Published online 17 February 2009 *Nucleic Acids Research, 2009, Vol. 37, No. 7* **2065–2069**
doi:10.1093/nar/gkp056

SURVEY AND SUMMARY

RDML: structured language and reporting guidelines for real-time quantitative PCR data

Steve Lefever[1], Jan Hellemans[1,*], Filip Pattyn[1], Daniel R. Przybylski[2], Chris Taylor[3], René Geurts[4], Andreas Untergasser[4] and Jo Vandesompele[1], on behalf of the RDML consortium

[1]Center for Medical Genetics, Ghent University Hospital, Ghent, Belgium, [2]Bio-Rad Laboratories, Inc., Hercules, CA, USA, [3]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK and [4]Laboratory of Molecular Biology, Department of Plant Science, Wageningen University, The Netherlands

Received October 7, 2008; Revised December 31, 2008; Accepted January 20, 2009

ABSTRACT

The XML-based Real-Time PCR Data Markup Language (RDML) has been developed by the RDML consortium (http://www.rdml.org) to enable straightforward exchange of qPCR data and related information between qPCR instruments and third party data analysis software, between colleagues and collaborators and between experimenters and journals or public repositories. We here also propose data related guidelines as a subset of the Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) to guarantee inclusion of key data information when reporting experimental results.

INTRODUCTION

Real-time quantitative PCR (qPCR) is the gold standard method for accurate and sensitive quantification of nucleic acid sequences. Over 25 000 papers indexed by PubMed refer to qPCR, demonstrating the success of this technology.

Real-time qPCR requires dedicated instruments that are able to quantify amplification products in real-time during each cycle. Each instrument has its own data collection software for display of measured values as amplification plots and melting curves, and for calculation of quantification cycle values (Cq). Of note, these values present only the first step in a long path of data processing. To this purpose, some instrument suppliers include basic tools for relative or absolute quantification. In general, these tools do not fulfill the researchers' needs for more specialized analysis; hence, numerous independent tools have been developed by academics (geNorm, qBase, DART, qCalculator, qPCR-DAMS, Q-gene and REST) as well as companies (qBasePlus, StatMiner, GenEx and SoFar) to fill the gap. The major obstacle for such tools, and in general for the exchange of all qPCR data, is the lack of a common format for qPCR data. The diversity of qPCR data formats has also hindered the publication of experimental data and its submission to public repositories as is currently done for microarray experiments (1–3).

The demand for a standardized qPCR data format was recognized several years ago and a first standardization initiative was presented at the international qPCR meeting in Weihenstephan, 2005. It was followed by an open discussion lasting 2 years. Recently, several thousand qPCR users were invited by the editor of the Gene Quantification portal (http://www.gene-quantification.com) to publicly review that draft standard. A number of research institutions and companies declared their support for this initiative, resulting in the creation of an international consortium to accelerate the development of a standard for qPCR data. This standard is to be known as RDML (Real-time PCR Data Markup Language). The consortium is organized into a key developer group, a member community and a large group of supporters. Interested individuals can join the consortium to declare their support for RDML (supporters) or to provide feedback and suggestions on developments (member community). The RDML project is supported by companies providing real-time PCR instruments and reagents (Bio-Rad Laboratories, Roche, Applied Biosystems) or reagents only (Eurogentec, Primer Design), developing PCR data analysis software (Biogazelle, LabonNet and MultiD Analysis) and organizing qPCR courses (Biogazelle, TA

*To whom correspondence should be addressed. Tel: +32 9 3320158; Fax: +32 9 3326549; Email: info@rdml.org or jan.hellemans@ugent.be

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

TAA Biocenter), as well as a number of opinion leaders, core facility managers and researchers active in the field of qPCR, including the developers of geNorm, qBase, RTprimerDB and REST. At the time of writing, one company (Biogazelle) and two academic projects (Primer3Plus and RTprimerDB) have committed themselves to provide RDML compatibility in their software at short notice.

