An RT-qPCR Data Analysis Platform

Thomas Krause[a], Elena Jolkver[a], Sebastian Bruchhaus[a], Michael Kramer[b] and Matthias Hemmje[c]

[a]University of Hagen, Faculty of Mathematics and Computer Science, Hagen, Germany
[b]ImmBioMed Business Consultants GmbH & Co. KG, Pfungstadt, Germany
[c]Research Institute for Telecommunication and Cooperation (FTK), Dortmund, Germany

Abstract

Gene expression data produced by RT-qPCR instruments is becoming increasingly important in laboratory diagnostics. The evaluation and management of this data often involve manual steps, as current software does not map the entire laboratory process. Workflow management systems (WMS) offer flexibility and expandability and can thus take into account individual requirements in laboratories. In addition to the mere evaluation of individual gene expression data, the requirements also include, for example, the linking with medical data, the generation of reports, the storage for meta-analyses, and the archiving of all relevant data. In addition, regulatory requirements such as those of the "In Vitro Diagnostic Regulation" (IVDR) must be taken into account for all of these requirements. This paper proposes a conceptual architecture consisting of a WMS for processing data, as well as an ecosystem of different components for the more advanced requirements. It also outlines a planned prestudy as the next step towards the implementation of such a system.

1. Introduction and Motivation

Genomics is an elementary branch of biology that deals with the genetic material of organisms (the "genome"). The subfield of genetics deals with individual sections of the genome, the "genes" and their interactions. Genes are information carriers that describe the structure of proteins, which are produced in a cell to carry out elementary functions.

"Gene expression" describes the mechanism, by which the information of a gene is transferred into protein. The extend and pattern of which genes (from the multitude of genes in the genome) are being "expressed" in a cell at a certain time point defines the pattern and degree of cellular protein production and thus the potential involvement of these proteins in cellular physiology and pathophysiology. The degree of the expression of a given gene depends on the cell type as well as its activation and differentiation status. Finally, gene expression is subject to external influences.

Therefore, the measurement of gene expression in cells can provide important diagnostic information and as such is being developed into a promising part of medical diagnostics being performed by appropriately equipped medical laboratories.

The processing of gene expression data in laboratories is supported by IT systems. Such IT systems aim to achieve a high degree of automation in order to reduce sources of error, increase throughput, reduce turnaround times, and allow for monitoring of the entire process under aspects of quality assurance. In this context, special regulatory challenges apply, which are specified, e.g. in the EU

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland
thomas.krause@fernuni-hagen.de (T. Krause); elena.jolkver@studium.fernuni-hagen.de (E. Jolkver); sebastian.bruchhaus@fernuni-hagen.de (S. Bruchhaus); m.kramer@immbiomed.de (M. Kramer); mhemmje@ftk.de (M. Hemmje)
0000-0003-4912-1703 (T. Krause); 0000-0003-4711-6708 (E. Jolkver); 0000-0002-7783-2636 (S. Bruchhaus); 0000-0001-8293-2802 (M. Hemmje)

Regulation for In Vitro Diagnostics (IVDR) [1], in specific DIN/ISO documents [2], or in the so-called "MIQE guidelines" [3].

1.1. Reverse-Transcription Quantitative PCR

Within a cell, the genes are defined sections of the DNA sequence. These can be "read" and transferred into messenger RNA ("mRNA") by certain enzymes. This process is called "transcription." Only in a second step, the mRNA sequence is being "translated," into an amino acid sequence (the protein).

The extent to which genes are transcribed and translated varies from cell-type to cell-type as well as from the differentiation/activation status of the respective cell type. The latter not only depends on internal factors within the respective cell type but also a multitude of factors in the cellular micro-environment, whether they stem from endogenous processes within the body or environmental factors (like food ingredients, environmental toxins, or physical factors, like irradiation).

The strength of gene expression can be measured by determining how many transcription products are currently present as mRNA in a sample at a specific time point. Various methods have been established for this purpose. One of these methods is reverse-transcription quantitative PCR ("RT-qPCR")[1].

