BIOINFORMATICS

Vol. 30 ECCB 2014, pages i631–i638
doi:10.1093/bioinformatics/btu473

Microarray R-based analysis of complex lysate experiments with MIRACLE

Markus List[1,2,3,*,†], Ines Block[1,2,†], Marlene Lemvig Pedersen[1,2], Helle Christiansen[1,2], Steffen Schmidt[1,2], Mads Thomassen[1,3], Qihua Tan[3,4], Jan Baumbach[5] and Jan Mollenhauer[1,2]

[1]Lundbeckfonden Center of Excellence in Nanomedicine NanoCAN, [2]Molecular Oncology, Institute of Molecular Medicine, [3]Human Genetics, Institute of Clinical Research, [4]Epidemiology, Biostatistics and Biodemography, Institute of Public Health and [5]Department of Mathematics and Computer Science, University of Southern Denmark, 5000 Odense, Denmark

ABSTRACT

Motivation: Reverse-phase protein arrays (RPPAs) allow sensitive quantification of relative protein abundance in thousands of samples in parallel. Typical challenges involved in this technology are antibody selection, sample preparation and optimization of staining conditions. The issue of combining effective sample management and data analysis, however, has been widely neglected.

Results: This motivated us to develop *MIRACLE*, a comprehensive and user-friendly web application bridging the gap between spotting and array analysis by conveniently keeping track of sample information. Data processing includes correction of staining bias, estimation of protein concentration from response curves, normalization for total protein amount per sample and statistical evaluation. Established analysis methods have been integrated with *MIRACLE*, offering experimental scientists an end-to-end solution for sample management and for carrying out data analysis. In addition, experienced users have the possibility to export data to R for more complex analyses. *MIRACLE* thus has the potential to further spread utilization of RPPAs as an emerging technology for high-throughput protein analysis.

Availability: Project URL: http://www.nanocan.org/miracle/

Contact: mlist@health.sdu.dk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Reverse-phase protein arrays are typically nitrocellulose-covered glass slides on which crude lysates of tissue samples or treated cell lines are spotted. Each single slide can carry several thousand spots. Only small amounts of lysate in the range of a few cell equivalents are required for each spot. Consequently, several hundred slide copies can be created at minimal sample consumption, each of which can be interrogated with a different protein-specific antibody. This allows high-throughput measurement of the relative abundance of proteins in up to several thousand samples. Parallel processing of large sample numbers discerns RPPAs from forward phase arrays, where probes are immobilized on a slide, and mass spectrometry, which are suitable for analysis of many proteins in small sample numbers.

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

The field of RPPAs has shown steady growth since its introduction in 2001 (Paweletz *et al.*, 2001) (Fig. 1). Several studies applied this technology successfully to protein and signaling pathway analyses in cancer (Leivonen *et al.*, 2009; Sevecka and MacBeath, 2006; Uhlmann *et al.*, 2012; York *et al.*, 2012), as well as for cancer subtype classification and prognosis of disease progression (Gonzalez-Angulo *et al.*, 2011; Sonntag *et al.*, 2014; Wiegand *et al.*, 2014). The relevance of RPPA data for multi-OMICS and high-throughput projects is also highlighted by its inclusion into the Cancer Genome Atlas, where, for instance, measurements of 171 antibodies for >400 samples of breast cancer patients are available (Atlas, 2012).

1.1 Challenges

Experimental challenges involved in this technology, such as antibody selection, sample preparation and optimization of staining conditions, have been addressed successfully in the past (Hennessy *et al.*, 2010; Mannsperger *et al.*, 2010b; Mircean *et al.*, 2005; Spurrier *et al.*, 2008). The subsequent image analysis can be handled by established methods and software developed for traditional microarrays, such as the commercial tool MicroVigene®. RPPA-specific challenges arise in the further processing of the quantified signal, which, in a first step, should be corrected for bias introduced through uneven staining (Neeley *et al.*, 2012). Another concern is the dynamic range of signal detection, which can be described as a sigmoidal curve due to limitations in sensitivity in the lower range and signal saturation in the upper range (Tabus *et al.*, 2006). Through adjusting each sample for the total protein amount *a priori*, measurement is possible in the linear range of this curve. However, this is often not feasible for high-throughput experiments, due to the trade-off between large sample numbers, feasibility and costs. To overcome this problem, samples are typically spotted multiple times in a dilution series to cover a broad dynamic range of protein concentrations, where each sample gives rise to a response curve. In a process called quantification, an estimate of relative protein abundance is created by merging these values (Supplementary Fig. S1). The resulting relative concentration estimates still need to be normalized for the total amount of protein, before they can be evaluated statistically. Statistical analysis comprises assessing significance of relative differences in protein concentration estimates, as well as their correlation between slides and