The RDML consortium is active in the development of (i) appropriate terminology, (ii) guidelines on minimum information for biological and biomedical investigations: MIQE (Minimum Information for Publication of Quantitative Real-Time PCR Experiments) (Bustin *et al.*, submitted for publication) and (iii) a flexible and universal data file structure with tools to create, process and validate RDML files. The ultimate goal of RDML and MIQE is to enable straightforward and transparent data exchange between qPCR instruments and third party data analysis software, between colleagues and collaborators—independent of the instruments or software being used—and between experimenters and journals or public repositories (Figure 1). All relevant information about the RDML project is available at http://www.rdml.org/.

TERMINOLOGY

For the efficient exchange of data it is important to speak the same language and to agree upon a common terminology. Due to the scope of the RDML format we will not discuss each term, but rather focus on those elements for which multiple names are in use, elements that can be interpreted in different ways, or whose intended role cannot be fully intuited from the name.

Sample

Sample can be used to refer to different inputs: tissue biopsy, cell culture, RNA extract, cDNA, cDNA dilution, etc. Depending on the interpretation of a sample, data may be processed in a different way (e.g. technical versus biological replicates). In RDML, sample refers to the nucleic acid material that is being added to the PCR reaction mix. As a consequence, technical replicate samples should contain the same name (reactions are performed on the same material), and biological replicates should contain different names (the nucleic acids derived from the different biological replicates are not the same).

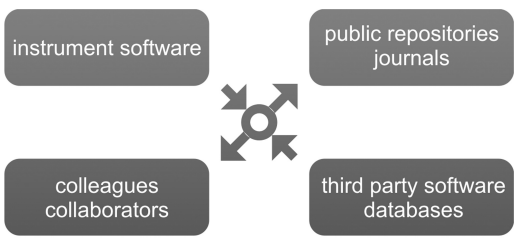

Figure 1. Real-time PCR data exchange with RDML. RDML acts as a facilitator of, and common language for exchange of qPCR data and related information between users, software programs and databases.

Target

Universal term for the nucleic acid sequence to be amplified (including but not limited to genes). We did not use the term gene because it cannot be used for intergenic sequences and it does not allow discrimination between different target sequences of the same gene.

Cq

Depending on the real-time instrument, either threshold cycle (Ct), crossing point (Cp) or a take-off point (Top) are used to refer to the same quantification cycle value (*Cq*): the fractional PCR cycle at which the target is quantified in a given sample.

Reaction

Depending on the real-time instrument a reaction corresponds to a well in a microtiter plate, a glass rotor capillary or a microfluidic reaction volume.

Run

Generic name for a plate, rotor or other physical form containing the data from one single PCR run.

Experiment

An experiment is a collection of runs that need to be analyzed as a single data set.

RDML file structure

Apart from a common terminology, we also developed a standard file structure to create a universal real-time PCR markup language. The RDML standard is based on XML (eXtensible Markup Language), an extensible language especially created to facilitate the sharing of data across different information systems, making it the perfect language in which to implement this standard.

RDML was constructed to accommodate the storage of data from multiple experiments. As can be seen in the simplified overview in Figure 2, the RDML schema basically consists of seven element types at root level, namely the blocks: documentation, ID, sample, target, experimenter, thermal cycling conditions and experiment. The checklist information according to MIQE (Bustin *et al.*, submitted for publication) is mainly stored in *documentation* elements. The *ID* elements contain multiple identifiers supplied by databases or repositories in which the file can be stored. All samples used in the different experiments and relevant information about them are added as *sample* elements. The same applies for the *target* elements that contain information about the genes and other target sequences. A list of experimenters who contributed to one or more experiments can be saved in *experimenter* elements while the *thermal cycling conditions* elements hold PCR programs.

