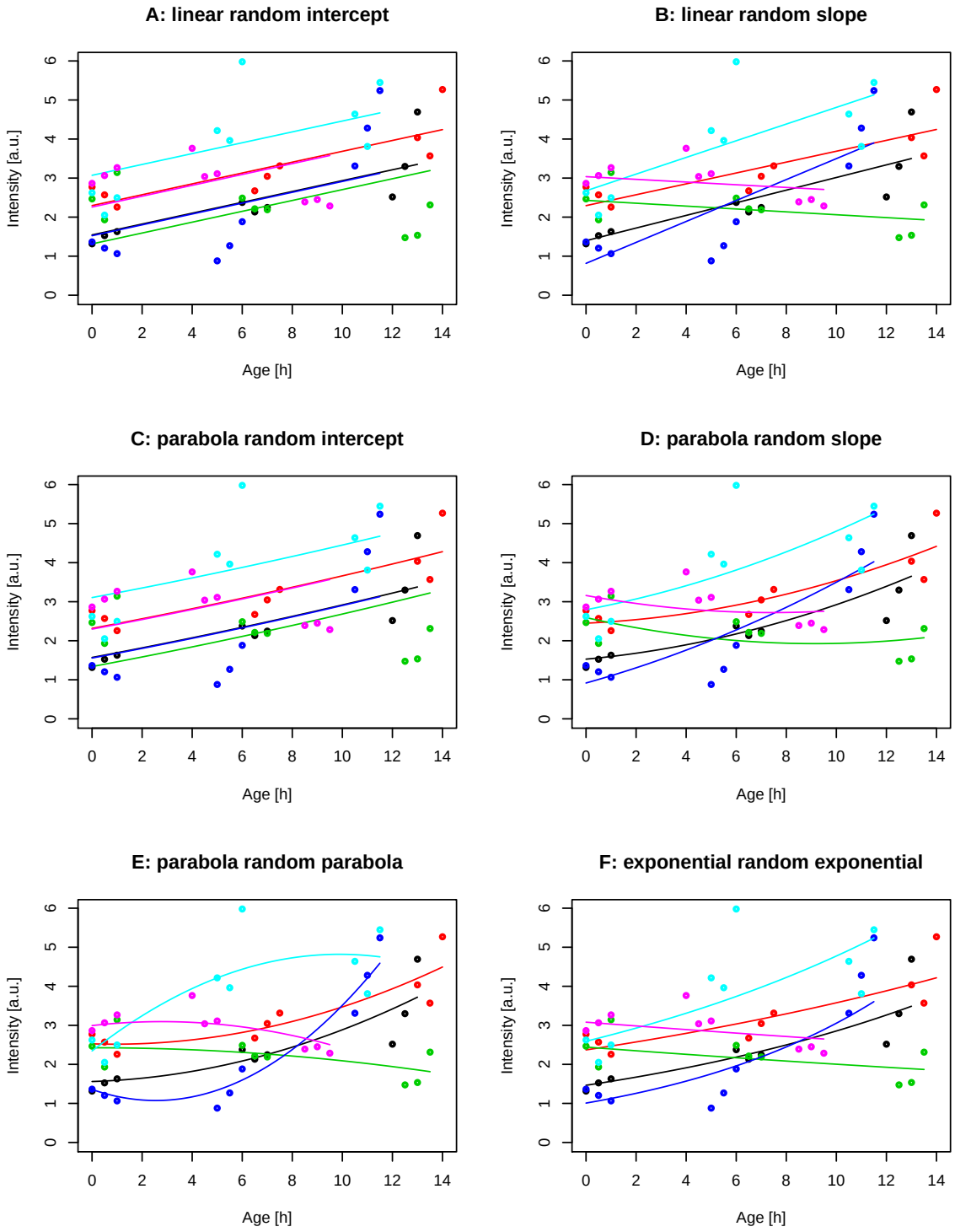

Figure 13: NanogVENUS normalized intensity. Intensities I and expectation lines of LMM for six randomly choosen cells.

4.1.1 Scatter diagrams

Estimated total noise $\widehat{\epsilon_0}_{ij}$ is the difference of the observerd intensityI_{ij} to the expectation line and estimated signal $\widehat{I'}_{ij}$ is the value on the expectation line (see figure 30 to 32). With scatter diagrams (see figures 14 to 17) the homogeneity of variance of the total noise can be compared with the homoscedasticity and heteroscedasticity of different noise in toydata (see figure 12).

In the majority of the diagrams the data points are obviously bound within two parallel lines (e.g. figure 17 B). This can be explained because the observed intensities are all positive. Therefore the estimated total noise, which is the difference between intensity and signal, can not be smaller than the negative signal values. This leads to the lower bound of the residuals. The upper bound is defined by the upper limit in data cleaning rules for the intensities, which is 10 in raw and 100 in normalized intensity datasets.

The estimated total noise is not homoskedastic for all cases. Heteroskedasticity seems to be more prevalent in Oct4VENUS, more in raw intensity data and more in the random intercept LMMs (see figure 14 to 17). Diagrams indicate an even more complex noise behavior in Oct4VENUS residuals versus predicted than only multiplicative plus additive noise. No big differences between random parabola and random exponential LMMs residuals can be seen.

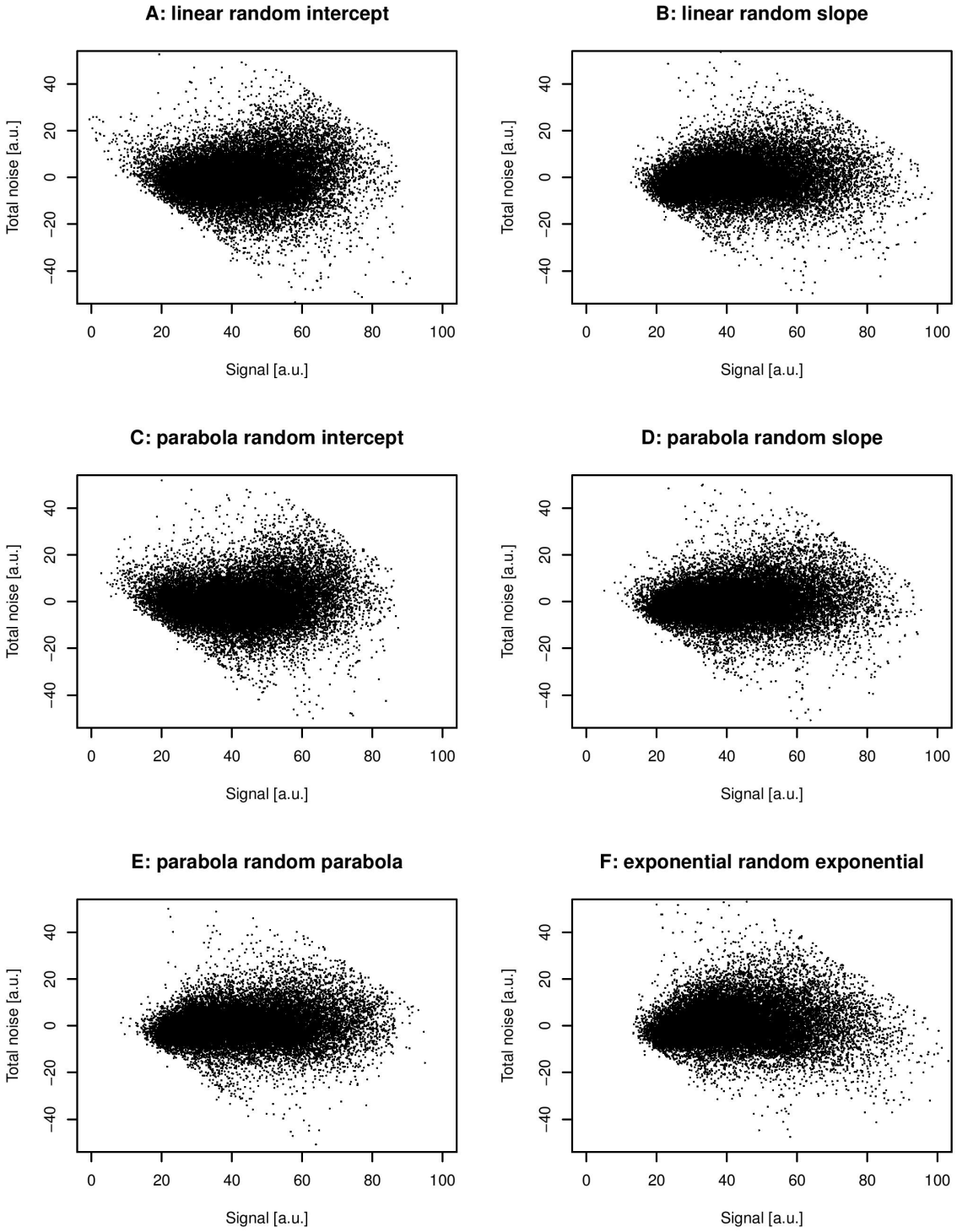

Figure 14: **NanogVENUS raw intensity.** Estimated total noise $\widehat{\epsilon_{0ij}}$ versus estimated signal $\widehat{I'}_{ij}$ for 23 887 observations from 2034 cells using six different LMMs.

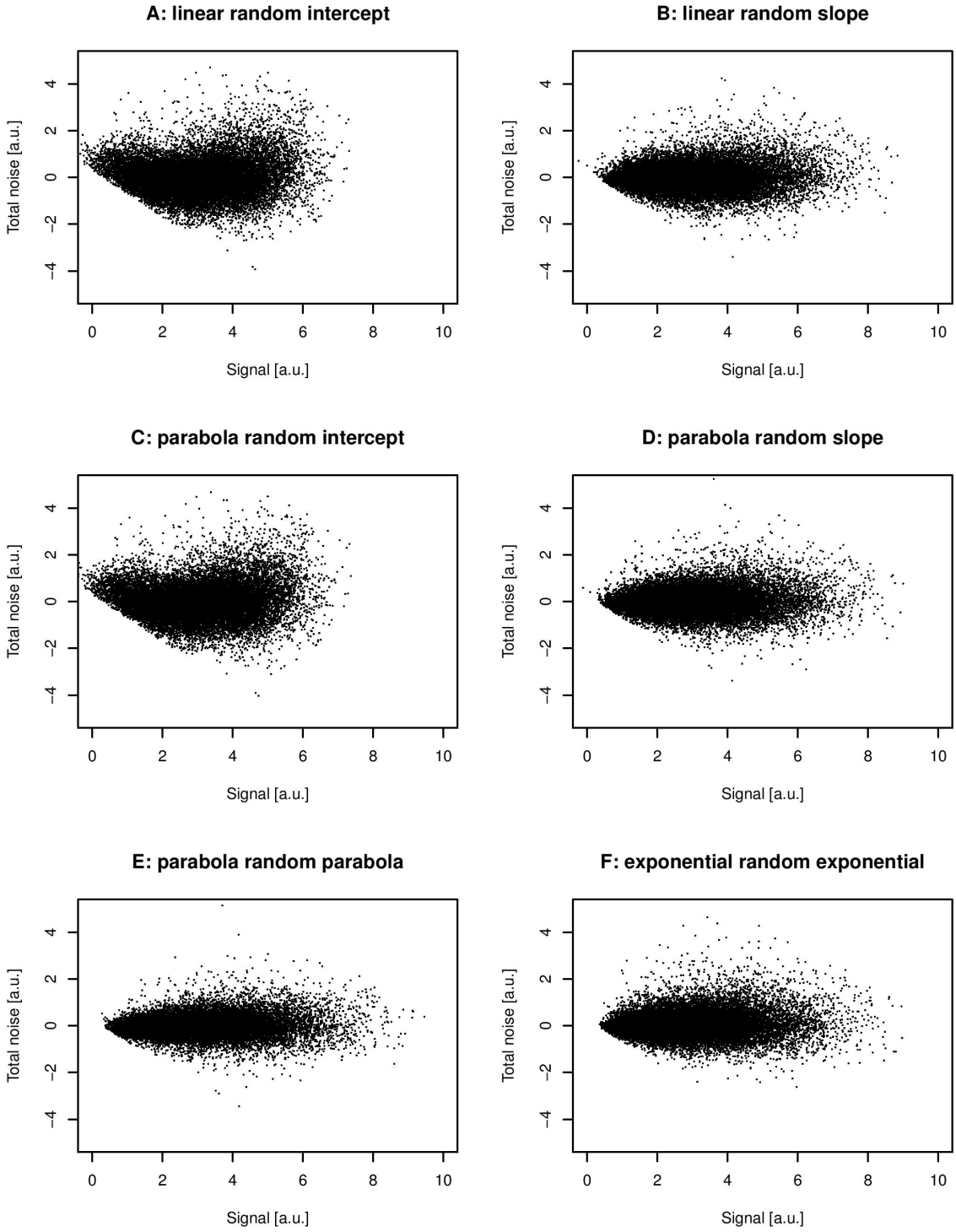

Figure 15: **NanogVENUS normalized intensity.** Estimated total noise $\widehat{\epsilon_{0ij}}$ versus estimated signal $\widehat{I'}_{ij}$ for 22 317 observations from 1850 cells using six different LMMs.

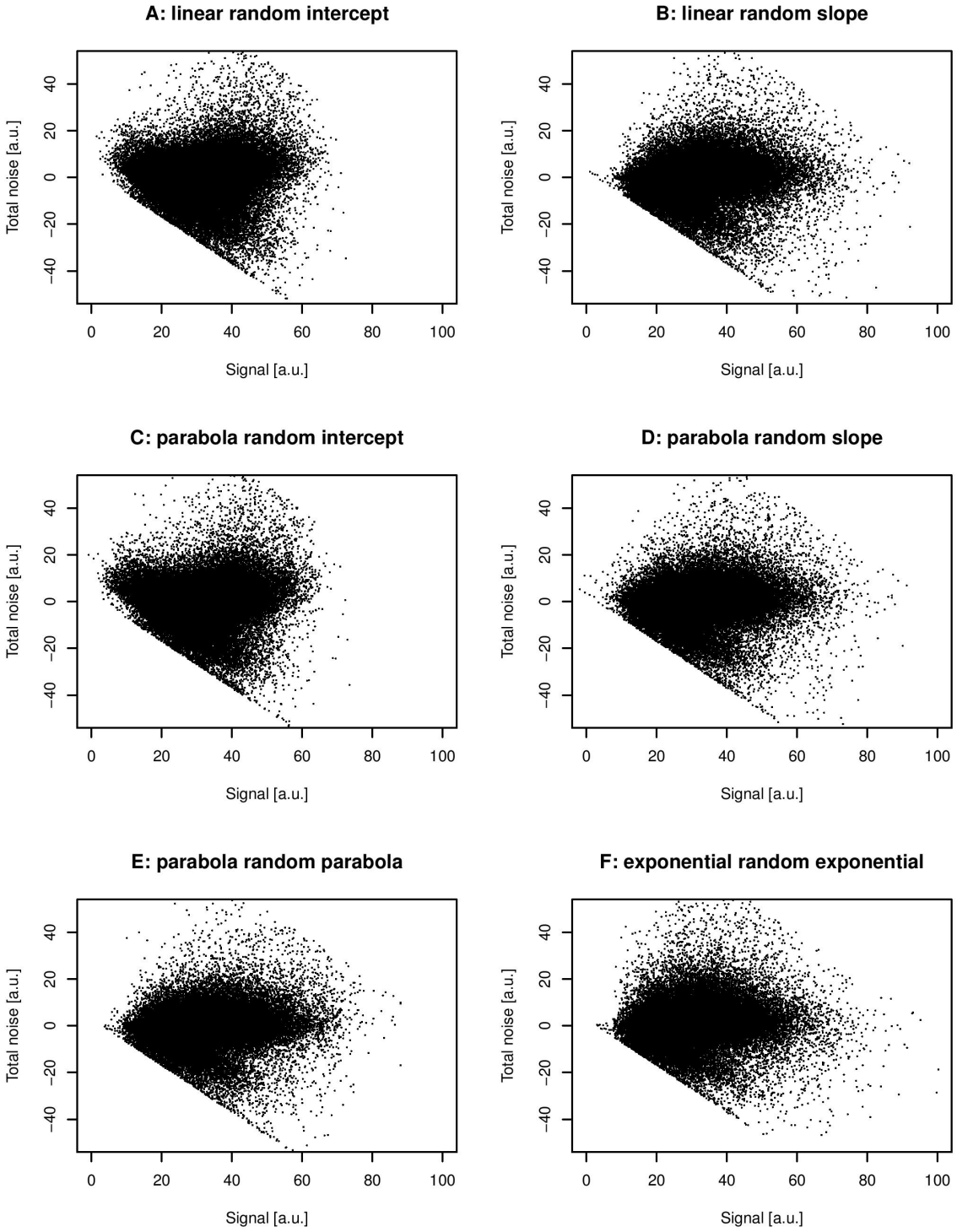

Figure 16: **Oct4VENUS raw intensity.** Estimated total noise $\widehat{\epsilon_{0}}_{ij}$ versus estimated signal $\widehat{I'}_{ij}$ for 57 857 observations from 2891 cells using six different LMMs.