© The Author 2014. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

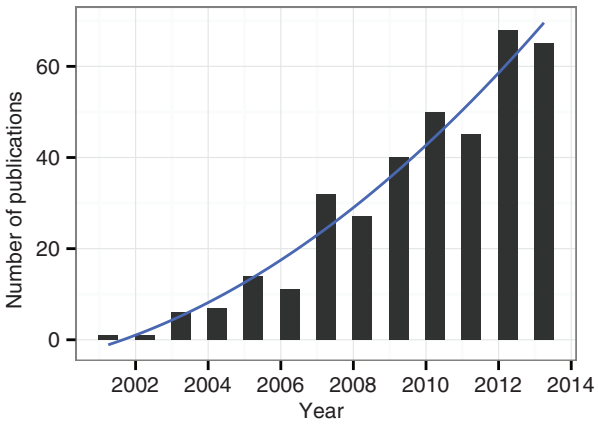


Fig. 1. Number of RPPA publications per year. Data were extracted for 2001–2013 from PubMed, using the search term 'reverse phase protein' OR 'reverse-phase protein'

other types of readout. Finally, the large number of samples involved in each experiment poses a significant challenge in terms of sample management.

1.2 Quantification

Mircean *et al.* (2005) proposed a linear model for merging the signal originating from individual response curves. One drawback of this approach is that a linear model cannot deal with samples close to saturation or close to the detection limit. Consequently, Tabus *et al.* (2006) compared a variety of parametric models and found that a logistic model was most suitable to reflect the sigmoidal shape of the response curve. Furthermore, a joint response curve based on all samples increased the confidence of the model fit, as similar chemistry can be assumed for all samples. Hu *et al.* (2007) showed that a more flexible non-parametric model yields more accurate estimates at the cost of robustness. Finally, Zhang *et al.* (2009) proposed a simplified robust parametric model called serial dilution curve based on the Sips model for DNA binding. In contrast to other parametric models, this method is based on meaningful and intuitive parameters like the detection limit, the dilution factor and the saturation level.

1.3 Normalization

If the total protein amount of each sample is not determined *a priori*, protein levels have to be normalized, to guarantee a meaningful comparison between samples. This can be achieved by either normalizing to a slide stained for total protein with, for instance, Sypro Ruby (Leivonen *et al.*, 2009) or Fast Green (Loebke *et al.*, 2007), or by using additional antibody stainings. For the latter approach, one can rely on either a selection of 'housekeeping' proteins that are assumed to be constantly expressed or on incorporating the entire panel of antibody-generated signals, where all antibodies are first centred and scaled before the median value is used for normalization. This so-called median loading normalization has been improved by Neeley *et al.* (2009) in a method called variable slope normalization. Here, a correction factor is included to take into account that additional bias arises due to independent slide measurements.

Finally, Neeley *et al.* (2012) also proposed an additional normalization step called surface adjustment that is applied to the raw data. As customary for all microarray data, the background signal is determined for each spot and subtracted from the foreground signal. This approach, however, does not correct for signal bias due to uneven antibody staining, which is an issue specific to RPPA technology. Positive control spots on the slide can be utilized for creating a smoothing surface mirroring the staining bias. A correction factor can then be calculated from a generalized linear model for each individual spot.

1.4 Sample tracking

A single RPPA experiment may comprise thousands of samples distributed over large slide sets. Precise sample tracking is a challenge that grows with the number, size and complexity of the RPPA experiments. To date, the only documented solution to address this critical issue is an integrated platform called RIMS (Stanislaus *et al.*, 2008), which provides features for uploading and annotating sample information, data visualization, correlation and pathway analysis. Notably, the authors also propose a XML standard called RPPAML to overcome the lack of a data exchange format for RPPA data and a standardized annotation. Unfortunately, however, none of the project URLs are accessible (last access attempt March 20, 2014), indicating that the project is no longer under active development and has not been adapted by the community. Being implemented for the commercial software *MATLAB*, *RIMS* also lacks integration of R methods commonly used for RPPA analyses. Finally, *RIMS* only supports sample tracking at the slide level and not at the level of the plate formats that form the basis for all experimentation. This leaves the most difficult step of sample tracking to the user: Samples are taken up by an extraction head configured to generate a slide in multiple extractions, thereby producing a complex spotting pattern that does not permit researchers to locate their samples in a straight-forward fashion.

1.5 Existing solutions

Implementations for both, parametric and non-parametric methods are available through the R packages *SuperCurve* (Hu *et al.*, 2007) and *RPPanalyzer* (Mannsperger *et al.*, 2010a). A major challenge in analysing RPPA data is, however, that end users are often not familiar with R. *SuperCurve* overcomes this problem partly by offering an tcl/tk-based graphical user interface, making both analysis and experimental design more accessible. Sample management on a larger scale, however, is neglected. *RIMS* addresses some of these issues, but does not cover more complex data analysis (see Table 1).

1.6 Microarray R-based analysis of complex lysate experiments (MIRACLE)

This motivated us to develop *MIRACLE*, a comprehensive and user-friendly open-source web application providing an end-to-end solution covering experimental design, sample tracking, data processing, normalization, as well as visualization and statistical analysis of the results. *MIRACLE* conveniently keeps

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

Table 1. Features of *MIRACLE* and other open-source solutions for processing RPPA data

Feature	SuperCurve	RPPanalyzer	RIMS	MIRACLE
Platform	R, tcl/tk	R	MATLAB, PHP	R, Grails
GUI[a]	✓	✗	✓	✓
Plate layouts	✗	✗	✗	✓
Plate readouts	✗	✗	✗	✓
Slide layouts	✗	✗	✗	✓
Virtual spotting	✗	✗	✗	✓
Visualization	✓	✓	✓	✓
Surface adjustment	✓	✗	✗	✓
Quantification	✓	✓	✗	✓
Normalization	✗	✓	✗	✓
Significance	✗	✗	✗	✓
Correlation	✗	✓	✗	✓
Timecourse analysis	✗	✓	✗	✗
Network analysis	✗	✗	✓	✗

[a]Graphical user interface.

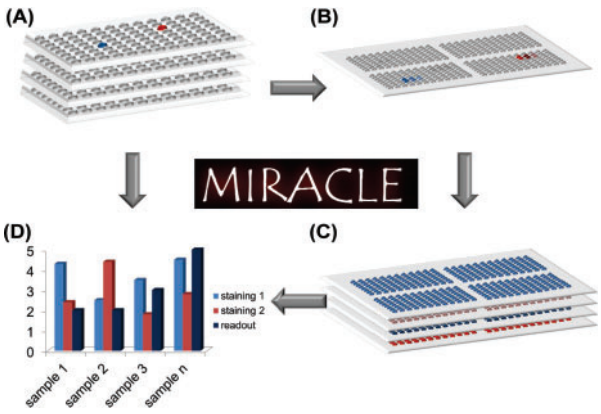


Fig. 2. Schematic exemplary work-flow of RPPA construction and analysis via *MIRACLE*. (**A**) Larger sample sets are stored in multiple source plates, with individual and partly complex sample information. Optionally primary plate readout data can be included. (**B**) The individual sample lysates are diluted (indicated by the colour gradient) in a first reformatting step and subsequently spotted onto slides in a customized pattern. (**C**) Multiple array copies are generated and stained with different antibodies, adding to the sample tracking demands. (**D**) An array scanner yields signal intensities for all spots, which need to be further processed to obtain the final results. *MIRACLE* offers a user interface for data analysis under automated mapping of samples on the RPPAs to plate- and slide-based data

track of sample information, starting with the source plates, throughout array generation and down to the signal data, in a process called virtual spotting.

MIRACLE allows biological researchers without any knowledge of R to process and analyse RPPA data efficiently, grasping back to established methods by interacting directly with R in the background. This interface will also allow future methods to be added in a straight-forward fashion. Results are directly visualized and can be investigated interactively with regards to statistical significance, as well as to correlation to primary plate based readout data.

MIRACLE is designed with user approachability in mind, but also supports R data analysts by offering a convenient data export/import interface with R. While the data analysis part of *MIRACLE* is particularly laid out for handling RPPA data, the sample management functionality is suitable for any kind of customized array design.

With its deep integration of sample management and data analysis, *MIRACLE* separates itself from existing solutions that only cover parts of the RPPA work-flow shown in Figure 2. See Table 1 for an overview of existing solutions and *MIRACLE*.

2 SYSTEM AND METHODS

2.1 Sample management

2.1.1 Plate and slide layouts In a typical RPPA experiment, lysate samples are stored in 96-well or 384-well microtiter plate, before they are subjected to microarray generation using a bioarrayer, -printer or -spotter. Already at this stage, *MIRACLE* supports experimental design by offering an interactive web interface for creating so-called plate layouts. To avoid cryptic and long sample names, several layers of information can be included, for instance regarding cell material, treatments, applied compounds, lysis conditions, etc. (Fig. 3). A lot of this information is shared by samples and is therefore redundant. *MIRACLE* stores information in relational databases, where each layer corresponds to a single property, keeping the data concise through use of ids and mapping tables (Supplementary Fig. S2). By relying on linked tables of a relational database, changes of layout properties are immediately available to all samples and experiments, thereby ensuring data consistency and comparability in the analyses.

Similar to plate layouts, users can also define slide layouts, in which sample properties for each individual spot of the slide can be edited using the aforementioned sample layers. The format of the slide is determined through specifying the number of rows and columns, as well as blocks where applicable, e.g. when using spotters, where each pin of the extraction head gives rise to a different block.

2.1.2 Virtual spotting As previously mentioned, it is not always trivial to determine the location of a sample on the slide, since a large number of samples originate from different plates. Furthermore, repetitive spotting of samples with various dilutions and varying number of depositions per spot has to be considered, as well as the format of the extraction head. *MIRACLE* addresses this issue in a feature called virtual spotting, where previously created plate layouts are combined with information about the operation mode of the spotter, such as format of the extraction head, column or row-wise extraction, top-to-bottom or left-to-right spotting, to determine the final layout. The selection and order of the plates can be manipulated via drag-and-drop and for each plate individual extractions can be excluded. If a so-called deposition pattern is used, in which

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

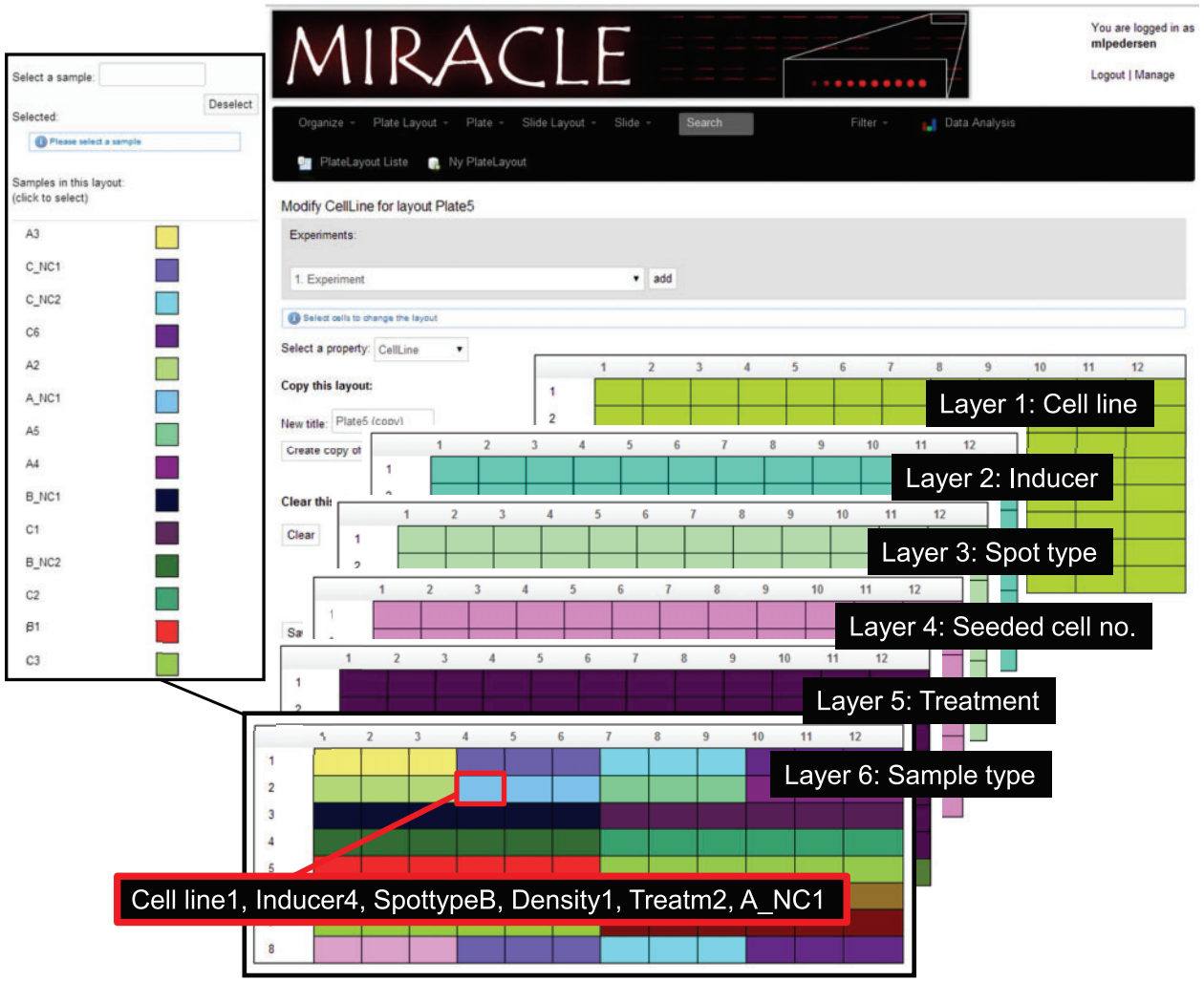


Fig. 3. Sample management of plate layouts illustrated for an exemplary 96-well plate layout. Several layers of information are accessible

several adjacent spots originate from the same sample, but are spotted with varying depositions, the layout can be simplified. Because these samples are otherwise identical, the respective columns of the layout can be merged.

2.1.3 Projects and experiments *MIRACLE* offers a quick search field to locate sample information quickly using full text search. However, to keep experimental data organized, projects and experiments can be created, linked to layouts and subsequently be used for filtering.

