7

Analysis of MLPA Data Using Novel Software Coffalyser.NET by MRC-Holland

Jordy Coffa[1,2] and Joost van den Berg[3]
[1]*MRC-Holland*
[2]*Free University Amsterdam*
[3]*Berg IT solutions*
The Netherlands

1. Introduction

Genetic knowledge has increased tremendously in the last years, filling gaps and giving answers that were inaccessible before. Medical genetics seeks to understand how genetic variation relates to human health and disease (National Center for Biotechnology Information, 2008). Although genetics plays a larger role in general, the knowledge of the genetic origins of disease has increased our understanding of illnesses caused by abnormalities in the genes or chromosomes, offering the potential to improve the diagnosis and treatment of patients. Normally, every person carries two copies of every gene (with the exception of genes related to sex-linked traits), which cells can translate into a functional protein. The presence of mutant forms of genes (mutations, copy number changes, insertion/deletions and chromosomal alterations) may affect several processes concerning the production of these proteins often resulting in the development of genetic disorders. Genetic disease is either caused by changes in the DNA of somatic cells in the body or it is inherited, e.g. by mutations in the germ cells of the parents.

Genetic testing is "the analysis of, chromosomes (DNA), proteins, and certain metabolites in order to detect heritable disease-related genotypes, mutations, phenotypes, or karyotypes for clinical purposes (Holtzman et al, 2002). In order to make this suitable for routine diagnostics dedicated, affordable, fast, easy-to-interpret and simple-to-use genetic tests are necessary. This allows scientists to easily access information that for instance can be used to: confirm or rule out a suspected genetic condition or help determine a person's chance of developing or passing on a genetic disorder. Several hundred genetic tests are currently in use, and more are being developed (Sequeiros et al, 2008). The Multiplex Ligation-dependent Probe Amplification (MLPA) is a PCR-based technique, which allows the detecting of copy number changes in DNA or RNA. MLPA can quantify up to 50 nucleic acid sequences or genes in one simple reaction, with a resolution down to the single nucleotide level (Schouten et al., 2002) needing only 20 ng of DNA. The MLPA procedure itself needs little hands on work allowing up to 96 samples to be handled simultaneously while results can be obtained within 24 hours. These properties make it a very efficient technique for medium-throughput screening of many different diseases in both a research and diagnostic settings (Ahn et al, 2007).

Over a million of MLPA reactions were performed last year worldwide but researchers are still concerned with the application of tools to facilitate and improve MLPA data analysis on large, complex data sets. MLPA kits contain oligo-nucleotide probes that through a biochemical reaction can produce signals that are proportional to the amount of the target sequences present in a sample. These signals are detected and quantified on a capillary electrophoresis device producing a fragment profile. The signals of an unknown sample need to be compared to a reference in order to assess the copy number. Profile comparison is a matter of professional judgment and expertise. Diverse effects may furthermore systematically bias the probe measurements such as: quality of DNA extraction, PCR efficiency, label incorporation, exposure, scanning, spot detection, etc., making data analysis even more challenging. To make data more intelligible, the detected probe measurements of different samples need to be normalized thereby removing the systematic effects and bringing data of different samples onto a common scale.

Although several normalization methods have been proposed, they frequently fail to take into account the variability of systematic error within and between MLPA experiments. Each MLPA study is different in design, scope, number of replicates and technical considerations. Data normalization is therefore often context dependent and a general method that provides reliable results in all situations is hard to define. The most used normalization strategy therefore remains the use of in-house brew analysis spreadsheets that often cannot provide the reliability required for results with clinical purposes. These sheets furthermore do not provide easy handling of large amounts of data and file retrieval, storage and archival needs to be handled by simple file management systems. We therefore set out to develop software that could tackle all of these problems, and provide users with reliable results that are easy to interpreter.

In this chapter we show the features and integrated analysis methods of our novel MLPA analysis software called Coffalyser.NET. Our software uses an analysis strategy that can adapt to fit the researcher objectives while considering both the biological context and the technical limitations of the overall study. We use statistical parameters appropriate to the situation, and apply the most robust normalization method based on the biology and quality of the data. Most information required for the analysis is extracted directly from the MRC-Holland database, producer of the MLPA technology, needing only little user input about the experimental design to define an optimal analysis strategy. In the next section we review the MLPA technology in more detail and explain the principles of MLPA data normalization. Then in section 3, we describe the main features of our software and their significance. The database behind our software is reviewed in section 4 and section 5 explains the exact workflow of our program reviewing the importance and methodology of each analysis step in detail. In the final section, we summarize our paper and present the future directions of our research.

2. Background

MLPA data is commonly used for sophisticated genomic studies and research to develop clinically validated molecular diagnostic tests, which e.g. can provide individualized information on response to certain types of therapy and the likelihood of disease recurrence. The most common application for MLPA is the detection of small genomic aberrations, often accounting for 10 to 30% of all disease-causing mutations (Redeker et al., 2008). In case of the very long DMD gene –involved in Duchenne Muscular Dystrophy— exon deletions and

duplications even account for 65-70% of all mutations (Janssen et al., 2005). Since MLPA can detect sequences that differ only a single nucleotide, the technique is also widely used for the analysis of complicated diseases such as congenital adrenal hyperplasia and spinal muscular atrophy, where pseudo-genes and gene conversion complicate the analysis (Huang et al., 2007). Methylation-specific MLPA has also proven to be a very useful method for the detection of aberrant methylation patterns in imprinted regions such as can be found with the Prader-Willi/Angelman syndrome and Beckwith-Wiedemann syndrome (Scott et al., 2008). The MS-MLPA method can also be used for the analysis of aberrant methylation of CpG islands in tumour samples using e.g. DNA derived from formalin-fixed, paraffin-embedded tissues.

MLPA kits generally contain about 40-50 oligo-nucleotide probes targeted to mainly the exonic regions of a single or multiple genes. The number of genes that each kit contains is dependent on the purpose of the designed kit. Each oligo-probe consists of two hemi-probes, which after denaturation of the sample DNA hybridize to adjacent sites of the target sequence during an overnight incubation. For each probe oligo-nucleotide in a MLPA kit there are about 600.000.000 copies present during the overnight incubation. An average MLPA reaction contains 60 ng of human DNA sample, which correlates to about 20.000 haploid genomes. This abundance of probes as compared to the sample DNA allows all target sequences in the sample to be covered. After the overnight hybridization adjacent hybridized hemi-probe oligo-nucleotides are then ligated using a ligase enzyme and the ligase cofactor NAD at a slightly lower temperature than the hybridization reaction (54 °C instead of 60 °C). The ligase enzyme used, Ligase-65, is heat-inactivated after the ligation reaction. Afterwards the non-ligated probe oligonucleotides do not have to be removed since the ionic conditions during the ligation reaction resemble those of an ordinary 1x PCR buffer. The PCR reaction can therefore be started directly after the ligation reaction by adding the PCR primers, polymerase and dNTPs. All ligated probes have identical end sequences, permitting simultaneous PCR amplification using only one primer pair. In the PCR reaction, one of the two primers is fluorescently labeled, enabling the detection and quantification of the probe products.

The different length of every probe in the MLPA kit then allows these products to be separated and measured using standard capillary fragment electrophoresis. The unique length of every probe in the probe mix is used to associate the detected signals back to the original probe sequences. These probe product measurements are proportional to the amount of the target sequences present in a sample but cannot simply be translated to copy numbers or methylation percentages. To make the data intelligible, data of a probe originating from an unknown sample needs to be compared with a reference sample. This reference sample is usually performed on a sample that has a normal (diploid) DNA copy number for all target sequences. In case the signal strengths of the probes are compared with those obtained from a reference DNA sample known to have two copies of the chromosome, the signals are expected to be 1.5 times the intensities of the respective probes from the reference if an extra copy is present. If only one copy is present the proportion is expected to be 0.5. If the sample has two copies, the relative probe strengths are expected to be equal. In some circumstances reliable results can be obtained by comparing unknown samples can to reference samples by visual assessment, simply by overlaying two fragment profiles and comparing relative intensities of fragments (figure 1).

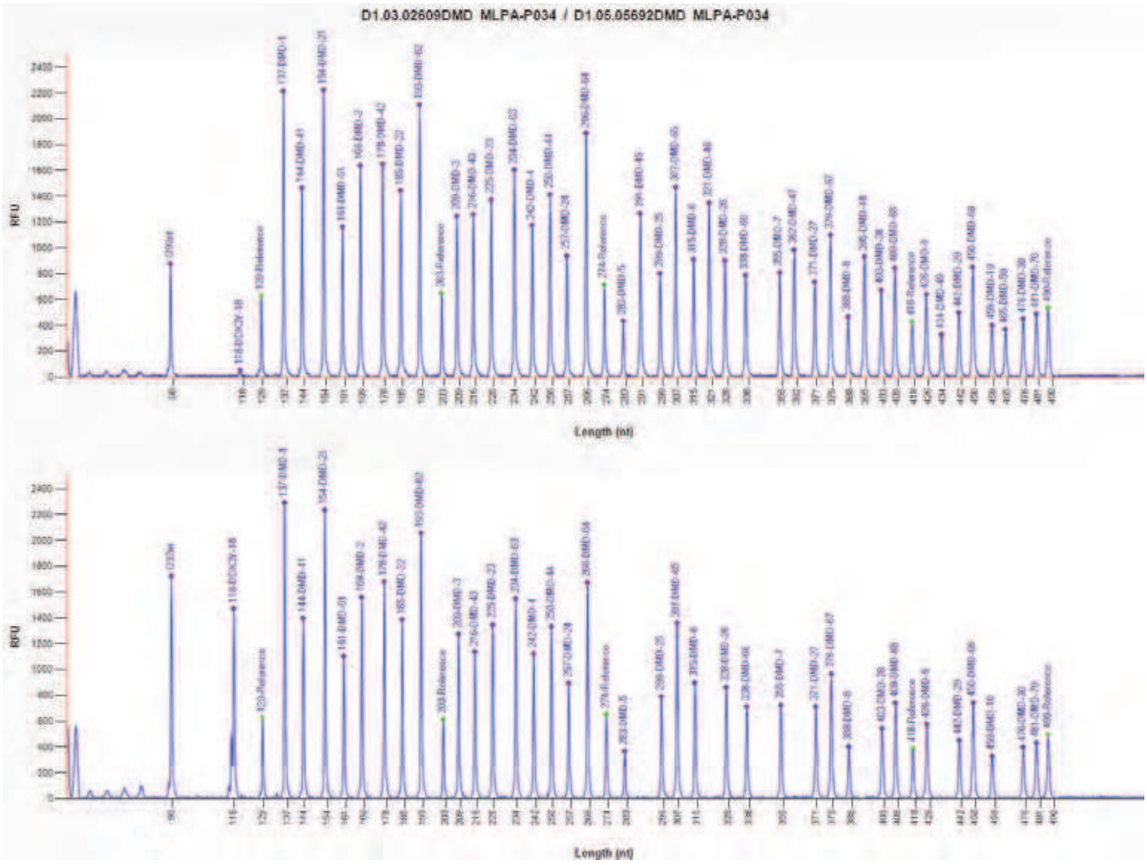


Fig. 1. MLPA fragment profile of a patient sample with Duchenne disease (bottom) and that of a reference sample (top). Duchenne muscular dystrophy is the result of a defect in the DMD gene on chromosome Xp21. The fragment profile shows that the probe signals targeted to exon 45-50 of the DMD gene have a 100% decrease as compared to the reference, which may be the result of a homozygous deletion.

