Cellular Oncology 30 (2008) 323–335
DOI 10.3233/CLO-2008-0428
IOS Press

MLPAnalyzer: Data analysis tool for reliable automated normalization of MLPA fragment data*

Jordy Coffa [a,c,**], Mark A. van de Wiel [a,b,d], Begoña Diosdado [a], Beatriz Carvalho [a], Jan Schouten [c] and Gerrit A. Meijer [a]

[a] *Tumour Profiling Unit & Micro Array Facility, Department of Pathology (Tumor Profiling Unit), VU University Medical Center, Amsterdam, The Netherlands*
[b] *Department of Biostatistics, Department of Pathology, VU University Medical Center, Amsterdam, The Netherlands*
[c] *MRC-Holland, Amsterdam, The Netherlands*
[d] *Department of Mathematics, VU University, Amsterdam, The Netherlands*

Abstract. *Background*: Multiplex Ligation dependent Probe Amplification (MLPA) is a rapid, simple, reliable and customized method for detection of copy number changes of individual genes at a high resolution and allows for high throughput analysis. This technique is typically applied for studying specific genes in large sample series. The large amount of data, dissimilarities in PCR efficiency among the different probe amplification products, and sample-to-sample variation pose a challenge to data analysis and interpretation. We therefore set out to develop an MLPA data analysis strategy and tool that is simple to use, while still taking into account the above-mentioned sources of variation.

Materials and methods: MLPAnalyzer was developed in Visual Basic for Applications, and can accept a large number of file formats directly from capillary sequence systems. Sizes of all MLPA probe signals are determined and filtered, quality control steps are performed, and variation in peak intensity related to size is corrected for. DNA copy number ratios of test samples are computed, displayed in a table view and a set of comprehensive figures is generated. To validate this approach, MLPA reactions were performed using a dedicated MLPA mix on 6 different colorectal cancer cell lines. The generated data were normalized using our program and results were compared to previously performed array-CGH results using both statistical methods and visual examination.

Results and discussion: Visual examination of bar graphs and direct ratios for both techniques showed very similar results, while the average Pearson moment correlation over all MLPA probes was found to be 0.42. Our results thus show that automated MLPA data processing following our suggested strategy may be of significant use, especially when handling large MLPA data sets, when samples are of different quality, or interpretation of MLPA electropherograms is too complex. It remains, however, important to recognize that automated MLPA data processing may only be successful when a dedicated experimental setup is also considered.

Keywords: MLPA, copy number estimation, ratio calculation, MLPA analysis, gene amplification, aCGH, Coffalyzer

1. Introduction

Over the last decade, advances in cytogenetics and molecular biology have allowed us to determine critical copy number aberrations in the pathogenesis of genetic syndromes and cancer. However, technologies used to study these genomic copy number alterations often require specific lab resources, large amounts of samples and low resolution of the methods used may not allow identifying the causal genes

*Availability: The algorithm is implemented in VBA and runs in the environment of Microsoft Office 2003. The MLPAnalyzer is available at: http://www.mlpa.com/coffalyser. Contact: Coffalyser@mlpa.com and GA.Meijer@vumc.nl. Supplementary information: Supplementary data are available at: http://www.mlpa.com/.

**Corresponding author: Jordy Coffa, MSc, VU University Medical Center, 1007 MB Amsterdam, The Netherlands. Tel.: +31 20 4444852; Fax: +31 20 4442964; E-mail: Coffa@mlpa.com.

located at these aberrant chromosomal regions. Multiplex ligation-dependent probe amplification (MLPA) is a polymerase chain reaction (PCR)-based approach, sufficiently sensitive, reproducible and sequence-specific to allow the relative quantification of 50 different target sequences in a single reaction, requiring only 20 ng of human DNA [1].

In MLPA, each oligo-probe consists of two hemi-probes which hybridize to adjacent sites of the target sequence. Adjacent hybridized hemi-probe oligonucleotides are ligated, permitting subsequent amplification. All ligated probes have identical end sequences, permitting simultaneous PCR amplification using only one primer pair. Due to the different length of every probe in the probe mix, each probe gives rise to an amplification product of a unique size between 130 and 480 bp [1]. These products can be separated and measured using standard capillary fragment electrophoresis. Amplification of these probes is proportional to the amount of the target sequences present in a sample. The measured intensity of fluorescence of each probe is then compared to the intensity of the same probe performed on normal human DNA to determine its relative copy number, which then is presented as a copy number ratio.

The advantages of MLPA puts this technology forward as an alternative to the more costly array-comparative genomic hybridization (arrayCGH) and fluorescent *in situ* hybridization (FISH) techniques used to inspect aberrant chromosomal regions. This has been illustrated in several studies that aimed to detect gene copy number changes [2–7]. However, despite the previously mentioned advantages, management of large numbers of samples and probes, dissimilarities in PCR efficiency among the different probes, and sample to sample variation still pose a challenge to data analysis and results interpretation. While part of these aspects may be tackled during the experimental procedure, others can be undertaken during data analysis. The aim of this study is therefore to develop a robust and standardized MLPA normalization strategy that takes into account structural sample to sample variation and differences in PCR efficiencies, and implement this into a software tool for rapid handling and interpreting of large amounts of MLPA data. Additionally, this software includes a visualization and database tool, providing an easier interpretation and storage of MLPA results, and sample handling protocols to further optimize MLPA results. The software runs in a Microsoft Excel XP, 2003 or 2007 (© 2006 Microsoft Corporation) environment, is easy to implement, have a user-friendly interface, is freely available and usable for any MLPA probe mix.

2. Material and methods

2.1. Materials

We investigated six colorectal cancer cell lines with a panel of 46 MLPA probes, targeted to chromosomal areas related to colon cancer progression [27]. DNA from colorectal cancer cell lines HT29, SW116, RKO, SNU4, HCT116 and COLO320, with arrayCGH data available, was collected at the VU University Medical Center (Amsterdam, The Netherlands).

Commercially available human genomic DNA was used as reference (Promega Corporation, Madison, Wisconsin, USA).

2.2. DNA isolation

Genomic DNA from cell lines was extracted using the Puregene DNA isolation kit (Biozym, Landgraaf, The Netherlands) according to the manufacturers' recommendations.

2.3. MLPA

All MLPA reactions were performed according to the standard MLPA reaction protocol [1]. All MLPA reactions were performed in triplicate. Each MLPA experiment included five samples from normal human DNA which were spread through the sample plate for normalization reasons. We generated a dedicated colon cancer MLPA mix, which targeted genes located on gained chromosomal arms previously described to be involved in colorectal cancer [27,28]. The MLPA probe mix included eleven probes targeting genes located on chromosome 8, 12 probes targeting genes on chromosome 13 and 16 probes targeting genes on chromosome 20. Eight reference probes for normalization purposes were added targeting chromosomal arms 2p, 4q, 12p and 16p. To measure DNA input, four concentration control fragments ($CCF_{1–4}$) and a ligation- dependent fragment (CCF_5) were also added to the MLPA mix. Separation, detection and quantification of the MLPA probe products were performed on a CEQ8000 capillary sequence system (Beckman Coulter, Fullerton, USA), using 1 μl of MLPA product mixed with 0.3 μl of size standard (08095 CEQ™ DNA Size Standard Kit-600) and 32 μl of formamide.