2.2 Data processing

2.2.1 Slides Slide layouts can be linked to an arbitrary number of slides, which correspond to the copies created during spotting. For each slide, additional information such as a barcode, the antibody that was used for staining and scanner settings, such as the wavelength of the readout can be specified. Three types of files can be uploaded, including the output file from the scanner containing signal intensities, an image of the slide and an experimental protocol.

2.2.2 Supported file formats The experimental protocol can be of any file type (e.g. doc, pdf or txt), while for images the most common file types, such as jpg, png and tiff, are supported. *MIRACLE* processes each image into a zoomable format for visual detection of quality issues, such as clogged tips, scratches or uneven stainings. With regards to the array scanner output, *MIRACLE* is not limited to certain file types, but has a flexible system supporting import of comma, semicolon, tab-separated or Microsoft Excel ® files without requiring a specifc format.

2.2.3 Processing raw data After successfully reading the scanner file, *MIRACLE* will offer to add all spots to the database. During this process, the signal of each spot is linked to the sample information stored in the slide layout. Subsequently, users can create heatmaps to visualize the data to detect quality problems. One example are block shifts introduced by the scanner software that can then be corrected for (Supplementary Fig. S3).

2.2.4 Plates and readouts Similar to how slides can be added to slide layouts, plates can be added to plate layouts, where

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

additional information, such as plate and well type, barcode and replicate number are stored. If a readout is performed before plates are subjected to spotting, for example, fluorescence-based measurement of cell viability or colourimetric analysis of total protein amount, *MIRACLE* allows for adding these results for each plate, utilizing the aforementioned file upload mechanism.

2.3 Data analysis

When sample management and data processing are complete, users can begin with data analysis. After selecting a slide layout, the user is presented with a list of all slides linked to this layout. In case the slide layout was created through virtual spotting, readouts linked to the source plates are also shown. By starting the analysis, the user will be forwarded to an R-based web application called *Rmiracle*, which will automatically begin to fetch the selected slides and readouts from the database.

2.3.1 Processing of raw signal *Rmiracle* offers data analysis in several steps (Fig. 4):

- A heatmap for visual inspection and correction of block shifts (Supplementary Fig. S3).
- Positive control spots can be used to correct for uneven staining using the method proposed by Neeley *et al.* (2012).
- If a dilution series has been spotted, a quantification method can be selected for merging these samples. *Rmiracle* currently supports *SuperCurve*, as well as implementations of a logistic model (Tabus *et al.*, 2006), serial dilution curve (Zhang *et al.*, 2009) and a non-parametric model (Hu *et al.*, 2007).
- Slides can be normalized for total protein amount by selecting between median loading, variable slope (Neeley *et al.*, 2009) or housekeeping normalization. For the latter, one or several of the slides have to be marked as loading controls.
- Significance of relative sample differences can be assessed by selecting a sample reference for performing Dunnett's test (Hothorn *et al.*, 2008).

2.3.2 Protein concentration estimates and sample grouping With the above settings, *Rmiracle* computes protein concentration estimates that allow assessment of relative differences between samples. To this end, we grasp back to the multi-layer sample information model of *MIRACLE* to group samples. Users can select horizontal and vertical grouping categories, for example cell-lines tested or treatments applied, which will then be reflected by different facets of a bar plot. Users can also select an additional category called 'fill' for separating bars by colour to achieve a third grouping dimension to compare, for instance, replicates with different numbers of depositions. The results are also shown in tabular form, including a download option, and are further accompanied by diagnostic plots specific for the selected quantification method. Figure 5 depicts the user interface of the *Rmiracle* analysis. Beyond data visualization and computation of protein concentration estimates, the analysis comprises additional features introduced below.

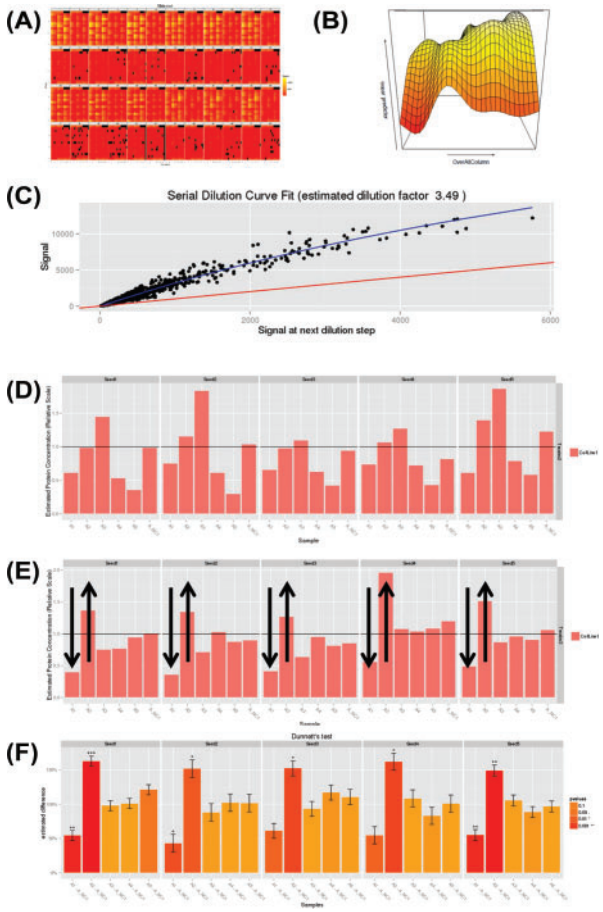


Fig. 4. Processing of raw data in Rmiracle: Signal intensities are displayed in heatmaps (**A**) for visual inspection. Subsequently, a surface adjustment based on control spots may be performed to correct for uneven staining (**B**). In case of serial sample dilutions, quantification can be applied (**C**) to obtain a single protein concentration estimate (**D**). Furthermore, data can be normalized to negative controls (A_NC1) and total protein amount, e.g. using protein data from separate slides, which enables the identification of effector samples (**E**, depicted by arrows). Significance of the sample differences is finally assessed in comparison to a selected negative control by applying Dunnett's test (**F**)

2.3.3 Comparison across slides and readouts The comparison tab provides a bar plot (Fig. 5F), in which an average is calculated for the previously selected colour fill category, since colours are here reserved for comparing protein concentration estimates across slides. It is also possible to include plate readouts.

2.3.4 Correlation An important aspect of quality control is signal correlation. In the correlation tab (Fig. 5G), all pairwise Pearson correlation coefficients are calculated for both, protein concentration estimates and raw signal intensities, and presented as a heatmap. Correlation is also shown between slides and plate readouts. This can be an important factor, e.g. in case a plate-based readout provides information about the total protein amount and should therefore correlate with the RPPA signal used for normalization.

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014



Fig. 5. Analysis using Rmiracle: A heatmap visualizes different slide properties, such as signal intensities (**A**). Users can change various parameters, for example inclusion of surface adjustment, selection of methods for quantification and normalization for total protein amounts. Specific samples can be selected and grouped based on different properties of the data set (**B**). Depending on the quantification method, diagnostic plots are shown (**C**). The results are displayed in an interactive table and in a bar plot (**D** and **E**). A global overview of protein concentration estimates is available for all slides and plate-based readouts (**F**). Pearson correlation coefficients are calculated for all slides, as well as for plate-based readouts (**G**). Significance is assessed by comparing sample groups to negative controls through Dunnett's test (**H**)

2.3.5 Significance The significance of relative differences between samples or between samples and a control is of great interest for experimental researchers. Traditionally, t-tests are used to obtain the necessary *P*-values, often neglecting multiple comparisons correction and issues arising from low replicate numbers. To address these issues, MIRACLE applies Dunnett's test (Fig. 5H), which is a t-statistic based multiple comparison method comparing each sample with a pre-defined control. In contrast to other methods, the variance is pooled across all samples, thereby dealing with low replicate numbers (Hothorn *et al.*, 2008).

2.3.6 Import and export Convenient import functions allow experienced R users to download RPPA data directly from *MIRACLE* by specifying ids or barcodes, respectively. R methods to process or visualize these data are directly available, allowing data analysts to perform deeper analysis not covered by the proposed work-flow. Each slide, as well as the resulting protein concentration estimates can be downloaded as tab- or comma-separated file, in which all layout information is included.

3 IMPLEMENTATION

3.1 *MIRACLE* web application

MIRACLE was built using *Grails* (http://grails.org/), a Groovy/Java based web application framework that allows for rapid development with a convention over configuration approach. *Grails* provides web application critical functionality through industry-proven projects and plug-ins. This includes, for instance, *Hibernate* for abstracting data access by modelling database tables through java classes and *Apache Lucene* (http://lucene.apache.org/core/) for efficient database search. Using hibernate allows MIRACLE to operate with any JDBC compatible SQL database, such as Microsoft ® SQL Server or Oracle ® MySQL. Data export to R is realized through a web service, in which efficient conversion between database content and JSON objects is facilitated using *Jackson* (https://github.com/FasterXML/jackson). Furthermore, the *SpringSecurity* project (http://projects.spring.io/spring-security/) limits access with a role-based user model. In case data should be accessible to users without an account, *MIRACLE* provides an alternative access

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

model through universal unique identifiers called security tokens. These are generated automatically for each slide and plate readout. In the current version (v. 0.8), all data are accessible to all users. With the next release, we will change this such that data are private for each user unless selected otherwise. To efficiently deal with large image files in *MIRACLE*, we created the *Grails OpenSeaDragon* plug-in (http://grails.org/plugin/open-seadragon) for generating and displaying pyramide representations of slide images in the Microsoft ® deep zoom format.

3.2 *Rmiracle* R package

All R functions have been wrapped in the R open-source package *Rmiracle*. This includes user and security token based authentication for downloading data from *MIRACLE*, methods for RPPA data processing, e.g. surface adjustment, various methods for quantification and normalization, as well as methods for visualization and statistical analysis using functionality implemented in the R package *multcomp*.

In addition, all of these features are accessible through two web applications developed directly on top of R utilizing *Shiny* (http://www.rstudio.com/shiny/). Both *Shiny* applications are included in the *Rmiracle* package and can be used independent of *MIRACLE* via uploading files in a *MIRACLE* compliant format (see Suppl. File 1 for an example).

4 DISCUSSION

RPPAs are a promising technology that finds growing application in both, basic and clinical research. While many of the challenges of this technology are similar to those of traditional microarrays, RPPA-specific challenges arise and have to be addressed. In our efforts to adapt this technology as secondary readout to high throughput genome-wide RNAi screens, we identified a lack of a comprehensive tool incorporating all necessary tasks, such as experimental design, sample tracking and data analysis. To fill this gap, we developed *MIRACLE*, a web-based tool with deep integration of R for efficient data analysis.

Using a database-driven web application, such as *MIRACLE*, for sample management offers a number of advantages. Due to relational tables, all data are kept in a concise and consistent format, where changes and updates are automatically propagated. In contrast to file-based storage solutions, no experimental information is lost upon turnover of laboratory staff and no problems arise from cryptic and inconsistent sample terminology. Collaboratively creating experimental data is significantly more convenient in web-based applications, as concurrency issues, such as file locks, can be avoided. In addition, all information can be located quickly, using full text search, which additionally increases efficiency.

With regards to sample management, both the *SuperCurve* package and *RIMS* provide a graphical user interface for specifying slide layouts, but it does not address sample tracking from plate to slide level and does not allow for multiple levels of sample information. Moreover, they lack the virtual spotting and layout editing features of *MIRACLE* that enable researchers to enter all sample-related information already on the plate level and *before* the complexity of the layout is increased by the array generation. This saves a significant amount of time and effectively avoids mistakes due to manual data processing.

The R packages *SuperCurve* and *RPPanalyzer* provide experienced R users with a wealth of options to analyse RPPA data. The results of these methods are relative protein concentration estimates. A logical next step could be to investigate how significant relative differences in protein levels are and how well results correlate, e.g. between slides used for normalization or between individual slides and plate-based readout. Only *RPPanalyzer* reports on slide to slide correlation (Mannsperger *et al.*, 2010a). Significance analysis is not part of any existing solution. Moreover, replicates are typically merged during data processing to increase confidence of the results, but thereby making them unavailable for subsequent significance analysis. In contrast, *Rmiracle* processes replicates individually and offers a comprehensive evaluation of significance and correlation. A number of analysis methods, such as serial dilution curve (Zhang *et al.*, 2009) have been published as R code, but have not been adapted to a user-friendly format, thereby limiting their application for experimental researchers. Experienced R users, on the other hand, need the flexibility of the R environment to perform deeper analysis of the data. *Rmiracle* strives to serve both target groups by incorporating a broad number of published methods on RPPA data analysis in both, command line and web interface. Additionally, *Rmiracle* can be used completely independently of the *MIRACLE* web application, requiring only a local installation of R.

Notably, the web application *RIMS* followed similar goals as *MIRACLE*, but did not include scenarios where more sophisticated data processing, e.g. quantification and normalization of the signal intensities, is necessary (Neeley *et al.*, 2009). Moreover, *RIMS* is not actively developed or available at the moment, stressing the need for a solution like *MIRACLE*.

4.1 Outlook

While *MIRACLE* has been designed to handle RPPA data, the sample tracking issues addressed here are in general common for researchers constructing customized microarrays, allowing adaption of *MIRACLE* to serve other array formats.

RPPA data are particularly suited for unraveling complex protein signaling mechanisms. Therefore, we plan to integrate *MIRACLE* with suitable tools for network and pathway analysis.

The minimum information about a micorarray experiment (MIAME) standard (Brazma *et al.*, 2001) and platforms like the gene expression omnibus (GEO) (Edgar *et al.*, 2002) offer an effective method for standardized data exchange of gene expression array data. In GEO, users can export data to R or utilize a web-based application called *GEO2R* (https://www.ncbi.nlm.nih.gov/geo/geo2r/) for basic analysis.

MIRACLE has the potential to deliver similar features to the RPPA community. We therefore intend to continue development towards an RPPA web portal for published data. In addition, we envision that *MIRACLE* could serve as a framework for comparing the performance of various methods for RPPA data processing.

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

5 CONCLUSION

In recent years, the application of RPPA technology has matured considerably. Along with this progress, suitable computational methods have been developed to address issues in data processing. To further promote acceptance of this technology, fully integrated tools like *MIRACLE* are indispensable. Furthermore, it can be expected that standardization of RPPA data in a common framework can substantially aid the development of novel algorithms and allow better integration at the level of network biology and other multi-OMICS data. It is our hope that *MIRACLE* will attract contributions from both, users and developers, which will help to strengthen the entire field. To this end, we have established a github repository (https://github.com/NanoCAN/MIRACLE) and established a demo application (http://www.nanocan.org/miracle/demo) containing biological sample data. Further documentation and a step-by-step user guide are available online (Supplementary File 2).

ACKNOWLEDGEMENT

We would like to thank Prof. Torben A. Kruse for valuable advice.

Funding: This work was supported by the Lundbeckfonden grant for the NanoCAN Center of Excellence in Nanomedicine, the Region Syddanmarks ph.d.-pulje and Forskningspulje, the Fonden Til Lægevidenskabens Fremme and co-financed by the INTERREG 4 A-program Syddanmark-Schleswig-K.E.R.N. with funds from The European Regional Development Fund.

Conflict of Interest: none declared.

REFERENCES

Atlas,T.C.G. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

Brazma,A. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.

Edgar,R. *et al.* (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.

Gonzalez-Angulo,A.M. *et al.* (2011) Functional proteomics can define prognosis and predict pathologic complete response in patients with breast cancer. *Clin. Proteomics*, **8**, 11.

Hennessy,B.T. *et al.* (2010) A technical assessment of the utility of reverse phase protein arrays for the study of the functional proteome in non-microdissected human breast cancers. *Clin. Proteomics*, **6**, 129–151.

Hothorn,T. *et al.* (2008) Simultaneous inference in general parametric models. *Biom. J.*, **50**, 346–363.

Hu,J. *et al.* (2007) Non-parametric quantification of protein lysate arrays. *Bioinformatics*, **23**, 1986–1994.

Leivonen,S.-K. *et al.* (2009) Protein lysate microarray analysis to identify microRNAs regulating estrogen receptor signaling in breast cancer cell lines. *Oncogene*, **28**, 3926–3936.

Loebke,C. *et al.* (2007) Infrared-based protein detection arrays for quantitative proteomics. *Proteomics*, **7**, 558–564.

Mannsperger,H. *et al.* (2010a) RPPanalyzer: Analysis of reverse-phase protein array data. *Bioinformatics*, **26**, 2202–2203.

Mannsperger,H.A. *et al.* (2010b) RNAi-based validation of antibodies for reverse phase protein arrays. *Proteome Sci.*, **8**, 69.

Mircean,C. *et al.* (2005) Robust estimation of protein expression ratios with lysate microarray technology. *Bioinformatics*, **21**, 1935–1942.

Neeley,E.S. *et al.* (2009) Variable slope normalization of reverse phase protein arrays. *Bioinformatics*, **25**, 1384–1389.

Neeley,E.S. *et al.* (2012) Surface adjustment of reverse phase protein arrays using positive control spots. *Cancer Informatics*, **11**, 77–86.

Paweletz,C.P. *et al.* (2001) Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. *Oncogene*, **20**, 1981–1989.

Sevecka,M. and MacBeath,G. (2006) State-based discovery: a multidimensional screen for small-molecule modulators of EGF signaling. *Nat. Methods*, **3**, 825–831.

Sonntag,J. *et al.* (2014) Reverse phase protein array based tumor profiling identifies a biomarker signature for risk classification of hormone receptor-positive breast cancer. *Transl. Proteomics*, **2**, 52–59.

Spurrier,B. *et al.* (2008) Reverse-phase protein lysate microarrays for cell signaling analysis. *Nat. Protoc.*, **3**, 1796–1808.

Stanislaus,R. *et al.* (2008) RPPAML/RIMS: a metadata format and an information management system for reverse phase protein arrays. *BMC Bioinformatics*, **9**, 555.

Tabus,I. *et al.* (2006) Nonlinear modeling of protein expressions in protein arrays. *IEEE Trans. Signal Process.*, **54**, 2394–2407.

Uhlmann,S. *et al.* (2012) Global microRNA level regulation of EGFR-driven cell-cycle protein network in breast cancer. *Mol. Syst. Biol.*, **8**, 570.

Wiegand,K.C. *et al.* (2014) A functional proteogenomic analysis of endometrioid and clear cell carcinomas using reverse phase protein array and mutation analysis: protein expression is histotype-specific and loss of ARID1A/BAF250a is associated with AKT phosphorylation. *BMC Cancer*, **14**, 120.

York,H. *et al.* (2012) Network analysis of reverse phase protein expression data: Characterizing protein signatures in acute myeloid leukemia cytogenetic categories t(8;21) and inv(16). *Proteomics*, **12**, 2084–2093.

Zhang,L. *et al.* (2009) Serial dilution curve: a new method for analysis of reverse phase protein array data. *Bioinformatics*, **25**, 650–654.

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014