It may however not be feasible to obtain reliable results out of such a visual comparison if:

1. The DNA quality of the samples and references is incomparable.
2. The MLPA kit contains probes targeted to a number of different genes or different chromosomal regions, resulting in complex fragment profiles
3. The data set is very large, making visual assessment very laborious.
4. The DNA was isolated tumor tissue, which often shows DNA profiles with altered reference probes

To make (complex) MLPA data easier understandable unknown and reference samples have to be brought on a common scale. This can be done by normalization, the division of multiple sets of data by a common variable in order to cancel out that variable's effect on the data. In MLPA kits, so called reference probes are usually added, which may be used in multiple ways in order to comprise a common variable. Reference probes are usually are targeted to chromosomal regions that are assumed to remain normal (diploid) in DNA of applicable samples. The results of data normalization are probe ratios, which display the balance of the measured signal intensities between sample and reference. In most MLPA studies, comparing the calculated MLPA probe ratios to a set of arbitrary borders is used to

recognize gains and losses (González, 2008). Probe ratios of below 0.7 or above 1.3 are for instance regarded as indicative of a heterozygous deletion (copy number change from two to one) or duplication (copy number change from two to three), respectively. A delta value of 0.3 is a commonly accepted empirically derived threshold value for genetic dosage quotient analysis (Bunyan et al. 2004). To get more conclusive results probes may be arranged according to chromosomal location as this may reveal more subtle changes such as those observed in mosaic cases.

3. Key features

3.1 Support wide range of file format

Our software is compatible with binary data files produced by all major capillary electrophoresis systems including: ABIF files (*.FSA, *.AB1, *.ABI) produced by Applied Biosystems devices, SCF and RSD files produced by MegaBACE™ systems (Amersham) and SCF and ESD files produced by CEQ systems (Beckman). We can also import fragment lists in text or comma separate format, produced by different fragment analysis software programs such as Genescan (Applied Biosystems), Genemapper (Applied Biosystems), CEQ Fragment analysis software (Beckman) and Genetools. Raw data files are however preferred since they allow more troubleshooting and quality check options as compared to size called fragment lists. Next to this, raw and analyzed data are then stored in a single database and more advanced reports can be created.

3.2 Optimized peak detection / quantification method for MLPA

All applied algorithms in our software are specifically designed to suit MLPA or MLPA-like applications. We designed an algorithm for peak detection and quantification specifically for MLPA peak patterns. Most peak detection algorithms simply identify peaks based on amplitude, ignoring the additional information in the shape of the peaks. In our experience, 'true' peaks have characteristic shapes, and including fluorescence of artifacts may introduce ambiguity into the analysis and interpretation process. Our algorithm has the ability to differentiate most spurious peaks and artifacts from peaks that originate from a probe product. We differentiate a number of different peak artifacts, such as: shoulder peaks, printout spikes, dye artifacts, split peaks, pull-up peaks, stutter peaks and non-template additions. It is often difficult to identify the correct peaks due to appearance of nonspecific peaks in the vicinity of the main allele peak. Our algorithm is therefore optimized to discriminate the different artifacts from the probe signals by usage of minimum and maximum threshold values on the peak -amplitude, -area, -width and -length. Next to this, it may also recognize split and shoulder -peaks by means of shape recognition, making correct identification of probe signals even more reliable. Following peak detection, quantification and size calling, our software allows one or more peaks to be linked to the original MLPA probe target sequence. This pattern matching is greatly simplified as compared to other genotyping programs and additionally provides a powerful technique for identifying and separating signal due to capillary electrophoresis artifacts. Our software may employ three different metrics to reflect the amount of probe fluorescence: peak height, peak area and peak area including its siblings. Peak siblings are the peak artifacts that are created during the amplification of the true MLPA products but have received an alternative length. To determine which metric should be used for data normalization, our program uses an algorithm that compares the signal level of each metric

over the reference probes in all samples, and compares this to the amount of noise over the same signals. The metric that has the largest level signal to noise is then used in the following normalization steps.

3.3 Performances and throughput

After a user logs in, analysis of a complete experiment can be performed in two simple steps: the processing of raw data and the comparison of different samples. Depending on the analysis setup and type of computer, the complete analysis may be completed in less than a minute for 24 samples. Our software can also make use of extra cores running in a computer, multiplying the speed of the analysis almost by two for each core. Because of problems arising from poor sample preparations, presence of PCR artifacts, irregular stutter bands, and incomplete fragment separations, a typical MLPA project requires manual examination of almost all sample data. Our software was designed to eliminate this bottleneck by substantially minimizing the need to review data. By creating a series of quality scores to the different processes users can easily pinpoint the basis for the failed analysis. These scores include quality assessment related to: the sample DNA, MLPA reaction, capillary separation and normalization steps (figure 6). The quality of each step can fall roughly into three categories.

1. High-quality or green. The results of these analysis steps can be accepted without reviewing.
2. Low-quality or red. These steps represent samples with contamination and other failures, which render the resulted data unsuitable to continue with. This data can quickly be rejected without reviewing; recommendations can be reviewed in Coffalyser.NET and used for troubleshooting.
3. Intermediate-quality or yellow. The results of these steps fall between high- and low-quality. The related data and additional recommendations can be reviewed in Coffalyser.NET and used to optimize the obtained results.

When the analysis is finished the results can be visualized in a range of different display and reporting options designed to meet the requirement of modern research and diagnostic facilities. Results effortlessly can be exported to all commonly used medical report formats such as: pdf, xls, txt, csv, jpg, gif, png etc.

3.4 Reliable recognition of aberrant probes

Results interpretation of clinically relevant tests can be one of the most difficult aspects of MLPA analysis and is a matter of professional judgment and expertise. In practice, most users only consider the magnitude of a sample test probe ratio, comparing the ratio against a threshold value. This criterion alone may often not provide the conclusive results required for diagnosing disease. MLPA probes all have their own characteristics and the level of increase or decrease that a probe ratio displays that was targeted to a region that contains a heterozygous gain or loss, may differ for each probe. Interpretation of normalized data may even be more complicated due to shifts in ratios caused by sample-to-sample variation such as: dissimilarities in PCR efficiency and size to signal sloping. Other reasons for fluctuations in probe ratios may be: poor amplification, misinterpretation of an artifact peak/band as a true probe signal, incorrect interpretation of stutter patterns or artifact peaks, contamination, mislabeling or data entry errors (Bonin et al., 2004). To make result interpretation more reliable our software combines effect-size statistics and statistical interference allowing users

to evaluate the magnitude of each probe ratio in combination with it's significance in the population. The significance of each ratio can be estimated by the quality of the performed normalization, which can be assessed two factors: the robustness of the normalization factor and the reproducibility of the sample reactions.

During the analysis our software estimates the reproducibility of each sample type in a performed experiment by calculating the standard deviation of each probe ratio in that sample type population. Since reference samples are assumed to be genetically equal, the effect of sample-to-sample variation on probe ratios of test probes is estimated by the reproducibility of these probes in the reference sample population. These calculations may be more accurate under circumstances where reference samples are randomly distributed across the performed experiment. Our program therefore provides an option to create a specific experimental setup following these criteria, thereby producing a worksheet for the wet analysis and a setup file for capillary electrophoresis devices. DNA sample names can be selected from the database and may be typed as a reference or test sample, positive control or negative control. This setup file replaces the need for filling in the sample names in the capillary electrophoresis run software thereby minimizing data entry errors.

To evaluate the robustness of the normalization factor our algorithm calculates the discrepancies computed between the probe ratios of the reference probes within each sample. Our normalization makes use of each reference probe for normalization of each test probe; thereby producing as many dosage quotients (DQ) as there are references probes. The median of these DQ's will then be used as the definite ratio. The median of absolute deviations between the computed dosage quotients may reflects the introduced mathematical imprecision of the used normalization factor. Next, our software calculates the effect of both types of variation on each test sample probe ratio and determines a 95% confidence range. By comparing each sample's test probe ratio and its 95% confidence range to the available data of each sample type population in the experiment, we can conclude if the found results are significantly different from e.g. the reference sample population or equal to a positive sample population. The algorithm then completes the analysis by evaluating these results in combination with the familiar set of arbitrary borders used to recognize gains and losses. A probe signal in concluded to be aberrant to the reference samples; if a probe signal is significantly different as from that reference sample populations and if the extent of this change meets certain criteria. The results are finally translated into easy to understand bar charts (figure 2) and sample reports allowing users to make a reliable and astute interpretation of the results.

3.5 Advanced data mining options

The database behind our software is designed in SQL and is based on a relational database management system (RDBMS). In short this means that data is stored in the form of tables and the relationship among the data is also stored in the form of tables. Our database setup contains a large number of subtraction levels, not only allowing users to efficiently store and review experimental sample data, but also allowing users to get integrative view on comprehensive data collections as well as supplying an integrated platform for comparative genomics and systems biology. While all data normalization occurs per experiment, experiments can be organized in projects, allowing advanced data-mining options enabling users to retrieve and review data in many different ways. Users can for instance review multiple MLPA sample runs from a single patient in a single report view. Results of multiple MLPA mixese may be clustered together, allowing users gain more confidence on

any found results. The database can further handle an almost unlimited number of specimens for each patient, and each specimen can additionally handle an almost unlimited number of MLPA sample runs. To each specimen additional information can be related such as sample type, tissue type, DNA extraction method, and other clinical relevant data, which can be used for a wide range of data mining operations for discovery purposes. Some of these operations include:

1. Segmenting patients accurately into groups with similar health patterns.
2. Evidence based medicine, where the information extracted from the medical literature and the corresponding medical decisions are key information to leverage the decision made by the professional.

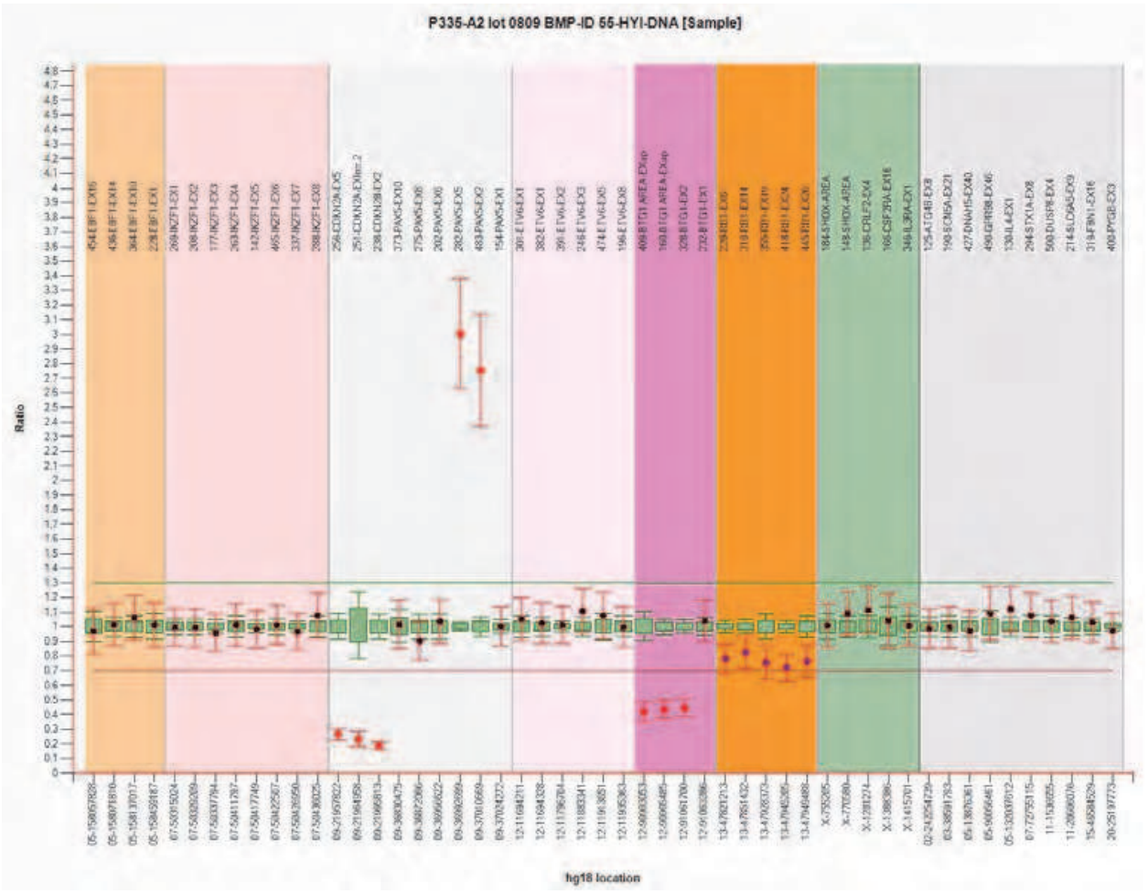


Fig. 2. Ratio chart of the results of a tumor sample analyzed with the P335 MLPA kit. Red dots display the probe ratios and the error bars the 95% confidence ranges. The orange box plots in the background show the 95% confidence range of the used reference samples. Map view locations are displayed on the x-axis and ratio results on the Y-axis. The red and green lines at ratio 0.7 and 1.3 indicate the arbitrary borders for loss and gain respectively. The displayed sample contains several aberrations and extra caution with interpretation is needed due to normal cell contamination.