2.4. Data analysis, the algorithm

2.4.1. Notation

$X_{L,i}$: Measurement of length of a peak for sample i.
$X_{S,i}$: Measurement of intensity of a peak for sample i.
PB_j^{Min}: Minimal set (bin) length for probe j.
PB_j^{Max}: Maximal set (bin) length for probe j.
$SP_{i,j}$: Measurement for sample i and test probe j.
$LSP_{i,j}$: Log converted measurement for sample i and test probe j.
$RP_{h,j}$: Measurement for reference h and test probe j.
$LRP_{h,j}$: Log converted measurement for reference h and test probe j.
$SC_{i,z}$: Measurement for sample i and reference probe z.
$LSC_{i,z}$: Log converted measurement for sample i and reference probe z.
$RC_{h,z}$: Measurement for reference h and reference probe z.
$LRC_{h,z}$: Log converted measurement for reference h and reference probe z.
$\overrightarrow{LSC_i} = (LSC_{i,1}, \ldots, LSC_{i,h})$.
$CCF_{i;r,m}$: Measurement for sample i or reference h and concentration control fragment m.
$DQ_{i,j}$ = Dosage quotient for test probe j for sample i.
m: Number of control probes.
w: Number of all probes.
n: Number of sample runs.
k: Number of reference runs.
q: Total number of probes.

2.4.2. Size calling and data filtering

Peak signals were related to a probe when the length of a peak insert 'was' between ± 2 bp of a set actual probe length (bin set) and when the intensity of the peak was larger than 3% of the sum of all peak signals in a sample run:

$$\left(X_{L,i} > PB_j^{\min} \;\&\; X_{L,i} < PB_j^{\max}\right) \;\&\; X_{S,i} > \frac{\text{Sum}(X_{S,1}, \ldots, X_{S,j})}{100}.$$

2.4.3. Signal-to-noise determination

To determine whether to use the probe peak height or peak area as the basis for normalization, the signal-to-noise ratio (SNR) was determined for both metrics. First for both metrics, the median signal of each reference probe as well as its standard deviation over the samples was computed. Next, the SNR was assessed by dividing the average of these median signals by the standard deviation pooled over the reference probes:

$$SNR = \frac{\sum_{i=1}^{n} [\text{med}(\overrightarrow{LSC_i})]/q}{\sqrt{\sum_{i=1}^{n} \sigma^2(\overrightarrow{LSC_i})/q}}.$$

2.4.4. Sample concentration control

The DNA concentration was assumed to be too low when one third of the median of the signal intensities of the amplification products of MLPA CCFs of 64 ($CCF_{I,1}$), 70 ($CCF_{I,2}$), 76 ($CCF_{I,3}$) and 82 ($CCF_{I,4}$) was greater than the signal intensity of the fifth control band of 92 ($CCF_{I,4}$) bp:

$$\frac{\text{med}(CCF_{i,1}, CCF_{i,2}, CCF_{i,3}, CCF_{i,4})}{3} > CCF_{i,5}.$$

2.4.5. Correcting for probe-wise bias by pre-normalization

Correction factors for probe specific biases were computed for all reference runs by dividing the actual probe signal through its predicted signal ($RP_L(pred) = a + b * X_L$), based on the least of squares method [14]. Our explanatory variable was the probe length (X_L), whereas our response variable was the probe signal (RP). The final probe-wise correction factors were determined by taking a median of the calculated values over all reference runs:

$$\text{Correction factor } P_{L,j} = \text{med}\left(\frac{RP_{j,1}}{RP_{L,j,1}(pred)}, \ldots, \frac{RP_{j,k}}{RP_{L,j,k}(pred)}\right).$$

This correction factor was then applied to all runs to reduce the effect of probe bias due to particular probe properties on the forthcoming regression-type normalization:

$$RP'_{i,j} = \frac{RP_{i,j}}{P_{L,j}},$$

$$SP'_{i,j} = \frac{SP_{i,j}}{P_{L,j}}.$$

2.4.6. Correcting for tailing effects

For every run, prior to normalization, the amount of slope of signals was approached by a function where

the log two transformed pre-normalized signals of all probes (or only the reference probes) were regressed linearly on the probe lengths:

$$LRP'_{i,j}(pred) = a + b * X_{L,j},$$
$$LSP'_{i,j}(pred) = a + b * X_{L,j}.$$

Slope (b) and intercept (a) were determined by using the least squares method after taking out large outliers, using a Monte-Carlo like simulation, adapted to search for expected probe deviations (gains/losses). The probe length-related normalization value was obtained by calculating the distance of each signal to the determined regression line:

$$LRP^*_{i,j} = LRP'_{i,j} - LRP'_{i,j}(pred),$$
$$LSP^*_{i,j} = LSP'_{i,j} - LSP'_{i,j}(pred).$$

2.4.7. Determining the reference

Reference signals values for each probe were determined by calculating the average signal over all reference runs, after correction for probe bias and slope.

$$\overrightarrow{LRP} = \frac{1}{k}\sum_{h=1}^{k} LRP^*_{h,j}.$$

2.4.8. MLPA normalization

We used every MLPA probe set as a reference probe for normalization to produce an independent ratio ($DQ_{j,z}$). The cell line COLO320 was furthermore normalized using all probes for normalization to demonstrate the importance of the chosen normalization factor on the MLPA results:

$$DQ_{i,j,z} = [LSP^*_{i,j} - LSC^*_{i,z}] - \left[\overline{LRP}^*_j - \overline{LRC}^*_z\right].$$

The median of all produced ratios was taken as the final probe ratio. The $DQ_{i,j}$ for a test probe in a sample is therefore:

$$DQ_{i,j} = \text{med}(DQ_{i,j,1}, \ldots, DQ_{i,j,m}).$$

2.4.9. Ratio confidence interval

The quality of the normalization constant (set reference probes) was assessed by calculating the median of absolute deviations (MAD) of each independent reference probe DQs to the final median ratio:

$$MAD_{i,j} = \text{med}^m_{z=1}\left(|DQ_{i,j,z} - DQ_{i,j}|\right).$$

Next, the intrinsic variation of each probe was computed by normalizing each separate reference run back to the defined reference signals. The standard deviation of the mean of each probe over the reference runs was estimated by:

$$\sigma RP_j = \frac{1}{\sqrt{k}}\sigma(LRP^*_{1,j}, \ldots, LRP^*_{k,j}).$$

The final standard deviation of each calculated ratio per MLPA probe in a sample was determined by the sum of the MAD value ($MAD_{i,j}$) and the intrinsic probe variation (σRP_j). MAD values were first converted to standard deviations by multiplying with 1.4826 [20] and divided by the square root of the number of reference probes. Both standard deviations were then pooled by squaring each standard deviation and taking the square root of the sum of both:

$$\sigma_{i,j} = \sqrt{(1.4826 * MAD_{i,j}/\sqrt{z})^2 + \sigma RP_j^2}.$$

The average and/or median value over all calculated MAD values in a run can furthermore be calculated and used as an indication for the quality of the normalization for that run.

2.5. MLPA to array CGH correlation

DNA copy number ratios measured by MLPA and BAC array CGH were compared for the cell lines studied. The array used contained approximately 2500 DNA clones evenly spread across the whole genome, with an average resolution of 1.4 Mb. Image acquisition, analysis and data extraction were performed as previously described [19]. The MLPAnalyzer automatically searched for the array CGH clone closest to the map view location (MV35, NCBI) of each MLPA probe. Next, ratio results for each MLPA probe and its corresponding arrayCGH clone over all cell lines were correlated by computing the Pearson moment correlations [14].