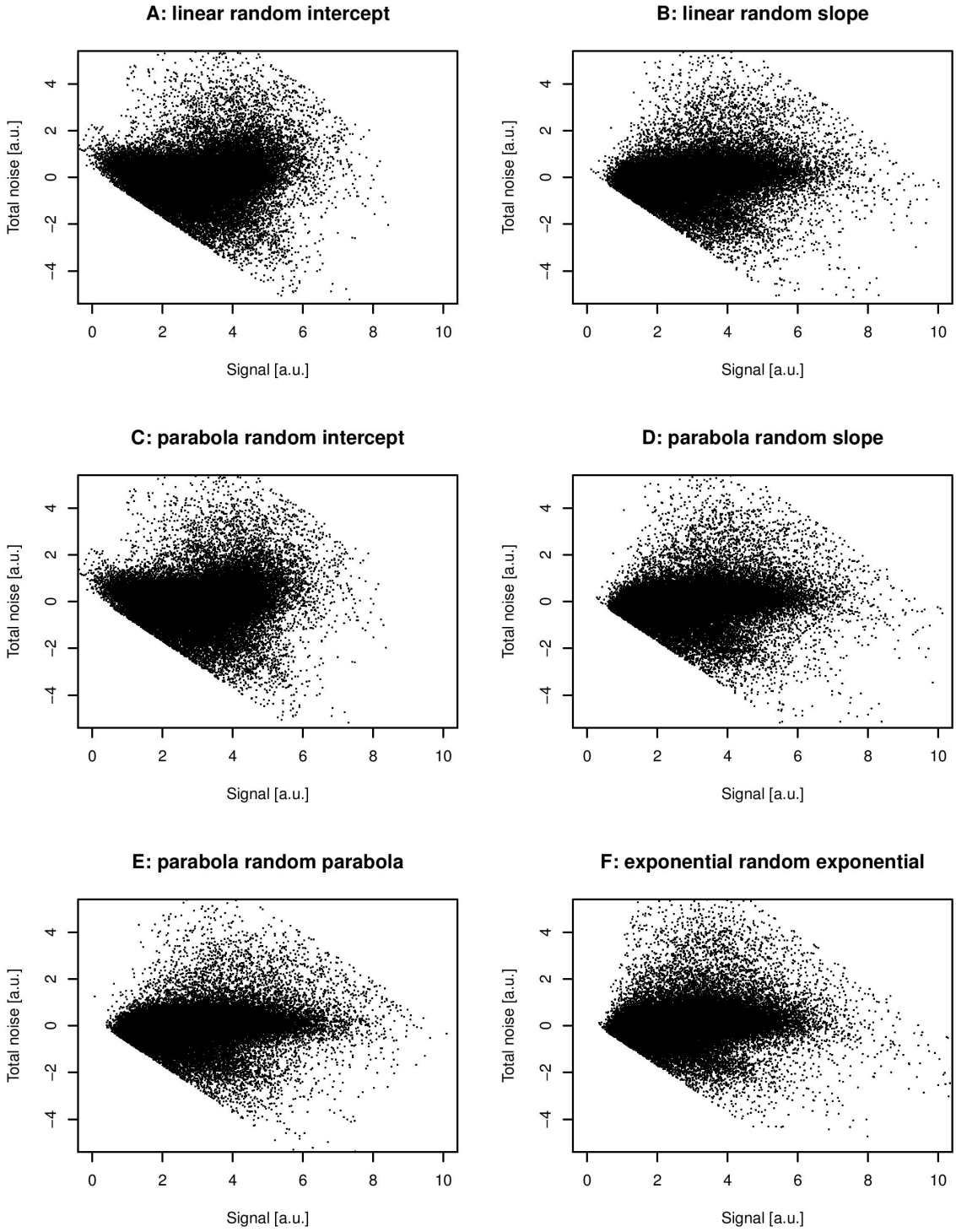

Figure 17: **Oct4VENUS normalized intensity.** Estimated total noise $\widehat{\epsilon_{0ij}}$ versus estimated signal $\widehat{I'}_{ij}$ for 57 185 observations from 2888 cells using six different LMMs.

4.1.2 Autocorrelation

With the estimated ACF (see chapter 3.3.2) the assumption of uncorrelated error noise terms is investigated. Raw and normalized intensity as well as NanogVENUS and Oct4VENUS show similar distribution of the estimated ACF coefficients (see figures 18 to 21).

Due to manual adjustment of the segmentation in NanogVENUS datasets, the majority of NanogVENUS cells has nine intensity observations only: three observations in the beginning, three observations in the middle and three observations at the end of the cell cycle (see figures 7 to 4). This is problematic for the interpretation of the ACF, because the timepoints of observations are not balanced then. Lag 1 is a shift of one observation which can be 30 minutes or 4 hours in NanogVENUS datasets.

ACF coefficients for lag 1 estimated for model A and C are mainly positive. The boxes, which range from the lower quartile to the upper quartile, does not include zero for three of the four datasets. Only for NanogVENUS raw intensity dataset the box includes zero for these models. For models B, D and F the medians for lag 1 for all data sets are positive again, but the boxes do include zero. For model E the median for lag 1 in the NanogVENUS datasets is negative and for Oct4VENUS datasets is positive and the boxes for all datasets include zero.

Thus, the assumption of uncorrelated residuals seems to be violated in the models A and C. The ACF analysis for the other models however are consistent with this assumption. The reason is that models A and C have no random effect age, which is of course positively correlated with the order of the observations. The missing explained variation of the random effect age is interpreted as autocorrelation.

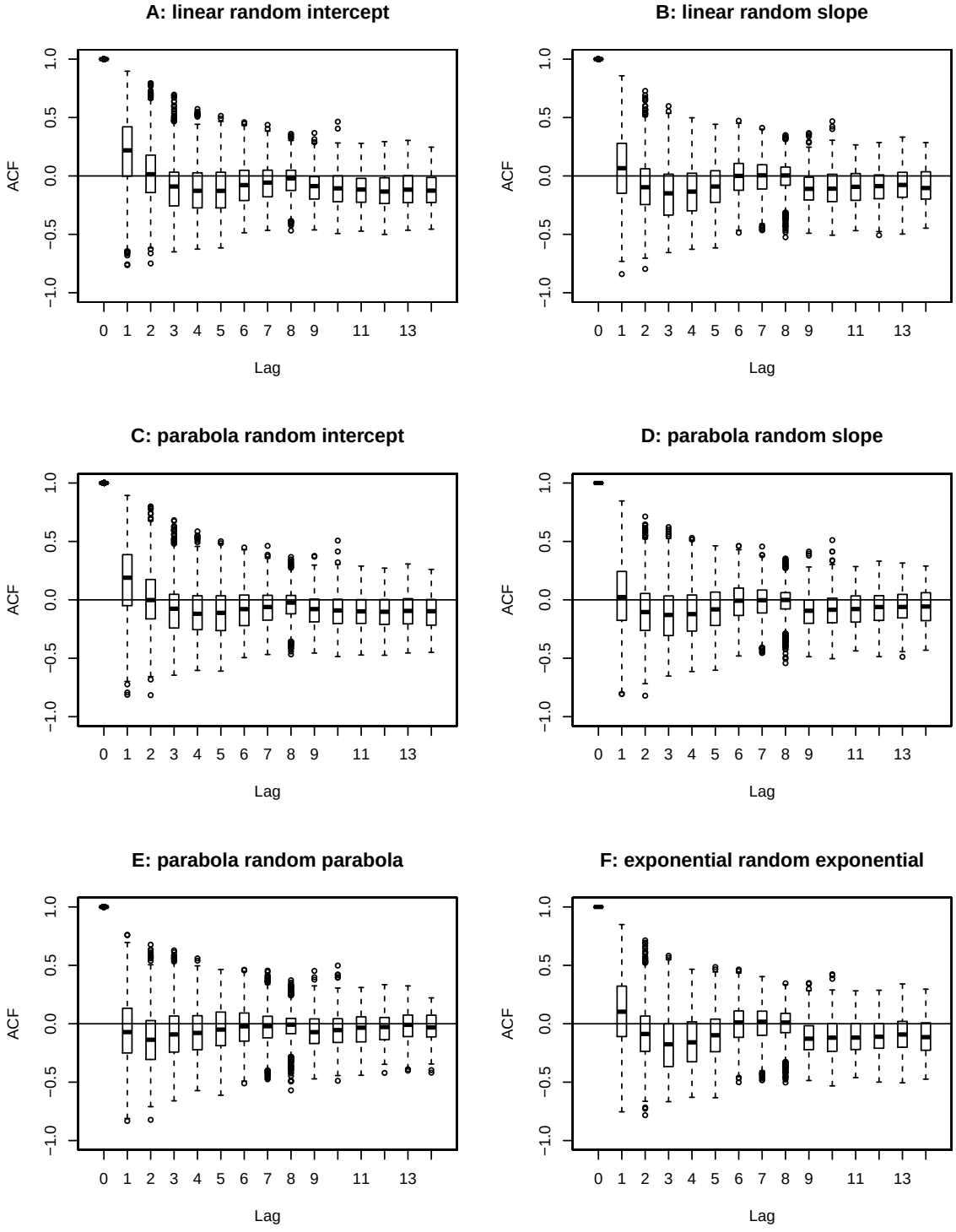

Figure 18: **NanogVENUS raw intensity.** Boxplot over estimated autocorrelation functions for the estimated total noise (number cells = 2034). The residuals are from six different LMMs.

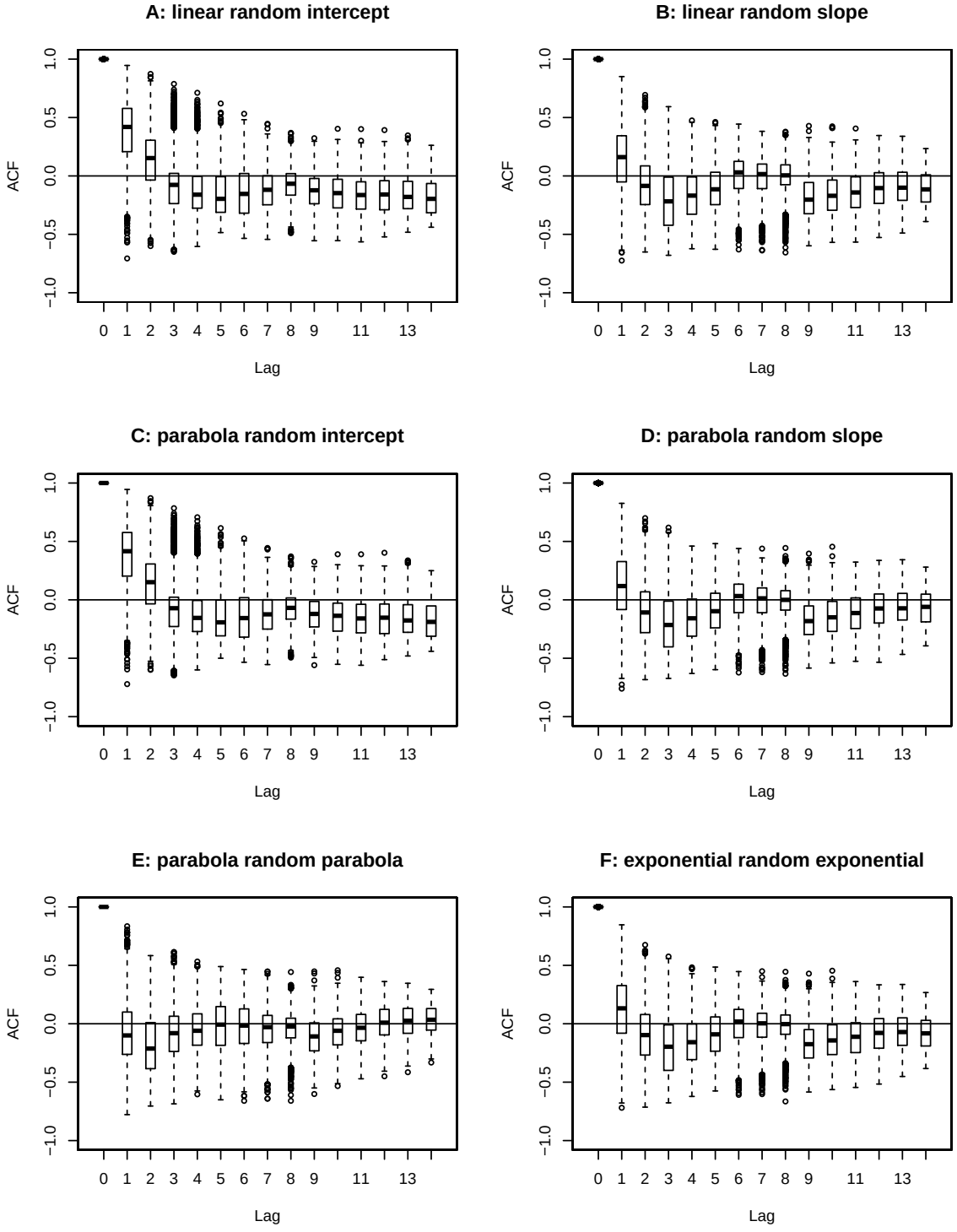

Figure 19: **NanogVENUS normalized intensity.** Boxplot over estimated autocorrelation functions for the estimated total noise (number cells = 1850). The residuals are from six different LMMs.

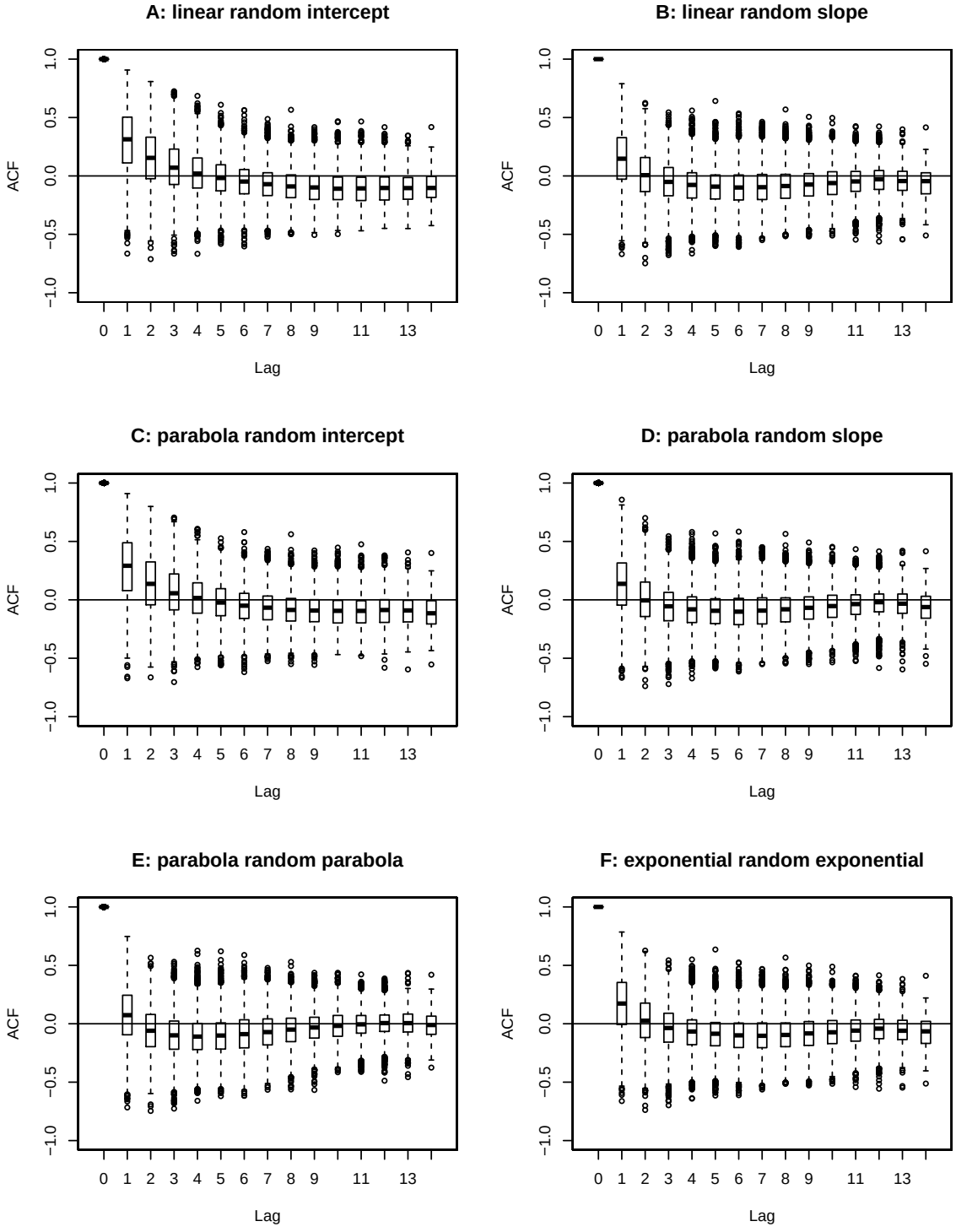

Figure 20: **Oct4VENUS raw intensity.** Boxplot over estimated autocorrelation functions for the estimated total noise (number cells = 2891). The residuals are from six different LMMs.