3. Non-parametric tests (distribution-free) used to compare two or ore independent groups of data.
4. Classification methods that can be used for predicting medical diagnosis.

4. About the database

4.1 Client server database model

Our software uses a SQL client–server database model to store all project/experiment-related data. The client-server model has one main application (server) that deals with one or several slave applications (clients). Clients may communicate to a server over the network, allowing data sharing within and even beyond their institutions. Even though this system may provide great convenience e.g. for people who are working on a single project but are working on different locations, both client and server may also reside in the same system. Having both client and server on the same system has some advances over running both separately: the database is better protected and both client and server will always have the same version number. In case an older client will try to connect to a server that has a newer version number, the client needs to be updated first. A client does not share any of its resources, but requests a server's content or service function. Clients therefore initiate communication sessions with servers that await incoming requests. When a new client is installed on a computer it will implement a discovery protocol in order to search for a server by means of broadcasting. The server application will then answer with its dynamic address that resolves any issues with dynamic IP addresses.

4.2 User access

In addition to serving as a common data archive, the database provides user authentication, robust and scalable data management, and flexible archive capabilities via the utilities provided within Software. Our database model acts in accordance with a simple legal system, linking users to one or multiple organizations. Each user receives a certain role within each organization to which certain right are linked. These rights may for instance include denial of access to certain data but may also be used to deny access to certain parts of the program. These same levels may also be applied on project level. Projects will have project administrators and project members. The initial project creators will also be the project administrators who are responsible for user management of that project.

4.3 Sessions

As soon as a user makes a connection with the server a session will be started with a unique identifier. Subsequent made changes by any user will be held to this identifier, in order to keep track of the made changes. This number is also used to secure experiment data when in use; this ensures no two users try to edit essential data simultaneously (data concurrency). When a user logs in on a certain system, all previously open session of that user will be closed. Every user can thus only be active on a single system. On closing a session, either by logout or by double login all old user locks will disappear.

4.4 Data retrieval and updates

In our software is equipped with MLPA sheet manager software, allowing users to obtain information about commercial MLPA kits and size markers directly from the MRC-Holland database. Next to this, the sheet manager also allows users to create custom MLPA mixes.

The sheet manager software can be used to check if updates to any of the MLPA mixes are available. The sheet manager can further carry out automatic checks for updates at the frequency you choose, or it can be used to make manual checks whenever you wish. It can display scheduled update checks and can work completely in the background if you choose. With just one click, you can check to see if there are new versions of the program, or updated MLPA mix sheets. If updates are available, you can download them quickly and easily. In case some MLPA mixes are already in use, users may choose to hold on to both the older version and updated versions of the mix or replace the older version.

5. Coffalyser.NET workflow

Figure 3 shows the graphical representation of the workflow of our software. After creating an empty solution, users can add new or existing items to the empty solution by using the "add new project" or "add new experiment" command from the client software context menu. By creating projects, users can collect data of different experiments in one collection. Next, data files can then be imported to the database and linked to an experiment. Users then need to define for each used channel or dye stream of each capillary (sample run) what the contents are. Each detectable dye channel can be set as a sample (MLPA kit) or a size marker. Samples may further be typed as: MLPA test sample, MLPA reference sample, MLPA positive control, or MLPA digested sample. The complete analysis of each MLPA experiment can be divided in 2 steps: raw data analysis and comparative analysis. Raw data analysis includes all independent sample processes such as: the recognition and signal determination of peaks in the raw data streams of imported data files, the determination of the sizes of these peaks in nucleotides and the process of linking these peaks to their original probe target sequences. After raw data analysis is finished, users can evaluate a number of quality scores (figure 6), allowing users to easily assess the quality of the produced fragment data for each sample. Users may now reject, accept and adjust sample types before starting the comparative analysis. During the comparative part of the analysis several normalization and regression analysis methods are applied in order to isolate and correct the amount of variation that was introduced over the repeated measured data. Found variation that could not be normalized out of the equation is measured and used to define confidence ranges. The software finally calculates the variation of the probes over samples of the same types, allowing subsequent by classification of unknown samples. After the comparative analysis is finished, users may again evaluate a number of quality scores this time concerning the quality of different properties related to the normalization. The users can finally evaluate the results by means of reporting and visualization methods.

5.1 Import / export of capillary data

Importing data is the process of retrieving data from files to the SQL Server™ (for example, an ABIF file) and inserting it into SQL Server tables. Importing data from an external data source is likely to be the first step you perform after setting up your database. Our software contains several algorithms to decode binary files from the most commonly used capillary electrophoresis devices (see paragraph 2.1). Capillary devices usually store measurements of relative fluorescent units (RFU) and other related data that is collected during fragment separation in computer files encoded in binary form. Binary

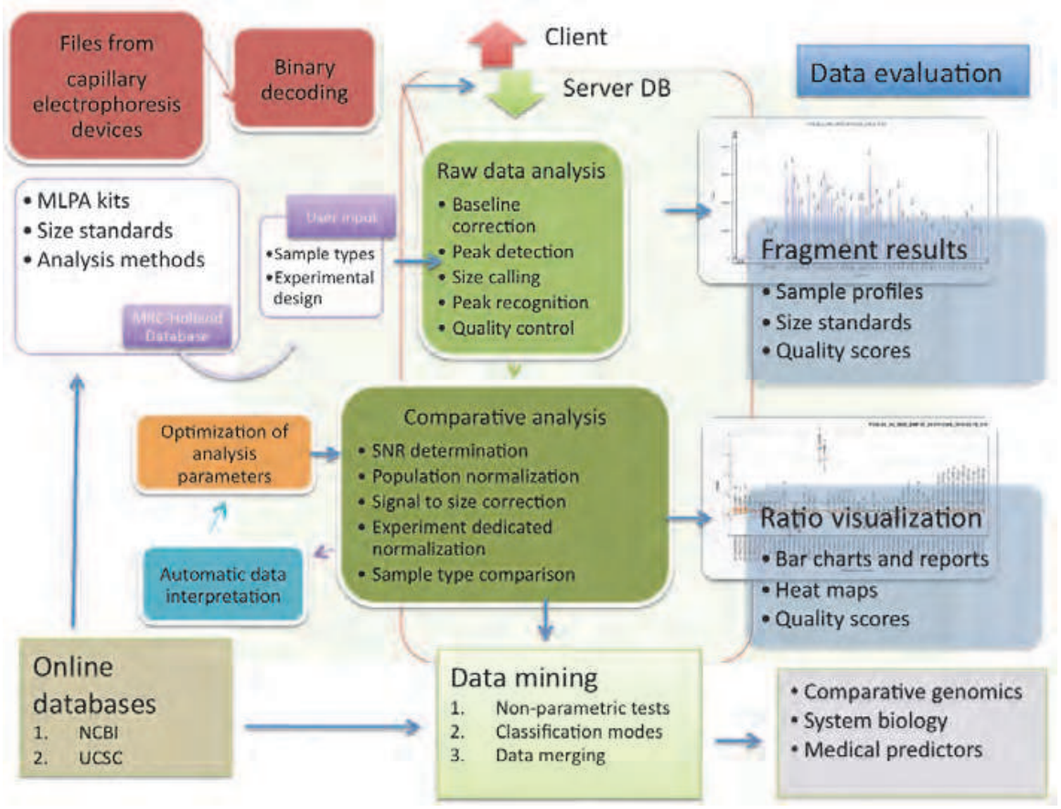


Fig. 3. Schematic overview of the Coffalyser.NET software workflow.

files are made up as a sequence of bytes, which our program decodes back into lists of the different measurements. The most important measurement being the laser induced fluorescence of the covalently bound fluorescent tags on the probe products and the size marker. The frequency at which these measurements occur depends on the type of system. A complete scan will always check all filters (or channels) and result in one data point. Almost all capillary systems are able to detect multicolor dyes permitting the usage of an internal size marker providing a more accurate size call than the usage of external size marker. Multicolor dyes may also permit the analysis of loci with overlapping size ranges, thus allowing multiple MLPA mixes to be run simultaneously in different dye colors.

After data has been imported into your SQL Server database, users can start the analysis. Users can choose to analyze the currently imported data or data that was imported in the past or a combination of both. Due to the relative nature of all MLPA data, it is recommended to analyze data within the confinements of each experiment. There do exist circumstances in which better results may be obtained by applying older collected reference data but one should use these options with caution.

Exporting data is usually a less frequent occurrence. Coffalyser.NET therefore does not have standard tools to export raw capillary data but rather depends on the provided tools and features of the SQL server. The data may be exported to a text file and then be read by third party applications such as Access or Microsoft Excel, which can then be used to view or manipulate the data.

5.2 Raw data analysis

5.2.1 Baseline correction

When performing detection of fluorescence in capillary electrophoresis devices it is some times the case that spectra can be contaminated by fluorescence. Baseline curvature and offset are generally caused by the sample itself and little can be designed in an instrument to avoid these interferences (Nancy T. Kawai, 2000). Non-specific fluorescence or background auto fluorescence should be subtracted from the fluorescence obtained from the probe products to obtain the relative fluorescence as a result of the incorporation of the fluorophore. The baseline wander of the fluorescence signals may cause problems in the detection of peaks and should be removed before starting peak detection. Our software corrects for this baseline by applying two times a median signal filter on the raw signals. First, the signals of the first 200 data points of each dye channel are extracted and its median was calculated. Then for every 200 subsequent data points till the end of the data stream, the same procedure is carried out. These median values are then subtracted from the signal of the original data stream to remove the baseline wander, resulting in baseline 1. This corrected baseline 1 is then fed as input for a filter that calculates the median signal over every 50 subsequent data points. These median values are then subtracted from all the signals that are below 300 RFU (for ABI-devices) on baseline 1, resulting in baseline 2. This second baseline is often necessary due to the relatively short distance between the peaks that derive from probe products with only a few nucleotides difference. By applying this second baseline correction solely on the signals that are in the lower range of detection, even peaks that reside close to each other may reside back to zero-signal, without subtracting too much fluorescence that originates from the probe products. Program administrators can modulate the default baseline correction settings, and also may store different defaults for each used capillary system.

