Published online 17 June 2016

Nucleic Acids Research, 2016, Vol. 44, No. 13 ***5995–6018***
doi: 10.1093/nar/gkw545

SURVEY AND SUMMARY

Toward reliable biomarker signatures in the age of liquid biopsies - how to standardize the small RNA-Seq workflow

Dominik Buschmann[1,2], Anna Haberberger[1], Benedikt Kirchner[1], Melanie Spornraft[1], Irmgard Riedmaier[3,4], Gustav Schelling[5] and Michael W. Pfaffl[1,*]

[1]Department of Animal Physiology and Immunology, TUM School of Life Sciences Weihenstephan, Technical University of Munich, Weihenstephaner Berg 3, 85354 Freising, Germany, [2]Institute of Human Genetics, University Hospital, Ludwig-Maximilians-University Munich, Goethestraße 29, 80336 München, Germany, [3]Department of Physiology, TUM School of Life Sciences Weihenstephan, Technical University of Munich, Weihenstephaner Berg 3, 85354 Freising, Germany, [4]Eurofins Medigenomix Forensik GmbH, Anzinger Straße 7a, 85560 Ebersberg, Germany and [5]Department of Anesthesiology, University Hospital, Ludwig-Maximilians-University Munich, Marchioninistraße 15, 81377 München, Germany

Received April 12, 2016; Revised May 31, 2016; Accepted June 3, 2016

ABSTRACT

Small RNA-Seq has emerged as a powerful tool in transcriptomics, gene expression profiling and biomarker discovery. Sequencing cell-free nucleic acids, particularly microRNA (miRNA), from liquid biopsies additionally provides exciting possibilities for molecular diagnostics, and might help establish disease-specific biomarker signatures. The complexity of the small RNA-Seq workflow, however, bears challenges and biases that researchers need to be aware of in order to generate high-quality data. Rigorous standardization and extensive validation are required to guarantee reliability, reproducibility and comparability of research findings. Hypotheses based on flawed experimental conditions can be inconsistent and even misleading. Comparable to the well-established MIQE guidelines for qPCR experiments, this work aims at establishing guidelines for experimental design and pre-analytical sample processing, standardization of library preparation and sequencing reactions, as well as facilitating data analysis. We highlight bottlenecks in small RNA-Seq experiments, point out the importance of stringent quality control and validation, and provide a primer for differential expression analysis and biomarker discovery. Following our recommendations will encourage better sequencing practice, increase experimental transparency and lead to more reproducible small RNA-Seq results. This will ultimately enhance the validity of biomarker signatures, and allow reliable and robust clinical predictions.

INTRODUCTION TO BIOMARKERS AND LIQUID BIOPSIES

The importance of biomarkers in molecular diagnostics is undisputed. A valid biomarker should be able to reveal a specific biological trait or a measurable change, which is directly associated with a change in the physiological condition of an organism. At the molecular and cellular levels, analysis of gene expression changes is the first step of exploration for any regulatory activity. Activating early response genes is a very dynamic process, allowing the organism to rapidly adapt to external or internal stimuli (1,2). Thus, gene expression profiling is the technique of choice to discover and identify transcriptional biomarkers that describe these changes affecting cells, tissues or the entire organism (3,4). Accessing this molecular information via biomarkers in tiny biopsies is a common procedure for many malignancies, but sampling tissues can be costly, painful and potentially impose additional risks on the patient (5). The readout of transcriptional biomarker signatures from minimally invasive sampling methods is therefore highly valued (6). Sampling patient biofluids, such as blood, urine, sweat, saliva or milk in liquid biopsies is currently being thought of

*To whom correspondence should be addressed. Tel: +49 8161 713511; Fax: +49 8161 713539; Email: michael.pfaffl@wzw.tum.de

as a crucial next step in biomarker research and molecular or clinical diagnostics (7).

The existence of extracellular DNA has been acknowledged for decades, and finds applications ranging from oncology to prenatal diagnostics (8,9). In 2005, the first study indicating the importance of microRNAs (miRNAs) in tumor diagnosis and monitoring was published (10). Ever since, the dysregulation of miRNAs in diseased tissues has gained significant prominence and expanded to an interest in extracellular miRNA as reflections of the malignant or dysfunctional alterations. The easy accessibility by blood sampling and remarkable stability of circulating miRNAs make them promising candidates in biomarker discovery. Numerous diseases and disorders, such as tumors, cardiovascular diseases, multiple sclerosis and liver injury have now been associated with altered extracellular miRNA profiles (11). Still, levels of circulating miRNA are presumably non-specific, and few overlapping reports of studies on the same disease have been published, possibly due to technical or methodological inconsistencies (12). Furthermore, miRNA levels seem to be associated with a wide range of conditions and outcomes in cancer research (13). It has therefore been hypothesized that changes in the profile of circulating miRNAs indicate a general state of disease or inflammation and rather derive from a non-specific response to the disease than the malady itself (14).

To date, gene expression profiling is the approach of choice for detecting diagnostic and prognostic biomarkers, or predicting drug safety. Reverse transcriptase quantitative real-time polymerase chain reaction (RT-qPCR) is considered the gold standard for exact and valid gene expression measurements, either for mRNA or small RNA specimens (15). More recently, digital PCR has emerged as a powerful and sensitive technique for absolute quantification of DNA molecules without the need for external calibration curves. Since RNA is converted into cDNA with varying efficiency, however, its applicability for RNA quantification is limited mostly by the reverse transcription (RT) reaction, which might lead to a skewed representation of initial RNA (16). Nowadays, the discovery and identification of potential new transcriptional biomarkers by RNA sequencing (RNA-Seq) is the holistic state of the art technique. The evaluation and validation of miRNA biomarkers by small RNA-Seq is now routinely being adopted for the identification of physiological or dysregulated miRNAs. Nevertheless, the subsequent validation of identified biomarker signatures by RT-qPCR is mandatory (17–19). But there is a lack of consensus regarding optimal methodologies or technologies for miRNA detection in liquid biopsies, their subsequent quantification and standardization strategies when different sequencing technologies or platforms, and library preparation chemistries are used.

Goal of this review

In this review we present a standardization procedure to discover and validate new biomarkers from liquid biopsies with focus on the entire small RNA-Seq workflow - from experimental design, sample stabilization, RNA extraction and quality control to library preparation, next generation RNA sequencing and all steps of small RNA-Seq data analysis, including validation and interpretation (Figure 1 and 2). Our goal is to point out the importance of experimental standardization and validation (20). The review will explain why and where problems in the small RNA-Seq workflow arise, discuss the real bottlenecks, and how one can resolve or at least circumvent them. We want to improve the quality of small RNA-Seq results by optimizing and standardizing the entire quantification procedure to receive better and more reproducible results. As a broader goal, the outcome of this expression profiling should result in valid biomarker signatures in order to make better predictions in molecular diagnostics. The review should follow the 'general MIQE and dMIQE idea' as published earlier, describing optimization strategies in the qPCR and dPCR workflow (21,22). Following our recommendations will encourage better experimental sequencing practice, lead to more reproducible results, and hence allow unequivocal interpretation of small RNA-Seq results. In summary, the outcome of miRNA analysis in liquid biopsies should be more reliable and valid for future predictions.

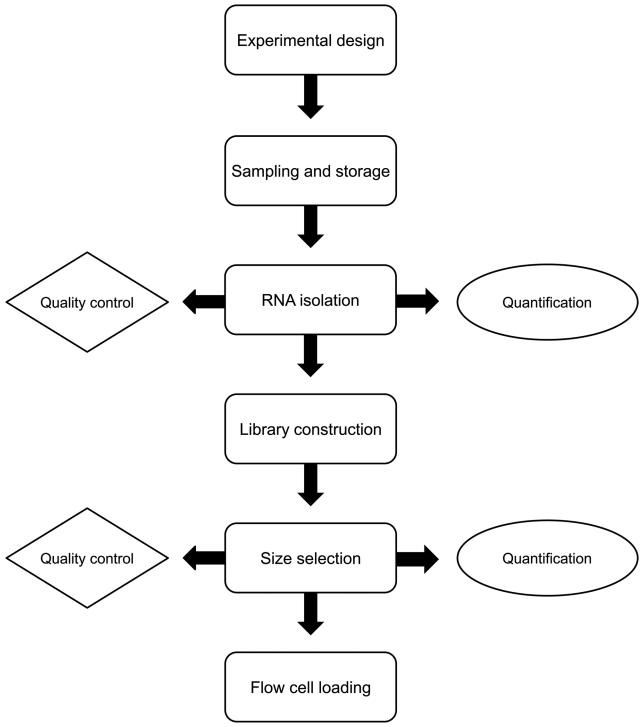

Figure 1. An overview of the small RNA library preparation workflow.

PRE-NGS AND PRE-PCR - THE SAMPLING BIAS

Experimental design and replication

The first step in planning a small RNA expression experiment is to set up a meaningful experimental design, including a reasonable number of replicates on the biological as well as technical level. Both biological and technical replicates have their place in biomarker discovery using RNA sequencing experiments. Biological replicates are crucial to correct for endogenous variability between experimental groups in order to ultimately draw generalized bi-

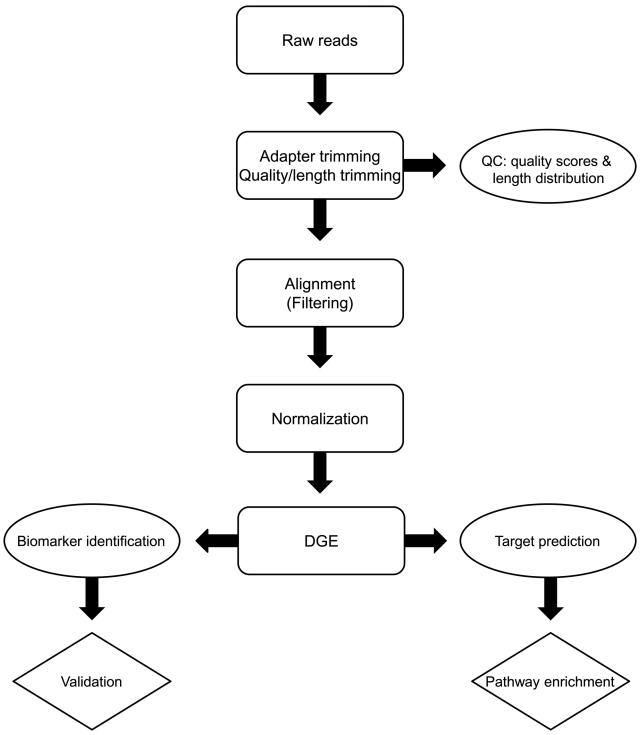

Figure 2. An overview of the small RNA-Seq data analysis workflow.

ological conclusions, whereas technical replicates can help assess the man-made bias introduced by the entire experimental setup and sequencing process itself.

Replicates in reality mean either biological replicates, representing the number of real individuals per experimental group, or technical replicates, repeated measurements of a biological sample with the goal of reducing technical noise. Technical replicates can be further subdivided and introduced either on the level of extraction, RT reactions per sample, sequencing depth, or the number of technical replicates in the sequencing step. Regarding biological sample size, one has to consider inter-individual genetic variation within the studied population. Within the human population, genetic variation is elevated in contrast to highly standardized and inbred animal models. This high variation in human populations as such is based on various factors. Human study groups can be standardized by age, weight or sex, but never by genetic background or lifestyle habits, which might have a remarkable impact on gene expression (nutrition e.g. coffee, alcohol, nicotine consumption, daily rhythms, sleep, stress and more). Regarding domesticated animals such as cattle or pigs, genetic variation is intermediate due to controlled reproduction with a limited number of male semen donors. Laboratory animals, including mice, rats or insects show very low genetic variation within one highly standardized and inbred animal strain. Genetic variation in cell-culture is dependent on the kind of cells used. Primary cell cultures from distinct, genetically different donors show species-specific biological variance, while the largely used permanent cell lines derived from one clone or one individual are genetically identical, and show no biological variance at all.

Different researchers already dealt with the question whether a higher number of biological replicates or a higher sequencing depth leads to better outcomes in RNA-Seq experiments. Increasing sequencing depth results in a higher number of reads, and thereby increases statistical power for the detection of differential gene expression (23). Hart *et al*. and Liu *et al*. concluded that a sequencing depth of 10 million (10M) reads is sufficient for mRNA expression analysis, and that increasing sequencing depth over 10M reads does not improve statistical power significantly. Both publications, however, stated that increasing the number of real biological samples significantly enhances statistical power of the experiment (24,25). Therefore, a higher number of biological samples is preferable over deeper sequencing. Previous reports further suggested including at least three biological replicates per group, depending on the inherent biological variation (26). For experiments involving samples with higher variability, such as human biofluids or specimens from diseased patients, even more replicates might be needed to correctly assess differential gene expression without detecting false-negative reads from biological noise. When biological variability is low, increasing replicates renders statistical power to the experiment. It has been shown that increasing the number of biological replicates in RNA-Seq experiments from two to five facilitates the detection of differential gene expression, but extensive biological replication to improve statistical power is still not utilized in most experiments (27). A recent publication on RNA-Seq additionally reported that experiments entailing only three biological replicates severely lack power to detect the majority of differentially expressed genes, and are only suited to identify transcripts with major fold changes (28). Increased replication markedly improved the correct assessment of differential expression. The authors suggested including at least 12 replicates in order to detect more than 90% of all truly differentially expressed genes.

Technical replicates are useful to characterize the technical variation of an experiment. In general, variability between technical replicates derives from the random sampling nature of sequencing and matches a Poisson distribution (29). Even though it can therefore be accounted for in downstream statistical analysis, some genes are known to deviate from Poisson sampling and thus falsely increase intra-group variability (30).

Another point that has to be considered is whether technical replicates are generally mandatory in gene expression analysis. Liu *et al*. stated that RNA-Seq shows a high reproducibility, concluding that technical replicates are not necessary (25). Due to fairly high library preparation and sequencing costs, technical replicates in RNA-Seq are mostly not realized. Comparing this with replicates in RT-qPCR analysis, Tichopad *et al*. investigated the effect of replicates on different levels of the RT-qPCR quantification workflow (31). The authors stated that replicates in qPCR are not essential, because inhibiting molecules should have been removed before qPCR takes place. The bias introduced by RT was multiple times higher compared to qPCR, wherefore RT replicates are reasonable and necessary (32).

To summarize, inter-individual or biological variation seems to have the highest impact, therefore replicates in biological samples are advisable. The authors recommend in-

troducing replicates early in the quantification workflow by including as many biological replicates as possible (31).

Tissue and RNA sampling and storage

When working with cell-culture, RNA sampling and later storage is unproblematic due to working directly in the molecular biology laboratory in a clean and non-contaminated environment. In contrast, if one collects diagnostic samples in the field, the sampling process cannot be performed under clean and safe laboratory conditions. Therefore, optimal tissue preservation, and thus total RNA preservation and stabilization, are essential points in the experimental workflow. Widely used methods include snap freezing tissues in liquid nitrogen, formalin fixation or storing tissues in RNAlater (Life Technologies), a solution that preserves tissue RNA from degradation, 'freezes' the RNA profile and allows storage of conserved tissue for several hours or days at room temperature.

In clinical research, tissue conservation is routinely performed by formalin fixation and paraffin embedding. Those formalin-fixed paraffin-embedded (FFPE) samples are advantageous for a number of downstream applications. But RNA analysis in FFPE samples is problematic because RNA is cross-linked and partly degraded. RNA extracted from FFPE is thereby of lesser quality for further expression studies (33,34). Due to the growing field of RNA expression analysis in clinical samples, the generation of biobanks for non-fixed frozen tissue is coming into focus (35). Therefore, snap freezing in liquid nitrogen or storage in RNA-preserving agents are the preferred methods for conservation of intact RNA. Publications dealing with the influence of both methods on RNA integrity concluded that high quality RNA can be extracted from tissues conserved in both ways (35,36). The method of tissue fixation has to be planned in detail for each individual experiment according to the given preconditions, which for instance means whether working with liquid nitrogen is generally possible.

RNA quality and RNA integrity

Good RNA quality and high RNA integrity are of great importance in any quantitative gene expression measurement. Degradation of RNA by RNAses, freezing and thawing, UV-light or heat leads to RNA fragmentation. Any RNA degradation influences the results obtained by quantitative downstream applications (37). Several methods to measure RNA integrity and quality exist. Most methods are based on high resolution agarose gel electrophoresis, monitoring the intensity of the major ribosomal 18S and 28S bands. High RNA quality is indicated by a 28S:18S ratio around 2.0 (37). In the past, results of RNA degradation relied on vague human interpretation of the agarose gel image. Nowadays, there are fully automated methods allowing digital interpretation and automatic estimation of the RNA integrity results. With those systems, minor amounts of RNA are labeled with an intercalating dye, and RNA is separated according to its molecular weight using capillary electrophoresis in a microfluidic device. By measuring laser-induced fluorescence detection, the retention time of RNA molecules is displayed in an electropherogram. Applying digital data analysis software, the 18S and 28S ribosomal intensity peaks in an electropherogram are automatically analyzed by a specific algorithm, and a numerical RNA quality score is calculated, whereby a score of 10 indicates intact RNA and a score of 1 completely degraded RNA. It should be mentioned, however, that the concept of RNA Integrity Number (RIN) values is optimized for total RNA profiles from higher eukaryotes, which inherently limits its applicability for studies on other species. Since RIN calculation is majorly based on ribosomal RNA subunit peaks, researchers working with samples differing from the prototypical mammalian RNA need to pay close attention to potentially shifted ribosomal bands. Integrity analysis of plant RNA is further complicated by the presence of additional chloroplast-derived ribosomal RNA that could be recognized as a degradation product and thus falsely lead to lower RIN values. Still, the assessment of RNA quality by measuring RIN has been successfully applied to a variety of non-mammalian organisms such as plants and bacteria (38–40). Although the importance of RNA integrity on downstream applications is well established, even excellent RIN values do not guarantee experimental success since they are unable to report the potential presence of contaminants that might inhibit further RNA processing.

An alternative way of determining transcript integrity is the so-called 3′/5′assay which is based on the quantification of mRNAs at the 3′-end and at the 5′-end. The ratio of the two fractions indicates the mRNA degradation status of the sample (41). The assay, however, is more labor-intensive and has another weakness due to unbalanced RT efficiency at the 3′-end and at the 5′-end.

There are various publications confirming the importance of high RNA quality for mRNA expression profiling studies using microarray and RT-qPCR assays (42–44). For RNA-Seq experiments, high quality RNA is of great importance as well. Degraded RNA leads to decreased quality of RNA-Seq data (45). Particularly the 3′ bias observed in degraded RNA has been shown to have an impact on the quality of RNA-Seq experiments (46). Feng *et al*. developed an algorithm that calculates an RNA quality parameter—the mRIN number— for each sample by quantifying the 3′ bias of read coverage for each measured gene (46).

Small RNAs include the highly prominent miRNAs, which are proven to show higher stability compared to longer RNAs, in particular mRNAs. Due to their short length they are less susceptible to RNA degradation by RNAses (47). The impact of RNA quality on small RNA-Seq has not been evaluated up to now, but it is well known that a high level of RNA degradation in a sample leads to a seemingly increased percentage of small RNAs due to degradation products. It is therefore likely that with decreasing RNA quality, short fragments are included in the sequencing library more frequently, and could thereby lead to a higher number of ambiguous hits after data mapping. The impact of RNA quality on miRNA quantification by SYBR green-based RT-qPCR was shown previously: decreasing RNA quality/integrity is correlated with an increasing Cq value (47).

Circulating RNA and microvesicles

Circulating RNAs are the preferred target in liquid biopsies, and are therefore highly accessed in molecular diagnostics. The RNA, mainly small RNA, found in cell-free blood plasma/serum is either packaged in microvesicles (e.g. exosomes, apoptotic bodies), associated to lipoproteins such as HDL (high-density lipoprotein) particles, or bound by stabilizing proteins (48). Circulating miRNAs are partly bound to proteins such as Argonaute 2 and lipoproteins, which contributes to their enhanced stability (49,50). A seminal paper published in 2007 reported functional miRNA encapsulated in extracellular vesicles (mainly exosomes) secreted by human and murine mast cell lines (51). Soon thereafter, additional reports described the applicability of extracellular vesicular miRNA as biomarkers in blood (52,53). The term circulating miRNA thus has to be used with caution, since it does not state whether the RNA is bound or encapsulated. Circulating or microvesicle-derived RNAs have already shown to be promising diagnostic biomarker for various diseases such as cardiovascular diseases or different kinds of cancer (54).

The composition of circulating vesicles reflects the physiological and pathological status of a patient, and is therefore of considerable diagnostic interest (55). Extracellular vesicles act as a protective shield and delivery vehicle for RNA, and are a treasure trove of easily accessible biological information. Both vesicular RNA and protein were shown to be potential targets for biomarker research (56). Even though considerable advances have been made in the field of extracellular vesicles, there is still no universal consensus on vesicle nomenclature (57). Despite inconsistent terminology, many researchers consider exosomes, the smallest class of extracellular vesicles, as a newly discovered and important mediator in intercellular communication. Since most circulating miRNAs derive from blood or endothelial cells and the contribution of diseased cells is arguably low, exosomes might provide a sampling fraction enriched in tissue-specific biomolecules (14).

There are numerous protocols and commercially available kits for the isolation of extracellular vesicles and extraction of circulating RNAs, in majority from human blood. Principles for isolating vesicles from biofluids include, among others, ultracentrifugation, precipitation, size exclusion chromatography, ultrafiltration, immunopurification and microfluidic approaches (58–61). While differential ultracentrifugation in conjunction with density gradient centrifugation is still considered the gold standard in vesicle isolation and generally yields preparations of high purity, it is labor-intensive, time-consuming and requires substantial sample material, rendering it unsuitable for many clinical and diagnostic applications. Choosing an appropriate isolation method for the particular study has been a topic of extensive debate, and multiple investigations have provided insights into the suitability of respective methods (62–65). Even though most methods were found to be able to isolate extracellular vesicles from various biofluids, yield and purity often differ substantially. Similarly, isolation methods also impact downstream applications: profiles of mRNA (66), miRNA (67) and vesicular protein (68) were shown to vary depending on the respective isolation. Generating pure isolates is complicated by both the complexity of biofluids and the tremendous heterogeneity of extracellular vesicles that even within a particular size range present various subpopulations with different molecular constitution (69,70). Although time-consuming, density gradient centrifugation is highly efficient in removing contaminating proteins and protein complexes, leading to reasonably pure vesicle preparations (62). Polymer-based precipitation methods, on the other hand, require less hands-on time, but suffer from co-isolating non-vesicular contaminants and residual precipitation reagents that can interfere with downstream processing and reduce the vesicle's biological activity (71). Recently, size exclusion chromatography has emerged as a less tedious alternative able to generate vesicles of purity comparable to density gradient-based methods, albeit with low throughput and yield (61,71). Excellent in-depth comparisons of methods for isolating extracellular vesicles from various biofluids can be found elsewhere (63,64,72). Regardless of the particular isolation approach, extraction of RNAs from liquid biopsies is well established. Measuring their concentration is nevertheless challenging due to low concentrations in biofluids. New advances in both sequencing and vesicle research, including careful optimization and standardization of techniques and protocols, will certainly foster progress toward highly specific biomarker signatures.

Blood sampling

In molecular diagnostics, blood is the primary and most important matrix for RNA expression analysis. In humans, minimally invasive sampling is of great advantage, hence blood is the matrix of choice for so-called liquid biopsies. Different and highly standardized methods and kit systems are available for the extraction of high quality RNA from blood, including total circulating RNA and microvesicular RNA. Which sampling system is applicable depends on the particular sample type (whole blood or only a cellular fraction, e.g. white blood cells, red blood cells or platelets), or whether cell-free circulating RNAs of interest are obtained from plasma or serum. For conservation of whole blood for RNA expression analysis, integrated systems for RNA degradation protection and freezing of the current RNA profile are available; namely the PAXgene System (PreAnalytix) and the Tempus System (Life Technologies). Both allow storage of whole blood samples at room temperature for several days or frozen for months without losing RNA quality. For both systems, dedicated kits are commercially available for extraction of RNA longer than 200 nt, or extraction of total RNA including small RNAs (<20 nt). Häntzsch *et al*. and Nikula *et al*. compared the two conservation systems and concluded that both result in high quality RNA samples (73,74). LeukoLock (Life Technologies) allows the extraction of leukocyte RNAs. Within this system, leukocytes are collected in a filter, and RNA is fixed using RNAlater which allows storage and extraction of high quality RNA from white blood cells (75–77).

Quantification of minimal amounts of RNA

The quantification of minimal amounts of total RNA from biopsies or microvesicle isolates is challenging. The de-

tection limit of conventional photometric RNA quantification methods is around 2 ng/μl (78). Due to diminished specificity in the lower concentration range, absorption and therefore quantification is mostly unspecific, because DNA contaminations cannot be distinguished from RNA. Fluorescence-based quantification methods use a fluorescent dye that specifically intercalates or associates with RNA, enabling precise quantification down to as little as 1 pg/μl (78). This method is based on conversion of the fluorescence signal of an unknown sample to a standard curve created from samples with known concentration. The Bioanalyzer 2100 small RNA assay (Agilent Technologies) also allows quantification of small RNAs, especially miRNAs. As mentioned above, this method is only valid in samples of high RNA quality and reasonable RNA quantity. It might result in false positive signals due to contamination of measured small RNA by RNA degradation products with ongoing RNA degradation (47). Due to very low concentrations in RNA samples extracted from plasma or microvesicles, fluorescence-based methods are preferable for small RNA-Seq studies.

How to improve RNA extraction

The extraction of extracellular small RNAs from serum, plasma or other biofluids such as urine or saliva is challenging due to low RNA concentrations. Using carriers to increase RNA output is helpful, whereby glycogen, yeast tRNA, or MS2 phage RNA are widely used. Due to potential interference of biological carrier RNAs with downstream applications, glycogen is the carrier of choice. The use of glycogen increases total RNA yield using most commercially available small RNA extraction kits (79,80). When establishing an extraction method, spiking starting material with known quantities of artificial or exogenous ribonucleotides, so-called spike-in controls, and quantifying their recovery is an easy way to assess the efficacy and reproducibility of the respective approach. Spike-in controls for miRNA extraction are, for example, artificial short RNAs in the length range of miRNAs or miRNA extracts from other species, such as *Caenorhabditis elegans*. Indeed, Burgos *et al*. (81) optimized RNA extraction from human cerebrospinal fluid by measuring the recovery of three previously spiked-in *C. elegans* miRNAs and found significant variation between commercially available kits, and even within technical replicates (81). It is recommended to add spike-in controls directly to the extraction buffer instead of adding it to the plasma or serum sample due to the presence of RNAses in biological samples, which might lead to degradation of the spike-in miRNA (79,80). Spike-in controls can be easily quantified by RT-qPCR in order to determine extraction recovery rate, and appropriately normalize resulting expression data (79). Furthermore, such spike-in controls are also useful to test the efficiency of the RT reaction step or to control for qPCR inhibitors.

LIBRARY PREPARATION - THE RT AND LIBRARY PREPARATION BIAS

The biases based in library preparation

Ultra-high-throughput sequencing allows global sequence profiling of the small RNA transcriptome. To this end, transcriptional targets need to be converted into sequencing libraries, entailing molecular modifications to make targets suitable for the small RNA-Seq chemistry. This pre-sequencing library preparation, however, introduces technical bias into the fine-tuned transcriptional screening and *de novo* discovery of transcripts (82).

In this chapter, we examine critical steps in preparing sequencing libraries from total RNA, and highlight the challenge of creating them in high quality. For the implementation of Next-Generation Sequencing (NGS) of small RNAs, the main task is to convert native small RNAs into sequencable molecules while minimizing technical bias. Preparing small RNA for expression profiling requires multiple enzymatic manipulation steps. These typically include sequential adaptor ligations to both ends of small RNAs, RT, and PCR-based amplification. The 3′-adaptor ligation introduces primer binding sites for first strand cDNA synthesis. The PCR step specifically enriches functional small RNAs with adaptors on both ends, and permits multiplexing through introducing unique barcodes to each sample. Ultimately, a size selection step ensures that only fragments pertaining to small RNAs are included in the final library. In the interest of comparing datasets generated in multiple RNA-Seq experiment with minimal distortion, the problem of pre-sequencing bias needs to be addressed according to the idea of the widely accepted MIQE guidelines (21). Previously, published experimental data showed that using identical starting RNA led to entirely different results concerning small RNA expression ratios due to the implementation of different library preparation strategies (83). Surprisingly, the choice of sequencing platform contributed little to the reported differences (Spearman's $\rho = 0.79$–0.95). Library replicates to test for reproducibility yielded comparable results ($\rho = 0.84$–0.99), indicating that data distortion was likely caused by differences inherent to cDNA construction protocols.

Bias resulting from low RNA input

Besides the quality of extracted total RNA (as discussed above), RNA quantity available for the particular experiment is crucial for successfully generating high-quality sequencing libraries. Various sample types such as plasma, serum or urine contain limited concentrations of small RNA due to lack of cellular material, which complicates library preparation. However, several efficient and sensitive methods for preparing libraries from sparse input material address this problem (84,85). Generally, it is recommended to use RNAs of similar quality and quantity for each sample within an experiment (54,86). Additionally, the capture efficiency of small RNAs from cell-free samples might be limited: Kim *et al*. reported that biological samples with low RNA concentration lack GC poor or highly structured miRNAs when extracting with the phenol/guanidine isothiocyanate reagent Trizol (Thermo Fisher Scientific) (87).

They hypothesized that small RNAs base pair with longer RNA species acting as carrier molecules, and thus compensate their limited capacity to precipitate in RNA extraction. Small RNAs with low GC content and stable secondary structures might interact with carriers less efficiently, reducing their representation in RNA preparations. For samples with low total RNA content, such as a small number of cells or biofluid specimens, the availability of longer RNAs that serve as carriers might be limiting the efficient recovery of this specific fraction of small RNAs. In order to minimize this bias, they recommended to avoid Trizol extractions, or to only compare samples with similar concentrations of total RNA. It was additionally suggested to stabilize RNA–RNA interactions by adding $MgCl_2$ in an attempt to equalize the extraction efficiency of all small RNA species.

The challenge of adapter and barcode ligation

Since the ligation step introduces the largest bias in RNA-Seq results, several studies investigated the effect of ligating 5′- and 3′-adapter or barcodes (88–93). Hafner *et al*. concluded that ligation efficiency depends on the sequence and secondary and tertiary self-structure of miRNAs and/or miRNA/adapter products (94). To reduce ligation bias, many researchers suggest using randomized adaptor pools containing various adapter sequences adjacent to the ligation junction (89,91,93,95). A recently published follow-up paper, however, observed that it is not necessary to design the randomized region near the ligation junction (96). Instead, this might complicate identification of the end of a miRNA sequence with an unknown sequence directly attached to it. Furthermore, the authors found out that miRNAs prefer to ligate to adapters with which they can form a particular structure, whereas the primary sequence is not the main contributor to ligation bias. Even better results can be achieved when the 5′- and 3′-adapter have complementary regions. The only commercial kit employing a similar strategy is the new NEXTflex Small RNA-Sequencing Kit (Bioo Scientific). It uses randomized sequences at the ligation site in massive concentrations to present small RNAs their optimal adapter. According to recent work by Baran-Gale *et al*., the NEXTflex protocol has shown a great reduction in bias and the best differential expression correlation to RT-qPCR (97).

Barcodes are very short distinct sequences which can be introduced in the sequence of interest to enable distinction of multiple samples at the same time and in the same lane of a flow cell. To enable multiplexing, a variety of barcode sets are commercially available (e.g. Illumina TruSeq Small RNA Library Preparation Kit: 48 unique indexes, New England Biolabs NEBNext® Multiplex Small RNA Library Prep Set for Illumina: 24 unique indexes, Bioo Scientific NEXTflex™ Illumina Small RNA-Sequencing Kit v3: 48 unique indexes). Depending on the library preparation kits used, barcodes can be introduced at three points in the library preparation: (i) during adapter ligation (94), (ii) during RT (89) or (iii) during PCR (98). Beside the fact that barcoding is a very useful tool, it causes technical bias by influencing the ligation efficacy, RT efficiency and PCR amplification (92,98). The above findings about the strong impact of base compositions in the core adapter sequence prove that it is crucially important to include barcodes only during RT or later in PCR (89,92,96). When carefully designing the library preparation strategy, it is therefore highly recommended to avoid barcode sequences near primer annealing sites, and to include barcodes only downstream of ligation reactions. It is, however, well described that multiple-template PCR amplification can result in sequence-dependent amplification bias due to template differences (18,89,99). In order to measure the PCR amplification bias resulting from barcodes, Van Nieuwerburgh *et al*. designed a new strategy named post-amplification ligation-mediated (PALM) barcoding, where the ligation of barcodes occurs after PCR without further purification of the library. No bias was observed when comparing PALM with Illumina's TrueSeq miRNA protocol, which introduces barcodes during the PCR step (98).

RNA modifications lead to ligation and RT bias

A simultaneous library construction for all small RNA species is challenging because of their different modified ends. Small RNAs possess different 5′- and 3′-modifications depending on their classes (e.g. miRNA or piRNA) and species origins (e.g. mammals, insects, or plants). While miRNAs in mammals carry a 2′-OH-modification at the 3′-end, many mammalian piRNAs or plant-derived miRNAs feature a 2′-O-methyl group on the ribose at the 3′-end (100,101). This may influence the efficiency of enzymes involved in ligation and cDNA synthesis. To minimize bias, it is important to notice that polyadenylation-based libraries are less suited for 2′-O-methylated RNAs. RNA tailing with poly(A) or poly(C) is significantly less efficient for modified 3′-ends, which might conceivably lead to the underrepresentation or even absence of some RNA species in cDNA libraries (82).

In ligation-based libraries, the ligation efficiency of RNAs with 2′-O-methyl groups can be significantly improved by a longer incubation time, reduced temperature, and the use of T4 RNA Ligase2 instead of T4 RNA Ligase1 (102,103).

Choosing appropriate enzymes for the RT step can also tone down the bias because of their known sensitivity to 2′-O-methyl groups. It is recommended to use avian myeloblastosis virus RTase or murine leukemia virus RTase to prevent favoring the transcription of some RNAs over others (103).

PCR amplification bias in library preparation

The efficiency of PCR amplification depends on the base composition of different types of templates, type of polymerase, PCR buffer composition, and potential presence of any inhibitory substances (104). It is well known that a varying GC-content is associated with unequal PCR amplification efficiencies and leads to template-specific preferences (105–108). To avoid that RNAs with high GC-content remain under-represented, one can perform an optimized PCR program with an extended initial denaturation time of 3 min and subsequent melt cycles of 80 s (109). Furthermore, choosing an appropriate polymerase will not only minimize GC-bias, but also narrow the length dis-

Table 1. Crucial steps and recommendations for small RNA sampling and library preparation

Step	To consider	Recommendation
Experimental design and replication	Type and number of samples Outcome of interest Variance within samples	Employ sufficient replication for question at hand Favor biological replicates over technical ones
Sequencing depth	Outcome of interest Replication	For a rough snapshot of gene expression or analysis of high-level transcripts, lower coverage is sufficient Sequencing depth needs to be increased for analysis of rare transcripts
Sampling and storage	Sampling environment Sample type Embedding/fixation Freezing/storage	Keep sampling conditions as clean as possible Choose an appropriate sampling system for the particular sample type Use agents to preserve and stabilize RNA Freeze samples as quickly as possible and store at appropriate temperature
RNA extraction	Quantity of input material Type of extraction kit Use of a carrier	Carefully optimize the method of extraction for the particular type and quantity of starting material Carrier material might be considered to increase small RNA yield
Total RNA	Expected yield and quantification system Quality of extracted RNA	Opt for fluorescence-based quantification of extracted RNA Check RNA quality and integrity by capillary electrophoresis
Addition of adapter	Type of RNA (e.g. miRNA, piRNA) modified ends	Be aware of ligation biases For small RNAs with modified 3′-ends avoid poly(A) or poly(C)-based approaches or modify protocol accordingly
Reverse transcription	Type of enzyme Introduction of barcodes	Choose appropriate enzyme for given experimental conditions Introduce barcodes during PCR
PCR amplification	Necessity Type of enzyme Number of cycles	Choose pre-amplification strategy based on the quantity of starting material Opt for high fidelity polymerases with low error rates Perform as few PCR cycles as possible
Size selection	Appropriate size range Precision of selection system	Select for cDNA fragments that reflect the size of the RNA of interest High-resolution gel electrophoresis to effectively separate small RNA species
Library purity and quantification	Contamination with adapter dimers Accurate quantification for precise flow cell loading	Assess library purity by capillary electrophoresis Quantify library by fluorimetric assays or qPCR/dPCR
Quality control	Quality and purity of samples at each step of the workflow	Control for sample quality throughout workflow: purity and integrity of initial sample, extracted RNA, cDNA library before and after size selection

tribution of generated PCR products. Several PCR polymerases such as Kapa HiFi (Kapa Biosystems) or AccuPrime Taq DNA Polymerase High Fidelity (Life Technologies) are recommended because of their ability to amplify difficult templates with higher efficiency and lower error rates (109,110). It was furthermore demonstrated that it is of high importance to select a suitable polymerase/buffer system, which can significantly reduce the PCR-mediated bias. In an attempt to optimally amplify DNA sequencing libraries, Dabney and Meyer tested 10 commercially available DNA polymerase/buffer systems and recommended the Herculase II Fusion enzyme as the best performer (107). Generally, it is recommended to use as few PCR cycles as possible for library amplification, and to compare only technical or biological replicates with the identical number of PCR cycles, since PCR noise accumulates with higher cycle number (110).