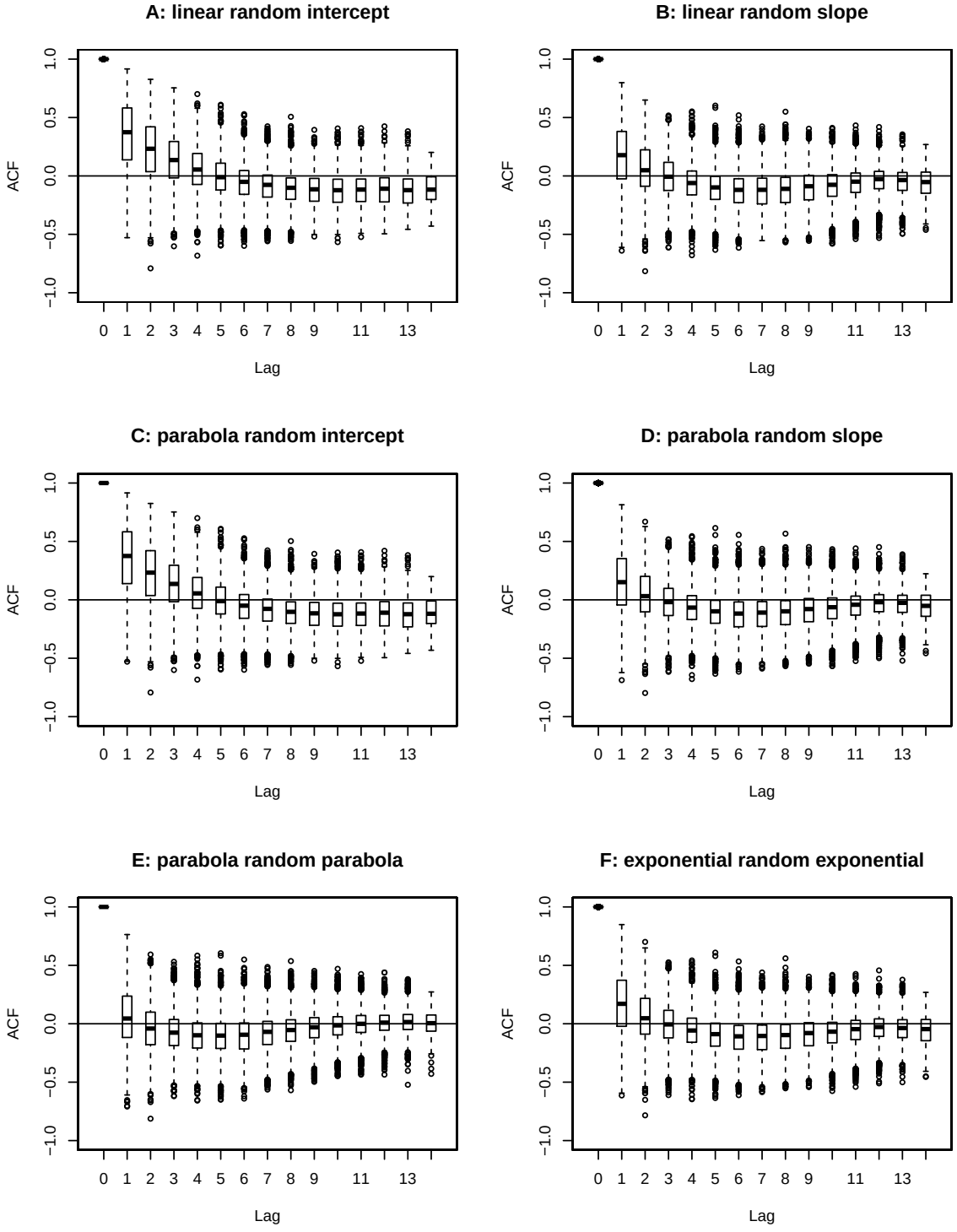

Figure 21: **Oct4VENUS normalized intensity.** Boxplot over estimated autocorrelation functions for the estimated total noise (number cells = 2888). The residuals are from six different LMMs.

4.1.3 Model comparison

We use adjusted R-square $\bar{R}^2$ for model comparison (see equation 4). For model E with fixed parabola and random parabola the $\bar{R}^2$ value is highest for all four datasets compared to the other models (see table 7) and therefore we choose here model E to estimate signal $I^{'}$ and noise ϵ_0.

	A	B	C	D	E	F
NanogVENUS raw intensity	0.646	0.705	0.674	0.720	0.735	0.678
NanogVENUS normalized intensity	0.721	0.839	0.722	0.849	0.888	0.845
Oct4VENUS raw intensity	0.490	0.581	0.507	0.590	0.617	0.554
Oct4VENUS normalized intensity	0.588	0.674	0.588	0.677	0.713	0.655

Table 7: Adjusted R-square $\bar{R}^2$ for the six LMMs of the for data sets.

However the majority of cells in NanogVenus datasets have intensity observations only in the beginning, in the middle and at the end of the cell cycle. Thus, LMMs with random parabola might overfit the observed NanogVENUS intensities and the noise is underestimated. Therefore model F with fixed and random exponential is included in the following evaluations of the noise parameters as a second model.

4.2 Toydata strategy

To verify the new derived methods to estimate noise level and copy numbers, toydata are used. Key numbers from mother cells of the NanogVENUS normalized intensity dataset are taken to simulate fluorescence intensities (see figures 22 and 9 C). Mean value is $\bar{I}_i = 4.1$ and standard deviation is $sd(I_i) = 1.5$.

In the methods to estimate (relative) noise parameters or copy numbers the absolute value of I_i is not importand. The relative distribution may influence the outcome to a certain extent. To avoid negative intensities, Gamma distribution is used to simulate cell intensity I_i' for toydata. To make it simpler, the standard deviation is taken to be 1/3 of the mean value. So the used parameter for Gamma distribution are shape factor $k = 9.0$ and scale factor $\theta = \frac{1}{9}\bar{I}_i$ ($mean = k\theta$ and $sd = \sqrt{k}\theta$).

For simulation of mitosis event data, copy numbers in mother cell n_i are drawn from a Gamma distribution (rounded to integer value). These n_i proteins are distributed stochastically equally to either daughter cell with a binomial distribution with size n_i and probability $p = 0.5$. The copy number of the second daughter cell n_{2i+1} is the difference of copy number of mother cell and first daughter cell: $n_{2i+1} = n_i - n_{2i}$. Mul-

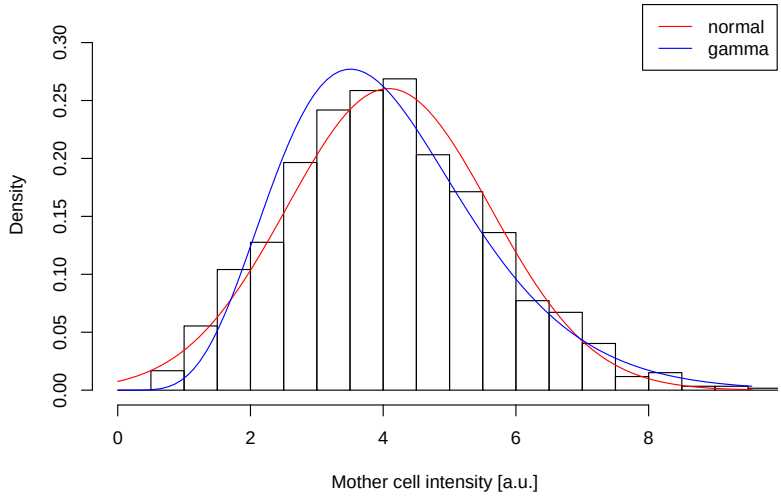

Figure 22: Histogram of NanogVENUS mother cell normalized intensity after data cleaning (N=1191). The red and blue lines show the corresponding densities of the normal and Gamma distributions.

tiplication of the copy numbers with conversion factor ν gives signal values, applying multiplicative and additive noise gives intensity values.

The parameter are choosen that in mitosis toydatasets the mean intensity of mother cell is $\bar{I'}_i = 1.5$. The mean intensity of mother and daughter cells is $\bar{I'} = \frac{1}{3}(\bar{I'}_i + \bar{I'}_{2i} + \bar{I'}_{2i+1}) = \frac{1}{3}(1.5 + 1.5/2 + 1.5/2) = 1.0$. Please note: $\bar{I'} = \bar{I}$. So with a conversion factor $\nu = 10^{-4}$ the mean copy number of daughters and mother cells is $\bar{n_0} = 1/\nu = 10^4$. Moreover the additve noise parameter σ_a is then equal to the relative additve noise parameter $\sigma_a/\bar{I'}$, which avoids confusion. Saying in this report the mean copy number in mitosis events is 1000 implies that the mean of the mother cell copy numbers is 1500 and the mean of daughter cell copy numbers is 750.

4.3 Estimation of noise parameters

4.3.1 Validation with toydata

4.3.1.1 Likelihood approach

To estimate σ_m and σ_a we use the log-likelihood derived in chapter 3.4.3.1, equation 6 with the estimated signal and noise values from the LMMs

$$l(\sigma_m^2, \sigma_a^2|\epsilon_\mathbf{0}) =$$
$$= \sum_{i=1}^{n} \ln \left(\int_{-\infty}^{I'+\epsilon_0} \frac{1}{2\pi\sigma_m\sigma_a \left(\epsilon_{0_i} - \epsilon_a + I'\right)} \exp\left(-\frac{\left(\ln\left(\frac{\epsilon_{0_i}-\epsilon_a}{I'}+1\right) + \frac{\sigma_m^2}{2}\right)^2}{2\sigma_m^2} - \frac{\epsilon_a^2}{2\sigma_a^2} \right) d\epsilon_a \right) .$$

The log-likelihood l can be maximized numerically for σ_m^2 and σ_a^2 to get the maximum likelihood estimator for the noise parameters.

R-function *integrate()* [20] is used for the numerical integration. This is an adaptive quadrature of functions of one variable over a finite or infinite interval. Maximization of the log-likelihood is done with the function *optimize()*, which uses a combination of golden section search and successive parabolic interpolations. We use the R-software with function *fdHess()* of the library *nlme* [21] to calculate the confidence intervalls numerically with equation 8 and 9.

$\widehat{\sigma}_m$	$\widehat{\sigma}_a$
0.0995	0.2009
0.0989	0.1995
0.0988	0.2004
0.1031	0.1980
0.1013	0.2003

Table 8: Results validation likelihood method to estimate noise parameter σ_m and σ_a. Each toydataset consists of $N = 50\ 000$ observations with parameter $\sigma_m = 0.1$ and $\sigma_a = 0.2$.

Five toydatasets with $N = 50\ 000$ observations are created with multiplicative log-normal distributed noise with $\sigma_m = 0.1$, additve normal distributed noise with $\sigma_a = 0.2$ and mean intensity $\bar{I}_i = 1.0$. Maximization of the log-likelihood (see equation 6) leads to the maximum likelihood estimates $\widehat{\sigma}_m$ and $\widehat{\sigma}_a$ (see table 8). 95% confidence intervall for these estimators are quite narrow ($< 10^{-5}$). Maximum likelihood estimation of the

parameters following equation 6 is shown in table 8. The deviation of the maximum likelihood estimator to true parameter is smaller 10^{-2} and might be caused by random effects while drawing the toydatasets. The 95% confidence intervall (see equation 8 and 9) spans a region smaller 10^{-5}.

It can be concluded that the likelihood approach of equation 6 to estimate multiplicative and additive noise parameters leads to reliable results.

4.3.1.2 Variance approach

With signal $I_i^{'}$ and total noise ϵ_{0_i} estimated in LMMs, the estimated coefficients from a linear model $\epsilon_{0_i}^2 = \beta_0 + \beta_1 I_i^{'2} + \epsilon_i$ lead to estimatior for the noise parameters σ_m and σ_a (see chapter 3.4.3.2, equations 11 and 12)

$$\widehat{\sigma}_a = \sqrt{\widehat{\beta}_0}$$
$$\widehat{\sigma}_m = \sqrt{\ln\left(\widehat{\beta}_1 + 1\right)}$$

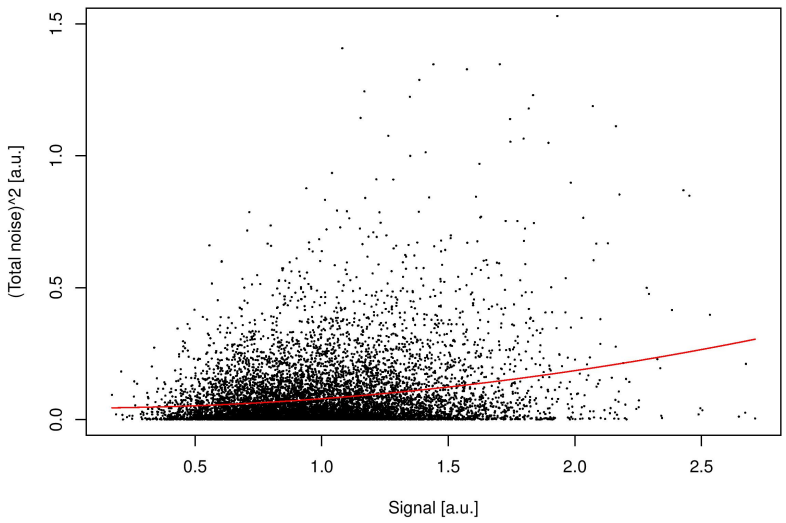

Figure 23: Example of one toydataset and the calculation of the noise parameters σ_m and σ_a with the variance approach. X-axis shows signal $I_i^{'}$, y-axis the square of the total noise $\epsilon_{0_i}^2 = (I_i - I_i^{'})^2$. The red line is the regression line of the linear model, $\epsilon_{0_i}^2 = \beta_0 + \beta_1 I_i^{'2} + \epsilon_i$. From the estimated coefficients of this linear model, estimates of the noise parameter can be derived.

Different values for noise parameter are investigated for the validation, $\sigma_m \in (0.1; 0.2)$ and $\sigma_a \in (0.1; 0.2) * \widehat{I'}$ with the mean signal $\widehat{I'} = 1.0$. 1000 toydatasets each with $N = 10\,000$ observations are generated and analysed. For every toydataset noise parameters σ_m and σ_a are estimated and median and central 95% quantil are derived (see table 9).

	σ_m	σ_a	$\widehat{\sigma}_m$		$\widehat{\sigma}_a$	
1	0.10	0.10	0.100	[0.093, 0.107]	0.100	[0.093, 0.107]
2	0.10	0.20	0.100	[0.088, 0.112]	0.200	[0.193, 0.206]
3	0.20	0.10	0.199	[0.188, 0.212]	0.101	[0.071, 0.120]
4	0.20	0.20	0.200	[0.186, 0.216]	0.200	[0.183, 0.213]

Table 9: Result empirical testing of variance algorithm to estimate noise parameters. Median and its 95% confidence interval. 1000 toydataset are created for each parameter combination with mean intensity $\overline{I_i} = 1.0$.

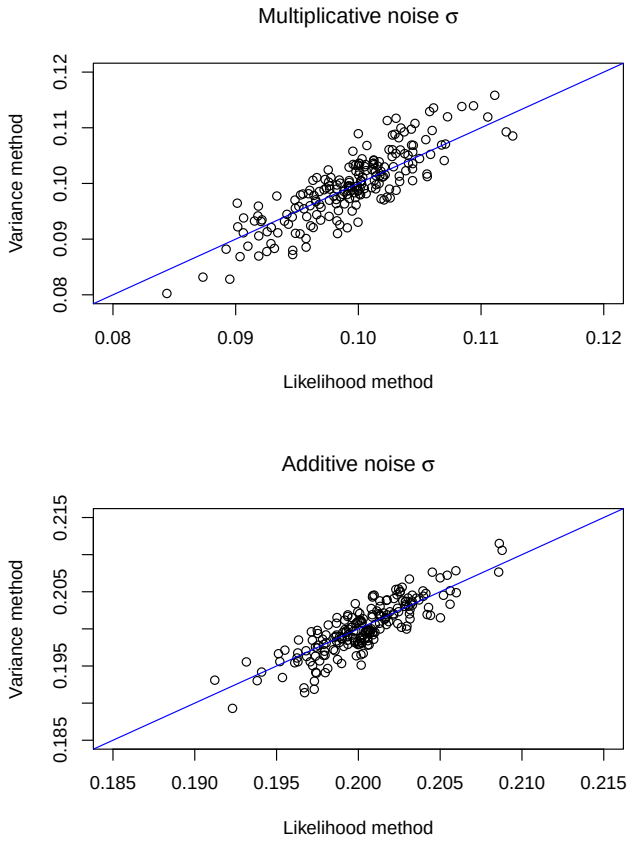

Figure 24: Comparison of toydata results for two methods. 200 toydatasets with $\sigma_m = 0.1$, $\sigma_a = 0.2$ and N=10 000 observations are analysed with both methods. The blue line is the angle bisector.