5.2.2 Peak detection

In capillary-based MLPA data analysis, peak detection is an essential step for subsequent analysis. Even though various peak detection algorithms for capillary electrophoresis data exist, most of them are designed for detection of peaks in sequencing profiles. While peak detection and peak size calling are very important processes for sequencing applications, peak quantification is not so important. Due to the relatively nature of the MLPA data, peak quantification is particularly important and has a large influence on the final results. Our peak detection algorithm exists of two separate steps; the first step exists of peak detection by comparison of the intensities of fluorescent units to set arbitrary thresholds and shape recognition, the second step exist of filtering of the generated peak list by relative comparison. Program administrators can modulate the peak detection algorithm thresholds, which make use of the following criteria:

1. Detection/Intensity threshold:
   This threshold is used to filter out small peaks in flat regions. The minimal and maximal peak amplitudes are arbitrary units and default values are provided for each different capillary system.
2. Peak area ratio percentage:
   Peak area is computed as the area under the curve within the distance of a peak candidate. Peak area ratio percentage is computed as the peak area divided by the total

amount of fluorescence times one hundred. The peak area ratio percentage of a peak must be larger than the minimum threshold and lower than the maximum set threshold.

3. Model-based criterion:
   The application of this criterion can consists of 3-4 steps:
   - Locate the start point for each peak: a candidate peak is recognized as soon as the signal increases above zero fluorescence.
   - Check if the candidate peak meets minimal requirements: the peak signal intensity is first expected to increase, if the top of the peak is reached and the candidate peak meets the set thresholds for peak intensity and peak area ratio percentage, then the peak is recognized as a true peak.
   - Discarding peak candidates: if the median signal of the previous 20 data points is smaller then the current peak intensity or if the current peak intensity returns to zero.
   - Detect the peak end: the signal is usually expected to drop back to zero designating the peak end. In some cases the signal does not return to zero, a peak end will therefore also be designated if the signal drops at least below half the intensity of the peak top and if the median signal of the 14 last data points is lower than the current signal.
4. Median signal peak filter:
   The median peak signal is calculated by the percentage of intensity of each peak as opposed to the median peak signal intensity of all detected peaks. Since the minimum and maximum thresholds are dependent on detected peaks, this filter will be applied after an initial peak detection procedure based on the criteria point 1-3.
5. Peak width filter:
   After peak end points have been identified, the peak width is computed as the difference of right end point and left end point. The peak width should be within a given range. This filter is also applied after an initial peak detection procedure.
6. Peak pattern recognition:
   This method is only applied for the size marker channel, and involves the calculation of the correlation between the data point of the peak top of the detected peak list (based on the criteria point 1-5) and the expected lengths of the set size marker. In case the correlation is less than 0.999, the previous thresholds will be automatically adapted and peak detected will be restarted. These adaptations mainly include adjustment of minimal and maximal threshold values.

5.2.3 Peak size calling

Size calling is a method that compares the detected peaks of a MLPA sample channels against a selected size standard. Lengths of unknown (probe) peaks can then be predicted using a regression curve between the data points and the expected fragment lengths of the used size standard, resulting in a fragment profile (figure 4). Coffalyser.NET allows the use of 2 different size-calling algorithms:

1. Local least squares method
2. 1st, 2nd or 3rd order least squares

The local least squares method is the default size calling method for our software. It determines the sizes of fragments (nucleotides) by using the local linear relationship

between fragment length and mobility (data points). Local linearity is a property of functions that have graph that appear smooth, but they need not to be smooth in a mathematical sense. The local linear least squares method makes use of a function that is only once differentiable at a point where it is locally linear. Different from the other methods, this function is not differentiable, because the slope of the tangent line is undefined. To solve the local linear function our algorithm first calculates the intercept and coefficient for each size marker point of the curve by use of a moving predictor. A local linear size of 3 points provides three predictions for each point along its curve that is surrounded by at least 2 points. The average intercept and coefficient are then stored for that point. Points at the beginning and the end of the curve will receive a single prediction, since they do not have any surrounding known values. The coefficient (β) and intercept (α) are calculated by solving the following equations 1 and 2.

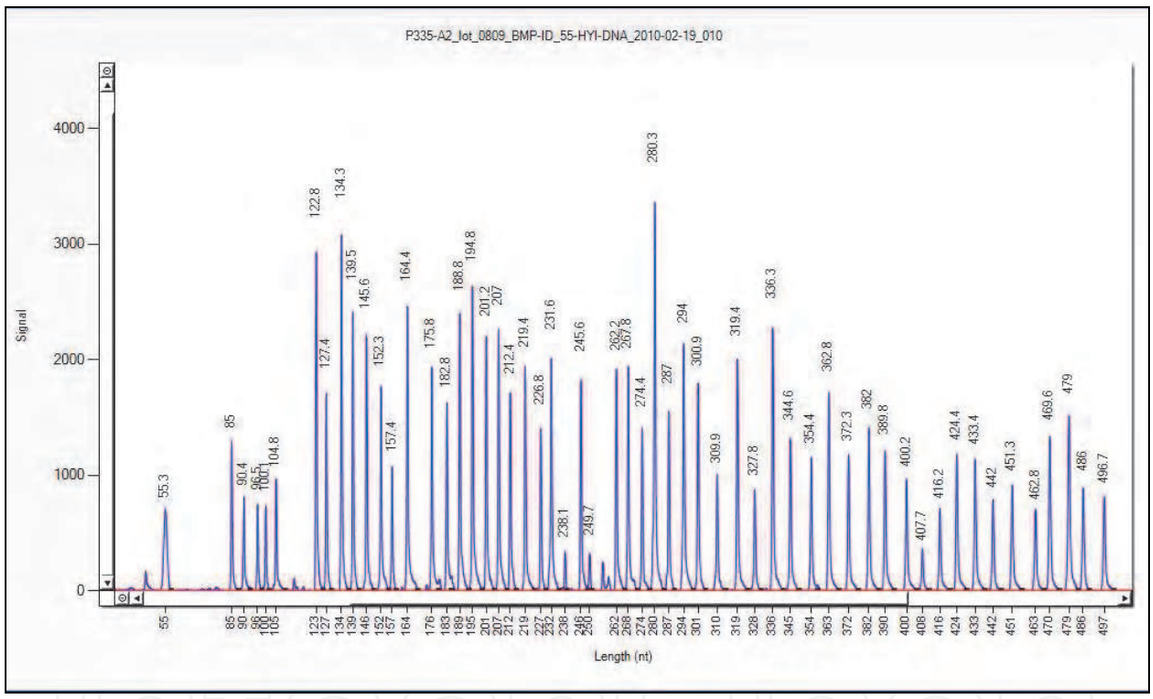


Fig. 4. MLPA fragment length profile displaying the lengths of all detected peaks from a sample. Peak lengths were determined by comparison of the data against a GS500-ABI size marker and determination of the length using the local least squares method.

$$\beta = \left(\frac{\sum X_i Y_i - \frac{1}{n} \sum X_i \sum Y_i}{\sum X_i^2 - \frac{1}{n} \left(\sum X_i \right)^2} \right) \tag{1}$$

$$\alpha = \left(\overline{Y} - \left(\beta * \overline{X} \right) \right) \tag{2}$$

E.g. if we use a size marker that has 15 known points and a local linear size of 3 points, the coefficient and intercept of point 5 will be calculated by equation 3 and 4.

$$\beta_5 = \frac{1}{3}\sum \frac{\sum X_{3-5}Y_{3-5} - \frac{1}{3}\sum X_{3-5}\sum Y_{3-5}}{\sum X_{3-5}^2 - \frac{1}{3}\left(\sum X_{3-5}\right)^2}; \frac{\sum X_{4-6}Y_{4-6} - \frac{1}{3}\sum X_{4-6}\sum Y_{4-6}}{\sum X_{4-6}^2 - \frac{1}{3}\left(\sum X_{4-6}\right)^2}; \frac{\sum X_{5-7}Y_{5-7} - \frac{1}{3}\sum X_{5-7}\sum Y_{5-7}}{\sum X_{5-7}^2 - \frac{1}{3}\left(\sum X_{5-7}\right)^2} \quad (3)$$

$$\alpha_5 = \frac{1}{3}\sum \left(\overline{Y} - \left(\beta_{3-5} * \overline{X}\right)\right); \left(\overline{Y} - \left(\beta_{4-6} * \overline{X}\right)\right); \left(\overline{Y} - \left(\beta_{5-7} * \overline{X}\right)\right) \quad (4)$$

To calculate the length of an unknown fragment our algorithm uses the calculated coefficient and intercepts calculated over the surrounded size marker peaks above and one below its peak. Each unknown point will be predicted twice where after the average value will be stored for that peak. If we wish to predict the length (Y) of an unknown fragment (X) of which the data point of the peak top is in between the data points of known fragments 5 and 6, we need to solve equation 5.

$$Y = \frac{1}{2}\sum \alpha_5 + \beta_5 * X; \alpha_6 + \beta_6 * X \quad (5)$$

5.2.4 Peak identification

Once all peaks have been size called, the profiles must be aligned to compare the fluorescence of the different targets across samples, an operation that is perhaps the single most difficult task in raw data analysis. Peaks corresponding to similar lengths of nucleotides may still be reported with slight differences or drifts due to secondary structures or bound dye compounds. These shifts in length make a direct numerical alignment based on the original probe lengths all but impossible. Our software uses an algorithm that automatically considers what the same peaks are between different samples, allowing easy peak to probe linkage. This procedure follows a window-based peak binning approach, whereby all peaks within a given window across different samples are considered to be the same peak (figure 5). Our software algorithm follows four steps: reference profile analysis, applying and prediction of new probe lengths, reiteration of profile analysis and data filtering of all samples.

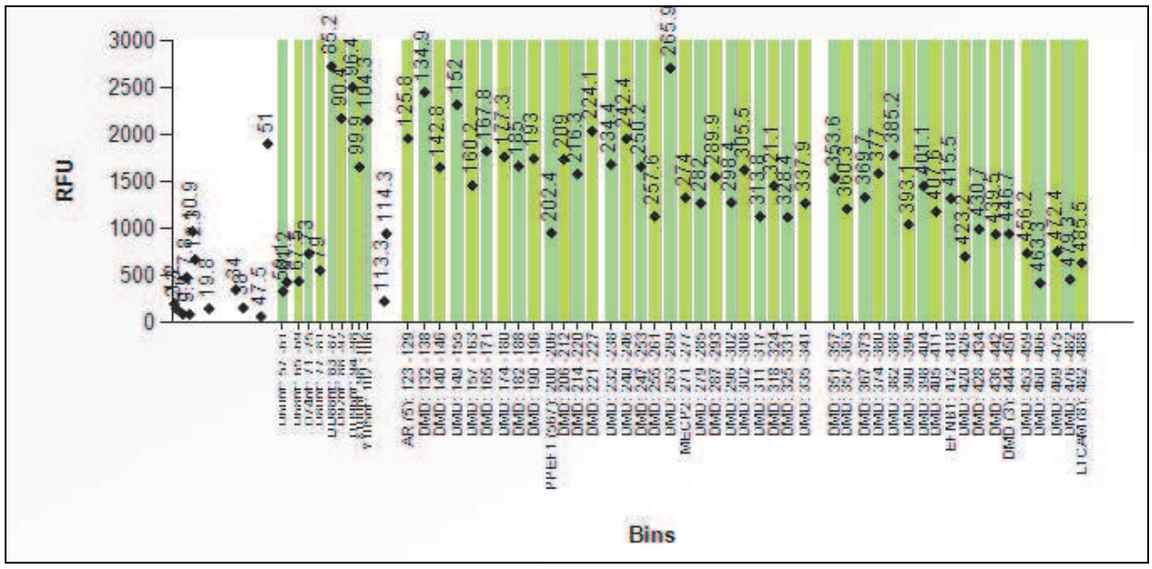


Fig. 5. Visualization of the collection of bins for a MLPA mix (x-axis) and the signal intensities in relative fluorescent units for detected peaks of a sample (y-axis).