For RT-qPCR, mRNA is first transcribed back into DNA ("reverse transcription") and then amplified in several cycles by polymerase chain reaction (PCR). By using sequence-specific so-called "primers" the amplification process is restricted to a specific target sequence corresponding to the gene of interest. By using fluorescent indicators ("probes"), the increase of DNA can be measured quantitatively during PCR cycling, i.e. in "real-time." This real-time approach is in contrast to so-called "end-point methods" where the DNA is only detected after completing a certain number of amplification rounds which makes them unsuitable for optimal expression quantification.

1.2. Delta Delta Cq method

RT-qPCR instruments process several samples and genes in parallel in different test tubes ("wells"). With each amplification cycle, the fluorescent signals of the wells are recorded. Plotted as a curve the fluorescent signal is exponential if a sample contains the target sequence as each cycle approximately doubles the amount of DNA. In the first cycles, the signal is still so weak that it is masked within the method-related background noise. Only when a certain threshold value is exceeded does the signal stand out clearly from the background noise. The corresponding cycle number at which the signal exceeds the threshold is known as the C_q value [5].[2] The higher the number of initial copies in a sample, the fewer cycles are needed to exceed the threshold and the lower the associated C_q value. The C_q value thus represents a measurable unit that correlates with gene expression.

Unfortunately, C_q values cannot be directly compared between different runs, samples, genes, instruments, or laboratories, as it is also dependent on many other factors. E.g. the measured value depends on how diluted the sample is or even how effective the amplification process is, which in turn depends on the equipment and reagents used. For better comparability, the measured C_q values can first be compared to a control gene for which the degree of activation is known and as constant as possible. Such genes are often termed "housekeeping genes." Both genes can be observed in parallel for the same sample. Taking the difference between the C_q value calculated for the control gene and

[1] Sometimes misunderstandings arise from the fact "rt" could stand for "real-time" or for "reverse transcription." In fact, the method combines both. To avoid confusion the term "RT-qPCR" should be used [4] per MIQE guidelines. The term RT-PCR should be used to describe PCR after reverse transcription (i.e. starting from (m)RNA) and not real-time PCR. Adams has provided a recent educational article on the aspects of "RT-PCR, qPCR, and RT-qPCR" [4].

[2] The symbol C_t is also often found in the literature, the MIQE guidelines [3] recommend the use of C_q though

the C_q value for the target gene, the ΔC_q value is obtained, which expresses whether the target gene has stronger or weaker expression [5]. The value thus determined can be better compared across different samples and runs on the same instrument.

From a scientific and diagnostic perspective, it may be interesting to compare the ΔC_q values for the gene of interest with control samples to determine differences in gene expression. For example, to test whether a particular drug affects gene expression and to what extent it does so, a treated sample and an untreated sample can each be compared with a control sample. To do this, one again calculates a difference between the ΔC_q values of the samples under investigation for the target genes and the ΔC_q values in the control sample for the same target genes. The results are $\Delta\Delta C_q$ values expressing relative enhancement or attenuation of gene expression compared to the control [5]. This method is therefore referred to as the $\Delta\Delta C_q$ method.

Since the C_q values and the delta values are logarithmic (each increment by one corresponds to a halving of the gene sequence under study), it is often desirable in practice to bring these values back to a linear scale. For this purpose, the conversion $2^{-\Delta\Delta C_q}$ can be used to determine the so-called "fold change" value [5]. On this linear scale, a value of 1 then graphically corresponds to "no change in gene expression of the target gene" compared to the control sample, while, for example, a value of 2 would express a doubling compared to the control sample.

The fold change formula above assumes that the PCR efficiency of the genes of interest as well as the control gene is the same. A more accurate calculation can be performed, when the individual efficiencies are measured and taken into account. The formula also assumes a single control gene. The use of only one control gene is however insufficient for most use cases [6]. To address both shortcomings the formula

$$\frac{E_{goi}^{\Delta C_{q,goi}}}{\sqrt[n]{\prod_z^n E_{ref_z}^{\Delta C_{q,ref_z}}}}$$

should be used, where $\Delta C_{q,ref_z}$ are the ΔC_q values of the control genes and E refers to the PCR efficiencies of the target and control genes respectively. Several algorithms have been developed to select the best reference genes from a set of candidates [7].

1.3. Processing Pipelines

Data analysis for genomics can involve a multitude of computer programs written in different programming languages. A flow of data between these programs orders them, typically forming a directed acyclic graph (DAG) of processing steps. The input for each step is an output of either its antecedent step or the original data source. The data is funneled through a metaphorical processing pipeline. Standardized data formats accomplish a separation of concerns between the individual steps. Such steps may be regarded as functional units or black boxes. They are the building blocks of an abstract automated workflow.

For example, a simple pipeline for qPCR might consist of a step that reads in a file in a particular format and transforms it into a uniform schema. A second step could then analyze the data based on this uniform schema and produce a result. The first processing step might later be replaced with another supporting a different data source and input format.

The considerable effort of orchestrating and documenting such tasks has prompted the development of many Workflow Management Systems (WMS) [8]. They are used for scalable batch processing of recurring interdependent tasks on-premise and in the cloud. WMSs handle monitoring, logging, dependency resolution, and deployment of workflows. Furthermore the abstract, highly regular DAG

structure of workflows reduces complexity. This facilitates the transparency and reproducibility of analytical results. Workflow definition languages like CWL and OpenWDL [9] allow versioning, publishing, and sharing of workflows. Some WMSs are specifically intended for the analysis of biomedical data, e.g. the Galaxy platform [10] for which some RT-qPCR tasks have already been developed [11].

WMS can be used to establish traceability, accountability, and data integrity. They log different kinds of metadata for documentation and monitoring purposes. Monitoring, in particular, is necessary for corrective and preventive actions (CAPA). These are part of the quality management requirements in the IVDR and the EU's proposed rules on artificial intelligence [12].

1.4. Requirements for Analysis Software

To understand the field of qPCR diagnostics and the real world requirements for analysis software in laboratories, a board-certified physician for laboratory medicine ("clinical pathologist"), head of a clinical pathology lab active in indirect patient care, and head of a laboratory consulting firm, Professor Michael Kramer, was invited to co-author this paper and to provide expert knowledge on the field.

Based on this practical experience, an initial review of relevant literature, legal frameworks, and industry guidelines, several requirements were identified that are essential for efficient use of qPCR analysis software in a diagnostic setting:

- Where possible, tasks should be automated to save time, increase throughput, and reduce errors
- All basic user interactions with the system should happen through graphical user interfaces as laboratory staff cannot be assumed to be IT experts
- Data processing and analysis should be extendable as well as customizable to react to new needs
- Integration with existing Laboratory Information Management System (LIMS) should be possible
- Basic analysis of RT-qPCR data should support relative quantification analysis (e.g. $\Delta\Delta C_q$ method [5])
- Relative quantification analysis should support multiple control genes as well as variable PCR efficiencies
- Extensibility should allow additional analysis methods such as standard curve and copy-number-variation [13]
- Analysis runs need to be reproducible
- Results need to be stored in a way to allow subsequent additional analyses
- Quality control is essential [14] and required by legal frameworks
- All results should be easily archivable, data formats must remain readable even after many years
- Legal requirements and industry guidelines like the IVDR and MIQE should be followed where possible

The last point is especially important. To be able to use the solution for diagnostic purposes in the European Union the IVDR has to be taken into account starting next year. Among many other requirements, the IVDR mentions reproducibility, reliability, and performance as key targets for software. It also requires solutions to be developed according to the state of the art and taking into account the software lifecycle, risk management, IT security, verification, and validation. Besides technical requirements, it demands several administrative measures such as naming a person responsible for compliance with suitable professional experience.

2. State of Art

qPCR instruments are supplied with basic software packages that enable the measurement of fluorescent changes, calculations of quantification cycle (C_q) values, and relative quantity determination [15]. Not all devices support the researcher with additional features, like qPCR efficiency correction, normalization to multiple reference genes, averaging and statistical tests. Therefore, several instrument-independent tools have been developed, and most of them can roughly be categorized as software running under Windows, web-based tools, or packages for the R computing environment [16].

For the determination of relative expression levels, the C_q value is first calculated from the raw fluorescence values. For this, several pre-processing steps might be needed like noise reduction, curve smoothing, removal of outliers, normalization, and curve fitting. Based on C_q, the different gene expression levels need to be normalized on one or multiple references and the final $\Delta\Delta C_q$ value determined. The ability of the software to handle error propagation in these steps as well as the subsequent support in statistical analysis and graphical visualization is of high importance for lab staff lacking statistical experience. Ideally, the software supports all of the aforementioned steps and is compliant with MIQE guidelines. Such an all-rounder is the qpcR package for R [17], which, combined with R's rich statistical and graphical package infrastructure can provide all required analyses. However, the usage of R requires some proficiency in scriptwriting, which can impose an initial barrier for less programming-savvy researchers. Out of the various tools developed for qPCR analyses, only a few support multiple of the aforementioned calculations.

QPCR [18] comprises a parser to import generated data from qPCR instruments and includes technical and biological replicate handling, incorporation of gene-specific efficiency, normalization using single or multiple reference genes, inter-run calibration, and fold change calculation. Moreover, the application supports the assessment of error propagation throughout all analysis steps and allows conducting statistical tests on biological replicates. Results can be visualized in customizable charts and exported for further investigation.

qBase [7] contains several algorithms for reference gene selection, normalization, quality control, inter-run-calibration, as well as copy number variation analysis. For reference gene selection and normalization, it offers the geNorm algorithm. geNorm ranks candidate genes by their stability and then picks a suitable number of genes to be used [6]. For normalization, a geometric averaging of these selected reference genes is then used. Other tools specifically for the selection of reference genes include BestKeeper [20] and NormFinder [21]. Both are available as plugins for Microsoft Excel.

While the analysis and normalization procedures offered by these tools are useful, they often only represent only a small part of the overall workflow from instrument setup to result visualization and beyond. Software included with the instruments is also usually proprietary so that they cannot be easily modified or used with other instruments [16]. Most tools seem to be tailored for research purposes and not for diagnostics, requiring several manual steps instead of automatization and integration. Usage for diagnostics in the European Union would also require compliance with IVDR which as of today is not available for the majority of tools.

3. Prestudy Design and Research Questions

Based on the perceived lack of freely available software that can be used for diagnostic purposes, considering especially the upcoming hardened IVDR legislation, it seems worthwhile to design, implement and certify such a system on an open-source base, to allow easier collaboration between

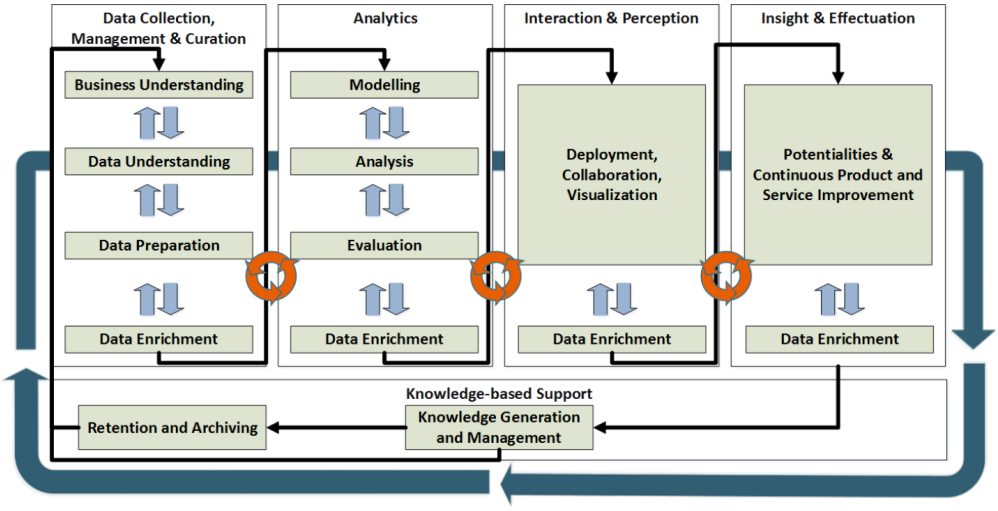

Figure 1: CRISP4BigData Reference Model

research teams and to reduce the time needed for new research to be used in diagnostics. While some basic requirements, existing software, and legal requirements have been outlined, further investigation into the current processes and resulting requirements seems necessary before implementation is possible. To support this, a prestudy is currently being planned that should answer the following research questions:

- What does a typical analysis process for RT-qPCR data look like in a laboratory?
- Which manual activities can be automated?
- What software already exists to automate the entire process or individual parts of it?
- What regulatory and technical requirements must be considered?
- How must a system be structured to meet the requirements and enable a high degree of automation?

4. Proposed Conceptual Architecture

While a detailed technical design and implementation require the prestudy that was outlined, it is possible to outline a conceptual architecture on a high level based on the known requirements so far, which can then be further detailed in the prestudy. For this goal the CRISP4BigData [22] reference model shown in fig. 1 was used as a basis. CRISP4BigData combines the Cross-industry Standard Process for Data Mining (CRISP-DM) with the IVIS4BigData model [23]. Figure 2 shows the conceptual architecture derived from the requirements and mapped to the reference model.

At the core of the architecture lies a WMS that allows a dynamic processing workflow where each step and the workflow as a whole can be customized. As described previously there are already well-known, good working tools for some parts of the qPCR workflow. Rebuilding these from scratch would be very time-consuming. Using a configurable workflow allows the reuse of these tools.

The processing workflow that is shown in fig. 2 has been separated into distinct phases. These phases map directly to the CRISP4BigData reference model and should be seen as a general template for a typical qPCR analysis process that can be customized by the laboratory. In practice, they will consist of one or more concrete processing tasks executed in sequence or parallel inside the WMS.

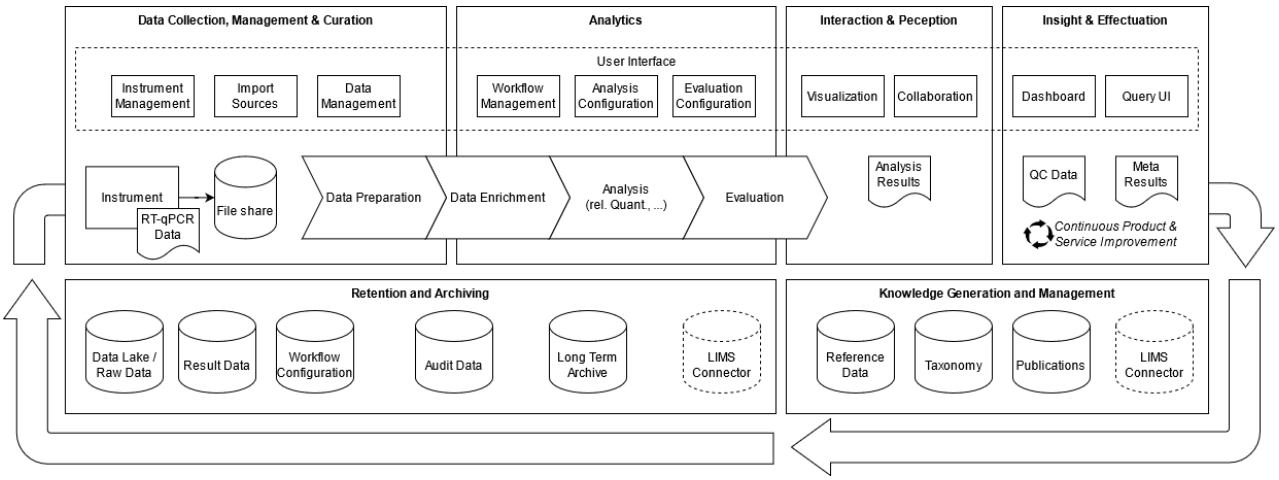

Figure 2: Conceptual Architecture

The workflow starts after data created by the instrument is read automatically from a file share or uploaded manually from the user interface. Analysis settings can be specified at upload time or pre-configured. During import, the data is transformed into an internal unified data structure that allows subsequent processes to work independently of the original data format (part of the "Data Preparation" phase). Data from previous runs is aggregated when required to allow for inter-run calibration and further augmented by loading reference data or computing intermediate results required by subsequent analysis in the "Data Enrichment" phase. The "Analysis" phase contains the analysis steps such as determining the relative quantification amounts. In the last phase ("Evaluation") the analysis results are interpreted and transformed into desired outputs.

The artifacts created by the workflow are available for visualization and collaboration via a user interface. The same user interface is also available to administer the system as a whole, which can be seen in the upper part of fig. 2. Long-term insights from several runs as well as quality control measures and monitoring can be found in the "Insights & Effectuation" layer. New Knowledge discovered at this level, like suitable reference genes for analysis, can be stored for future use as part of the "Knowledge Generation and Management" layer.

Lastly, all input data, output data, intermediate results, and metadata such as the processing parameters used are persisted to enable reproducibility of the results at any time later. Legal requirements regarding storage and deletion periods are ensured by technical and organizational measures. These data stores are found at the bottom of the diagram as part of the "Retention and Archiving" layer. For long-term storage, suitable standards such as OAIS [24] are used to ensure future readability. For integration with existing LIMS, a connector is displayed in the same layer to accommodate the need for synchronization between the systems.

Some of the aspects mentioned above such as reproducibility are also part of IVDR compliance. Using a layered architecture with individual components can also help to provide other aspects such as reliability and performance because components can be used to parallelize workloads or to provide redundancy. The prestudy will aim to address additional points of the IVDR.

5. Evaluation

The conceptual architecture can be evaluated on a preliminary basis by comparing the identified requirements with the proposed solution. Further evaluation is planned at a later stage during the

prestudy.

The requirement of a graphical user interface has been addressed in the architecture by including it as an explicit component. The desire for flexible and configurable processing and integration with existing software is fulfilled by using a WMS as the core of the solution. This also partially addresses the subsequent points for the specific types of analysis to be used as the workflow approach allows any kind of analysis to run in principle. Furthermore, it was proposed to use existing tools and algorithms where possible inside the workflow which would allow the mentioned analysis methods to be executed. WMSs also facilitate reproducibility as the concrete settings used in a workflow are easily saved to allow re-execution. Storage for later analysis and for archiving is part of the architecture. For archival a specific format was suggested. Lastly, compliance with the IVDR was discussed. The specific requirements for compliance are a complex topic and need to be evaluated separately.

References

[1] The European Parliament and the Council of the European Union, In vitro diagnostic regulation: IVDR, 2017. URL: http://data.europa.eu/eli/reg/2017/746/2017-05-05.

[2] DIN, Medizinische laboratorien - anforderungen an die qualität und kompetenz, 2014.

[3] S.A. Bustin, V. Benes, J.A. Garson, J. Hellemans, J. Huggett, M. Kubista, R. Mueller, T. Nolan, M.W. Pfaffl, G.L. Shipley, J. Vandesompele, C.T. Wittwer, The MIQE guidelines: Minimum information for publication of quantitative real-time PCR experiments, Clinical Chemistry, 55 2009 611–622. doi:10.1373/clinchem.2008.112797.

[4] G. Adams, A beginner's guide to RT-PCR, qPCR and RT-qPCR, The Biochemist, 42 2020 48–53. doi:10.1042/BIO20200034.

[5] K.J. Livak, T.D. Schmittgen, Analysis of relative gene expression data using real-time quantitative PCR and the 2(-delta delta c(t)) method, Methods, 25 2001 402–408. doi:10.1006/meth.2001.1262.

[6] J. Vandesompele, K. de Preter, F. Pattyn, B. Poppe, N. van Roy, A. de Paepe, F. Speleman, Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes, Genome Biology, 3 2002 RESEARCH0034. doi:10.1186/gb-2002-3-7-research0034.

[7] J. Hellemans, G. Mortier, A. de Paepe, F. Speleman, J. Vandesompele, qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data, Genome Biology, 8 2007 R19. doi:10.1186/gb-2007-8-2-r19.

[8] J.M. Perkel, Workflow systems turn raw data into scientific knowledge, Nature, 573 2019 149–150. doi:10.1038/d41586-019-02619-z.

[9] J. Vivian, A.A. Rao, F.A. Nothaft, C. Ketchum, J. Armstrong, A. Novak, J. Pfeil, J. Narkizian, A.D. Deran, A. Musselman-Brown, H. Schmidt, P. Amstutz, B. Craft, M. Goldman, K. Rosenbloom, M. Cline, B. O'Connor, M. Hanna, C. Birger, W.J. Kent, D.A. Patterson, A.D. Joseph, J. Zhu, S. Zaranek, G. Getz, D. Haussler, B. Paten, Toil enables reproducible, open source, big biomedical data analyses, Nature Biotechnology, 35 2017 314–316. doi:10.1038/nbt.3772.

[10] B. Grüning, J. Chilton, J. Köster, R. Dale, N. Soranzo, M. van den Beek, J. Goecks, R. Backofen, A. Nekrutenko, J. Taylor, Practical computational reproducibility in the life sciences, Cell Systems, 6 2018 631–635. doi:10.1016/j.cels.2018.03.014.

[11] N. Zanardi, M. Morini, M.A. Tangaro, F. Zambelli, M.C. Bosco, L. Varesio, A. Eva, D. Cangelosi, PIPE-t: A new galaxy tool for the analysis of RT-qPCR expression data, Scientific Reports, 9 2019 17550. doi:10.1038/s41598-019-53155-9.

[12] European Parliament and Council, Proposal for a regulation laying down harmonised rules on artificial intelligence (artificial intelligence act), 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

[13] C.A. Heid, J. Stevens, K.J. Livak, P.M. Williams, Real time quantitative PCR, Genome Research, 6 1996 986–994. doi:10.1101/gr.6.10.986.

[14] S.C. Taylor, K. Nadeau, M. Abbasi, C. Lachance, M. Nguyen, J. Fenrich, The ultimate qPCR experiment: Producing publication quality, reproducible data the first time, Trends in Biotechnology, 37 2019 761–774. doi:10.1016/j.tibtech.2018.12.002.

[15] S. Bustin, A. Bergkvist, T. Nolan, In silico tools for qPCR assay design and data analysis, Methods in Molecular Biology (Clifton, N.J.), 760 2011 283–306. doi:10.1007/978-1-61779-176-5_18.

[16] S. Pabinger, S. Rödiger, A. Kriegner, K. Vierlinger, A. Weinhäusel, A survey of tools for the analysis of quantitative PCR (qPCR) data, Biomolecular Detection and Quantification, 1 2014 23–33. doi:10.1016/j.bdq.2014.08.002.

[17] A.-N. Spiess, qpcR: Modelling and analysis of real-time PCR data, 2018. URL: https://CRAN.R-project.org/package=qpcR.

[18] S. Pabinger, G.G. Thallinger, R. Snajder, H. Eichhorn, R. Rader, Z. Trajanoski, QPCR: Application for real-time PCR data management and analysis, BMC Bioinformatics, 10 2009 268. doi:10.1186/1471-2105-10-268.

[19] What makes qbase+ unique?, 18.09.2018. URL: https://www.qbaseplus.com/features.

[20] M.W. Pfaffl, A. Tichopad, C. Prgomet, T.P. Neuvians, Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper–excel-based tool using pair-wise correlations, Biotechnology Letters, 26 2004 509–515. doi:10.1023/B:BILE.0000019559.84305.47.

[21] C.L. Andersen, J.L. Jensen, T.F. Ørntoft, Normalization of real-time quantitative reverse transcription-PCR data: A model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets, Cancer Research, 64 2004 5245–5250. doi:10.1158/0008-5472.CAN-04-0496.

[22] K. Berwind, M.X. Bornschlegl, M.A. Kaufmann, M. Hemmje, Towards a cross industry standard process to support big data applications in virtual research environments, in: Proceedings of the Collaborative European Research Conference (CERC) 2016, 2016.

[23] M.X. Bornschlegl, F.C. Engel, R. Bond, M.L. Hemmje, eds., Advanced visual interfaces. Supporting big data applications: AVI 2016 workshop, AVI-BDA 2016, bari, italy, june 7-10, 2016, revised selected papers, Springer International Publishing, Cham; s.l., 2016. doi:10.1007/978-3-319-50070-6.

[24] CCSDS Secretariat, Reference model for an open archival information system (OAIS). Recommended practice, 2012. URL: https://public.ccsds.org/pubs/650x0m2.pdf.