3. Results (analysis strategy)

3.1. Experimental design

MLPA is a technique where relative signal changes are being measured. Thus, a sample run alone will not provide the information needed to estimate the copy number changes without a reference run to compare

to. Multiple reference runs are furthermore needed to estimate the reproducibility of each MLPA probe in a distinct experimental setup. Preferably, multiple standardized reference samples should be tested at multiple time points during the course of the experiments and randomly spread through the MLPA sample plate to cover any technical experimental variation. We performed 5 MLPA reactions on reference samples in each MLPA experiment. Test samples were performed in triplo and the average of the three independently analyzed results provided our final results.

3.2. MLPAnalyzer features and implementation

MLPAnalyzer is written in Visual Basic for Applications (VBA), an implementation of Visual Basic which is built into all Microsoft Office applications. VBA is closely related to Visual Basic, but runs from within the host application (Microsoft Excel XP, 2003 or 2007 (© 2006 Microsoft Corporation)), rather than as a standalone application. The data flow of MLPAnalyzer is briefly outlined in Fig. 1. All discussed algorithms are implemented in the MLPAnalyzer and can be accessed through an easy to handle user form.

To increase the MLPA data flow, the MLPAnalyzer directly accepts raw size called data files in *.txt or *.CSV format exported from most common capillary electrophoresis systems. The fragment analysis software provided with the electrophoresis system should thus be used for size calling of the separated MLPA products. Size calling and electropherogram visualization of ABIF data files, from the ABI-310 and ABI-3100 series can also be performed by the MLPAnalyzer, omitting the usage of any other software. After importing, all data will be filtered from background signals and a DNA concentration, ligation check, and a signal count is performed automatically. Following these quality checks, the actual normalization commences which deals with the dissimilarities among different probes and sample to sample variation by performing: signal-to-noise determination, probe bias correction, slope correction, data normalization and ratio confidence determination. Individual sample charts and reports are created afterwards, as well as overall project ratio and statistical results. All results are directly stored on disk as new Excel files, results can be accessed from within the program itself, or by exploring the created results files.

3.3. Managing large sample numbers

Management of large batches of MLPA samples is mainly problematic because MLPA fragment data needs individual visual run inspection and manual filtering of background signals. The MLPAnalyzer automatically filters all fragment data and simultaneously performs all necessary quality checks, significantly increasing the data stream.

3.3.1. Automated data filtering

Software developed for DNA fragment analysis, such as Genescan®, Genotyper®, Genemapper®, Peak Scanner software® (Applied Bio systems, Foster City, CA), CEQ 2000® and 8000® (Beckman Coulter, Fullerton, USA) are typically used for peak-detection, size calling and intensity quantification. All these programs produce data files containing peak length, peak height and peak area of every MLPA probe. These different programs allow for a number of different size calling methods and the use of different size standard markers, which consequently gives the probes a deviation length from the original probe size.

Currently, most MLPA users export measurements of fluorescent units for all peaks into Excel spreadsheets. Probe signals and background signals are then separated manually from each other in Excel. In the MLPAnalyzer, the user can adjust the actual probes sizes and subsequently import and filter the raw data files automatically, continuing only with the MLPA probe specific signals, either the complete peak area or largest peak height signal which are stored for further analysis.

3.3.2. Signal-to-noise determination

In practice MLPA users usually choose to use the peak height as a metric for normalization, which has some practical advantages [9]. To be more methodically sound, the signal-to-noise ratio (SNR) can be calculated for both metrics as MLPAnalyzer does. MLPAnalyzer then continues with the metric having the lowest signal-to-noise ratio providing the most accurate results.

3.3.3. DNA concentration check

All MLPA kits contain concentration control fragments (CCF) which can be recognized by the presence of 4 fragments at regular distances whose lengths always co-vary together. Amplification products of MLPA CCFs of 64, 70, 76 and 82 bp probes will be prominent if the amount of sample DNA is very low. In contrast, the fifth control band of 92 bp is ligation-

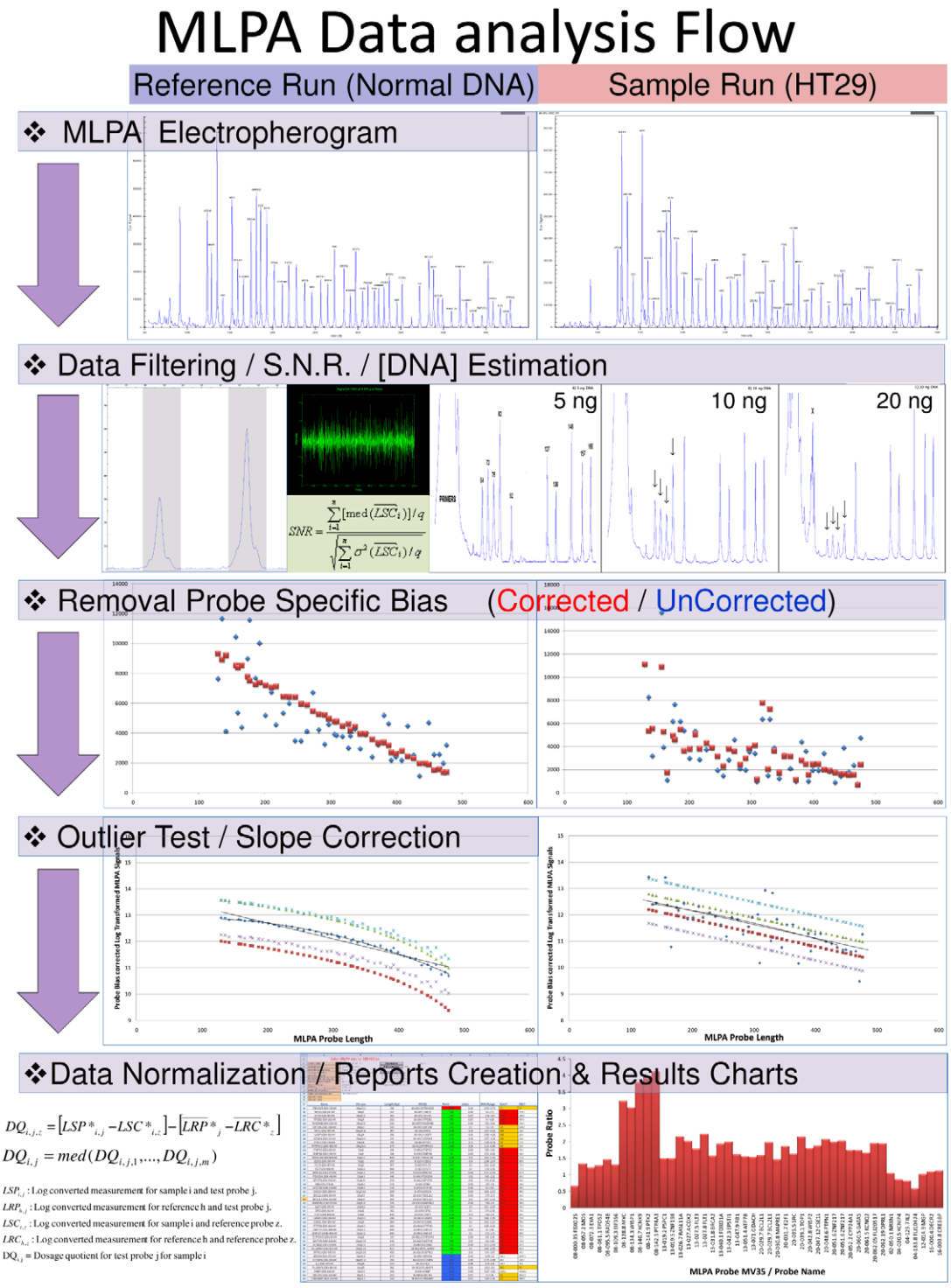

Fig. 1. MLPAnalyzer system data flow. MLPA relative fluorescent signals (y-axis) ordered on probe length (x-axis) in the different stages of processing for a reference (left) and a test sample (right). The complete MLPA process consists of: the MLPA reaction, capillary electrophoresis, data filtering, signal-to-noise determination, quality control, probe bias correction, slope correction, normalization, ratio confidence determination and the creation of plots and reports.