The crucial task in data binning is to create a common probe length reference vector (or bin). In the first step our algorithm applies a bin set that searches for all peaks with a length closely resembling that of the design length of that probe. Next, the largest peak in each temporary bin is assumed to be the real peak descending from the related probe product. To create a stable bin, we calculate the average length over all real peaks of all used reference samples. If no reference samples exist, the median length over all collected real peak from all samples will be used. Since some probes may have a large difference between their original and detected length the previously created results may often not suffice. We therefore check if the length that we have related to each probe is applicable in our sample set. We do this by calculating how much variation exists over collected peaks length in each of the previous bins. If the variation was too large (standard deviation > 0.2) or no peak at all was found in any of the bins, the expected peak length for that probe will be estimated by prediction. The expected probe peak lengths may be predicted by using a second-order polynomial regression on using the available data of the probes for which reproducible data was found. Even though a full collection of bins is now available, the lengths of the probe products that were predicted may not be very accurate. The set of bins for each probe in the selected MLPA mix will therefore be improved by iteration of the previous steps. The lengths provided for the bins are now based on the previously detected or predicted probe product lengths allowing a more accurate detection of the real probe peaks. Probes that were not found are again predicted and a final length reference vector or bin is constructed for each probe. This final bin set can be used directly for data filtering but may also be edited manually in case the automatically created bin set may not suffice.

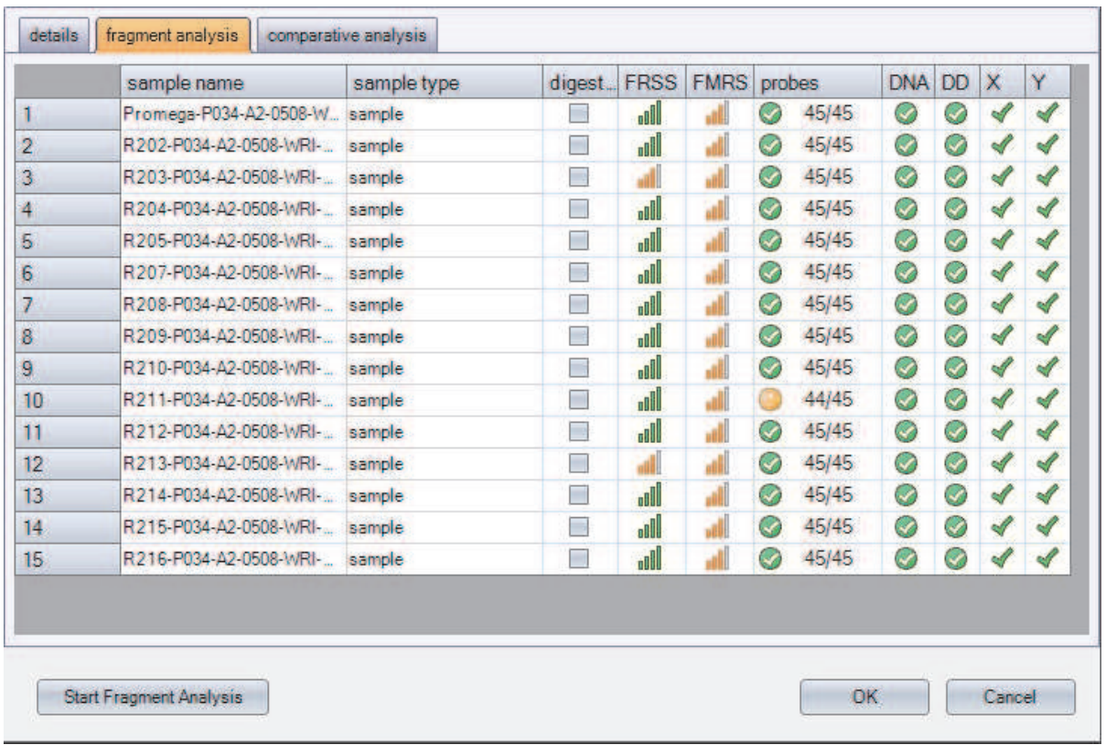
Data filtering is the actual process where the detected fragments of each sample are linked with gene information to a probe target or control fragment. Our algorithm assumes that peaks within each sample that fall within the same provided window or bin and have sufficient fluorescence intensity are the same probe (figure 4). Our algorithm is also able to link more than one peak to a probe within one sample. The amount of fluorescence of each probe product may then be expresses the peak height, peak area of the main peak and the summarized peak area of all peaks in a bin. An algorithm can then be used to compare these metrics and decide which should optimally be used as described at 3.2, alternatively users may set a default metric. The summarized peak area may reflect the amount of fluorescence best if peaks are observed that show multiple tops which all originate from the amplification of the same ligation product. Such peaks may be observed if:

1. Too much input DNA is added the amplification reaction and the polymerase was unable to complete the extension for all amplicons (Clark, J. M. 1988).
2. Peaks were discovered which are one base pair longer than the actual target due to non-template addition.
3. The polymerase was unable to complete the adenine addition on all products that resulted in the presence of shoulder peaks or +A/-A peaks (Applied Biosystems, 1988).

5.2.5 Raw data quality control

In the final step of the raw data analysis the software performs several quality checks and translates this into simple scores (figure 6).

These quality checks are the result of a comparison of sample specific properties such as: baseline height, peak signal intensity, signal to size drop, incorporated percentage of primer etc., to expected standards specific for each capillary system. Several quality checks are furthermore performed using the control fragments providing information about the used



	sample name	sample type	digest..	FRSS	FMRS	probes	DNA	DD	X	Y
1	Promega-P034-A2-0508-W...	sample				45/45				
2	R202-P034-A2-0508-WRI-...	sample				45/45				
3	R203-P034-A2-0508-WRI-...	sample				45/45				
4	R204-P034-A2-0508-WRI-...	sample				45/45				
5	R205-P034-A2-0508-WRI-...	sample				45/45				
6	R207-P034-A2-0508-WRI-...	sample				45/45				
7	R208-P034-A2-0508-WRI-...	sample				45/45				
8	R209-P034-A2-0508-WRI-...	sample				45/45				
9	R210-P034-A2-0508-WRI-...	sample				45/45				
10	R211-P034-A2-0508-WRI-...	sample				44/45				
11	R212-P034-A2-0508-WRI-...	sample				45/45				
12	R213-P034-A2-0508-WRI-...	sample				45/45				
13	R214-P034-A2-0508-WRI-...	sample				45/45				
14	R215-P034-A2-0508-WRI-...	sample				45/45				
15	R216-P034-A2-0508-WRI-...	sample				45/45				

Fig. 6. Coffalyser.NET screenshot. FRSS means fragment run separation score. FMRS means fragment MLPA reaction score. Probes, displays the number of found signals to the number of expected signals. The last columns display the quality of the DNA concentration and denaturation and the presence of the X and Y- fragments.

DNA itself as described before (Coffa, 2008). The quality scores allow users to easily find problems due to: the fragment separation process, MLPA reaction, DNA concentration or DNA denaturation. Users may then reject, accept and adjust sample types before starting the comparative analysis.

5.3 Comparative analysis

During the comparative part of the analysis we aim to isolate the amount of variation that was introduced over the repeated measured data and provide the user with meaningful data by means of reporting and visualization methods. The program is equipped with several normalization strategies in order to allow underlying characteristics of the different types of data sets to be compared. During normalization we bring MLPA data (probe peak signals) of unknown and reference samples to a common scale allowing easier understandable data to be generated. In MLPA, normalization refers to the division of multiple sets of data by a common variable or normalization constant in order to cancel out that variable's effect on the data. In MLPA kits, so called reference probes are usually added, which are targeted to chromosomal regions that are assumed to remain normal (diploid) in the DNA of all used samples.

Our algorithm is able to make use of the reference probes in multiple ways in order to comprise a common variable. In case a MLPA kit does not contain any reference probes, the common variable can be made out of probes selected by the user or the program will make

an auto-selection. After normalization the relative amount fluorescence related to each probe can be expressed in dosage quotients, which is the usual method of interpreting MLPA data (Yau SC, 1996). This dosage quotient or ratio is a measure for the ratio in which the target sequence is present in the sample DNA as compared to the reference DNA, or relative ploidy. To make the normalization more robust our algorithm makes use of every MLPA probe signal, set as a reference probe for normalization to produce an independent ratio ($DQ_{i,\,h,\,j,\,z}$). The median of all produced ratios is then taken as the final probe ratio ($DQ_{i,\,h\,j}$). This allows for the presence of aberrant reference signals without profoundly changing the outcome. If we want to calculate the dosage quotient for test Probe J of unknown Sample I as compared to t reference Sample H, by making use of reference Probes Z (1-n), we need to solve the equation 6.

$$DQ_{i,h,j} = med\left(\frac{\left[S_iP_j \,/\, S_iP_{z=1}\right]}{\left[S_hP_j \,/\, S_hP_{z=1}\right]}, \frac{\left[S_iP_j \,/\, S_iP_{z=2}\right]}{\left[S_hP_j \,/\, S_hP_{z=2}\right]}, \ldots \frac{\left[S_iP_j \,/\, S_iP_{z=n}\right]}{\left[S_hP_j \,/\, S_hP_{z=n}\right]}\right) \tag{6}$$

The data for each test probe of each sample ($DQ_{i,\,h,\,j}$) will be compared to each available reference sample (S_h=n), producing as many dosage quotients as there are reference samples. The final ratio ($DQ_{i,\,j}$) will then estimated by calculating the average over these dosage quotients. In case no reference samples are set, each sample will be used as reference and the median over the ratios be calculated.

5.3.1 Dealing with sample to sample variation

Each MLPA probe is multiplied during the amplification reaction with a probe specific efficiency that is mainly determined by the sequence of the probe, resulting in a probe specific bias. Even though the relative difference of these probes in signal intensity between different samples can be determined by normalization or visual assessment (figure 1), the calculated ratio results may not always be easy to understand by employing arbitrary thresholds only. This is mainly due to sample-to-sample variation or more specific, a difference in the amplification efficiency of probe targets between reference and sample targets. Chemical remnants from the DNA extraction procedure and other treatments sample tissue was subjected to, may allot to impurities that influence the *Taq* DNA polymerase fidelity. Alternatively target DNA sequences may have been modified by external factors, e.g. by aggressive chemical reactants and/or UV irradiation which may result in differences in amplification rate or extensive secondary structures of the template DNA that may prevent access to region of the target DNA by the polymerase enzyme (Elizatbeth van Pelt-Verkuil, 2008). An effect that is commonly seen with MLPA data is a drop of signal intensity that is proportional with the length of the MLPA product fragments (figure 7). This signal to size drop is caused by a decreasing efficiency of amplification of the larger MLPA probes and may be intensified by sample contaminants or evaporation during the hybridization reaction. Signal to size drop may further be influenced by injection bias of the capillary system and diffusion of the MLPA products within the capillaries.