The median of the estimated noise parameters $\widehat{\sigma}_m$ and $\widehat{\sigma}_a$ is equal within a range of less than 1% to the used true parameters of the toydata, σ_m and σ_a. The 95% confidence intervals for the medians are based on the 1000 toydatasets. The range of the interval is much bigger than the confidence interval of the likelihood method which are smaller 10^{-5} (see chapter 4.3.1.1). One reason might be the variation within the toydatasets due to random sampling.

The analysis with the likelihood method of the same 4 x 1000 toydatasets analysed with the variance method is computional demanding and from only 200 toydatasets with the same parameters ($\sigma_m = 0.1$, $\sigma_a = 0.2$ and N=10 000 observations) the noise parameters are estimated with both methods (see figure 24). The estimated parameter with the variance method are positive correlated with the estimator of the likelihood method (correlation $\widehat{\sigma}_m$ 0.86 and correlation $\widehat{\sigma}_a$ 0.85). This indicates that the variation within the toydatasets is a main contributor to the width of the confidence interval for the variance method.

Using the likelihood estimates of the noise parameter here as the true parameter within every single toydataset, the variance within the estimates from the variance method can be reduced by more than 70%. This indicates that the width of the corresponding confidence intervals would be only half without the variations within the randomly drawn toydatasets.

It can be concluded that the variance approach of equations 11 and 12 to estimate multiplicative and additive noise parameter leads to reliable results.

4.3.2 Application on mESC data

Noise parameter of the mESC datasets are estimated with likelihood and variance method using. Table 10 displays the results for the raw and normalized intensities fluorescence data from NanogVENUS and Oct4VENUS. For each dataset, residuals and signal estimators are calculated with two different LMMs, model E with fixed and random parabola and model F with fixed and random exponential (see chapter 4.1.3. For the maximum likelihood estimations only residuals with predicted values $\epsilon_{ij} > 0.2$ are used to avoid problems with divergent numeric integrals.

For the variance method the estimated parameter for the multiplicative noise $\widehat{\sigma}_m$ ranges from 0.102 to 0.188 and for the likelihood method from 0.111 to 0.211. The multiplicative noise is smallest for all datasets in model E (parabola) with range for $\widehat{\sigma}_m$ from 0.102

		variance method		likelihood method	
		$\widehat{\sigma}_m$	$\widehat{\sigma}_a$	$\widehat{\sigma}_m$	$\widehat{\sigma}_a$
NanogVENUS	E: parabola	0.123	4.725	0.129	4.370
raw intensity	F: exponential	0.141	5.857	0.165	4.230
NanogVENUS	E: parabola	0.102	0.264	0.111	0.226
normalized	F: exponential	0.121	0.367	0.147	0.254
Oct4VENUS	E: parabola	0.186	4.301	0.145	5.757
raw intensity	F: exponential	0.187	6.205	0.187	5.907
Oct4VENUS	E: parabola	0.185	0.356	0.167	0.412
normalized	F: exponential	0.188	0.544	0.211	0.383

Table 10: Estimation of noise parameters. Variance method and likelihood method are applied to the residuals from the two different LMMs.

to 0.185. The multiplicative noise does not change much with normalization process. In NanogVENUS datasets for model E the estimated multiplicative noise parameter with variance method are smaller than for model F, while for Oct4VENUS they are the same. This indicates an overfitting of the data with model E in NanogVENUS data sets (see chapter 4.1.3.

		$\widehat{\sigma}_a/\bar{I}$	
		variance	likelihood
NanogVENUS	E: parabola	0.109	0.101
raw intensity	F: exponential	0.135	0.097
NanogVENUS	E: parabola	0.088	0.075
normalized	F: exponential	0.122	0.085
Oct4VENUS	E: parabola	0.132	0.177
raw intensity	F: exponential	0.190	0.181
Oct4VENUS	E: parabola	0.118	0.137
normalized	F: exponential	0.181	0.127

Table 11: Estimation of relative additive noise parameter. The estimated additive noise parameter $\widehat{\sigma}_a$ (see table 10) are devided by mean intensity $\bar{I}$ (see table 3).

The estimated parameters for additive noise, $\widehat{\sigma}_a$, are very different for raw and normalized intensity data. For raw data it ranges from 4.230 to 6.205 and for normalized data it ranges from 0.226 to 0.544. The additive noise for Oct4VENUS is bigger than for NanogVENUS. In model F, $\widehat{\sigma}_a$ is higher compared to the model E except in Oct4VENUS normalized. The additive noise changes a lot with the normalization process, it is reduced by a factor between 11 and 19. The main reason for this reduction in $\widehat{\sigma}_a$ is that with normalization the mean intensity values, $\bar{I}$, is reduced from 43.4 to 3.0 in NanogVENUS, which is a factor 14, and from 32.6 to 3.0 in Oct4VENUS,

which is a factor 11 (see table 3). While reducing the intensities, the additive noise is reduced as well. For easier comparision the relative additive noise $\widehat{\sigma}_a/\bar{I}$ is shown in table 11. Mean intensities for the datasets are displayed in table 3. The range for relative additive noise is 0.075 to 0.190.

Overall, there is a good agreement between the estimators from the variance method and the estimators from the likelihood method in the NanogVENUS data. The agreement between variance method and likelihood method is smaller for Oct4VENUS datasets. This might be because Oct4VENUS noise is more irregular in the residuals versus predicted diagrams (see figure 14 to 17) which indicates that higher order of noise are included. Our models with muliplicative and additive noise only may be too simple for Oct4VENUS datasets. The segmentations in the NanogVENUS microscope images are manually re-adjusted which reduces image processing noise. This re-adjustment is not performed in Oct4VENUS images. So the quality of the Oct4VENUS datasets are lower and their noise is more irregular.

		95% CI $\widehat{\sigma}_m$	95% CI $\widehat{\sigma}_a$
NanogVENUS	E: parabola	[0.118, 0.129]	[4.411, 5.029]
raw intensity	F: exponential	[0.135, 0.148]	[5.528, 6.180]
NanogVENUS	E: parabola	[0.098, 0.106]	[0.250, 0.278]
normalized	F: exponential	[0.115, 0.126]	[0.347, 0.386]
Oct4VENUS	E: parabola	[0.179, 0.193]	[3.957, 4.599]
raw intensity	F: exponential	[0.178, 0.197]	[5.876, 6.478]
Oct4VENUS	E: parabola	[0.178, 0.192]	[0.321, 0.390]
normalized	F: exponential	[0.179, 0.195]	[0.511, 0.573]

Table 12: 95% confidence interval for estimated noise parameter with variance method. Based on bootstraping with 1000 drawings.

The accuracy of the parameter estimations can be rated with a 95% confidence interval. For the likelihood method the interval is estimated with the negative Hessian, also called information matrix (see chapter 3.4.3.1). For the variance method it is estimated via Bootstraping with 1000 draws (see table 12). The widths of the 95% confidence interval for the maximum likelihood estimators are quite small. They spans less than 10^{-4} of the estimated values for all multiplicative noise parameters and less than 10^{-2} of the estimated values for all additive noise parameters, and therefore they are not displayed here. The 95% confidence intervals for $\widehat{\sigma}_m$ and $\widehat{\sigma}_a$ for the variance method ranges around $\pm 10\%$ of the estimator value and even wider for $\widehat{\sigma}_a$ in Oct4VENUS normalized with model A.

4.3.3 Comparison of the two methods

Two methods are derived, a likelihood approach and a variance approach, to estimate multiplicative and additive noise parameters from single cell time-lapse fluorescence data. The likelihood method requires numerical integration and optimization. The variance method estimates a linear model and proved to be unbiased, reliable, very easy to implement and gives for our toydata sets and mESC datasets almost the same estimater as the maximum likelihood estimator. So both approaches are appropriate methods. Moreover the variance method is a quick and good alternative to likelihood method to estimate additive and multiplicative noise parameter in setting comparable to ours with 20 000 or more observations.

		$\widehat{\sigma}_m$	$\widehat{\sigma}_a/\bar{I}$
NanogVENUS	raw intensity	0.135	0.105
NanogVENUS	normalized	0.116	0.087
Oct4VENUS	raw intensity	0.186	0.179
Oct4VENUS	normalized	0.187	0.132

Table 13: Median of estimated multiplicative and relative additive noise parameters.

Finally we can conclude that the available NanogVENUS and Oct4VENUS time lapse fluorescence microscopy intensities carry considerable multiplicative and additive noise. The median of the estimated multiplicative log-normal noise parameter $\widehat{\sigma}_m$ for NangoVENUS is 0.14 for raw and 0.12 for normalized intensity data set and for Oct4VENUS they are 0.19 for raw and for normalized intensity data set. The median of the estimated relative additive normal noise parameter $\widehat{\sigma}_a/\bar{I}$ for NangoVENUS is 0.11 for raw and 0.09 for normalized intensity data set and for Oct4VENUS it is 0.18 for raw and 0.13 for normalized intensity data set (see table 13).

4.4 Estimation of copy number from mitosis events

4.4.1 Validation with toydata

4.4.1.1 Likelihood approach

To estimate ν, we use the integral derived in chapter 3.5.4.1, equation 25:

$$f_\Delta(\Delta_i) =$$
$$= \int_0^{I_i'} \int_0^{\infty} \int_0^{\infty} \frac{1}{2\pi^2\sqrt{2}\sigma_m^2\sigma_a\epsilon_{m_1}\epsilon_{m_2}\sqrt{\nu I_i'}} \cdot \exp\left(-\frac{2\left(I_{2i}' - \frac{1}{2}I_1'\right)^2}{\nu I_i'} - \frac{\left(\ln(\epsilon_{m_1}) + \frac{\sigma_m^2}{2}\right)^2}{2\sigma_m^2}\right)$$
$$\cdot \exp\left(-\frac{\left(\ln(\epsilon_{m_2}) + \frac{\sigma_m^2}{2}\right)^2}{2\sigma_m^2} - \frac{(\Delta_i - I_{2i}'(\epsilon_{m_1} + \epsilon_{m_2}) + I_i'\epsilon_{m_2})^2}{4\sigma_a^2}\right) d\epsilon_{m_1}\, d\epsilon_{m_2}\, dI_{2i}'$$

with the estimator for I_i' (see equation 26)

$$\widehat{I_i'} = \frac{1}{2}(I_i + I_{2i} + I_{2i+1} + \widehat{gain}).$$

Maximization of the log-likelihood in respect to parameters σ_m^2, σ_a^2 and ν gives the maximum likelihood estimators for the parameters (see equation 27).

$$l(\sigma_m^2, \sigma_a^2, \nu|\mathbf{\Delta}) = \sum_{i=1}^{N} \ln f_\Delta(\Delta_i)$$

For the 3-dimensional integration R-function *cuhre()* from package *R2Cuba* is used. This is a multidimensional numerical integration with a deterministic iterative adaptive algorithm [23][24]. Maximization of the log-likelihood is done with function *optimize()* [20].

observation	I_i	I_{2i}	I_{2i+1}
1	1.31174	0.63286	0.51201
2	1.88340	0.87850	1.18184

Table 14: Toydataset for upper limit verification. Intensity of mother cell I_i and daughter cells I_{2i}, I_{2i+1}.

The upper limits of the integral (see equation 25) for the multiplicative noise terms ϵ_{m_1} and ϵ_{m_2} is infinity. For the numerical calculation the upper limits must be real.

The dependancy of the integral on different upper limits is tested with two toydata observations with parameter $\nu = 0.001$, $\sigma_m = 0.1$ and $\sigma_a = 0.1$ (see table 14). In this toydataset the standard deviation of the multiplicative noise terms is $SD(\epsilon_m) = \sqrt{\exp(\sigma_m^2) - 1} = 0.1003$ and mean value $\bar{\epsilon}_m = 1.0$ A first choice for the upper limit for the integral is $\bar{\epsilon}_m + 5 \cdot SD(\epsilon_m) = 1.5$. This first choice is verified numerically and the integral value for different upper limits is calculated (see figure 25).

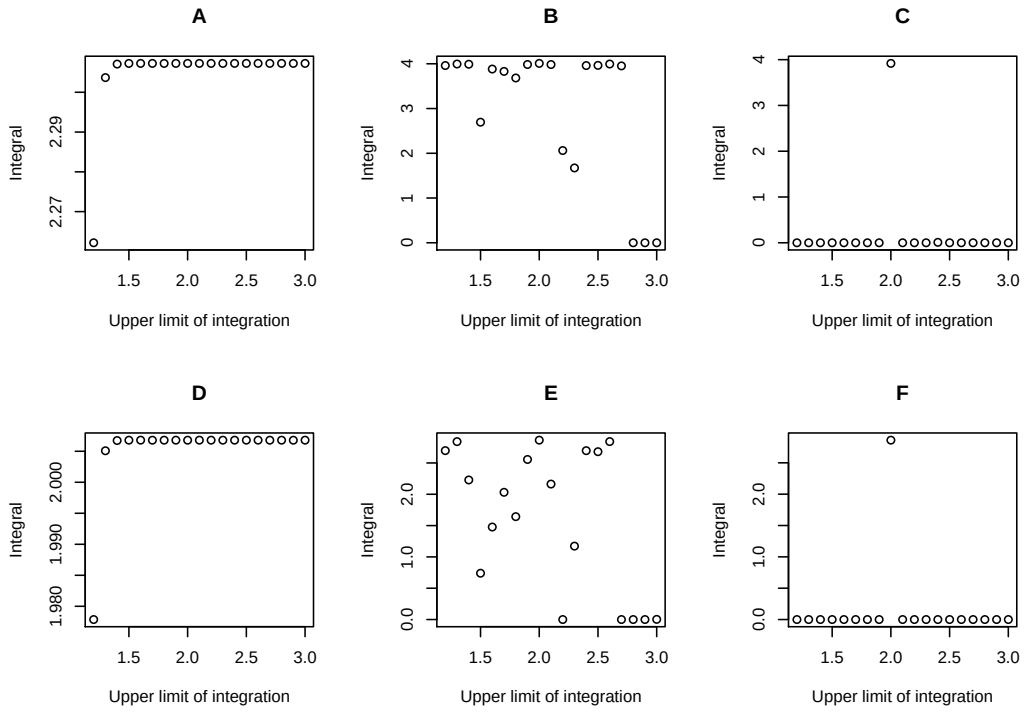

Figure 25: **Toydata integration.** Integral of equation 25 is solved numerically for different upper limits for ϵ_{m_1} and ϵ_{m_2} and different additive noise parameter σ_a. Parameter $\nu = 0.001$ and $\sigma_m = 0.1$ are constant for all figures. Figure A to C: observation 1 of the toydataset (see table 14) is used. Figure D to F: observation 2 of the toydataset (see table 14) is used. Figure A and D: $\sigma_a = 0.1$. Figure B and E: $\sigma_a = 0.01$. Figure C and F: $\sigma_a = 0.001$.

The integral (see equation 25) is evaluated for two different toydata observations (see table 14) with different upper limits for ϵ_{m_1} and ϵ_{m_2}, for $\nu = 0.001$ and $\sigma_m = 0.1$ and for different settings of parameter σ_a (see figure 25). The behavior of the integral is quite similar for the two different observations (see figure 25 A to C versus (see figure 25 D to F). For case A and D with additive noise parameter $\sigma_a = 0.1$ the calculated values for the integral seem stable for upper limits greater 1.5. For case B and E with $\sigma_a = 0.01$ the integral values fluctuates a lot for increasing upper limits. For case C and F the only integrals greater 10^{-3} are for upper limit is 2.0. A very similar result is

obtain if σ_m is modified. If ν is modified, the values for the integral are rather stable (see appendix 7.4 figure 33 and 34).

The adaptive algorithm used for the integration makes a grid in the integration range. For each gridpoint the function is evaluated and a integral value derived. If the grid points do not include the area of considerable probability mass, the integration fails. The mass of the function regarding ϵ_m is expected to be around 1, because it is log-normal distributed with mean is 1.0. The exact shape of the function, of course, depends on the observation. For small σ_a and σ_m the area of mass is quite narrow. It seems that in these cases with upper limit 2.0 and lower limit 0.0 we get at least one initial gridpoint in the area of considerable probability mass and the adaptive algorithm can starts its optimization process to get reliabe results. In the following an upper limit of 2.0 and a lower limit of 0.0 for the integrals of the multiplicative noise terms ϵ_{m_1} and ϵ_{m_2} is used for all observations. However it has not been shown yet that this integration limits are robust for all observations.

We use toydatasets with noise parameter σ_m and σ_a similar to the results obtained for our mESC data (see table 10 and 11) and different conversion factor ν to validate this method to estimate the copy numbers and the parameters (see table 15). Each toydataset consists of 1000 mitosis events.