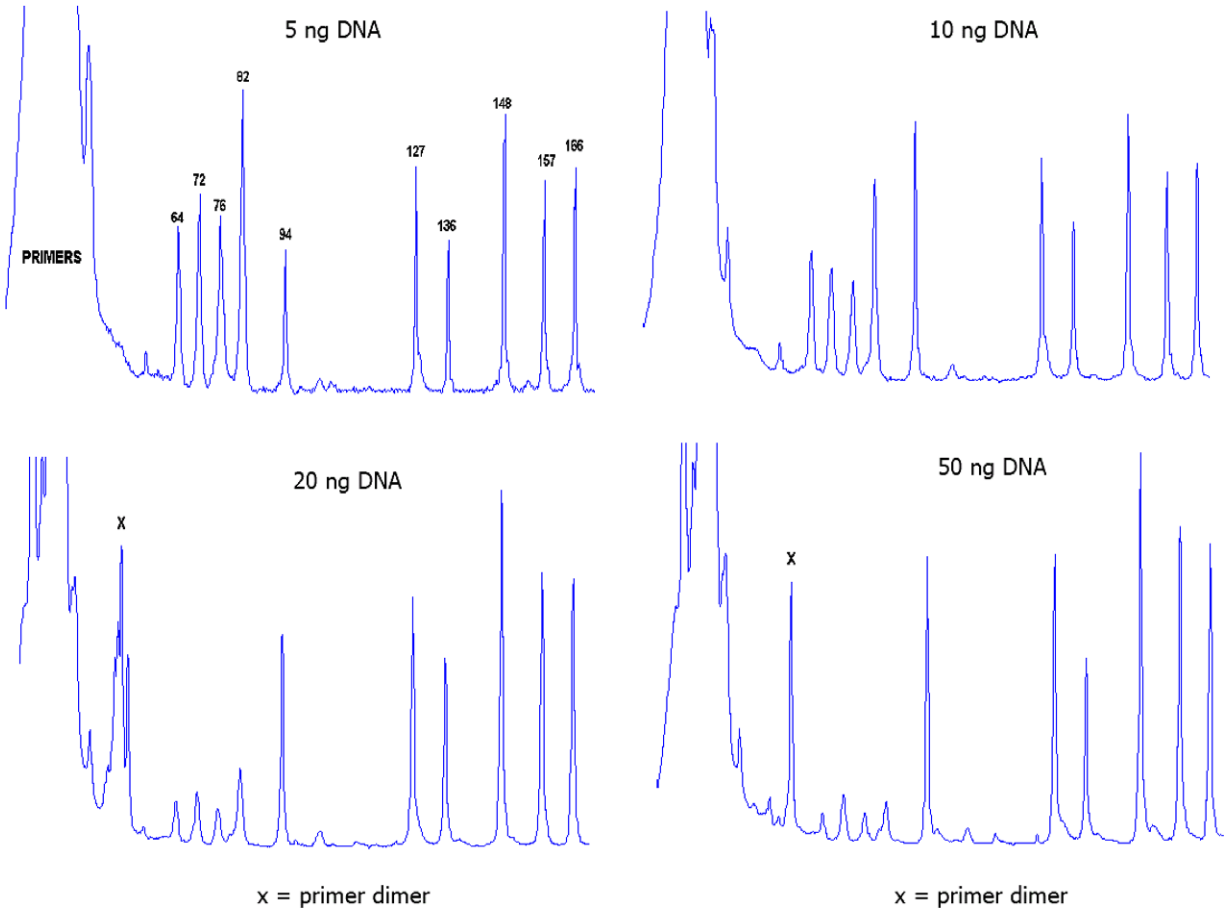

Fig. 2. Analysis of the concentration control probe fragment (64, 72, 76, 82 bp) relatively to the 92 ligation dependent probe indicates if sufficient DNA was available during the MLPA reaction to create reliable results.

dependent and should have a signal similar to most other MLPA amplification products (Fig. 2). MLPA CCFs are currently investigated by examination of the individual electropherograms. The MLPAnalyzer automatically detects and compares the concentration and ligation-dependent control fragment signals, providing a simple DNA concentration check, ensuring analysis of only reliable runs. Users will be notified when the DNA concentration was found to be too low, but may choose to normalize the data anyway if all probe signals are present.

3.4. Dealing with sample to sample variation

Sample to sample variation or more specific, structural differences between reference and sample runs causes normalization problems. Before commencing with the actual data normalization, each reference and sample run needs to be corrected for possible sloping artifacts. The MLPAnalyzer first pre-normalizes all data, dealing with the probe-wise biases, where after a regression analysis determines the amount of signal sloping and corrects for it.

3.4.1. Correcting for probe-wise bias by pre-normalization

Each MLPA probe is multiplied during the amplification reaction with a probe specific efficiency, mainly determined by the sequence of the probe, resulting in a probe specific bias. The extent of this bias will be estimated automatically by the MLPAnalyzer for each probe using the reference runs, assuming these were performed on normal human DNA. Optimally, all MLPA probes produce similar signal intensities, reflecting 2 genomic copies. A linear systematic probe length effect on the signal is expected, probe signal intensities can therefore be regressed linearly on the probe lengths, to determine the predicted signals. Correction factors for these probe specific biases are then

computed per reference run, by dividing the actual probe signal through its predicted signal, based on the least of squares method [14]. The final probe-wise correction factor is then determined by taking a median of the calculated values over all reference runs which are later applied to all runs taking out the probe bias effect for the coming slope normalization.

3.4.2. Correcting for tailing effects

Caused by a decreasing efficiency of amplification of the larger MLPA probes, a systematic probe length effect or size to signal drop can be seen in all MLPA electropherograms. This effect may be enhanced by sample contaminants or evaporation during the hybridization reaction. Signal sloping may further be influenced by injection bias of the capillary system and diffusion of the MLPA products within the capillaries. In addition, the peak heights will be more affected by diffusion of the MLPA products than peak areas, but are influenced less by the presence of shoulder peaks than peak areas. In conclusion, each separated MLPA run may exhibit a run specific signal decrease which needs to be adjusted, before normalization. In practice, MLPA data are rarely corrected for this effect that can induce false deletions of the longer probes.

The MLPAnalyzer allows several different options to correct for sloping effect in different situations. First, for every run the amount of sloping of signal intensities will be approached by a function where the log-transformed pre-normalized signals of the reference probes or all probes are regressed linearly on the probe lengths. Slope and intercept can then be determined using the least squares method (LS) or least of median squares method (LMS). Least of squares method is set as default and minimizes the error deviation which is more accurate when only small error values occur [14], prior to applying the LS method, large outliers are therefore identified and ignored in the subsequent calculation. For the LMS, the estimator must yield the smallest value for the median of squared residuals computed for the entire data set making this method very robust in the presence of outliers, but less precise than the LS. The LMS algorithm used in the MLPAnalyzer was adapted from Rousseeuw [10]. After determining the slope and intercept, signal intensities are predicted for each probe. The probe length-related values can then be obtained by calculating the distance of each log transformed pre-normalized signal to its predicted signal.