In order to minimize the amount of variation in and between reference and sample data and create a robust normalization strategy our algorithm follows 7 steps. By automatic interpretation of results after each step our algorithm can adjust the parameters used for the next step thereby minimizing the amount of error that may be introduced by the use of

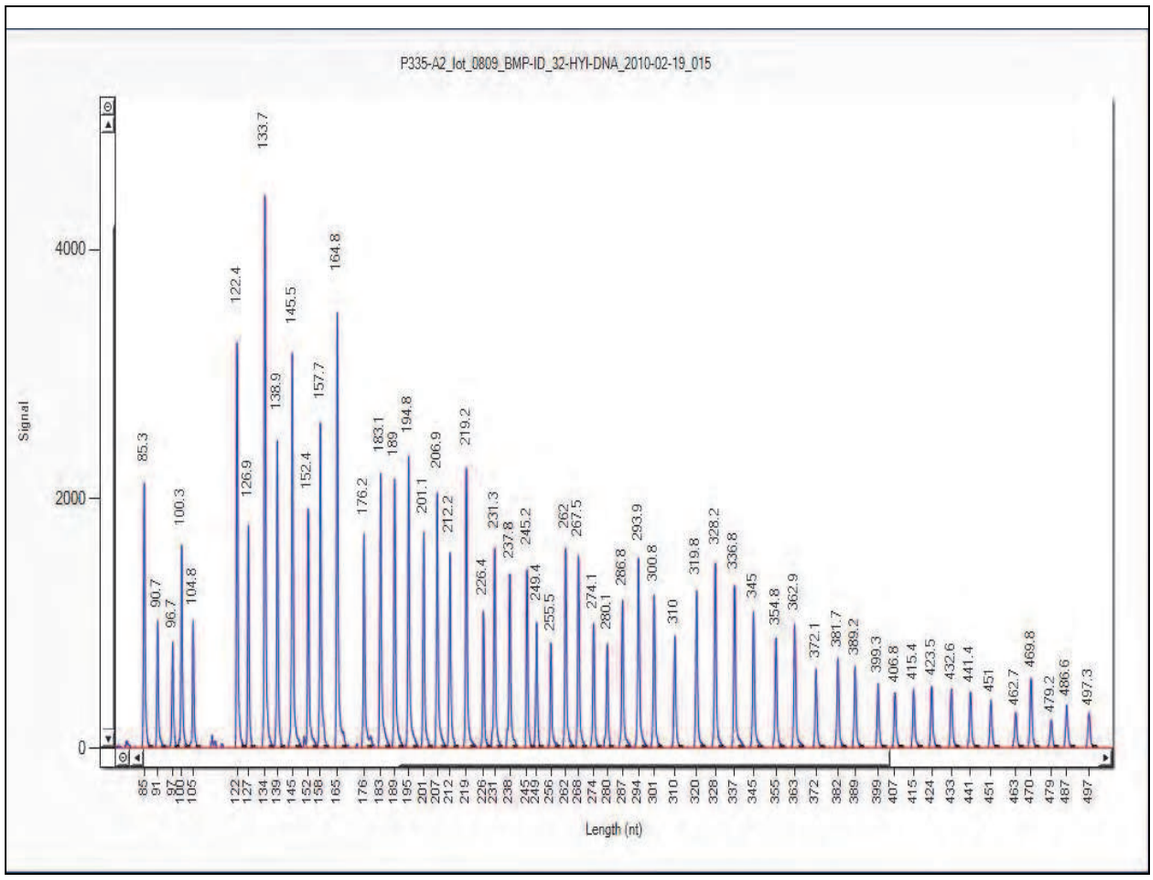


Fig. 7. MLPA fragment profile of a sample with a large drop in signal as size. This effect may have a similar result on the dosage quotients if not corrected for.

aberrant reference signals. The following 7 steps are performed in a single comparative analysis round:

1. Normalization of all data in population mode. Each sample will be applied as a reference sample and each probe will be applied as a reference probe.
2. Determination of significance of the found results by automatic evaluation using effect-size statistics and comparison of samples to the available sample type populations.
3. Measure of the relative amount of signal to size drop. If the relative drop is less than 12% a direct normalization will suffice, any larger drop will automatically be corrected by means of regression analysis (step 4-5).
4. Before correction of the actual amount of signal to size drop, samples are corrected for the MLPA mix specific probe signal bias. This can be done by calculating the extent of this bias in each reference run by regressing the probe signals and probe lengths using a local median least squares method. Correction factors for these probe specific biases are then computed by dividing the actual probe signal through its predicted signal. The final probe-wise correction factor is then determined by taking a median of the calculated values over all reference runs. This correction factor is then applied to all runs to reduce the effect of probe bias due to particular probe properties on the forthcoming regression normalization.

5. Next we calculate the amount of signal to size drop for every sample by using a function where the log-transformed probe bias corrected signals are regressed with the probe lengths using a 2nd order least squares method. Signals from aberrant targets are left out of this function, by applying an outlier detection method that makes use of the results found at step 2 as well as correlation measurements of the predicted line. The signal to size corrected values can then be obtained by calculating the distance of each log transformed pre-normalized signal to its predicted signal.
6. Normalization of signal to size corrected data in the user selected mode and determination of significance of the found results.

Our algorithm then measures the amount variation that could not be resolved in the final normalization to aid in results interpretation and automatic sample classification. To measure the imprecision of the normalization constant, each time a sample is normalized against a reference, the median of absolute deviations ($MAD_{i,h,j}$) is calculated between the final probe ratio ($DQ_{i,h,j}$) and the independent dosage quotients using each reference probe ($DQ_{i,h,j,z}$). The average of all collected $MAD_{i,j}$ values over the samples are then average to estimate the final amount of variation introduced by the imprecision of reference probes. Our algorithm estimates the final $MAD_{i,j}$ for each probe J in sample I and by equation 7.

$$MAD_{i,j} = \frac{1}{N}\sum_{z=1}^{N} med^{m}{}_{z=1}\left(\left|DQ_{i,h,j,z} - DQ_{i,h,j}\right|\right) \tag{7}$$

Since the final probe ratio ($DQ_{i,j}$) for each probe in each sample is estimated by the average over the dosage quotients ($DQ_{i,h,j}$) that were calculated using each reference sample (equation 8), the amount variation that was introduced over the different samples is estimated by calculating the standard deviation over these probe ratios (equation 9).

$$\sigma_{i,j} = \frac{1}{N}\sum_{h=1}^{N}\left(DQ_{i,h,j} - DQ_{i,j}\right)^{2} \tag{8}$$

$$DQ_{i,j} = \frac{1}{N}\sum_{h=1}^{N} \underset{z=1}{\overset{M}{Med}}\left(\frac{\left[S_i P_j / S_i P_z\right]}{\left[S_h P_j / S_h P_z\right]}\right) \tag{9}$$

Our algorithm then estimates the 95% confidence range of each probe ratio ($DQ_{i,j}$) of each sample by following 3 steps:

1. Conversion of the MAD values to standard deviations by multiplying with 1.4826 Albert, J. (2007)
2. Calculation of a single standard deviation for each probe ratio by combining the calculated value of step 1 with the standard deviation calculated over the reference samples by equation 9. This can be done by first converting both standard variations to variations by converting the values to the power of two. Then we sum up the outcome of both and take the square root.
3. Defining the limits of the confidence range by adding and subtracting a number of standard deviations of the final probe ratio ($DQ_{i,j}$) from equation 8.

Discrepancies on estimated dosage quotient by the used reference probes and/or reference samples may lead to an increase of the width of this confidence range, indicating a poor normalization. Since 95% is commonly taken as a threshold indicating virtual certainty (ZAR, J.H., 1984), our algorithm on default uses 1.96 standard deviations (equation 10) to calculate the confidence ranges for probe ratios.

$$DQ_{i,j}^{95\%} = +/-1.96 * \left(\sqrt{\left(\left(1.4826 * \left(MAD_{i,j}\right)\right)^2 + \left(\sigma_{i,j}\right)^2 \right)} \right) \quad (10)$$

5.3.2 Interpretation of the calculated dosage quotients

The previous sections explained how probe ratio are calculated and how our algorithm estimates the amount of introduced variation. In this section, we reflect on what those results mean for empirical comparison of users. To make data interpretation easier our program allows the use advanced visualization methods but also contains an algorithm allowing automatic data interpretation. Our algorithm compares the ratio and standard deviation of a test probe from a single sample to the behavior of that probe within a sub-collection of samples. This allows the program for instance to recognize if a result from an unknown sample is significantly different from the results found in the reference sample population. Alternatively, it may find if a sample is equal to a sample population, for instance a group of positive control samples. To make an estimation of the behavior of a probe ratio within a sample population, we calculate the average value and standard deviation for each probe over samples with the same sample type. In order to calculate the confidence range of probe J in for instance the reference sample population, we need to solve equation 11. N in this case refers to all probe ratio results ($DQ_{i,j}$) from samples that were defined in the normalization setup with the sample type: reference sample (h).

$$DQ_{ref\ j}^{95\%} = \overline{DQ_{i,j}} +/- 1.96 * \left(\frac{1}{N} \sum_{i=h}^{N} \left(DQ_{i,j} - \overline{DQ_{i,j}} \right)^2 \right) \quad (11)$$

Probe result of each sample are then classified in three categories, by comparison to the confidence ranges of available sample types. A probe result of a sample is either significantly different to a sample population, equal to a sample population or the result is ambiguous. To define if a probe result of an unknown sample is significantly different (>>*) to sample population, our algorithm employs 2 criteria:

1. The difference in the magnitude of the probe ratio, as compared to the average of that probe calculated over samples with the same sample type, needs to exceed a delta value of 0.3. In case an unknown sample is compared to the reference sample population, the average ratio for each probe is always approaches 1.
2. The confidence range of the probe of the unknown sample (equation 10) cannot overlap with the confidence range of that probe in a sample population (equation 11).

An unknown sample in classified to be equal (=) to the population of a certain sample type if:

1. The difference in the magnitude of the probe ratio, as compared to the average of that probe calculated over samples with the same sample type, is less than 0.3.
2. The probe ratio of the unknown sample falls within the confidence range of that probe in a sample population (equation 11).

Probe results that are ambiguous, consequently only meet one of the two criteria in order to characterize the result to be different or equal. Ambiguous probe results that do show a difference in the magnitude of the probe ratio, as compared to the average of that probe calculated over samples with the same sample type, but have overlapping 95% confidence ranges will be marked with an asterisk (*). In case the overlap of the confidence ranges is less than 50% the probe results will be marked with a smaller or greater than symbol plus asterisk (<* or >*). Ambiguous probe results that do not show a difference in the magnitude of the probe ratio, but do show a difference in confidence ranges may be displayed with a single and double smaller or greater than symbols, depending on the size of the difference.

COFFALYSER

Coffalyser MLPA sample report: P335-A2 lot 0809 BMP-ID 55-HYI-DNA

Sample type: Sample | Project: p335 | Experiment: p335 bmp

Performed by: Administrator | Machine: ABI | Report date: 10-8-2011 | Run date: 2-8-2011