The main part of an RDML file is formed by one or more *experiment* blocks; each block containing the data of one experiment. For each experiment, the actual data is organized and stored on a run-by-run basis using *run* elements. The *run* element is further subdivided into elements containing information about the experimental

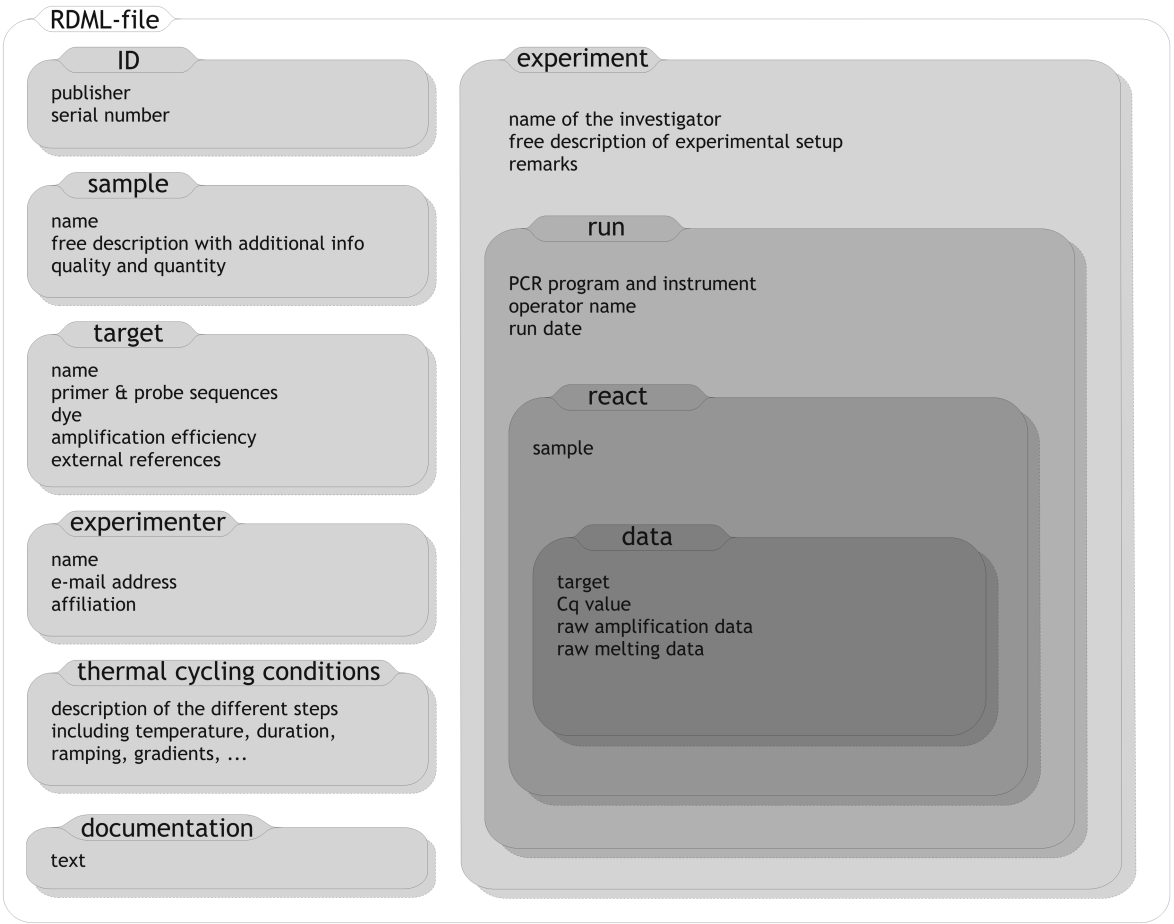

Figure 2. Schematic representation of the RDML XML schema.

setup, the IDs of the experimenters who participated and multiple *reaction* elements associated with a name and the ID of the sample analyzed in that reaction. Using an experimenter's ID that refers to the corresponding personal information in the top-level *experimenter* element, has the benefit of grouping similar information in one place, creating a more database like structure and easier transfer of the data into repositories; it may also make for more compact instance documents if the same information is referred to several times. The same principle was applied to the *thermal cycling conditions*, *sample* and *target* elements by mentioning only their corresponding ID in the *reaction* elements. To support the use of multiplex analysis, different *data* elements, containing quantification (and possibly raw) data can be created for each reaction.