3.5. MLPA normalization

Determination of a single ratio per test probe in a sample requires us to have a single reference signal value for each probe. Reference signal values for every probe in the MLPAnalyzer will therefore be determined by calculating the average (or median) of reference signal values in all reference runs, after correction for probe bias and slope.

Classical normalization methods, such as global normalization, have proven to be powerful [18]. However, because of the relative low numbers of targets in an MLPA experiment, these methods are also sensitive to bias. To circumvent this problem MLPAnalyzer makes use of the designated reference probes in the MLPA mixes. These reference probes are targeted to regions which are known to be diploid in both reference and test samples. Each reference probe signal will be used as a normalization constant, thereby determining the ratio (DQ) of each test probe between reference and test sample. The MLPAnalyzer may furthermore use each probe in the MLPA mix as a normalization constant, although it may be best to use both approaches. The median of all produced ratios, estimates the final probe ratio, or ploidy status of the sample test probe sequences in a MLPA mix. The robustness of the normalization thus depends on the number of reference probes, their chromosomal locations, and the origin of the samples that are being used. In general, when analyzing tumor DNA, more reference probes are required, than when germ line DNA from patients with cytogenetic disorders are tested, provided reference probes are carefully selected. It is recommended to have at least 8 reference probes in a probe set, which allows 3 reference probes to be aberrant without compromising the used normalization factor.

3.6. Dealing with MLPA probe variation

MLPA probes all have their own characteristics and therefore yield a level of variation which is unique for each probe. This probe variation may furthermore be influenced by sample contaminations and should therefore be measured within each MLPA experiment on the reference runs. Final probe ratios are furthermore influenced by the normalization factor used, which usually is a combination of different probes. By adding these two factors, the MLPAnalyzer provides a measurement of variation on all calculated probe ratios, which aids in results interpretation.

3.6.1. Ratio confidence interval

The quality of the normalization constant will be assessed by calculating the median of absolute deviations (MAD) of each independent DQ to the final median ratio per probe. The presence of multiple reference runs is used to calculate the intrinsic variation of each probe, by normalizing each separate reference run back to the defined reference signals, either the average or median of all. The final standard deviation of each calculated ratio per MLPA probe in a sample can then be determined by the sum of the MAD value and the intrinsic probe variation. 95% confidence intervals are furthermore computed by multiplying the final standard deviation according to the Student's t distribution.

The average and median value over all calculated MAD values will furthermore be computed per run, which will indicate the quality of the normalization for that run. Discrepancy on estimated DQ by the used reference probes will thus lead to an increase of the average/median MAD value, indicating a poor normalization.

3.7. Results ordering and visualization

Since chromosomal aberrations often span larger regions [27], ordering probe data by Map View locations (NCBI, Map view version 36) results in clustering of probes targeting the same region. Aberrations can be recognized more easily this way and probes targeting the same region may confirm each other's result. MLPA data should always be confirmed by replicate experiments. Results are finally displayed as bar graphs or XY plots, Map view locations can be displayed on the x-axis, and ratio results on the y-axis and the standard deviations are displayed as error-bars to enhance result interpretation (Fig. 4).

4. Results (MLPA to CGH correlation)

To validate MLPAnalyzer, we compared previously obtained array CGH results of cell lines: HT29, SW116, RKO, SNU4, HCT116 and COLO320 to MLPA results of the same cell lines. Both methods for DNA copy number profiling produced comparable ratios (Figs 3–5), confirming the validity of MLPA in combination with MLPAnalyzer. The average Pearson moment correlation coefficient over all probes was 0.42, indicating an average overall positive correlation. Notably, DNA copy number ratios obtained with MLPA were higher in most experiments than the CGH ratios (Fig. 4), as previously has been found by Postma et al. [2].

Most MLPA probes showed a Pearson moment correlation coefficient greater than 0.3 (70%), while some MLPA probes showed little or no correlation. The negative correlations found on the MYC, WISP1, CDX2, FLT3 and SRC are due to the high amplifications found in the COLO320 cell lines by MLPA, which were not found by CGH. Other low correlations, such as for the MLPA probes between 41.1–51.4 MB located on chromosome 13, may be explained by a large difference in target position between the probes and clone, which was for some as large as 8 Mb.

The COLO320 cell line MLPA data was furthermore analyzed twice using two different normalization strategies (Fig. 3); once all signals were normalized to the reference probes (MLPA C) and secondly against all probes (MLPA P). The large difference in results shows the importance of the chosen reference probes for the MLPA normalization. The reference probe normalization method assumes that the probes located at chromosomes 2, 4, 12 and 16 are normal, while in fact all but the probes targeting 4q were gained, as shown by the CGH results. This discrepancy among the reference probes causes in an increase of the probe related MAD values and thus an increase in the probe ratio related standard deviation. The average MAD value over the complete run was furthermore found to be higher than 0.8, indicating that the used reference probes were not optimal for normalization of this run.

The population normalization method uses a median of all probes as a normalization factor, showing the gains of the earlier chosen reference probes at chromosomes 2, 4p, 12 and 16, also confirmed by CGH. When the population normalization method was used, the average MAD value was also found to be higher than 0.8. This was however expected for population normalization, since changes in the used probes (all) are also expected. Final results for this cell line should thus be determined subjectively by evaluation of the ratio results of the different methods to the expected aberrations.

5. Discussion

MLPA analysis of small series of samples can easily be performed by visual inspection of the peak pattern of a patient superimposed over a peak pattern of a reference [3]. This can be adequate for genetic diseases where the MLPA probes target for the exons of a single gene and both samples and reference DNA are