D [nt]	Gene-Exon	Chr.band	Hg18 loc.	Probe Nr.	Height	Area	T-Area	Pre	Ratio	IRatio	[RPS]	[RPQ]	Stdev	[REF]	[Sam]	[Pos]	Width	E [nt]	d[nt]
59	Q60nt	-	-	-	0	0	0	0.00	0.00	0.00	0.00	0.00	0.00	-	-	-	0	0	0.0
66	Q68nt	-	-	-	0	0	0	0.00	0.00	0.00	0.00	0.00	0.00	-	-	-	0	0	0.0
73	Q74nt	-	-	-	0	0	0	0.00	0.00	0.00	0.00	0.00	0.00	-	-	-	0	0	0.0
79	Q80nt	-	-	-	0	0	0	0.00	0.00	0.00	0.00	0.00	0.00	-	-	-	0	0	0.0
85	DD88nt	-	-	-	1285	12957	12957	0.00	0.00	1.63	0.00	0.00	0.00	-	-	-	50	85.2	-0.1
91	Q92nt	-	-	-	789	8014	8014	0.00	0.00	1.00	0.00	0.00	0.00	-	-	-	49	90.4	-0.1
96	DD96nt	-	-	-	746	7663	7663	0.00	0.00	0.95	0.00	0.00	0.00	-	-	-	41	96.7	0.1
100	X100nt	-	-	-	733	7344	7344	0.00	0.00	0.93	0.00	0.00	0.00	-	-	-	41	100.1	-0.1
105	Y105nt	-	-	-	937	9527	9527	0.00	0.00	1.19	0.00	0.00	0.00	-	-	-	50	104.6	-0.1
Control fragments																			
454	EBF1-16	05q33.3	05-158057828	12527-13577	872	11996	11996	0.83	0.94	0.94	0.02	0.06	0.05	=	=	=	50	451.1	0.1
436	EBF1-14	05q33.3	05-158071810	13868-15386	1088	14794	14794	0.88	0.99	0.99	0.01	0.06	0.05	=	=	=	50	432.8	0.1
364	EBF1-10	05q33.3	05-158137017	14059-15657	1672	21110	21110	1.00	1.03	1.02	0.01	0.06	0.04	=	=	=	50	362.9	0.0
228	EBF1-1	05q33.3	05-158459187	12509-14269	1404	14904	14904	1.10	0.99	1.00	0.01	0.06	0.04	=	=	=	43	226.3	-0.1
Region: 05q					Region analysis:			Average	Median	Stdev	Min	Max							
Chr. band: 05q33.3 Size: 401359 nt ; 401 kb								*0.99*	*1.00*	*0.03*	*0.94*	*1.02*							
269	IKZF1-1	07p12.2	07-050315024	13877-15918	1866	20281	20281	1.04	0.97	0.95	0.01	0.06	0.04	=	=	=	50	267.3	0.0
208	IKZF1-2	07p12.2	07-050329206	14056-15654	2219	23996	23996	1.10	0.96	0.97	0.00	0.07	0.05	=	=	=	50	206.7	0.0
177	IKZF1-3	07p12.2	07-050337794	12536-13586	1879	19141	19141	1.10	0.93	0.93	0.03	0.06	0.05	=	=	=	48	176.1	0.0
263	IKZF1-4	07p12.2	07-050411797	13873-15917	1896	21431	21431	1.06	0.98	0.97	0.00	0.06	0.04	=	=	=	50	261.7	-0.1
142	IKZF1-5	07p12.2	07-050417749	13872-15390	2400	24318	24318	1.18	0.96	0.96	0.02	0.06	0.05	=	=	=	50	138.8	-0.1
465	IKZF1-6	07p12.2	07-050422507	14061-15923	678	9996	9996	0.85	0.98	1.00	0.01	0.06	0.05	=	=	=	50	462.9	0.2
337	IKZF1-7	07p12.2	07-050426950	13869-15387	2217	26815	26815	0.93	0.94	0.95	0.02	0.06	0.05	=	=	=	50	336.6	-0.2
288	IKZF1-8	07p12.2	07-050436025	12549-15921	1537	18376	18376	1.10	1.05	1.04	0.03	0.06	0.06	=	=	=	50	286.5	-0.1
Region: 07p					Region analysis:			Average	Median	Stdev	Min	Max							
Chr. band: 07p12.2 Size: 121001 nt ; 121 kb								*0.97*	*0.96*	*0.03*	*0.93*	*1.04*							
256	CDKN2A-5	09p21.3	09-021957822	10334-10871	225	**2161**	**3289**	**0.18**	**0.17**	**0.17**	0.00	0.01	0.01	<<*	<<*	?	**26**	255.3	0.0
251	CDKN2A-2a	09p21.3	09-021964958	10333-15916	297	2734	2734	**0.24**	**0.22**	**0.23**	0.01	0.01	0.01	<<*	<<*	?	**21**	249.2	0.0
238	CDKN2B-2	09p21.3	09-021995813	10337-15914	336	3298	3298	**0.20**	**0.18**	**0.18**	0.00	0.01	0.01	<<*	<<*	?	**33**	237.7	0.1
373	PAX5-10	09p13.2	09-036830475	12521-13571	1130	14355	14355	0.94	0.99	1.02	0.01	0.06	0.05	=	=	=	50	372.1	0.0
275	PAX5-8	09p13.2	09-036872066	14057-15919	1378	15004	15004	0.93	0.88	0.86	0.02	0.05	0.04	<<	=	=	50	273.9	0.0
202	PAX5-6	09p13.2	09-036956622	14055-15653	2158	23552	23552	1.16	1.01	1.00	0.03	0.06	0.05	=	=	=	50	201.1	0.1
282	PAX5-5	09p13.2	09-036992699	13870-15920	3261	40183	40183	**3.08**	**2.92**	**2.93**	0.05	**0.18**	**0.14**	>>*	>>*	>>*	50	280	0.1
483	PAX5-2	09p13.2	09-037010669	13876-15924	1448	21743	21743	**2.27**	**2.67**	**2.69**	0.07	**0.17**	**0.14**	>>*	>>*	>>*	49	**479.9**	**0.6**
154	PAX5-1	09p13.2	09-037024272	12501-13551	1719	17812	17812	1.18	0.97	0.97	0.03	0.06	0.05	=	=	=	50	152.3	-0.1
Region: 09p					Region analysis:			Average	Median	Stdev	Min	Max							
Chr. band: 09p13.2 - 09p21.3 Size: 15066 kb ; 15 Mb								*1.12*	*0.97*	*0.97*	*0.17*	*2.93*							
301	ETV6-1	12p13.2	12-011694211	14058-15656	1766	21049	21049	1.06	1.02	1.05	0.01	0.06	0.05	=	=	=	47	300.5	-0.1
382	ETV6-1	12p13.2	12-011694328	14060-15658	1367	17485	17495	0.94	0.99	1.01	0.01	0.06	0.04	=	=	=	50	381.6	0.0
391	ETV6-2	12p13.2	12-011796704	13875-15393	1171	15572	15572	0.92	0.98	1.00	0.01	0.06	0.04	=	=	=	50	389.3	0.1
246	ETV6-3	12p13.2	12-011883341	13874-15915	1803	19440	19440	1.18	1.08	1.05	0.01	0.07	0.05	=	=	=	49	245.1	0.0
474	ETV6-5	12p13.2	12-011913651	13871-15389	1291	18192	18192	0.90	1.05	1.08	0.03	0.07	0.06	=	=	=	50	469.9	0.2
196	ETV6-8	12p13.2	12-011935363	14054-15652	2557	27619	27619	1.12	0.97	0.98	0.02	0.06	0.05	=	=	=	50	194.6	-0.1
Region: 12p					Region analysis:			Average	Median	Stdev	Min	Max							
Chr. band: 12p13.2 Size: 241152 nt ; 241 kb								*1.03*	*1.03*	*0.04*	*0.98*	*1.08*							
409	BTG1 AREA-up	12q21.33	12-090903053	14117-15718	357	4900	4900	**0.37**	**0.41**	**0.41**	0.01	0.03	0.02	<<*	<<*	<<*	44	407	0.1
160	BTG1 AREA-up	12q21.33	12-090905485	14116-15717	1071	11192	11192	**0.51**	**0.42**	**0.43**	0.01	0.03	0.02	<<*	<<*	<<*	45	157.5	-0.2
328	BTG1-2	12q21.33	12-091061700	12553-13603	862	10513	10513	**0.43**	**0.43**	**0.43**	0.00	0.03	0.02	<<*	<<*	<<*	50	327.8	-0.3
232	BTG1-1	12q21.33	12-091063286	12542-15913	1946	22529	22529	1.13	1.01	0.98	0.02	0.06	0.05	=	=	=	50	231.3	0.1
Region: 12q					Region analysis:			Average	Median	Stdev	Min	Max							
Chr. band: 12q21.33 Size: 160233 nt ; 160 kb								***0.56***	***0.43***	*0.24*	*0.41*	*0.98*							
220	RB1-6	13q14.2	13-047821213	01782-01346	1900	20475	20475	0.86	0.76	0.76	0.01	0.05	0.04	<<	<<	>>	50	219	-0.1
310	RB1-14	13q14.2	13-047851432	01789-01353	975	12043	12043	0.82	0.80	0.82	0.02	0.05	0.04	<<	=	>>	48	309.8	-0.1
355	RB1-19	13q14.2	13-047928373	01792-15922	1115	13537	13537	0.71	0.73	0.73	0.02	0.05	0.04	<<	<<	=	45	354.6	0.0
418	RB1-24	13q14.2	13-047945205	01797-01360	684	9239	9239	**0.64**	**0.70**	0.70	0.01	0.04	0.03	<<	<<	>>	50	415.5	0.2
445	RB1-26	13q14.2	13-047949488	01799-01362	773	10665	10665	**0.66**	0.74	0.75	0.02	0.04	0.04	<<	<<	=	50	441.7	0.2
Region: 13q					Region analysis:			Average	Median	Stdev	Min	Max							
Chr. band: 13q14.2 Size: 128275 nt ; 128 kb								*0.75*	*0.75*	*0.04*	*0.70*	*0.82*							
184	SHOX-AREA-down	Xp22.33P	X-000755205	06293-06219	1577	**16612**	**17063**	1.12	0.96	0.96	0.02	0.06	0.05	=	=	=	38	183	-0.1
148	SHOX-AREA-down	Xp22.33P	X-000770580	05648-06218	2171	22014	22014	1.30	1.06	1.07	0.03	0.06	0.05	=	=	=	50	145.3	-0.1
136	CRLF2-4	Xp22.33P	X-001281274	13889-15427	3065	31230	31230	**1.34**	1.08	1.09	0.03	0.07	0.05	=	=	=	50	133.6	-0.1
166	CSF2RA-16	Xp22.33P	X-001388386	14052-15650	2381	24386	24386	1.21	1.01	0.99	0.03	0.06	0.05	=	=	=	49	164.7	-0.2
346	IL3RA-1	Xp22.33P	X-001415701	13907-15445	1279	15854	15854	0.96	0.98	0.96	0.01	0.06	0.04	=	=	=	50	344.9	-0.1
Region: Xp					Region analysis:			Average	Median	Stdev	Min	Max							
Chr. band: Xp22.33PAR Size: 660496 nt ; 660 kb								*1.01*	*0.99*	*0.05*	*0.96*	*1.09*							

Fig. 8. Part of a pdf report from a tumor sample analyzed with the P335 MLPA kit. The report shows clear aberrations at 9p21.3, 9p13.2 and 12q21.33. Less clear is the ratio of RB1, which displays a slight decrease in signal as opposed to the reference population, but doesn't surpass the threshold value, due to sample mosaicism.

5.3.3 Reporting and visualization

Automatic data interpretation cannot replace the specialist judgment of a researcher. Knowledge about the expected genetic defect of the target DNA and other sample information may be crucial. To assist the user with data interpretation, our software automatically sorts all probe results based on the last updated map view locations of the probes. Chromosomal aberrations often-span larger regions (M. Hermsen, 2002), which allow probes targeted to that region to cluster together by sorting. Our software can then make a single page PDF reports, containing a summary of all relevant data, probe ratios (figure 8), statistics, quality controls and charts (figure 2 & 4) of a single sample.