All elements of this data format are optional, making RDML very flexible and widely useful for a multitude of purposes from sharing information about samples to the exchange of raw measurement data. Additionally, the XML nature of RDML allows for straightforward extension with new elements or features to contain extra information if required in the future. A more detailed description and technical information is available on the RDML website (http://www.rdml.org) and on a public open source repository (http://sourceforge.net/projects/rdml/).

GUIDELINES ON MINIMUM INFORMATION

As indicated above, RDML files may contain richly annotated experimental data or just the Cq values for each reaction. RDML files can also be used to exchange sample and target information, experiment layouts or PCR programs. This flexibility has a potential drawback; it allows meaningless RDML files to be created. In addition, it is crucial that data acquisition, analysis and reporting become more transparent to allow reinterpretation and to guarantee compliance with quality standards (4). Therefore, following the example of the microarray community and their MIAME (Minimum Information About a Microarray Experiment) guidelines (1), we propose RDML guidelines as a subset of the Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE).

Due to the complexity and diversity of experiments in which qPCR is utilized, the scope of the RDML guidelines is limited to the PCR technology itself, which means

these can be easily integrated into other minimum information guidelines that focus on the wider experimental context, such as MIQE or MIAME. To coordinate this effort, the RDML consortium recently joined the MIBBI project (Minimum Information for Biological and Biomedical Investigations) (5). The minimum information guidelines have been kept minimal to facilitate the creation of RDML guidelines compliant data files that create the least demand on researchers' time, while requiring sufficient information for other researchers to interpret and reanalyze the data contained within a RDML guideline-compliant RDML file. We once again stress that these guidelines describe the *minimum* information, rather than the desirable *optimal* information; i.e. the information collected is only sufficient for a subset of all possible usage scenarios.

An RDML guidelines compliant data file should contain all measured data as well as information about the samples and targets being analyzed. In addition, data must be linked to samples and targets in an unequivocal way.

Data

A *Cq* value should be provided for each reaction well in the experiment. In multiplex experiments, results should be provided for every color being measured. As an alternative, or in addition, amplification data may be provided along with the settings that have been used to obtain *Cq* values. The latter approach is most relevant if amplification plots are used for other means than just the determination of *Cq* values (e.g. quality control or determination of reaction specific amplification efficiencies). These data must be linked to samples and targets in an unequivocal way.

Sample information

Each sample should contain a unique name or sample identifier to allow the unequivocal assignment of results to samples. Sample types (unknown, negative or positive control, standard) should be included as well because they determine how the data should be processed. For standard samples, absolute or relative input quantities should be given.

Target information

Each target should contain a unique name or target identifier to allow the unequivocal assignment of results to targets. Information should be included to give meaning to the targets. This can be an official gene symbol, a reference to RTPrimerDB, a commercial assay or a name that is further explained in a paper or in additional annotation fields. As for samples, information should be provided about the target type. Target types (target of interest, reference target) are required to allow proper calculation of normalized relative quantities.

TOOLS

Various web tools, implemented using PHP and JavaScript, have been made available by the RDML consortium to aid researchers in creating valid RDML files. These tools, together with the RDML XML schema and example data files, are freely available at http://www.rdml.org. In addition, programming libraries (for Java, C,.NET 2.0/3.5 and PHP) for reading and creating RDML in third party software are in preparation and will be made available via the webpage in the coming months.

RDML GENERATOR

Awaiting broad availability of third party RDML compliant software, we developed a web application for creating RDML files (http://www.rdml.org/chooseTool.php?new). General information concerning the experiment such as a list of experimenters, samples or targets can be supplied by uploading one or multiple files. The actual experimental data, with additional information about the run or the way the data have been analyzed, can be submitted in one of several formats. After data submission, users are allowed to add a new run to an existing experiment or to create additional experiments before the RDML generator compiles the data into a downloadable RDML file.

RDML VALIDATOR

A validator (http://www.rdml.org/validate.php) is available to check if an RDML file, whether created manually, using the online RDML generator or with third party software, is consistent with a certain version of the RDML schema. This tool reports possible errors by displaying messages that can be used to change and revalidate the RDML file until it complies with the RDML schema. For submission of real-time qPCR data to a peer-review journal, an additional validator schema is available to check whether the minimum information as specified in MIQE is contained in the RDML file.

DISCUSSION

RDML is intended to facilitate the exchange and storage of real-time qPCR data by means of a universal data standard that serves researchers and promotes the development of third party software. This article has summarized the rationale underpinning RDML, described its various parts, and has presented the web applications and other tools that have been developed to support it. We have also described the consensus across the consortium regarding the minimum information guidelines.

We have tried to provide support for any type of information that may be relevant in a qPCR experiment in order to address the needs of the widest possible range of users. However, we do realize that the current RDML format may not be seen as complete or optimal by all. We therefore welcome any feedback on the current version of RDML. All suggestions will be evaluated by their value for the qPCR community, and will be taken into account during development of future versions of RDML.

The public acceptance and implementation of these candidate standards (RDML and MIQE) is expected to have major benefits for all users of the qPCR technology. Once established, RDML will allow users to transfer qPCR data to the analysis software of their choice easily, regardless of the instrument used for the original measurement. An established data format standard will also promote the development of innovative third party software providing researchers with a wider array of tools. Researchers will be able to exchange their data with their colleagues or with collaborators in large multi-center studies using a mix of different instruments and data analysis tools. Furthermore, RDML will facilitate the development and operation of public repositories for expression data derived from qPCR experiments much like the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) for microarray data. This will become increasingly important as qPCR experiments scale up, with instruments generating more than 100 000 data points in a single run. A standard qPCR data format and common guidelines on minimal information along with standard-compliant databases will also greatly facilitate the submission of real-time qPCR data as supplemental data for papers allowing other researchers to reinterpret the data or to perform a meta-analysis increasing the value of such publications for all.

ACKNOWLEDGEMENTS

We would like to thank all the members of the RDML consortium for their support and constructive suggestions.

FUNDING

Netherlands Genomics Initiative Horizon Project (050-71-052 to A.U., R.G.); the UK Natural Environment Council's Environmental Bioinformatics Centre. (to C.T.); the Fund of Scientific Research Flanders (FWO) (to J.H., J.V); a postdoctoral grant grant of the Ghent University Special Research Fund (BOF) (to F.P.); a doctoral grant of the Ghent University Special Research Fund (BOF) (to S.L.). Funding for open access charge: BOF-UGENT.

Conflict of interest statement. D.R.P. is employed by Bio-Rad Laboratories (USA), a manufacturer of real-time PCR instruments. J.V. and J.H. are the founders of Biogazelle (Belgium), a developer of real-time PCR data-analysis software. The other authors have no conflicts of interests.

REFERENCES

1. Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.
2. Brazma,A. and Parkinson,H. (2006) ArrayExpress service for reviewers/editors of DNA microarray papers. *Nat. Biotechnol.*, **24**, 1321–1322.
3. Edgar,R. and Barrett,T. (2006) NCBI GEO standards and services for microarray data. *Nat. Biotechnol.*, **24**, 1471–1472.
4. Bustin,S.A. (2008) Molecular medicine, gene-expression profiling and molecular diagnostics: Putting the cart before the horse. *Biomarkers Med.*, **2**, 201–207.
5. Taylor,C.F., Field,D., Sansone,S.A., Aerts,J., Apweiler,R., Ashburner,M., Ball,C.A., Binz,P.A., Bogue,M., Booth,T. *et al.* (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.*, **26**, 889–896.