No of repeats	ν	σ_m	σ_a	n
5	10^{-2}	0.17	0.13	100
5	10^{-3}	0.17	0.13	1000
5	10^{-4}	0.17	0.13	10 000
5	10^{-5}	0.17	0.13	100 000

Table 15: Parameter for toydatasets for validation of likelihood method to estimate copy numbers.

For the optimizations the parameter σ_m, σ_a and ν are logarithmized ($log_{10}(\sigma_m)$, $log_{10}(\sigma_a)$, $log_{10}(\nu)$, see table 16) for more stable results. The starting values for the optimization are drawn from the uniform distribution over the optimization range (see table 17). The integral for one observations needs in our implementation 3 seconds computational time, so for the calculation of the log-likelihood with 1000 observations and one parameter set 50 minutes are needed. After 180 hours and up to 7 optimization steps for each toydataset the computing of the maximzation of the log-likelihood (see equation 27) in respect to the parameter σ_m, σ_a and ν was stopped. For each toydataset maximum likelihood estimators for the parameters are derived (see table 18).

	No	ν	σ_m	σ_a	$log_{10}(\nu)$	$log_{10}(\sigma_m)$	$log_{10}(\sigma_a)$
A	5	10^{-2}	0.17	0.13	-2	-0.77	-0.89
B	5	10^{-3}	0.17	0.13	-3	-0.77	-0.89
C	5	10^{-4}	0.17	0.13	-4	-0.77	-0.89
D	5	10^{-5}	0.17	0.13	-5	-0.77	-0.89

Table 16: Input parameter for toydatasets for validation of likelihood method to estimate copy numbers with logarithmized parameters.

	Optimization range		
	$log_{10}(\nu)$	$log_{10}(\sigma_m)$	$log_{10}(\sigma_a)$
A	[-4, 0]	[-1.27, -0.27]	[-1.39, -0.39]
B	[-5, -1]	[-1.27, -0.27]	[-1.39, -0.39]
C	[-6, -2]	[-1.27, -0.27]	[-1.39, -0.39]
D	[-7, -3]	[-1.27, -0.27]	[-1.39, -0.39]

Table 17: Parameter optimization range in likelihood method to estimate copy numbers with logarithmized parameters.

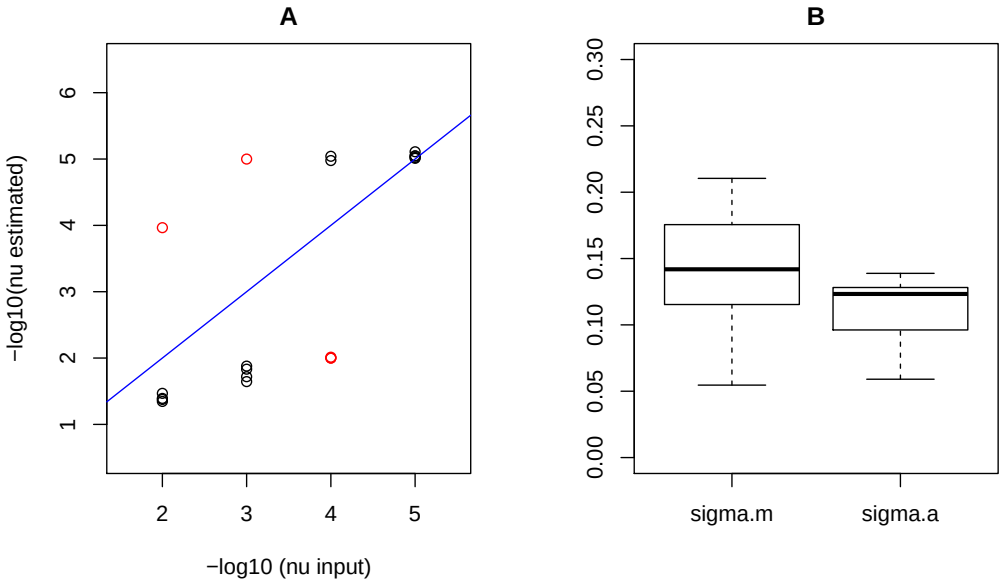

Figure 26: Toydata result from table 18. Figure A displays the negative log_{10} of the estimated conversion factors $\widehat{\nu}$. The blue line is the bisecting line. Red dots are the estimators on the boundary of the optimization range. Figure B shows boxplots of all estimated noise parameters $\widehat{\sigma}_m$ and $\widehat{\sigma}_a$. Input values are $\sigma_m = 0.17$, $\sigma_a = 0.13$.

ν	$\widehat{\nu}$		$\widehat{\sigma}_m$	$\widehat{\sigma}_a$
10^{-2}	$0.011 \cdot 10^{-2}$	#	0.177	0.139
10^{-2}	$3.402 \cdot 10^{-2}$		0.080	0.100
10^{-2}	$4.203 \cdot 10^{-2}$		0.118	0.059
10^{-2}	$4.057 \cdot 10^{-2}$		0.108	0.070
10^{-2}	$4.502 \cdot 10^{-2}$		0.055	0.078
10^{-3}	$13.219 \cdot 10^{-3}$		0.142	0.120
10^{-3}	$22.630 \cdot 10^{-3}$		0.102	0.096
10^{-3}	$0.010 \cdot 10^{-3}$	#	0.179	0.122
10^{-3}	$19.086 \cdot 10^{-3}$		0.140	0.096
10^{-3}	$14.700 \cdot 10^{-3}$		0.127	0.129
10^{-4}	$97.022 \cdot 10^{-4}$	#	0.142	0.125
10^{-4}	$99.986 \cdot 10^{-4}$	#	0.124	0.128
10^{-4}	$0.091 \cdot 10^{-4}$		0.173	0.130
10^{-4}	$99.986 \cdot 10^{-4}$	#	0.113	0.138
10^{-4}	$0.105 \cdot 10^{-4}$		0.194	0.127
10^{-5}	$0.946 \cdot 10^{-5}$		0.170	0.126
10^{-5}	$0.893 \cdot 10^{-5}$		0.174	0.120
10^{-5}	$0.901 \cdot 10^{-5}$		0.164	0.135
10^{-5}	$0.975 \cdot 10^{-5}$		0.177	0.127
10^{-5}	$0.777 \cdot 10^{-5}$		0.210	0.096

Table 18: Result validation of likelihood method to estimate copy numbers for toydatasets with different input values ν. Mean copy number is $n = 1/\nu$, because $\bar{I} = 1.0$. # indicates that $\widehat{\nu}$ lies on the boundary of the optimization range.

For five toydata sets the maximum likelihood estimators for ν are on the boundary of the optimization range. Two of them are on the lower boundary and three on the upper boundary (see table 18). For toydata sets with $\nu = 10^{-2}$ (copy number $n = 100$) all estimated $\widehat{\nu}$ (except that one, which lies on the boundary) are overestimated by a factor 3.4 to 4.5. For toydatasets with $\nu = 10^{-3}$ (copy number $n = 1000$) all estimated $\widehat{\nu}$ (except that, which lies on the boundary) are overestimated by a factor 13 to 23 and for toydatasets with $\nu = 10^{-4}$ (copy number $n = 10000$) all estimated $\widehat{\nu}$ (except those, which lie on the boundary) are underestimated by a factor 9.5 to 11. For toydatasets with $\nu = 10^{-5}$ (copy number $n = 100000$) all estimated $\widehat{\nu}$ are little underestimated by a factor 1.02 to 1.29. Graphical display of all $\widehat{\nu}$ indicates accumulations below 10^{-2} and at 10^{-5} (see figure 26).

The estimators for the noise parameter σ_m and σ_a are all estimated within an accuracy of factor 3. No noise parameter estimator is on the boundary of the optimization range. With this limited testcases it can not be rated proberly if the likelihood approach to estimate copy numbers works acceptably.

4.4.1.2 Variance approach

With a linear model (see chapter 3.5.4.2, equation 31)

$$(I_{2i} - I_{2i+1})_j^2 = \beta_0 + \beta_1 \widehat{I}'_i + \beta_2 \widehat{I}'^2_{i\,i_j} + \epsilon_j \,.$$

and coefficient comparison an estimator for the conversion factor ν can be derived (see equation 32)

$$\widehat{\nu} = \frac{\widehat{\beta_1}}{1 + \widehat{\beta_2}} \,.$$

Parameter	Values						
mitosis N	10^3	10^4	10^5	10^6			
ν	10^{-3}	10^{-4}	10^{-5}	10^{-6}			
σ_m	0.0	0.005	0.01	0.02	0.05	0.1	0.2
σ_a	0.0	0.005	0.01	0.02	0.05	0.1	0.2

Table 19: Parameters for toydatasets for testing the variance approach to estimate the conversion factor ν. Mean signal for mother cells for all cases is $\bar{I}'_i = 1.5$.

Toydata with various values for the number of mitosis events (observations) N, conversion factor ν, multiplicative and additive noise parameter σ_m, σ_a are used to validate the variance appoach to estimate ν (see table 19). The mean intensity for mother cells for all test cases is $\bar{I}'_i = \bar{I}_i = 1.5$, so the mean intensity for all cells, mother and daughters together, is $\bar{I} = \frac{1}{3}(\bar{I}_i + \bar{I_{2i}} + \bar{I_{2i+1}}) = \frac{1}{3}(1.5 + 0.75 + 0.75) = 1.0$. This means, that the mean copy number $\bar{n}$ for the mother cells and its daughter cells is the inverse of the conversion factor ν, $\bar{I} = \nu\bar{n}$. So the range for the mean copy number is $\bar{n} \in (10^3, 10^4, 10^5, 10^6)$. Moreover the parameter σ_a of the normally distributed additive noise can be interpreted as relative additive noise of the mean cell intensity $\sigma_a/\bar{I}$. For details on toydata generation see chapter 4.2. 1000 toydata sets for all combination of the parameters from table 19 are generated and its $\widehat{\nu}$ derived with variance method. Althogether 784 000 toydata sets and $\widehat{\nu}$ are used for this validation.

With density lines of the estimated $\widehat{\nu}$ for a combination of the parameters, it is tested, if the variance method gives unbiased estimates. Figure 27 shows the density lines of $\widehat{\nu}$ of the toydata sets with parameter $\nu = 10^-4$ ($\bar{n} = 10\ 000$, $N \in (10^3, 10^4, 10^5, 10^6)$ and σ_a and $\sigma_m \in (0.0, 0.005, 0.01, 0.02)$. The corresponding figures for other parameter sets look similar. Each diagram represents one parameter set, x-axis is the value of $\widehat{\nu}$. The dotted black lines indicate the true conversion factor ν. The different colors refer

to different amounts of mitosis events per toydataset N. All densitiy lines are symetric around the true conversion factor ν (dotted line). This shows that the variance method is unbiased.

With more mitosis events the width of the density peak gets smaller (black to red to green): The uncertainty of the estimations can be reduced with higher number of mitosis events. Next we investigate the 95% confidence intervals in respect to different parameter sets, σ_a, σ_m and N to see the capability of the method. Figure 28 shows the process of the empirical 95% confidence intervall of the estimated conversion factors in respect to the additive noise level for a subset of σ_a and σ_m input values. The corresponding figures for other noise parameter look similar. X-axis is relative additive noise parameter σ_a/I and the data are split for different multiplicative noise level σ_m. The light green bars indicate the area outside facor 10 accuracy for $\widehat{\nu}$: $\widehat{\nu} \notin (10\nu, 0.1\nu)$. To have a copy number estimate $\widehat{n} = I/\widehat{\nu}$ which 95% confidence interval is within an range of factor 10, the shown confidence intervalls in this figure must not touch the light green area.

Figure 29 finnally shows the result of all toydata in four heat diagrams. Similar to figure 28, the empirical 95% confidence intervals for the estimated conversion factors are analysed. The different grey colors indicate of how many mitosis events a dataset must consists of to get with 95% probability an estimate for copy number within an accuracy range of factor 10. E.g. for a copy number of 100 000, only small noise levels are tolerable ($\sigma_m, \sigma_a/I < 0.02$) and up to 1 million mitosis events are needed to estimate the conversion factor with 95% probalility with an accuracy of factor $\pm$ 10.

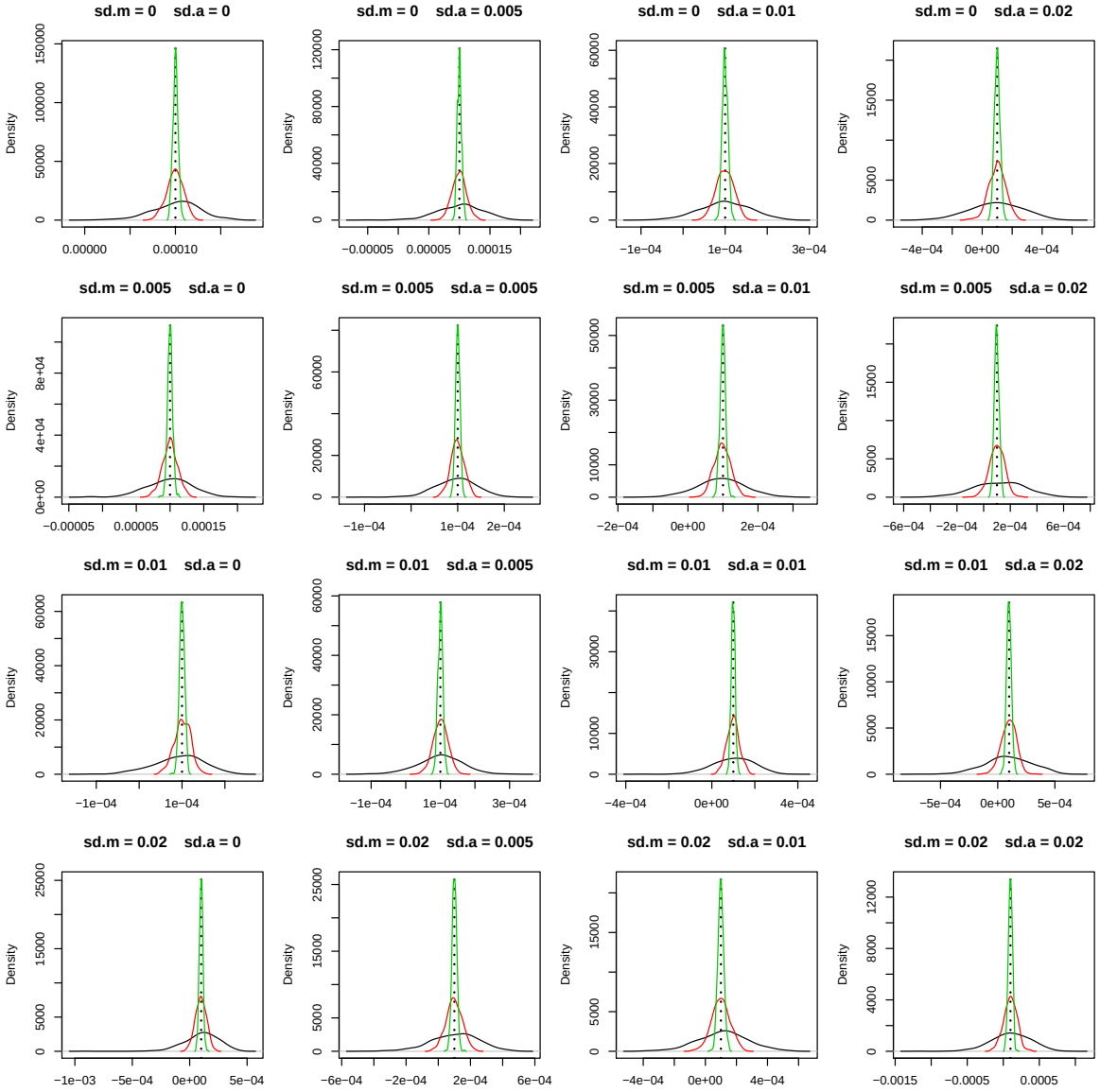

Figure 27: Estimated conversion factors $\widehat{\nu}$ with variance method on toydata. Mean copy number of proteins is 10 000. Density of 1000 toydata sets, black is for $N = 10^4$ events, red is for $N = 10^5$ events, green is for $N = 10^6$ events per toydata set. Dotted line indicates true conversion factor $\nu = 10^{-4}$.