	Gene	Positi...	Length	Order	P335-A2_lo...	P335-A2_lo...	P335-A2_lo...	P335-A2_lo...
46	EBF1 - 16	05-158...	454	001	0.96	0.97	1.02	0.99
44	EBF1 - 14	05-158...	436	002	0.98	0.96	1	0.94
36	EBF1 - 10	05-158...	364	003	1.13	1.1	1.08	1.08
24	IKZF1 - 1	07-050...	269	004	1.08	1.08	0.55	1.11
14	IKZF1 - 2	07-050...	208	005	0.99	0.99	0.51	1.03
9	IKZF1 - 3	07-050...	177	006	0.94	0.8	0.48	0.86
23	IKZF1 - 4	07-050...	263	007	1.04	1.02	0.57	1.11
4	IKZF1 - 5	07-050...	142	008	1	0.95	0.49	0.9
47	IKZF1 - 6	07-050...	465	009	1.03	0.88	0.49	0.98
33	IKZF1 - 7	07-050...	337	010	0.91	0.92	0.56	0.92
27	IKZF1 - 8	07-050...	288	011	1.13	1.14	0.54	1.14
22	CDKN2A - 4	09-021...	256	012	0.61	0.88	0.99	0.98
21	CDKN2A - 2a	09-021...	251	013	0.64	1.06	0.94	1.01
19	CDKN2B - 2	09-021...	238	014	0.14	0.9	0.95	1.04
37	PAX5 - 10	09-036...	373	015	1.24	1.11	1	0.96
25	PAX5 - 8	09-036...	275	016	1.05	0.94	1.09	0.88
13	PAX5 - 6	09-036...	202	017	1.01	0.88	1.05	0.87
26	PAX5 - 5	09-036...	282	018	1.11	0.94	1.17	0.95
49	PAX5 - 2	09-037...	483	019	1.13	0.99	1.09	1.09
6	PAX5 - 1	09-037...	154	020	1.01	0.93	0.97	1.03
29	ETV6 - 1	12-011...	301	021	0.46	0.93	1.12	1
38	ETV6 - 1	12-011...	382	022	0.55	1.02	1.06	0.98
39	ETV6 - 2	12-011...	391	023	0.59	0.95	1.01	1.02
48	ETV6 - 5	12-011...	474	024	0.53	1.02	1.01	1.08
12	ETV6 - 8	12-011...	196	025	1.07	1.01	1.07	1.01
41	BTG1 AREA	12-090...	409	026	1.13	1.05	0.98	1
7	BTG1 AREA - 29	12-090...	160	027	0.99	1.01	1	0.96
32	BTG1 - 2	12-091...	328	028	1.08	1.01	1.04	1.11
18	BTG1 - 1	12-091...	232	029	1.08	0.96	1.04	0.96
16	RB1 - 6	13-047...	220	030	0.97	0.94	0.97	1
30	RB1 - 14	13-047...	310	031	0.97	0.95	0.97	0.97
35	RB1 - 19	13-047...	355	032	0.99	0.93	0.95	0.94
42	RB1 - 24	13-047...	418	033	1.01	0.94	1.02	1
45	RB1 - 26	13-047...	445	034	1	0.99	0.97	0.95
10	SHOX-AREA	23-755...	184	035	1.01	0.97	1.04	1.49
5	SHOX-AREA	23-770...	148	036	1.07	1.05	1.05	1.5
3	CRLF2 - 4	23-001...	136	037	1.06	0.99	1.07	1.46
8	CSF2RA - 16	23-001...	166	038	1.06	1.05	1.11	1.66
34	IL3RA - 1	23-001...	346	039	0.93	0.92	1.13	1.43
17	EBF1 - 1	05-158...	226	040	1.01	0.84	0.96	0.86
20	ETV6 - 3	12-011...	244	041	0.53	1.03	0.94	0.98
1	REFERENCE (125 nt)	02-242...	125	REF001	0.9	0.9	0.93	0.98
11	REFERENCE (190 nt)	03-038...	190	REF002	0.91	0.96	1.02	0.98
43	REFERENCE (427 nt)	05-013...	427	REF003	0.93	0.92	0.95	0.87
50	REFERENCE (490 nt)	05-090...	490	REF004	1.11	1.08	1.05	1.11
2	REFERENCE (130 nt)	05-132...	130	REF005	1.07	1.04	0.95	1.02
28	REFERENCE (294 nt)	07-072...	294	REF006	1.01	1.15	1.62	1.09
51	REFERENCE (500 nt)	11-001...	500	REF007	1	1	0.94	0.95
15	REFERENCE (214 nt)	11-020...	214	REF008	1.03	1.04	1.04	1.09
31	REFERENCE (319 nt)	15-046...	319	REF009	0.98	0.9	0.99	0.95
40	REFERENCE (400 nt)	20-025...	400	REF010	1.02	1.01	1.1	1.01

Fig. 9. Screen shot of from a tumor sample analyzed with the P335 MLPA kit. Probe ratio results of targets estimated as significantly increased as opposed to the reference population are marker green; those estimated as significantly decreased are marked red.

Our software enables users further, to display MLPA sample results in large array of different chart types figure 2 & 4). Charts may all be exported to different formats such as: jpg, gif, tiff, png, bmp. The results of a complete experiment may be plot together in grids and heat map algorithms may be applied to provide users a simple overview (figure 9). These grids may be exported to file formats (XML, txt, csv) that may be opened in Microsoft Excel. Alternatively these grids may also be exported to PDF files or several imaging formats.

6. Conclusions and future research

In this chapter we showed the options and applied algorithms of our MLPA analysis software, called Coffalyser.NET. Our software integrates new technologies enhancing the speed, accuracy and ease of MLPA analysis. Recognition of aberrations is improved by companioning effect-size statistics with statistical interference allowing users to interpreter units of measurement that are meaningful on a practical level (L. Wilkinson, 1999), while also being able to draw conclusions from data that are subject to random variation, for example, sampling variation (Bickel, Peter J.; Doksum, Kjell A., 2001). Our software contains extensive methods for results reporting and interpretation. It may also provide an alternative to software such as: Applied BioSystems Genotyper® and GeneScan® or GeneMapper®software; LiCor's SAGA, MegaBACE® Genetic Profiler and Fragment Profiler. Compatible with outputs from all major sequencing systems i.e. ABI Prism®, Beckman CEQ and MegaBACE® platforms. Coffalyser.NET is public freeware and can be downloaded from the MRC-Holland website.

Using data-mining techniques such as support vector machines in the large volumes of data obtained by large-scale MLPA experiments, may serve as a powerful and promising mechanism for recognizing of results patterns, which can be used for classification. Our future directions therefore concentrate on developing novel methods and algorithms that can improve recognition of disease related probe ratio patterns optimizing results in terms of validity, integrity and verification.

7. References

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

Albert, J. (2007) Bayesian Computation with R. Springer, New York

Applied Biosystems. (1988). AmpFℓSTR® Profiler Plus™ PCR Amplification Kit user's manual.

Bickel, Peter J.; Doksum, et al. (2001). Mathematical statistics: Basic and selected topics. 1

Clark, J. M. (1988). Novel non-templated nucleotide addition reactions catalyzed by procaryotic and eucaryotic DNA polymerases. *Nucleic Acids Res* 16 (20): 9677–86.

Coffa, J. (2008). MLPAnalyzer: data analysis tool for reliable automated normalization of MLPA fragment data, Cellular oncology, 30(4): 323-35

Ellis, Paul D. (2010). The Essential Guide to Effect Sizes: An Introduction to Statistical Power, Meta-Analysis and the Interpretation of Research Results. United Kingdom: Cambridge University Press.

Elizatbeth van Pelt-Verkuil, Alex Van Belkum, John P. Hays (2008). Principles and technical aspects of PCR amplification.

González J. 2008. Probe-specific mixed model approach to detect copy number differences using multiplex ligation dependent probe amplification (MLPA), BMC bioinformatics, 9:261

Hermsen M., Postma C. (2002). Colorectal adenoma to carcinoma progression follows multiple pathways of chromosomal instability, Gastroenterology, 123 (1109-1119)

Holtzman NA, Murphy PD, Watson MS, Barr PA (1997). "Predictive genetic testing: from basic research to clinical practice". *Science (journal)* 278 (5338): 602–5.

Huang, C.H., Chang, Y.Y., Chen, C.H., Kuo, Y.S., Hwu, W.L., Gerdes, T. and Ko, T.M. (2007). Copy number analysis of survival motor neuron genes by multiplex ligation-dependent probe amplification. *Genet Med*. 4, 241-248.

Janssen, B., Hartmann, C., Scholz, V., Jauch, A. and Zschocke, J. (2005). MLPA analysis for the detection of deletions, duplications and complex rearrangements in the dystrophin gene: potential and pitfalls. *Neurogenetics*. 1, 29-35.

Kluwe, L., Nygren, A.O., Errami, A., Heinrich, B., Matthies, C., Tatagiba, M. and Mautner, V. (2005). Screening for large mutations of the NF2 gene. *Genes Chromosomes Cancer*. 42, 384-391.

Michils, G., Tejpar, S., Thoelen, R., van Cutsem, E., Vermeesch, J.R., Fryns, J.P., Legius, E. and Matthijs, G. (2005). Large deletions of the APC gene in 15% of mutation-negative patients with classical polyposis (FAP): a Belgian study. *Hum Mutat.* 2, 125-34.

Nakagawa, Shinichi; Cuthill, Innes C (2007). "Effect size, confidence interval and statistical significance: a practical guide for biologists". *Biological Reviews Cambridge Philosophical Society* 82 (4): 591–605

"NCBI: Genes and Disease". NIH: National Center for Biotechnology Information (2008).

Redeker, E.J., de Visser, A.S., Bergen, A.A. and Mannens, M.M. (2008). Multiplex ligation-dependent probe amplification (MLPA) enhances the molecular diagnosis of aniridia and related disorders. *Mol Vis.* 14, 836-840.

Schouten, J.P. (2002), Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Research, 20 (12): e57

Scott, R.H., Douglas, J., Baskcomb, L., Nygren, A.O., Birch, J.M., Cole, T.R., Cormier-Daire, V., Eastwood, D.M., Garcia-Minaur, S., Lupunzina, P., Tatton-Brown, K., Bliek, J., Maher, E.R. and Rahman, N. (2008). Methylation-specific multiplex ligation-dependent probe amplification (MS-MLPA) robustly detects and distinguishes 11p15 abnormalities associated with overgrowth and growth retardation. *J Med Genet.* 45, 106-13.

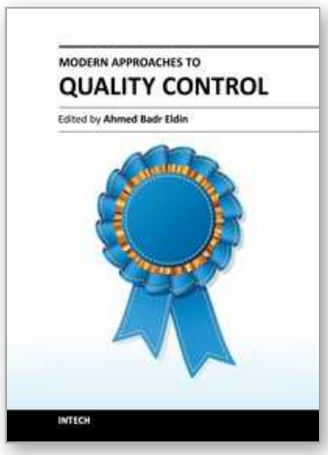
Sequeiros, Jorge; Guimarães, Bárbara (2008). Definitions of Genetic Testing EuroGentest Network of Excellence Project.

Taylor, C.F., Charlton, R.S., Burn, J., Sheridan, E. and Taylor, GR. (2003). Genomic deletions in MSH2 or MLH1 are a frequent cause of hereditary non-polyposis colorectal cancer: identification of novel and recurrent deletions by MLPA. *Hum Mutat.* 6, 428-33.

Wilkinson, Leland; APA Task Force on Statistical Inference (1999). "Statistical methods in psychology journals: Guidelines and explanations". *American Psychologist* 54: 594–604. doi:10.1037/0003-066X.54.8.594.

Yau SC, Bobrow M, Mathew CG, Abbs SJ (1996). "Accurate diagnosis of carriers of deletions and duplications in Duchenne/Becker muscular dystrophy by fluorescent dosage analysis". *J. Med. Genet.* 33 (7): 550–558. doi:10.1136/jmg.33.7.550.

Zar, J.H. (1984) *Biostatistical Analysis.* Prentice Hall International, New Jersey. pp 43–45



Modern Approaches To Quality Control

Edited by Dr. Ahmed Badr Eldin

ISBN 978-953-307-971-4
Hard cover, 538 pages
Publisher InTech
Published online 09, November, 2011
Published in print edition November, 2011

Rapid advance have been made in the last decade in the quality control procedures and techniques, most of the existing books try to cover specific techniques with all of their details. The aim of this book is to demonstrate quality control processes in a variety of areas, ranging from pharmaceutical and medical fields to construction engineering and data quality. A wide range of techniques and procedures have been covered.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Jordy Coffa and Joost van den Berg (2011). Analysis of MLPA Data Using Novel Software Coffalyser.NET by MRC-Holland, Modern Approaches To Quality Control, Dr. Ahmed Badr Eldin (Ed.), ISBN: 978-953-307-971-4, InTech, Available from: http://www.intechopen.com/books/modern-approaches-to-quality-control/analysis-of-mlpa-data-using-novel-software-coffalyser-net-by-mrc-holland

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen