Review and Research on Cancer Treatment
Volume 4, Issue 1 (2018) ISSN 2544-2147

Normalisation of microRNA qPCR results in lung cancer – strategies and problems

Marcin Kaszkowiak

Bartosz Szmyd

Dorota Pastuszak-Lewandoska

Ewa Brzeziańska-Lasota

The corresponding author: dorota.pastuszak-lewandoska@umed.lodz.pl, Dorota Pastuszak-Lewandoska, PhD ,

Department of Biomedicine and Genetics Chair of Biology and Medical Parasitology, Medical University of Lodz 251 Pomorska Street (C-5 building), 92-213 Lodz

Abstract: Lung cancer (LC) is one of the most common and death-causing carcinomas in the world. There is an urgent need for innovative diagnostic tools and treatment strategies. MicroRNAs have a great potential for being reliable LC biomarkers and therapy targets. qPCR is one of the best methods for evaluating their expression during research. However, it has to be performed in a very careful way. Any committed mistakes, especially during normalization of output data may lead to misleading conclusions. Data Normalisation (DS) is a crucial step, required for levelling different sample sizes, volumes of nucleic acids and other varying factors, allowing us to compare results with one another. There is no universal strategy for this process. Every known method (volume size, housekeeping genes, global mean, etc.) has many advantages and drawbacks. This article will discuss different approaches, their pros and cons, and suggest ways to minimize qPCR inaccuracy.
Keywords: qPCR, microRNA, normalization, quantile, global mean

1. Background

Lung cancer incidence and mortality has been steadily growing since 1930s [1]. Nowadays, it has become a global problem. LC was responsible for 13% of all cancer incidences worldwide and caused 160340 deaths in United States [1] in 2012, what made it the most common carcinoma and a leading cause of cancer deaths.

The prognosis of Lung cancer strongly depends on the moment of diagnosis. The 5-year survival rate ranges from 52% (local disease) to 4% (distant disease). It results in relatively high mortality, since only about 15% of LCs are diagnosed at early stages. [2] It raises an urgent need for reliable biomarkers, that would make diagnosing process faster and more accurate.

MicroRNAs are small strands of ribonucleic acid (built by about 23 nucleotides), regulating genes expression at the transcriptomic level. They affect mRNA by either repressing translation or inducing degradation [3]. Many recent studies suggest their significant role in cancer etiopathogenesis [4] and therefore consider them as promising biomarkers and therapeutic targets [5, 8]. That makes an accurate evaluation of microRNAs expression a crucial step for developing new drugs and diagnostic procedures.

2. Technical introduction

Real-time PCR is a commonly used and reliable technique for measuring microRNA relative expression. Among its advantages are sensitivity, high throughput and accuracy. However, as every method it is prone to errors. There are two types of variation in the obtained experimental data: biological (describing real molecular processes) and experimentally-induced (due to experiment handling), that needs to be excluded from further analysis. It is the purpose of data normalization. Reducing the influence of material's quantity, integrity, purity, etc. makes it possible to compare the results from different experiments [9].

3. Classic Normalization methods

3.1. Normalization to sample size

The simplest, but also the least used method is to carefully obtain the same sample volume (or mass, number of cells, etc.) form each patient. This approach may seem obvious and straightforward, but is extremely prone to errors. It is impossible to measure sample size precisely enough. Moreover, biological material often contains a different percentage of microRNA, what makes accurate analysis impossible [10].

3.2. Normalization to total sample RNA

$$\delta'_{ij} = \frac{\delta_{ij}}{r_i}$$

Formula 1. Normalization to total sample RNA. δ' – expression after normalization of *j-th* microRNA in *i-th* sample; δ - expression of *j-th* microRNA in *i-th* sample; r – total RNA in *i-th* sample

Another volume-orientated technique bases on quantification of all RNAs present in a sample. This measurement can be easily performed prior to reverse transcription. Unfortunately, it can be affected by poor material quality. Moreover, this method has two serious disadvantages. Firstly, it is unable to normalize variation formed in further steps of qPCR method. Secondly, it bases on the assumption that proportion if different RNA types (mRNA, rRNA, miRNA, etc.) is constant within all cells, which has been proven wrong [10].

3.3. Normalization to artificial molecule

$$\delta'_{ij} = \frac{\delta_{ij}}{\gamma_i}$$

Formula 2. Normalization to exogenous RNA. δ' – expression after normalization of *j-th* microRNA in *i-th* sample; δ – expression of *j-th* microRNA in *i-th* sample; γ – exteral molecule's expression in *i-th* sample.

This method uses an external molecule as a reference microRNA. Exogenous RNA (usually *C. elegans* cel-miR-39) of known quantity is added to a sample before performing Reverse Transcription. On the contrary to previously described technique, this method allows to reduce variation caused by further qPCR steps, but has no use in normalization of differences in samples' volume and quality [11]. Among its advantages there is also insensitivity to in-cell biological fluctuations. However, this may be considered as a drawback, since it raises a need for result validation [10]. Generation of suitable, not always commercially available external particles may also be a serious problem in smaller laboratories [10].

3.4. Normalization to endogenous reference microRNAs

$$\delta'_{ij} = \frac{\delta_{ij}}{\theta_i}$$

Formula 3. Normalization to "housekeeping microRNA". δ' – expression after normalization of *j-th* microRNA in *i-th* sample; δ – expression of *j-th* microRNA in *i-th* sample; θ – housekeeping microRNA expression in *i-th* sample

Endogenous control is the most popular and most universal tool for a qPCR data normalization [12]. Similarly to gene expression analysis, this method is based on the assumption, that some microRNAs' expression (the so called "housekeeping microRNAs") is close to constant in all samples and experimental conditions [13]. With this hypothesis, other expressions are adjusted using given formula. It allows to reduce variation caused by all stages of an experiment. Unfortunately, this technique has numerous limitations. There are no'100% universal' reference molecules. Many studies report that, on contrary to our assumption, some housekeeping microRNAs / genes expression can significantly vary depending on tissue type and experimental conditions [14, 19]. This makes proper selection of endogenous control a very complex problem and should be solved for each project individually in preliminary research. This step is extremely important, since it is necessary for correct interpretation of results [20, 21].

Many researchers suggest usage of more sophisticated version of this method – considering combination (geometric mean) of a few housekeeping microRNAs expression as stable and reliable control [13]. Candidates for mean arguments can be selected using bioinformatics methods.

$$\delta'_{ij} = \frac{\delta_{ij}}{\sqrt[n]{\prod_{k=1}^{n} \theta_{ki}}}$$

Formula 4. Normalization to multiple housekeeping microRNAs. δ' – expression after normalization of *j-th* microRNA in *i-th* sample; δ - expression of *j-th* microRNA in *i-th* sample; θ – *k-th* housekeeping microRNA expression in *i-th* sample, *n* – number of housekeeping microRNAs

The most commonly used molecules in endogenous microRNA expression normalization are small coding RNAs – RNU6A and RNU6B. Despite being considered as universal housekeeping genes [11], many studies report difficulties in their application – e.g., RNU6B is not applicable in circulating microRNA analysis [14]. RNUs are not microRNAs, which introduces bias due to their different expression profiles and biochemical character. Usage of microRNAs and their combinations as endogenous control is advised.

Among different miRNAs studied, miR-16 is highly and invariantly expressed in different tissues. Bioinformatic analysis has proven its high stability, especially in combination with miR-93. They are followed by miR-191, miR-106a, miR-17-5p and miR-25 [11].

4. Data-driven Normalization methods

4.1. Normalization to global mean

$$\delta'_{ij} = \frac{\delta_{ij}}{\sqrt[n]{\prod_{k=1}^{n} \varphi_{ki}}}$$

Formula 5. Normalization to global mean. δ' – expression after normalization of *j-th* microRNA in *i-th* sample; δ - expression of *j-th* microRNA in *i-th* sample; φ – *k-th* microRNA expression in *i-th* sample, *n* – number of microRNAs

Gold-standard workflow in qPCR data analysis consists of a pilot experiment in search for molecules with stable expression, selecting a few of them and using their expression combination for normalizing the results in further research. This approach often presents many problems and difficulties. Suitable endogenous controls are often either very hard, or even impossible to find.

In response to those difficulties, relatively new method has been developed – global mean normalization. It is based on the observation that in a large and unbiased group of microRNAs, the average expression can be used for normalization that would reduce variation from all steps of the experiment [9].

Studies suggest, that this method either outperforms or is as good as endogenous control in terms of stability. It also allows us to observe real and significant biological changes, that could be blurred when using other normalization techniques [22].

For usage in projects with analysis of smaller microRNA sets (i.e. 4 microRNAs), with help of bioinformatic tools, a few microRNAs or small RNAs controls can be selected that reassemble the mean expression value [22].

4.2. Quantile Normalization algorithm

It is simple, easy-applicable method that bases on simple statistical parameter – quantile. This technique is mainly used in normalization of microarray data [21], but there is no obstacle for usage in qPCR results. It bases on an assumption, that average distribution of gene expression levels in the cell remains constant [13].

The algorithm treats sample as n data points (number of microRNAs), sorts them in ascending/ descending order and transforms *i-th* row in all samples with its mean value. Those steps change expression levels while preserving rank-order (it is an origin of method's name – quantiles remain constant for each sample). The detailed step-by-step description is presented on Figure 1.

Create a matrix *M* with *k* rows (number of microRNAs in the study) and *p* columns (number of samples in a study). Fill this array with your experimental data (microRNA name and its expression).

Create a matrix *M'* by sorting each column in ascending order. It corresponds to a **quantile** distribution of each sample.

Iterate over rows and replace each column with the row's mean value. After this process normalized microRNAs expressions can be rearranged in the starting order.

Figure 1. Quantile normalization

There may occur situations, where all investigated microRNAs for each sample won't fit in one plate. Technical variation occurring between those plates can (and should) be reduced using same algorithm, considering plates as different samples. However, it is very important that examined microRNAs should be distributed randomly across plates. Afterwards, normal procedure may be performed.

4.3. Rank-Invariant Set Normalization algorithm

In "housekeeping" classic expression normalization method, there has to be taken an a priori hypothesis, that one (or more) microRNA /gene expression is constant across all samples. This approach was often proven wrong [14, 19]. In a dataset large enough, suitable microRNAs can be selected from the data itself, after the experiment [13]. Bioinformatic analysis allows us to find rank-invariant (having the same rank across all datasets) microRNAs across all samples, which are verified to be suitable for normalization purpose [23]. The detailed algorithm description is presented on Figure 2.

Create a matrix *M* with *k* rows (number of microRNAs in the study) and *p* columns (number of samples in a study). Fill this array with your experimental data (microRNA name and its expression).

Create a matrix *M'* by sorting each column in ascending order.

Select one microRNA set as a reference, regarding experiment's purpose and methodology. This can be one special patient, commercially available set, global mean or average, etc.

Compare each of *p* samples with *R* and find microRNAs that have the same rank in both datasets. Create a set of genes *S*, by intersecting the results of all pairwise comparisons.

Let α_i be the average expression of rank invariant microRNAs (stored in *S*) in *i-th* sample. Calculate $\beta_i = \frac{\alpha_i}{\alpha_R}$ ratio.

Multiplicate *i-th* column of *M* by β_i. Dataset contains normalized microRNA expression values.

Figure 2. Rank-Invariant Set Normalization

5. Conclusions

There is no universal, perfect normalization method. All techniques have many advantages and also drawbacks. While Data-Driven methods are more accurate, cheaper, universal, and easy to use in large, genom-wide experiments, they may be inaccurate, or even not-applicable in smaller projects. Classic methods, especially "Housekeeping microRNA normalization" are considered gold-standard for normalizing qPCR results. Among its advantages are low-price and easy usage even in single-gene research. They are, however, extremely prone to errors and may cause a serious misinterpretation of expression data.

The perfect normalization method has yet to be discovered. Nowadays, suitable technique should be chosen individually for each experiment, after cautious pros and cons consideration, literature study and public databases analysis. Careful selection is crucial for research to succeed.

Table 1. List of key features of each normalization method. Shading: Classic Normalization Methods, Data-Driven Normalization Methods

normalization method	+	-
sample size	Cheap Straightforward	Extremely prone to errors Impractical
total sample RNA	Easy to apply Reduces Reverse Transcription Bias	Prone to mechanical errors Based on wrong assumptions
artificial molecule	Insensitive to biological fluctuations Precise	Expensive Does not reduce variation created before Reverse Transcription step
housekeeping microRNAs	Reduces variation from all steps of experiment Popular Easily applicable and replicable	Prone to errors Not universal Based on wrong assumptions
global mean	Reduces variation form all steps of experiment Universal Easy to use Accurate Based on reasonable assumptions	Accuracy dependent on study size
quantile normalization	Reduces variation from all steps of the experiment Accurate Universal Straightforward Based on reasonable assumptions Normalizes cross-plate variation	Accuracy dependent on study size Requires random microRNA distribution in a dataset
Rank-invariant set	Combines advantages of housekeeping microRNAs method and data-driven techniques Applicable post-hoc Allows to identify suitable endogenous control	Accuracy dependent on study size

References

1. Mao Y., Yang D., He J., Krasna M. J., *Epidemiology of Lung Cancer*, Surgical Oncology Clinics of North America, 25 (2016), s. 439-445.
2. Liam C.-K., Andarini S., Lee P., Ho J. C.-M., Chau N. Q., Tscheikuna J., *Lung cancer staging now and in the future*, Respirology, 20 (2015), s. 526–534.
3. Wahid F., Shehzad A., Khan T., Kim Y. Y., *MicroRNAs: Synthesis, mechanism, function, and recent clinical trials*, Biochimica et Biophysica Acta - Molecular Cell Research, 1803 (2010), s. 1231-1243.
4. Leva G. Di, Garofalo M., Croce C. M., *microRNAs in cancer*, Annu Rev Pathol, 9 (2014), s. 287-314.
5. Berindan-neagoe I., Monroig P., Pasculli B., George A., Medicine T., Hatieganu P. I., Juan S., Rico P., Sciences P., *MicroRNAome genome: a treasure for cancer diagnosis and therapy*, 64 (2015), s. 311-336.
6. Zhao Q., Li P., Yu J. M. X., *MicroRNAs in Lung Cancer and Lung Cancer Bone Metastases: Biomarkers for Early Diagnosis and Targets for Treatment*, Recent Patents on Anti-Cancer Drug Discovery, 10 (2015) s. 182–200.
7. Kong X.-M., Zhang G.-H., Huo Y.-K., Zhao X.-H., Cao D.-W., Guo S.-F., Li A.-M., Zhang X.-R., *MicroRNA-140-3p inhibits proliferation, migration and invasion of lung cancer cells by targeting ATP6AP2.*, International journal of clinical and experimental pathology, 8 (2015), s. 12845-52.
8. Yang H., Tang Y., Guo W., Du Y., Wang Y., Li P., Zang W., et al., *Up-regulation of microRNA-138 induce radiosensitization in lung cancer cells*, Tumor Biology, 35 (2014), s. 6557-6565.

9. Vandesompele J., *From reference genes to global mean normalization Critical elements contributing to successful qPCR results*, (2012).
10. Huggett J., Dheda K., Bustin S., Zumla A., *Real-time RT-PCR normalisation; strategies and considerations*, Genes and Immunity, 6 (2005), s. 279-284.
11. Schwarzenbach H., Calin G., Pantel K., *Which is the accurate data normalization strategy for microRNA quantification?*, 61 (2016), s. 1333-1342.
12. Schwarzenbach H., Da Silva A. M., Calin G., Pantel K., *Data normalization strategies for microRNA quantification*, Clinical Chemistry, 61 (2015), s. 1333-1342.
13. Mar J. C., Kimura Y., Schroder K., Irvine K. M., Hayashizaki Y., Suzuki H., Hume D., Quackenbush J., *Data-driven normalization strategies for high-throughput quantitative RT-PCR*, BMC Bioinformatics, 10 (2009).
14. Xiang M., Zeng Y., Yang R., Xu H., Chen Z., Zhong J., Xie H., Xu Y., Zeng X., *U6 is not a suitable endogenous control for the quantification of circulating microRNAs*, Biochemical and Biophysical Research Communications, 454 (2014), s. 210-214.
15. Schmittgen T. D., i Zakrajsek B. A., *Effect of experimental treatment on housekeeping gene expression: validation by real-time, quantitative RT-PCR*, Journal of Biochemical and Biophysical Methods, 46 (2000), s. 69-81.
16. Thellin O., Zorzi W., Lakaye B., De Borman B., Coumans B., Hennen G., Grisar T., Igout A., Heinen E., *Housekeeping genes as internal standards: Use and limits*, Journal of Biotechnology, 75 (1999), s. 291-295.
17. Ullmannová V., Haskovec C., *The use of housekeeping genes (HKG) as an internal control for the detection of gene expression by quantitative real-time RT-PCR*, Folia biologica, 49 (2003), s. 211-216.
18. Tricarico C., Pinzani P., Bianchi S., Paglierani M., Distante V., Pazzagli M., Bustin S. A., Orlando C., *Quantitative real-time reverse transcription polymerase chain reaction: normalization to rRNA or single housekeeping genes is inappropriate for human tissue biopsies,* Analytical Biochemistry, 309 (2002), s. 293-300.
19. Dheda K., Huggett J. F., Bustin S. A., Johnson M. A., Rook G., i Zumla A., *Validation of housekeeping genes for normalizing RNA expression in real-time PCR*, BioTechniques, 37 (2004), s. 112-119.
20. Saviozzi S., Cordero F., Lo Iacono M., Novello S., Scagliotti G. V., Calogero A. R., *Selection of suitable reference genes for accurate normalization of gene expression profile studies in non-small cell lung cancer,* BMC Cancer, 6 (2006), s. 1-10.
21. Hellemans J., i Vandesompele J., *Selection of Reliable Reference Genes for RT-qPCR Analysis*, [w:] *Quantitative Real-Time PCR: Methods and Protocols*, (Biassoni R., i Raso A.) New York, NY: Springer New York, 2014, s. 19-26.
22. Mestdagh P., Van Vlierberghe P., De Weer A., Muth D., Westermann F., Speleman F., i Vandesompele J., *A novel and universal method for microRNA RT-qPCR data normalization*, Genome Biol, 10 (2009), s. R64.
23. Tseng G. C., *Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects*, Nucleic Acids Research, 29 (2001), s. 2549-2557.