MLPA Probe Name	MV35	SD	PPMC	Cell lines													
				HT29		HCT116		SNU4		SW116		RKO		COLO320			
				aCGH	MLPA C	aCGH	MLPA C	aCGH	MLPA C	aCGH	MLPA C	aCGH	MLPA C	aCGH	MLPA C	MLPA P
FBXO25-D01-310-M	08-000.35	0.07.	0.83	0.4	0.7	1	1.1	1	0.8	0.6	0.8	1	1.0	1.0	0.7	0.9
MOS-D01-247-M	08-057.2	0.08	0.52	0.9	1.4	0.9	1.0	0.9	1.1	1.1	1.1	1.6	1.4	0.9	0.6	1.0
EYA1-D10-193-M	08-072.3	0.03	0.45	0.9	1.3	1	1.4	0.9	1.0	1.3	1.0	2.1	1.4	1.0	0.6	0.9
TPD52-D01-202-M	08-081.1	0.05	0.68	1	1.4	1.3	1.4	0.9	1.0	1.1	1.0	1.4	1.6	1.1	0.5	0.8
RAD54B-D01-292-M	08-095.5	0.12	0.74	0.9	1.5	1.2	1.5	0.9	1.0	1	1.0	1.3	1.9	1.0	0.6	0.9
EIF3S6-D02-388-M	08-109.3	0.11	0.14	1	1.2	1.2	1.5	1	0.8	1.4	0.8	1.1	2.0	0.8	0.7	0.8
MYC-D02-157-M	08-128.8	0.04	-0.3	2.9	3.5	1.3	1.5	1	1.2	1.1	1.2	1.3	1.6	1.0	24.5	40.4
WISP1-D01-130-M	08-134.3	0.05	-0.32	2.9	3.5	1.3	1.5	1	1.3	1.2	1.3	1.7	1.6	83.1	0.6	1.0
KCNK9-D01-331-M	08-140.7	0.07	0.99	3.1	3.7	1.4	1.4	1	1.0	1.2	1.0	1.6	2.0	1.0	0.8	1.1
PTK2-D02-319-M	08-141.9	0.11	0.91	2.9	4.0	1.3	1.6	0.9	0.9	1.3	0.9	1.2	2.2	1.0	1.1	1.4
PTP4A3-D04-458-M	08-142.5	0.08	0.97	2.9	3.9	1.4	1.2	0.9	0.8	1.3	0.8	1.7	2.3	1.0	1.0	1.0
PSPC1-D01-218-M	13-019.2	0.04	0.87	1.7	1.8	1	0.9	0.9	1.1	1	1.1	1.1	0.9	0.9	0.7	1.0
ZNF198-D03-136-M	13-019.5	0.04	0.67	1.7	1.7	1	0.8	0.9	1.2	1	1.2	1.1	0.8	0.9	0.7	1.2
RASL11A-D01-409-M	13-026.7	0.09	0.45	1.2	2.1	1.1	1.0	1	1.0	1.1	1.0	1	1.2	1.0	0.5	0.6
CDX2-D01-307-M	13-027.4	0.08	-0.39	1.2	2.0	1	1.0	1	1.0	1	1.0	1.1	1.0	0.8	25.4	35.4
FLT3-D01-187-M	13-027.5	0.11	-0.35	1.2	2.0	1	0.9	1	1.1	1	1.1	0.8	0.9	0.8	20.5	33.3
FLT1-D11-466-M	13-027.8	0.07	0.73	1.2	1.9	1	1.1	1	1.0	1	1.0	0.7	1.2	0.8	0.6	0.7
BRCA2-D03-175-M	13-031.8	0.03	0.15	1.1	1.7	1	0.9	1.2	1.1	1.3	1.1	1	0.8	1.1	0.4	0.6
FOXO1A-D01-418-M	13-040.1	0.1	0.53	1.2	1.8	1	1.0	0.9	0.8	1.2	0.8	1.1	1.1	0.8	0.7	0.7
EPSTI1-D01-338-M	13-042.3	0.09	0.35	1.1	2.0	1	1.0	0.9	1.2	1	1.2	1.1	1.0	0.8	0.7	0.9
RB1-D11-274-M	13-047.9	0.06	0.38	1.1	1.7	1	0.9	0.9	1.1	1	1.1	1.1	1.0	0.8	0.7	1.0
ATP7B-D04-238-M	13-051.4	0.08	0.67	1.2	1.7	1.1	1.1	0.8	1.1	1	1.1	1	0.9	0.8	0.6	0.9
DACH-D01-265-M	13-071.0	0.09	0.34	1.1	1.6	0.9	0.9	0.8	1.1	1	1.1	1.1	0.8	0.8	0.7	0.9
BCL2L1-D01-160-M	20-029.7	0.04	0.11	1.6	1.7	1	0.9	1.5	1.8	0.8	1.8	0.9	1.3	1.2	1.1	1.7
BCL2L1-D04-382-M	20-029.7	0.13	0.77	1.6	1.9	1	1.0	1.5	1.4	1.1	1.4	1.4	1.7	1.2	1.0	1.3
MAPRE1-D01-178-M	20-030.8	0.04	-0.13	1	1.7	1	0.9	1.4	1.7	0.8	1.7	1.3	1.2	1.4	0.8	1.5
E2F1-D01-391-M	20-031.7	0.05	0.48	1.6	1.8	1	1.0	1.4	1.7	0.8	1.7	1.2	1.7	1.4	1.0	1.2
SRC-D01-142-M	20-035.5	0.04	-0.18	1	1.8	0.9	1.0	1.5	1.9	1.3	1.9	1.4	1.3	1.4	11.1	16.3
TOP1-D01-400-M	20-039.1	0.07	0.48	1.7	2.1	1	1.1	1.7	1.3	1.7	1.3	1.1	1.7	1.0	0.7	1.0
WISP2-D01-298-M	20-042.8	0.1	0.95	1.7	1.8	1.1	1.0	1.6	1.5	1.5	1.5	1.5	1.5	1.0	0.7	0.9
CSE1L-D01-354-M	20-047.12	0.07	0.49	1.4	1.8	1.1	1.0	1.5	1.3	1.6	1.3	1.6	1.4	1.0	1.0	1.3
PTPN1-D01-364-M	20-048.6	0.09	0.64	1.4	1.9	1	1.1	1.5	1.4	1.5	1.4	1.7	1.7	0.9	1.0	1.2
ZNF217-D02-445-M	20-051.6	0.06	0.68	1.6	1.8	1	1.0	1.5	1.3	1.8	1.3	1.6	1.7	1.0	1.0	1.0
ZNF217-D01B-450-M	20-051.6	0.06	0.67	1.6	1.8	1	1.0	1.5	1.6	1.8	1.6	1.1	1.7	1.0	1.0	1.1
CYP24A1-D01-211-M	20-052.2	0.05	0.71	1.6	1.8	1.1	1.0	1.5	1.7	1.8	1.7	1.6	1.2	1.0	0.8	0.8
GATA5-M01-436-M	20-060.5	0.1	0.54	1.6	1.6	1.1	0.7	1.5	1.0	1.2	1.0	1.6	1.5	1.4	1.0	1.1
KCNQ2-D13-229-M	20-061.5	0.07	0.74	1.6	1.8	1	0.7	1.3	1.7	1.2	1.7	1	1.4	1.3	0.8	1.2
FLJ20517-D02-427-M	20-062.05	0.08	0.41	1.6	1.8	1	0.8	1.3	1.9	0.9	1.9	1.1	1.7	1.0	1.1	1.1
OPRL1-D01-458-M	20-062.19	0.12	0.43	1.4	1.7	1	0.8	1.3.	1.5	0.9	1.5	0.9	1.6	1.0	0.9	1.0
NRXN1-D21-472-M	02-050.1	0.08	0.03	1.1	1.0	1	1.1	1.1	0.6	1.3	0.6	0.7	1.1	1.7	1.2	1.4
KCNIP4-D01-256-M	04-020.5	0.07	0.83	0.9	1.0	0.9	1.0	0.9	0.9	1	0.9	0.8	1.0	1.5	0.9	1.3
IL2-D01-373-M	04-123.7	0.12	-0.79	1	0.8	0.9	1.0	1.1	0.9	1	0.9	0.9	1.0	0.9	0.6	0.8
FLJ10474-D01-166-M	04-183.8	0.04	0.22	0.9	0.7	0.9	1.0	0.9	1.1	1	1.1	0.9	0.9	0.8	0.4	0.7
MGP-D01-346-M	12-014.9	0.07	0.1	0.9	1.0	0.9	1.1	1	1.0	1.7	1.0	0.9	1.3	1.9	1.0	1.2
DECR2-D02-283-M	16-000.4	0.09	0.71	1	1.1	0.8	0.4	1.1	1.2	1	1.2	1	1.0	1.2	1.0	1.4
CREBBP-D03-328-M	16-003.8	0.08	0.61	1	1.1	1	1.0	1.1	1.1	1	1.1	0.9	1.1	1.2	1.0	1.5

Fig. 3. Copy number changes for genes located at chromosome 8p, 8q, 13q, 20p, 20q and reference genes (chromosomes 2, 4, 12p and 16p) located throughout the genome as determined by MLPA and CGH of the cell lines: HT29, HCT116, Colo320, SNU4, SW116 and RKO. Copy number increases are depicted in green (>1.2); copy number decreases in red (0.8); and normal copy numbers in blue (0.8–1.2). The displayed map view locations may differ for the CGH results, as the closest available clone results to the MLPA probe locations were taken. MLPA C, refers to the normalization method, reference probes normalization, while MLPA P refers to a normalization method which uses all probes (population). The displayed standard deviations (SD) are the results of the reproducibility test, displayed PPMC values are the correlation results between each MLPA probe and the closest available raw CGH clone based on the 5 performed cell lines.