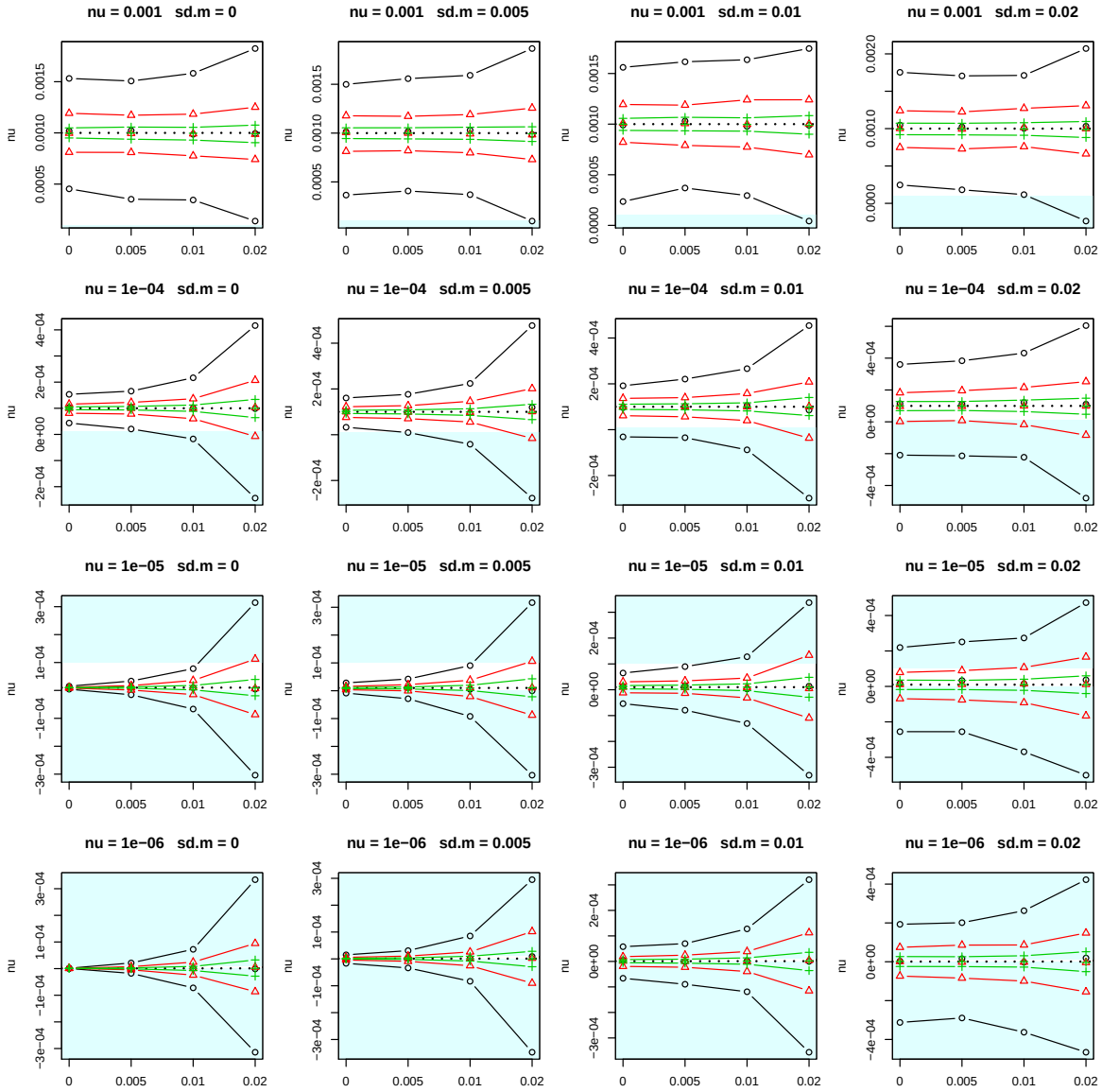

Figure 28: Estimated conversion factors $\widehat{\nu}$ with variance method. X-axis is σ_a and y-axis is upper and lower bound of central 95% confidence quantil of $\widehat{\nu}$. In each plot we show different numbers of events, N=10^4 (black), N=10^5 (red) and N=10^6 (green). The light blue area indicates values outside $(\nu/10, 10\nu)$.

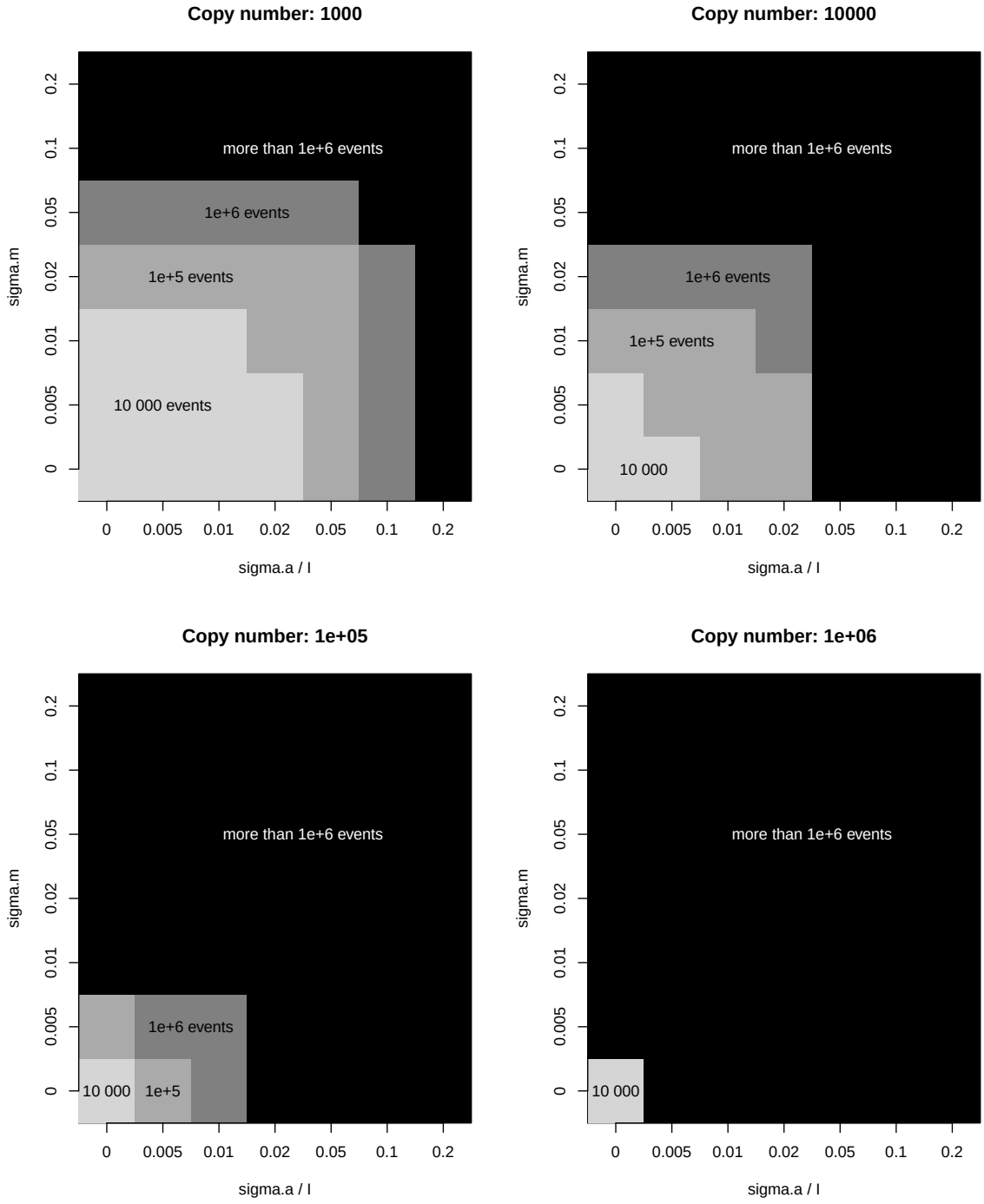

Figure 29: Toydata result: Estimated conversion factor with variance method. The different grey colours indicate the needed number of mitosis events to get an estimate for the copy number which 95% confidence interval is within a factor 10 of the true value.

4.4.2 Application on mESC data

4.4.2.1 Likelihood approach

Computational maximization of the log-likelihood (see equation 27) of the mESC datasets to estimate the fusion protein copy numbers is stopped after 180 hours and 6 optimization steps. For the NanogVENUS datasets the maximum likelihood estimates $\widehat{\nu}$ is 1.79 $\cdot 10^{-3}$ for the raw intensity data set and 1.37 $\cdot 10^{-2}$ for the normalized intensity data set (see table 20). For the Oct4VENUS datasets $\widehat{\nu}$ is 1.50 $\cdot 10^{-3}$ for the raw intensity data set and 6.66 $\cdot 10^{-3}$ for the normalized intensity data set. With the mean fluorescence intensity of the mESC datasets (see table 3) the estimated copy numbers $\widehat{n} = \bar{I}/\widehat{\nu}$ are for NanogVENUS 24308 (raw) and 218 (normalized) and for Oct4VENUS 21717 (raw) and 452 (normalized) (see table 20)

	$\widehat{\nu}$	$\widehat{\sigma}_m$	$\widehat{\sigma}_a/\bar{I}$	$\widehat{n}$
NanogVENUS raw	1.786 $\cdot 10^{-3}$	0.166	0.023	24308
NanogVENUS normalized	1.3717 $\cdot 10^{-2}$	0.107	0.047	218
Oct4VENUS raw	1.500 $\cdot 10^{-3}$	0.117	0.053	21717
Oct4VENUS normalized	6.655 $\cdot 10^{-3}$	0.109	0.030	452

Table 20: Result likelihood method mESC data.

For the NanogVENUS datasets the simultaneously estimated maximum likelihood noise parameters $\widehat{\sigma}_m$ (see table 20) are similar to the estimates from chapter 4.3, table 10. For the Oct4VENUS datasets $\widehat{\sigma}_m$ is 40% smaller as the results in chapter 4.3. The maximum likelihood estimator for relative additive noise $\widehat{\sigma}_a/\bar{I}$ here is up to a factor 5 smaller than the results in chapter 4.3.

The 95% confidence intervals for ν can be estimated with the approximation that the maximum likelihood estimator $\widehat{\nu}$ is normally distributed and the second derivate of the loglikelihood of the mitosis data (compare equation 8). The limits of the 95% confidence intervals for n are the inverses of the limits of the confidence interval for $\widehat{\nu}$ (see equation 33).

$$
\begin{aligned}
CI_{\nu 95\%} &= \widehat{\nu} \pm 1.96 \cdot \frac{1}{-\frac{\partial^2}{\partial^2 \nu^2} l(\widehat{\nu^2}, \widehat{\sigma_m}, \widehat{\sigma_a} | \mathbf{\Delta})} \\
CI_{n 95\%} &= \frac{1}{CI_{\nu 95\%}}
\end{aligned}
\tag{33}
$$

	$\widehat{\nu}$	95% CI
NanogVENUS raw	$1.786 \cdot 10^{-3}$	$[1.783 \cdot 10^{-3}, 1.789 \cdot 10^{-3}]$
NanogVENUS normalized	$1.372 \cdot 10^{-2}$	$[1.371 \cdot 10^{-2}, 1.373 \cdot 10^{-2}]$
Oct4VENUS raw	$1.500 \cdot 10^{-3}$	$[0.026 \cdot 10^{-3}, 2.637 \cdot 10^{-3}]$
Oct4VENUS normalized	$6.655 \cdot 10^{-3}$	$[6.653 \cdot 10^{-3}, 6.656 \cdot 10^{-3}]$

Table 21: 95% confidence intervals for conversion factor ν.

	$\widehat{n}$	95% CI
NanogVENUS raw	24 308	[24 271, 24 346]
NanogVENUS normalized	218	-
Oct4VENUS raw	21 717	[10 955, 1 235 074]
Oct4VENUS normalized	452	-

Table 22: 95% confidence intervals for copy numbers n. The 95% confidence interval for the normalized intensity datasets is smaller than 1 unit and therefore not displayed here.

The 95% confidence interval for the maximum likelihood estimation of the fusion protein copy numbers are pretty narrow for NanogVENUS raw and normalized intensity and for Oct4VENUS normalized intensity data sets. However for Oct4VENUS raw intensity data set the confidence interval ranges from 10 955 to 1.2 millions (see table 22).

4.4.2.2 Variance approach

The accuracy of the variance approach is highly depending on the noise magnitude. If there is too much noise, the accurcy of the coversion factor estimation is low.

	$\widehat{\nu}$	95% CI
NanogVENUS raw	4.339	[1.003, 7.587]
NanogVENUS normalized	-0.083	[-0.323, 0.084]
Oct4VENUS raw	-0.859	[-2.522, 0.588]
Oct4VENUS normalized	-1.412	[-2.516, -0.293]

Table 23: Variance method. Estimation of conversion factor $\widehat{\nu}$. The 95% confidence interval is based on 10 000 bootstrap samples.

Table 23 shows the result of the estimated conversion factors $\widehat{\nu}$ with the variance method and its 95% confidence interval. Negative conversion factor would indicate negative copy numbers, which does not make sense. For NanogVENUS raw intensity

we get a positive conversion factor $\widehat{\nu} = 4.34$. Interestingly this is almost exact a tenth of the mean intensity in this dataset of 43.4. This means the mean copy number would be 10. However the copy number does not change with normalization process and so we can not derive copy numbers for NanogVENUS here.

Figure 29 summarizes the results of the tests for the variance method with toydata. It shows the needed number of observed mitosis events to get an conversion factor estimation with an accaptable accuracy depending on the noise level and the expected copy number. For the available NanogVENUS and Oct4VENUS data with noise level σ_m and σ_a/I of about 0.15, more than 1 million observed mitosis events are needed to get with 95% probability an estimate with an accuracy of factor $\pm$ 10. In our datasets only 1191 to 1235 mitosis events are included (see table 2). This is too little to get a reliable estimate for conversion factor ν and copy number n with variance method.

4.4.3 Comparison of the two methods

A likelihood approach and a variance approach are derived and validated to estimate conversion factors from mitosis fluorescence data. Both algorithms have their strengths and weaknesses.

The likelihood method requires 3-dimensional numerical integration and optimization. For our mESC and toy datasets the computation was stopped after 180 hours. There are different possibilities to speed up the numerical optimizations process e.g. to integrate for a simple grid first and/or to approximate the integral on a new gridpoint with the help of the integrals and its derivates from the neighboring gridpoints. The validation with toydata does not show clearly the capability of the likelihood approach. For 5 out of 20 toydata sets the estimated $\widehat{\nu}$ are on the boundary of the optimization range and the other values are accumulated below 10^{-2} and around 10^{-5}. One weakness of the performed integrations are the fixed upper limits for the integrals for ϵ_{m1} and ϵ_{m2} at 2.0 and the fixed lower limits at 0.0. An evaluation of the integral for every observation separately to locate the area of considerable probability mass to get more reasonable integration limits may stabilize the integration and optimization step in the likelihood approach.

The variance method fits a linear model to estimate the conversion factor. The advantage of this unbiased method is its ease of use. Implementation is simple and calculation time is negligible. In three out of our four mESC datasets the estimated $\widehat{\nu}$ are negative which would imply negative copy numbers. This is outside of the reasonable range.

Only for NanogVENUS raw intensity data set ν is estimated to 4.3, which would imply a copy number of 10. Validation with toydata shows that for our datasets with the observed amount of multiplicative and additive nose more than 1 million observations are necessary to get copy number estimations with acceptable accuracy.

The likelihood approach estimates low mean copy numbers for the normalized intensity data sets (218 for NanogVENUS and 452 for Oct4VENUS) and relatively high mean copy numbers for the raw intensity data sets (24 308 for NanogVENUS and 21 717 for Oct4VENUS). This is remarkable because we expect that the true copy numbers do not change with the normalization process of the intensities. The estimates for the multiplicative noise parameter are similar whereas the estimates for relative additive noise parameter are smaller than the estimates from chapter 4.3. Because of the long computational time, it was not possible to rate the capability of the likelihood method in an appropriate manner. Further investigations shall be performed to understand this method better and its sensitivity to noise.

5 Discussion

In 2005 Rosenfeld et al. [14] proposed a method to estimate the conversion factor from single cell time-lapse fluorescence data. This method considers the whole lineage tree of cells with stopped fusion protein expression, but it can be adjusted for our mESCs, which express fusion proteins. A fundamental assumption is that the partitioning of the fusion proteins from mother to either daughter cell is homogeneous. Later Rosenfeld et al. [15] included additive noise and used a likelihood approach as well as a variance approach to derive their methods. However the combination of multiplicative plus additive noise is not covered with this method. Interestingly the authors We use both approaches to enhance this technique and include multiplicative plus additive noise.

Our datasets contain no indication against homogeneous partitioning of proteins during mitosis. Data analysis shows that the intensities are neither normally nor log-normally distributed. On average, the intensities of both daughter cells are higher than the intensity of their mother cell, which indicates a copy number gain during mitosis. We see many individual trends of the intensity during a cell cycle. In some cells the intensity increases steadily over time, in other cells the intensity remains almost the same or fluctuates. These individual trends must be included in the LMMs for the estimation of noise and signal. Models with random intercept only seem to be not sufficient as the ACFs for these models indicate autocorrelations. Therefore we recommend models with random parabola and random exponential effects in age. However, in our NanogVENUS datasets most of the cells have only intensity observations at the beginning, the middle and the end of its cell cycle. Choosing a LMM with individual parabola for theses cells will possible lead to over fittings and the noise will be underestimated.

In the literature, multiplicative and additive noise are described as main contributions in camera imaging [17]. In our data, signal and noise scatter diagrams show heteroscedasticity in the noise terms, which seems to be not only caused by multiplicative noise. We choose log-normally distributed multiplicative noise and normally distributed additive noise for our models. The two proposed methods to estimate additive and multiplicative noise parameters, the likelihood approach and the variance approach, led to the almost same result. Interestingly the variance approach proved to be a reliable and easy to implement method to estimate noise parameters. The noise is smaller for the NanogVENUS datasets, what might be because its segmentation has been manually readjusted. The multiplicative noise parameter σ_m is estimated to 0.12 and 0.13 for the NanogVENUS datasets and to 0.19 for the Oct4VENUS datasets. The relative additive noise parameter $\sigma_a/\bar{I}$ is estimated to 0.09 and 0.10 for the NanogVENUS datasets and

to 0.13 and 0.18 for the Oct4VENUS datasets. This means that the contribution from both noise terms, additive and multiplicative, are in the same range.

This noise must be considered for fusion protein copy number estimations. Western blot analyses indicate that the copy numbers for Nanog are 400 000 [13] or 1.5 millions [12] and 180 millions for Oct4. These copy numbers would cause a relative standard deviation due to the apportion process in the difference of the two daughter cell intensities $\sigma_{binom} < 1.6 \cdot 10^{-3}$ (see equation 14). This means the task in estimating copy numbers is to find traces of the protein partitioning during mitosis, which relative magnitude is around 10^{-3}, in observations, which noise has relative magnitude 10^{-1}.

The variance method to estimate fusion protein copy numbers gives unbiased estimates for the conversion factors in regime with multiplicative plus additive noise (see figure 29). The big advantage of this method is its ease of use and its fast computing. However the noise level in the mESC data used in this project is too high and the number of events too small. The experimental noise must be reduced at least by a factor 10 and the observed number of mitosis events must be increased at least by a factor 1000 to get reliable estimates. The second method uses maximum likelihood theorie to estimate fusion protein copy numbers from fluorescence intensities. For this it is necessary to solve and optimize numerically a 3-dimensional integral. The computation time for a data set with 1000 observations was more than 100 hours in a single kernel. The likelihood method estimates the multiplicative and additive noise parameter simultaneously with the copy numbers. The estimates for the multiplicative noise parameter here for the mitosis data are similar to the multiplicative noise parameters estimated for the longitudinal data which confirms this method.

Though the estimates for relative additive noise parameters for the mitosis data are much smaller than the estimates for relative additive noise parameters for the longitudinal data. This can be explained with the transcriptional noise, which describes the stochastic process of the gene expression. During mitosis, the transcription of NanogVENUS and Oct4VENUS might be reduced and the transcriptional noise is lower and smaller estimates of the noise parameters are obtained for the mitosis data. This would mean that the transcriptional noise is mainly an additive noise and the microscope measurement error is mainly a multiplicative noise.

We estimate the mean fusion protein copy numbers in NangoVENUS raw intensity to 24 308 and in NangoVENUS normalized intensity to 218 only. This discrepancy is also seen in the Oct4VENUS datasets with estimated copy numbers are 21 717 and 452 for raw intensity and normalized intensity data set respectively. This can not be

explained by biology, because both datasets, raw and normalized, are based on the same experiment with the same mESCs with the same copy numbers. However in the raw data the noise distribution seems to be more heterogeneous than in the normalized data and so our noise model is not correct then which may cause this difference in the estimation. Another reason might be that the normalization process changes the structure of observed intensities and the used assumptions such as linear relation between copy numbers and fluorescence intensities are valid only in the post processed intensity datasets. The small confidence interval for the copy numbers for normalized intensity data sets compared with the wider confidence intervals for the raw intensity data sets support the conclusion, that the normalized intensity data sets are more reliable. This would mean that the copy numbers of the fusion proteins are several hundreds only.

Anyway, the estimated copy numbers are less than the reported copy numbers from western blot analyses, which found 400 000 to 180 million Nanog and Oct4 proteins in mESCs [12] [13]. This difference might be explained with the actual partition process of the proteins from mother to daughter cells which may be affected by several mechanisms. Prior to partition the proteins might cluster to multimers or tack on the DNA or the concentration of different proteins might influence each other or a combination of all of them [25] and the assumption of homogeneous partitioning is not valid. For example if the fusion proteins cluster to multimers with 1000 proteins each and the partitioning from mother to daughter cell is homogeneous for the multimers (and not for the single proteins), the binomial model for the partitioning hold for multimers and not for single protein. This would mean that the actual copy number in mESC is the estimated copy number from our model multiplied with the size of the multimers. Therefore our analysis indicates that Nanog clusters prior mitosis to multimers with mean copy number is 2000 to 7000 and Oct4 clusters to multimers with mean copy number is 400 000.

6 Outlook

Single cell time-lapse fluorescence microscopy is an emerging technique and our methods to estimate fusion protein copy numbers may be applied to other datasets and cell types. However some refinements to the methods are advisable:

Further refinement on the likelihood approach may include an analysis of the integral function to find the area of considerable probability mass for each observation individually to define its integration limits. This can be done with an optimization to find for every observation the values for the integration variables for which the integrand becomes maximal. With the second derivatives, the Hessian matrix, the shape of the integrand and the area of considerable probability mass can be estimated with the approximation of an multivariate normal distribution. Then this area is used for the integration. Moreover with the covariance matrix of the multivariate normal distribution, which is defined by this Hessian matrix and its prefactor, which is defined by the maximal value of the integrand, the integral can be approximated to save computing time. The optimization over the three parameters, σ_m, σ_a and ν, may be carried out simultaneously and not one after another as it is done within this thesis. These steps shall stabilize the maximum likelihood estimation and reduce computation time.

Further data refinement are required to apply the variance approach successfully. Reduction of the measurement noise level and an increase of the observed number of mitosis events are necessary. Higher resolution in microscope imaging to reduce noise is technically possible and automated image processing might help to get higher number of observations. Moreover, reliable results with the variance method may be obtained for fusion proteins with lower copy numbers, for which the noise tolerance is higher because the relative difference caused by the partitioning process is higher.

The result of the likelihood method to estimate copy numbers indicates that the additive noise is low during mitosis events. Maybe a model with multiplicative noise only and without additive noise can be applied successfully to estimate copy numbers. We used the median of the last three intensity observations from the mother cell as the actual mother cell intensity and we used the median of the first three intensity observations from the daughter cells as the actual daughter cell intensities. Another possibility is to use the last respectively the first values, or the mean of the last respectively the first three intensity observations, or to estimate the actual mother and daughter cell intensities with the parameters from the LMMs.

Strictly speaking, our methods to estimate copy numbers, likelihood and variance ap-

proach, estimate the number of "units" for which homogeneous partitioning holds. If the fusion proteins cluster prior to mitosis, we estimate the number of multimers. However if the fusion proteins tack on the DNA, no relative difference would be caused by the partitioning process, because the DNA is divided absolutely homogeneous, and infinite copy numbers should be estimated with our methods. Knowing this, comparison of copy numbers derived from western blot or other techniques with copy numbers estimated with our methods may help to understand mechanism of protein behaviors during mitosis. Analysis of copy numbers on sub datasets with different mother cell intensities may show new protein behavior during mitosis.

7 Appendix

7.1 Density transformation

7.1.1 Transformation for univariate density functions

Suppose X is a continuous random variable and $y = g(x)$ strong monotonic with the inverse function $x = g^{-1}(y)$ and the continuous derivative $\frac{\partial g^{-1}(y)}{\partial y}$. Then the density function for the random variable $Y = g(X)$ is

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{\partial g^{-1}(y)}{\partial y} \right|.$$

7.1.2 Transformation for bivariate density functions

Suppose $\mathbf{X} = (X_1, X_2)^t$ is a bivariate random variable with joint density function $f_{X_1,X_2}(x_1, x_2)$. If $(x_1(y_1, y_2), x_2(y_1, y_2)) \in T_{X_1,X_2}$, the joint density function after transformation T, $\mathbf{Y} = (Y_1, Y_2)^t = T(X_1, X_2)$ is

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(x_1(y_1, y_2), x_2(y_1, y_2)) \cdot |J(y_1, y_2)|$$

while $J(y_1, y_2)$ is the Jacobi determinant

$$J(y_1, y_2) = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} \\ \frac{\partial x_1}{\partial y_2} & \frac{\partial x_2}{\partial y_2} \end{vmatrix}.$$

Please note, that bivariate density functions might only be transformed in another bivariate density functions. To reduce the dimensionality, one variable can be integrated out.

7.2 Variance and covariance

Following rules for variances and covariances are used [26]: X, Y, Z are random variables, a, b are constant factors.

$$var(X) = E\left([X - E(X)]^2\right) = E(X^2) - E^2(X)$$

$$var(X \pm Y) = var(X) + var(Y) \pm 2\, cov(X, Y)$$

$$var(XY) = E^2(X)\, var(Y) + E^2(Y)\, var(X) + var(X)\, var(Y)$$

$$cov(X, Y) = E([X - E(X)]\,[Y - E(Y)]) = E(XY) - E(X)E(Y)$$

$$cov(aX + Z + b, Y) = a \cdot cov(X, Y) + cov(Z, Y)$$

7.3 Estimation of copy number using cell size information

For the estimation of protein copy numbers it is assumed, that the proteins have the same chance to be allocated in either daughter cell (chapter 3.5.1). Teng and colleagues [16] do not think like this and looked at the cell areas and flourescence intensities and found positive correlations coefficients between 0.45 and 0.58 in their data when plotting the ratio of the daughter cell intensities to its mother cell intensities $y = f_{2i}/f_i$ against the ratio of its areas $x_i = A_{2i}/A_i$.

However we think, that the observed positive correlation between intensities and cell areas mainly are not due to the assumed inhomogenous partitioning process of the proteins but due to image capturing and image processing noise. Noise in segmentation leads to variation in cell areas and to variation in intensities as well. If the cell area is bigger due to noise, the intensity is probably higher as well which may cause the observed positive correlation between areas and intensities. Therefore we think that this model is not applicable for our datasets. However this method, which uses cell size information, is explained here because it is an interessting approach.

In this approach, each protein from the mother cell has not the same chance to be allocated in one of the two daughter cells. The chance is defined by the ratio of the size of the daughter cell $p = x_i$. So $P(y_i|x_i)$ is an binomial distribution with $p = x_i$. Like with the other techniques, a normal distribution with is used as a aproximation with mean $\mu = x_i$ and variance $\sigma^2 = \sigma_N^2$.

$$f_{Y_i}(y_i|x_i) = \sqrt{\frac{1}{2\pi\sigma_N}} e^{-\frac{y_i - x_i)^2}{2\sigma_N^2}}$$

Actually, the authors use a bivariate normal distribution for (x_i, y_i) with covariance matrix $\Sigma = \begin{pmatrix} \sigma_A^2 & 0 \\ 0 & \sigma_N^2 \end{pmatrix}$ and mean vector $\mu = \begin{pmatrix} 0.5 \\ x_i \end{pmatrix}$ to motify their assumptions. Because of the missing covariance ($\Sigma_{1,2} = 0$), this gives the same result as the univariate approach shown here.

Maximize the likelihood for σ_N

$$\widehat{\sigma_N^2} = \frac{1}{L} \sum_i^L (y_i - x_i)^2.$$

In the dataset from Teng and colleauges, the error in area partitioning is small ($\sim -3.5\%$). This might be the motivation to approximate the variance in copy number

with the binomial distribution with $p = 0.5$:
$\widehat{\sigma_N^2} = 1/N_0 p(1-p) = \nu/(4\, f_0)$, where N_0 and f_0 are mean protein number and intensities of the mother cells. Finally the conversion factor $\widehat{\nu}$ is given

$$\widehat{\nu} = 4 f_0 \frac{1}{L} \sum_i^L (y_i - x_i)^2 \qquad\qquad f_0 = \frac{1}{L} \sum_i^L f_i.$$

7.4 Additional figures

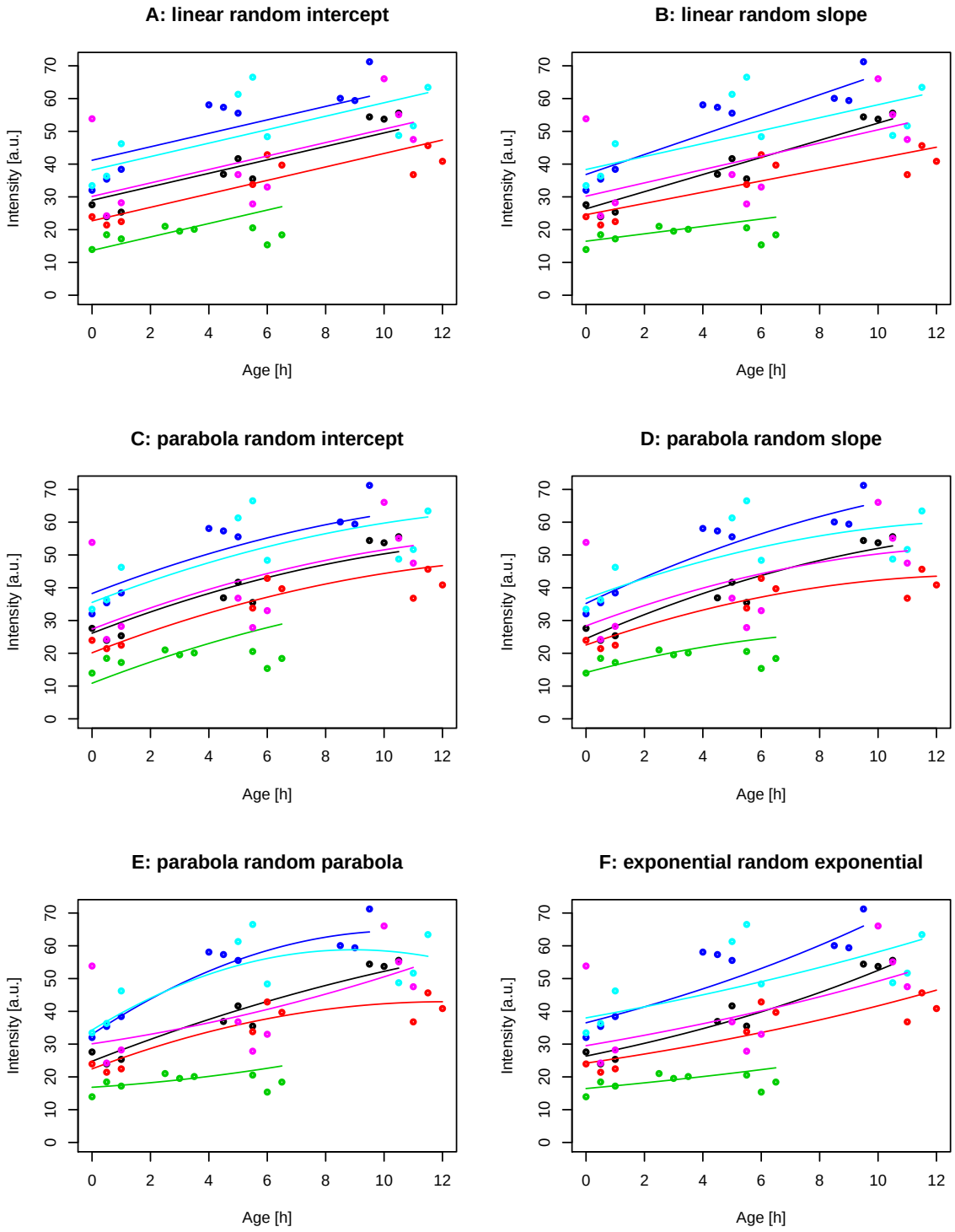

Figure 30: NanogVENUS raw intensity. Intensities I and expectation lines of LMM for six randomly choosen cells.

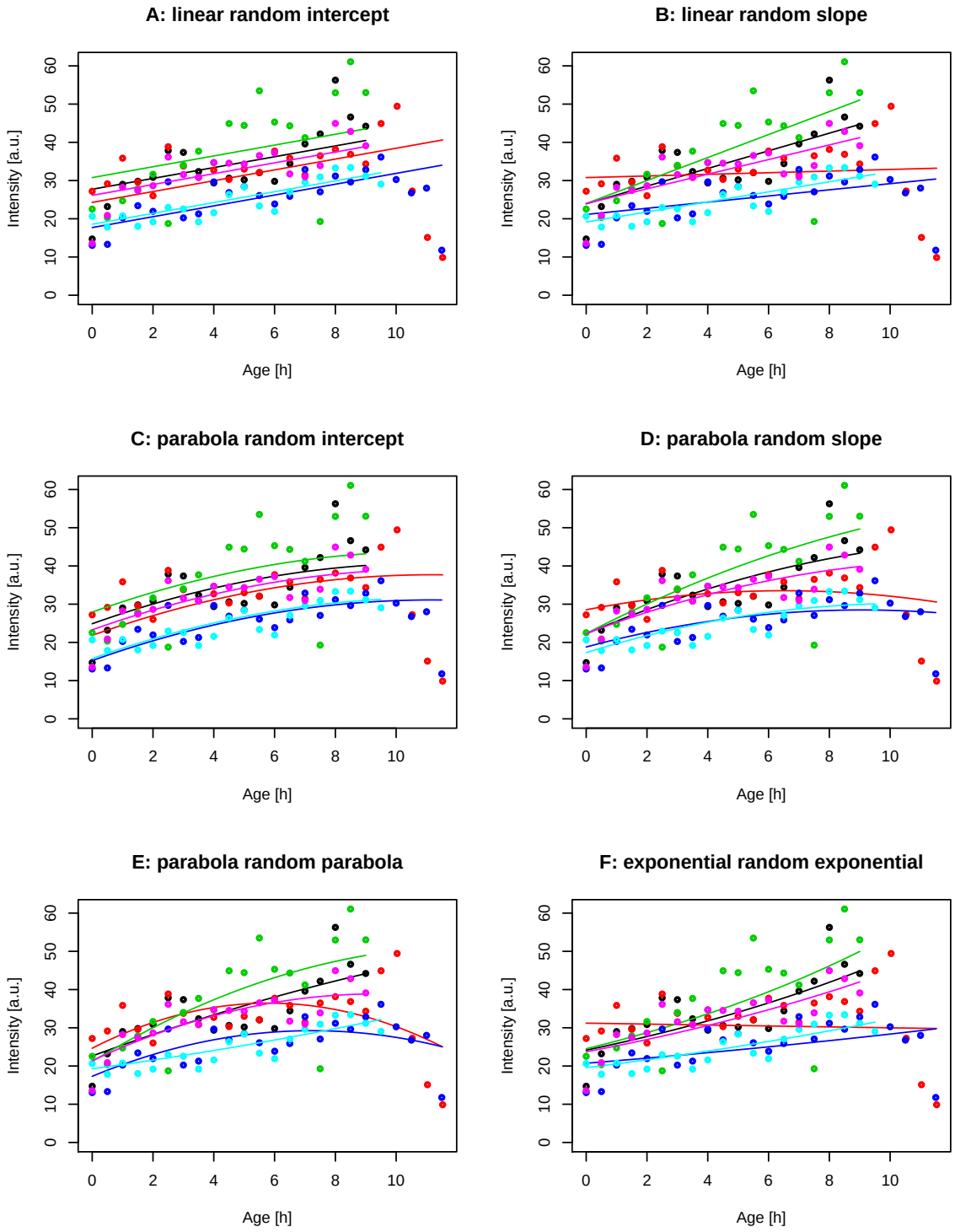

Figure 31: OctVENUS raw intensity. Intensities I and expectation lines of LMM for six randomly choosen cells.

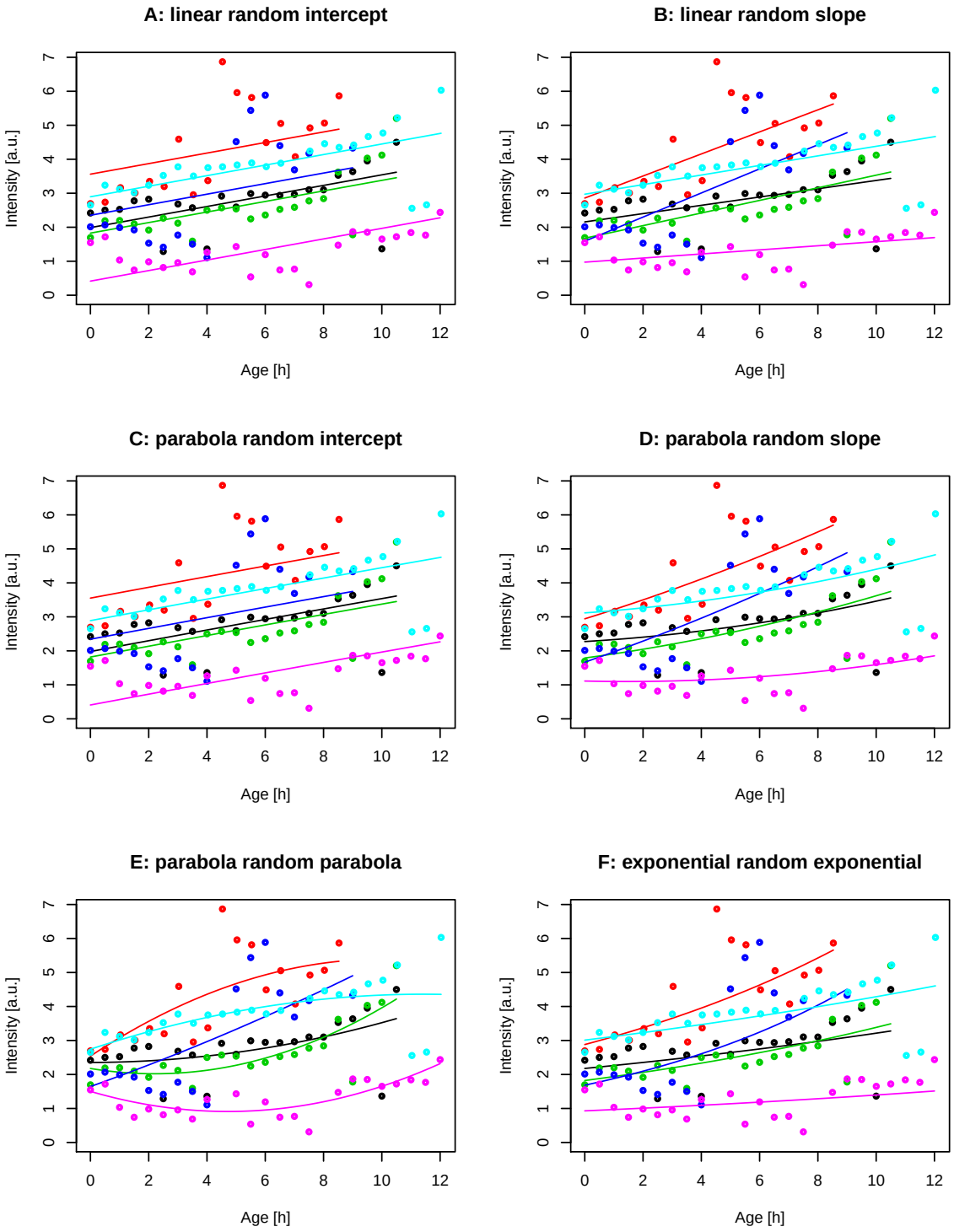

Figure 32: OctVENUS normalized intensity. Intensities I and expectation lines of LMM for six randomly choosen cells.

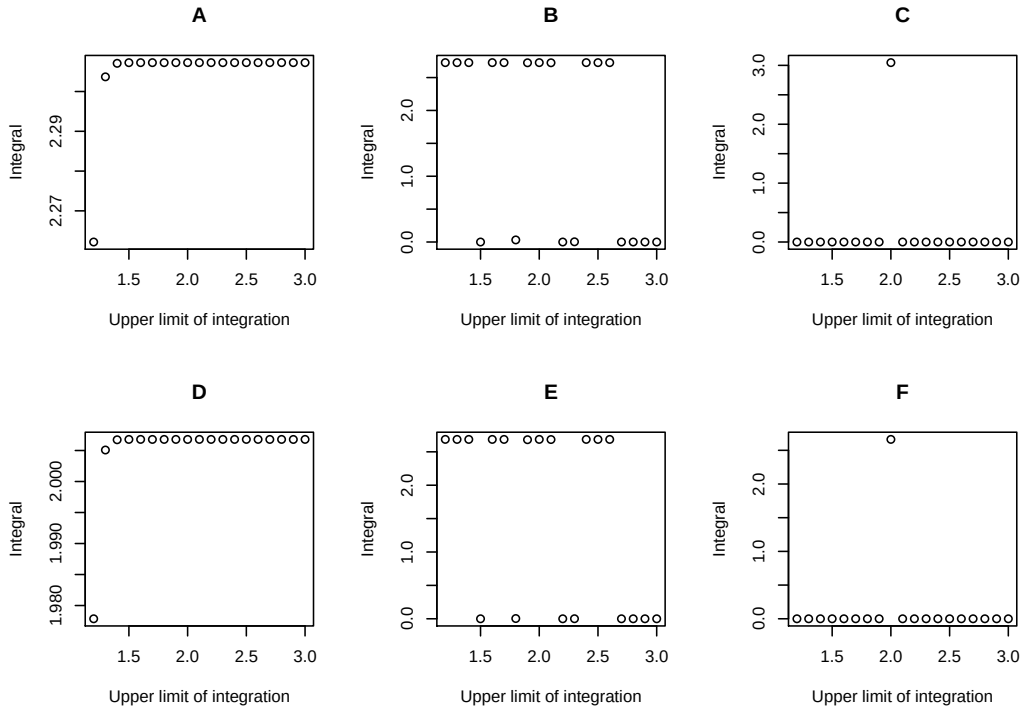

Figure 33: **Toydata integration.** Integral of equation 25 is solved numerically for different upper limits and different multiplicative noise parameter σ_m. $\nu = 0.001$, $\sigma_a = 0.1$. A-C: observation 1 of the toydata set is used, D-F: observation 2 of the toydata set is used (table (14)). A and D: $\sigma_m = 0.1$. B and E: $\sigma_m = 0.01$. C and F: $\sigma_m = 0.001$.

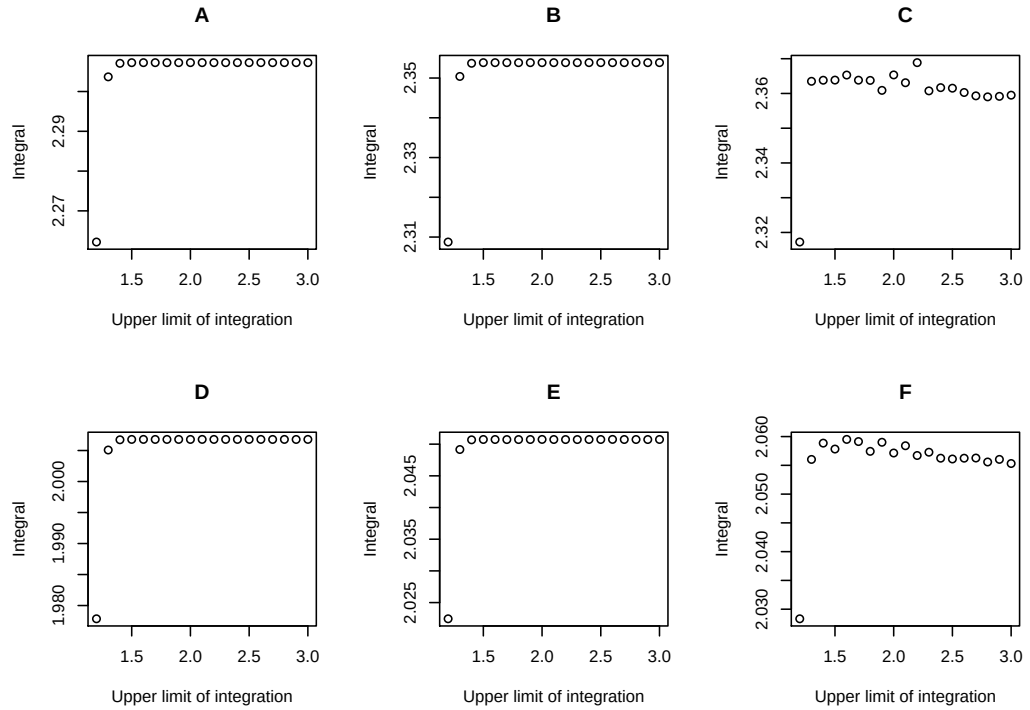

Figure 34: **Toydata integration.** Integral of equation 25 is solved numerically for different upper limits and different conversion faktor ν. $\sigma_m = 0.1$, $\sigma_a = 0.1$. A-C: observation 1 of the toydata set is used, D-F: observation 2 of the toydata set is used (table (14)). A and D: $\nu = 10^-3$. B and E: $\nu = 10^-4$. C and F: $\nu = 10^-5$.

References

[1] M. J. Evans and M. H. Kaufmann, "Establishment in culture of pluripotential cells from mouse embryos," *Nature*, vol. 292, pp. 154–156, July 1981.

[2] G. R. Martin, "Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells," *Proc Natl Acad Sci USA*, vol. 78 (12), pp. 7634–7638, December 1981.

[3] I. Chambers, "Cells of the ever young: Getting closer to the truth," *ScienceDaily*, June 2012.

[4] G. Pan and J. A. Thomson, "Nanog and transcriptional networks in embryonic stem cell pluripotency," *Nature*, vol. 17, pp. 42–49, Jan 2007.

[5] I. Glauche, H. Maria, and R. Ingo, "Nanog variability and pluripotency regulation of embryonic stem cells - insights from a mathematical model analysis," *PLoS ONE*, vol. 5, p. e11238, 06 2010.

[6] V. Chickarmane, C. Troein, U. Nuber, H. Sauro, and C. Peterson, "Transcriptional dynamics of the embryonic stem cell switch," *PLoS Computational Biology*, vol. 2(9), pp. 1080–1092, Sep 2006.

[7] J. Silva, I. Chambers, S. Pollard, and A. Smith, "Nanog promotes transfer of pluripotency after cell fusion," *Nature*, vol. 441, pp. 997–1001, June 2006.

[8] M. Strasser, F. Theis, and C. Marr, "Stability and multiattractor dynamics of a toggle switch based on a two-stage model of stochastic gene expression," *Biophys J.*, vol. 102, pp. 19–29, Jan 2012.

[9] G. JW, J. Krijgsveld, and A. Heck, "Quantitative proteomics by metabolic labeling of model organisms," *Mol Cell Proteomics*, vol. 9(1), pp. 11–24, Jan 2010.

[10] B. Schwanhausser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach, "Global quantification of mammalian gene expression control," *Nature*, vol. 473, pp. 337–342, May 2011.

[11] B. D. Lindenbach, M. J. Evans, A. J. Syder, B. Woelk, T. L. Tellinghuisen, C. C. Liu, T. Maruyama, R. O. Hynes, D. R. Burton, J. A. McKeating, and C. M. Rice, "Complete replication of hepatitis c virus in cell culture," *Science*, vol. 309, pp. 623–626, July 2005.

[12] A. Filipczyk *Personal communication*, April 2011.

[13] N. Mullin *Personal communication*, April 2011.

[14] N. Rosenfeld, J. Young, U. Alon, P. S. Swain, and M. B. Elowitz, "Gene regulation at the single-cell level," *Science*, vol. 307, pp. 1962–1965, 2005.

[15] N. Rosenfeld, T. J. Perkins, U. Alon, M. B. Elowitz, and P. S. Swain, "A fluctuation method to quantify in vivo fluorescence data," *Biophysical Journal*, vol. 91, no. 2, pp. 759–766, 2006.

[16] S. W. Teng, Y. Wang, K. Tu, T. Long, P. Mehta, N. Wingreen, B. Bassler, and N. Ong, "Measurement of the copy number of the master quorum-sensing regulator of a bacterial cell," *Biophysical Journal*, vol. 98, pp. 2024–2031, 2010.

[17] K. Irie, A. E. McKinnon, and I. M. Woodhead, "A technique for evaluation of ccd video-camera noise," *IEEE TCSVT*, vol. 1265, pp. 1–5, 2007.

[18] M. Schwarzfischer, C. Marr, J. Krumsiek, P. S. Hoppe, T. Schroeder, and F. J. Theis, "Efficient fluorescence image normalization for time lapse moviesl," *Proc. Microscopic Image Analysis with Applications in Biology*, 2011.

[19] L. Fahrmeir, T. Kneib, and S. Lang, *Regression: Modelle, Methoden und Anwendungen*. Springer Verlag, 2. ed., September 2009.

[20] R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[21] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and R Development Core Team, *nlme: Linear and Nonlinear Mixed Effects Models*, 2012. R package version 3.1-103.

[22] M. Lindstrom and D. Bate, "Nonlinear mixed effects models for repeated measures data," *Biometrics*, vol. 46, pp. 673–6879, 1990.

[23] T. Hahn, A. Bouvier, and K. Kieu, *R2Cuba: Multidimensional Numerical Integration*, 2012. R package version 1.0-7.

[24] J. Berntsen and T. O. Espelid, "An adaptive algorithm for the approximate calculation of multiple integrals," *ACM Transactions on Mathematical Software*, vol. 17(4), pp. 437–451, 1991.

[25] D. Huh and J. Paulsson, "Random partitioning of molecules at cell division," *Proc Natl Acad Sci U S A*, vol. 108, pp. 15004–15009, 2011.

[26] L. Sachs and J. Hedderich, *Angewandte Statistik, Methodensammlung mit R*. Springer Verlag, 12. ed., March 2006.

Ich versichere, dass ich diese Masterarbeit selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe.

25.09.2012 .....

Anton Hilger

Vielen Dank an Fabian Theis für die Möglichkeit, diese Masterarbeit in seiner Gruppe zu erarbeiten, an Carsten Marr für die sehr gute Betreuung, an Michael Schwarzfischer für die Bereitstellung der Daten und allen anderen Mitgliedern von CMB für geleistete Hilfe, fachliche Diskussionen und die gute Stimmung.