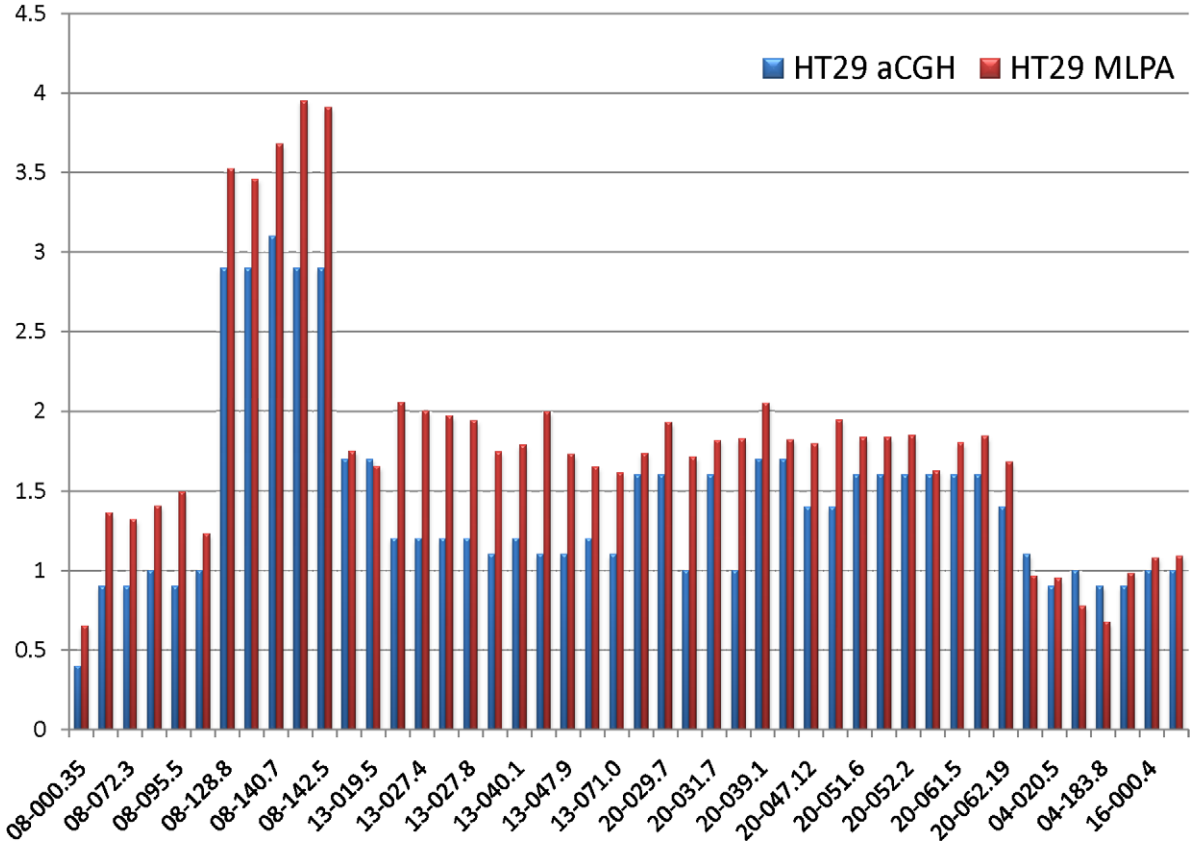

Fig. 4. MLPA result graph created by the MLPAnalyzer of a HT29 cell line (red bars). Map view locations (NCBI) are displayed on the x-axis and ratio results on the y-axis. Standard deviations of each MLPA probe are displayed as error-bars. CGH raw ratio results of the closest available clones are displayed in blue bars.

of comparable quality. More complex diseases, MLPA tests designed for multiple genomic regions, samples of different quality, and in general larger data sets, require more complex data analysis strategies. MLPA has proven to be a very robust technique [1], though final results can be biased: the environment, methods and instruments used, may introduce error. Here we report a detailed MLPA data preprocessing and analysis strategy which takes into account commonly observed forms of variation in MLPA, such as size to signal drop and idiosyncratic probe variation. All algorithms have been implemented in user friendly software providing a quick, free and accurate MLPA data analysis tool.

To obtain optimal MLPA results a good normalization technique alone is not enough. MLPA data analysis strategies must be contemplated on during experimental design. Key to reliable normalization lies in the design of the MLPA mix and in the choice and usage of test and references samples. Sample and reference DNA having a common origin, equal DNA concentrations and similar extraction procedures are more easy to compare, and will provide more accurate results. Furthermore, using multiple reference samples will provide information on reproducibility of the data obtained with MLPA probes in the set used and aid in judging significance of results.

After creating a well-designed experimental setup and performing all MLPA reactions, MLPA users may face a number of capillary sequence systems which can be used for fragment separation, peak-detection, size calling and intensity quantification. Since capillary sequence devices are commonly used for sequencing, fragment analysis of MLPA products usually is performed sub optimally. Optimal run settings are different for each device and should be determined empirically. The quality of each MLPA separation should be inspected by investigating the overall signal intensity, the level of signal sloping, the current during the run and the pattern of the size marker. Capillary fragment separation troubleshooting protocols are therefore also

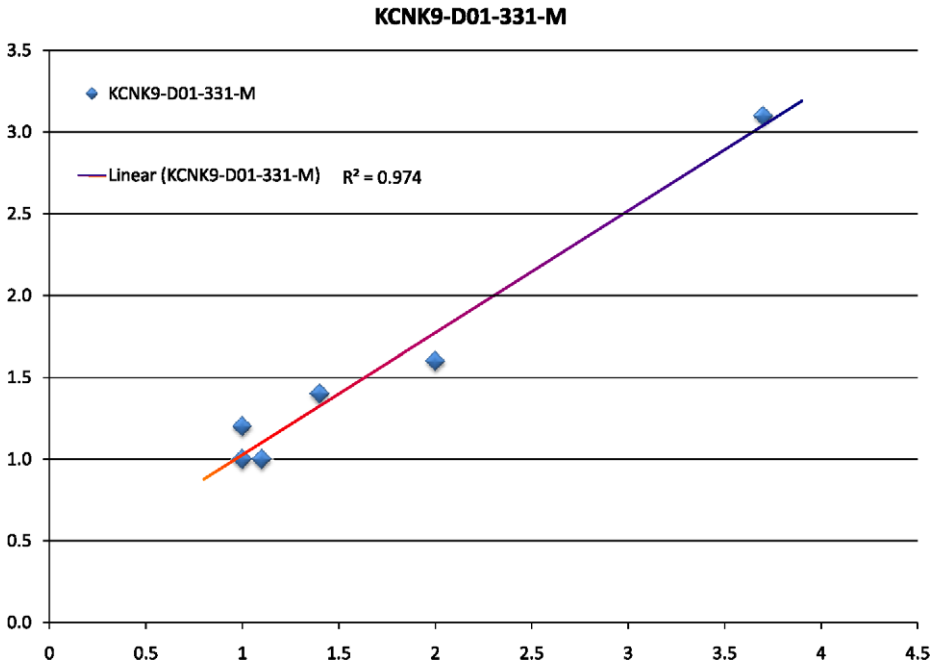

Fig. 5. Scatter plot displaying the MLPA ratio probe results against the closest array-CGH clone ratio results of the highest correlating probe, KCNK9. The Pearson moment correlation for this probe was 0.99.

included in MLPAnalyzer to optimize MLPA product separation and detection.

MLPAnalyzer has three different primary copy number analysis methods, designed for optimal normalization of MLPA data of kits designed for tumor samples, cytogenetic disease and mental retardation (for the latter two data is not shown). These analysis methods differ in the probe signals they use for slope correction and normalization, providing the most robust analysis method for the expected sample types. MLPA data processing is however context dependent and a single setting for general use is not the optimal approach. Normalization methods should be fine tuned using advanced settings depending on the MLPA kit used, sample type and expected pattern of aberrations.

The MLPA and arrayCGH results showed positive correlations for most probes evaluated. Yet, comparison of DNA copy number alterations measured by MLPA and arrayCGH is less straight forward than may be expected. The platforms differ in resolution (MLPA 60-mer probes, arrayCGH 100k BACs), the platforms do not fully align so data needed to be extrapolated, and dynamic range of MLPA is higher than that of arrayCGH. Visual and statistical examination of both results did however show that both methods produce alike results, indicating the success of the suggested analysis strategy.

Earlier versions of MLPAnalyzer have furthermore already been successfully used in larger data sets [2, 22–26]. In conclusion we found that the MLPAnalyzer is a useful tool for MLPA data processing, although users should be aware of the limitations and types of variation often found during the MLPA process. The total MLPA data analysis process can be performed in minutes, including all necessary quality control steps, DNA copy number calculations and visualizations. It features applicability for all available MLPA mixes and new mixes can easily be added. Even though the MLPAnalyzer currently runs in a Microsoft Excel environment, a stand-alone version is currently under construction, which will also support size calling of files coming directly from capillary sequence systems.

Acknowledgements

The authors thank the VUMC department's pathology and mathematics for providing support on developing the MLPAnalyzer ("Coffalyser") application and providing the samples used. This work was furthermore supported by MRC-Holland in Amsterdam and ZONMW.

References

[1] J.P. Schouten, Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification, *Nucl. Acids Res.* **30** (2002), e57.

[2] C. Postma, Chromosomal instability in flat adenomas and carcinomas of the colon, *J. Pathol.* **205** (2005), 514–521.

[3] J.J.P. Gille, Genomic deletions of MSH2 and MLH1 in colorectal cancer families detected by a novel mutation detection approach, *Br. J. Cancer* **87** (2002), 892–897.

[4] T.E. Buffart, DNA copy number changes at 8q11-24 in metastasized colorectal cancer, *Cell. Oncol.* **27** (2005), 57–65.

[5] S.M. Wilting, Increased gene copy numbers at chromosome 20q are frequent in both squamous cell carcinomas and adenocarcinomas of the cervix, *J. Pathol.* **10** (2006), 1002.

[6] K.K.S. Lai, Detecting exon deletions and duplication of the DMD gene using multiplex ligation dependent probe amplification (MLPA), *Clin. Biochem.* **39** (2006), 367–372.

[7] T. Lalic, Deletion and duplication screening in the DMD gene using MLPA, *Eur. J. Human Gen.* **13** (2005), 1231–1234.

[8] E.M. Northrop, Detection of cryptic subtelomeric chromosome abnormalities and identification of anonymous chromatin using quantitative multiplex ligation-dependent probe amplification (MLPA) assay, 2005.

[9] J.W. Ahn, Detection of subtelomere imbalance using MLPA: validation development of an analysis protocol, and application in a diagnostic centre, *BMC Med. Gen.* **8** (2007), 9.

[10] P.J. Rousseeuw and A.M. Leroy, *Robust Regression and Outlier Detection*, John Wiley & Sons, New York, 1987.

[11] H. Edelsbrunner and D.L. Souvaine, Computing median-of-squares regression lines and guided topological sweep, *J. Am. Statist. Assoc.* **85** (1990), 115–119.

[12] C. Postma, M.A. Hermsen and J. Coffa, Chromosomal instability in adenomas and carcinomas of the colon, *J. Pathol.* **205** (2005), 514–521.

[13] J. Carter and J. Li, Genomic expression array profiling of chromosome 20q amplicon in human colon cancer cells, *Ind. J. Human Gen.* **11** (2005), 128–134.

[14] F. Galton, Kinship and correlation, *North Am. Rev.* **150** (1890), 419–431.

[15] P.J. Rousseeuw, Least median-of-squares regression, *J. Am. Statist. Assoc.* **79** (1984), 871–880.

[16] D.M. Hawkin, The feasible set algorithm for least of median squares of squares regression, *Comput. Statist. Data Anal.* **16** (1993), 81–101.

[17] D.M. Mount, A practical approximation algorithm for LMS line estimator, *Comput. Statist. Data Anal.* **51** (2007), 2461–2486.

[18] F.B.L. Hogervorst and P.M. Nederlof, Large genomic deletions and duplications in the BRCA1 gene identified by a novel quantitative method, *Cancer Res.* **63** (2003), 1449–1453.

[19] D. Pinkel, High resolution analysis of DNA copy number variation using compartitive genomic hybridization to microarrays, *Nat. Gen.* **20** (1998), 207–211.

[20] J. Albert, *Bayesian Computation with R*, Springer, New York, 2007.

[21] A.M. Snijders, Assembly of microarrays for genome-wide measurement of DNA copy number, *Nat. Gen.* **29** (2001), 263–264.

[22] V. Gatta, Identification and characterization of different SHOX gene deletions in patients with Leri–Weill dyschondrosteosys by MLPA assay, *J. Human Gen.* **52** (2007), 21–27.

[23] J.S. Huang, Associations between VHL genotype and clinical phenotype in familial von Hippel–Lindau disease, *Eur. J. Clin. Invest.* **37** (2007), 492–500.

[24] S.K. Ma Edmond, Amplification, mutation and loss of heterozygosity of the EGFR gene in metastatic lung cancer, *Int. J. Cancer* **120** (2007), 1828–1831.

[25] V. Martinez-Glez et al., Multiplex ligation-dependent probe amplification (MLPA) screening in meningioma *Cancer Gen. Cytogen.* **173** (2007), 170–172.

[26] N.A. Anya et al., Molecular analysis of primary gastric cancer, corresponding xenografts, and 2 novel gastric carcinoma cell lines reveals novel alterations in gastric carcinogenesis, *Human Pathol.* **38** (2007), 903–913.

[27] M. Hermsen and C. Postma, Colorectal adenoma to carcinoma progression follows multiple pathways of chromosomal instability, *Gastroenterology* **123** (2002), 1109–1119.

[28] B. Carvalho and C. Postma, Multiple oncogenes at the chromosome 20q amplicon contribute to colorectal adenoma to carcinoma progression (submitted).