Library preparation of samples with limited starting material is challenging: researchers have to make a compromise between introducing PCR bias and not detecting lowly expressed transcripts that might not have been sufficiently amplified. Okino *et al.* recently presented a highly multiplexed pre-amplification approach that massively increases the abundance of target genes while keeping amplification bias at bay (111). Since gene expression patterns were maintained throughout up to 14 PCR cycles, analysis of pre-amplified samples yielded similar results to samples not undergoing pre-amplification. Gene expression profiling studies on low input samples might greatly benefit from such a distortion-free enrichment strategy. Recently, more sophisticated library preparation strategies to avoid PCR bias altogether were developed for both bulk and single cell analyses (112,113). By introducing unique molecular identifiers (UMI), researchers are able to detect absolute numbers of DNA or RNA molecules, since each nucleic acid in the

Table 2. Crucial steps and recommendations for small RNA-Seq data analysis

Step	To consider	Recommended tools or algorithms
Data pre-processing	Trimming adapters Removing short reads	Btrim, FASTX-Toolkit
Quality control	Library size and read distribution across samples Per base/sequence Phred score Read length distribution Assess degradation Check for over-represented sequences	Btrim, FASTX-Toolkit, FaQCs
Read alignment (Filtering)	Reference database or genome Annotation Mismatch rate Handling of multi-reads	Bowtie, BWA, HTSEQ, SAMtools, SOAP2
Normalization	Library sizes and sequencing depth Batch effects Read distribution Replication level	DESeq2, EdgeR, svaseq
DGE analysis	Data distribution Replication level False discovery rate	DESeq2, EdgeR, SAMSeq, voom limma
Target prediction of miRNAs / siRNAs	*Insilico* prediction or experimental validation Canonical and non-canonical target regulation	miRanda, miRTarBase, TarBase
Biomarker identification	Sensitivity Specificity Classification rate	DESeq2, Simca-Q, Numerous R packages: base, pcaMethods, Mixomics

starting material is tagged with a unique sequence during RT. After sequencing and mapping, UMI are counted to infer absolute copy numbers without including PCR duplicates in the analysis. Even though UMI-based library preparation has only been applied to mRNA sequencing so far, similar approaches might also be developed for small RNA-Seq in the future.

Gel size selection

The fragmentation of DNA by acoustic shearing, sonication or enzymatic digestion to attain the desired target length of 100–500 bp fragments is not necessary for sequencing small RNAs, which are usually considered to be shorter than 200 nt (110). For miRNA sequencing, fragment sizes of adaptor–transcript complexes and adaptor dimers hardly differ in size. An accurate and reproducible size selection procedure is therefore a crucial element in small RNA library generation. To assess size selection bias, Locati *et al*. used a synthetic spike-in set of 11 oligoribonucleotides ranging from 10 to 70 nt that was added to each biological sample at the beginning of library preparation (114). Monitoring library preparation for size range biases minimized technical variability between samples and experiments even when allocating as little as 1–2 % of all sequenced reads to the spike-ins. Potential biases introduced by purification of individual size-selected products can be reduced by pooling barcoded samples before gel or bead purification.

Since small RNA library preparation products are usually only 20–30 bp longer than adapter dimers, it is strongly recommended to opt for an electrophoresis-based size selection (110). High-resolution matrices such as MetaPhor™ Agarose (Lonza Group Ltd.) or UltraPure™ Agarose-1000 (Thermo Fisher Scientific) are often employed due to their enhanced separation of small fragments. To avoid sizing variation between samples, gel purification should ideally be carried out in a single lane of a high resolution agarose gel. When working with a limited starting quantity of RNA, such as from liquid biopsies or a small number of cells, however, cDNA libraries might have to be spread across multiple lanes. Based on our expertise, we recommend freshly preparing all solutions for each gel electrophoresis to obtain maximal reproducibility and optimal selective properties. Electrophoresis conditions (e.g. percentage of the respective agarose, buffer, voltage, run time, and ambient temperature) should be carefully optimized for each experimental setup. Improper casting and handling of gels might lead to skewed lanes or distorted cDNA bands, thus hampering precise size selection. Additionally, extracting the desired product while avoiding contaminations with adapter dimers can be challenging due to their similar sizes. Bands might be cut from the gel using scalpel blades or dedicated gel cutting tips. DNA gels are traditionally stained with ethidium bromide and subsequently visualized by UV transilluminators. It should be noted, however, that short-wavelength UV light damages DNA and leads to reduced functionality in downstream applications (115). Although the susceptibility to UV damage depends on the DNA's length, even short fragments of <200 bp are affected (116). For size selection of sequencing libraries, it is therefore preferable to use transilluminators that generate light with longer wavelengths and lower energy, or to opt for visualization techniques based on visible blue or green light which do not cause photodamage to DNA samples (117,118). In order not to lose precious sample material, size-selected libraries should always be handled in dedicated tubes with reduced nucleic acid binding capacity.

Precision of size selection and purity of resulting libraries are closely tied together, and thus have to be examined carefully. Contaminations can lead to competitive sequencing of adaptor dimers or fragments of degraded RNA, which reduces the proportion of miRNA reads. Rigorous quality control checkpoints and size selection steps are therefore

crucial. In order to assess length distribution and potential contaminations, it is recommended to use high sensitivity capillary gel electrophoresis assays. The size profile of final library preparation products is dictated by the initial small RNA's size distribution extended with respective sequencing adapters.

Library quantification and flow cell loading

Methods of quantitating final cDNA libraries are still highly debated in the field, and have a significant impact on the sequencing experiment since precise loading of flow cells is crucial for optimal cluster densities. Overloading results in overlapping clusters, reduced quality of reads, and ultimately diminishes the data output of the experiment (119). Low numbers of clusters, or underclustering, on the other hand, yields high-quality data, but a less-than-ideal output. Impurities in sequencing libraries not only skew library quantitation, but also affect cluster generation: shorter fragments such as adapter dimers cluster more efficiently and thus restrict clustering of target RNAs. Capillary gel electrophoresis is a useful tool to assess library integrity, insert size and contaminations, but detects both amplifiable and non-amplifiable molecules (120). Spectrophotometrical methods of nucleic acid quantification are not sensitive enough to precisely quantitate cDNA libraries, and suffer from also measuring single-stranded DNA and free nucleotides. Fluorometric assays such as PicoGreen (Thermo Fisher Scientific) or Qubit (Thermo Fisher Scientific) are more applicable due to increased sensitivity, and specifically quantify double-stranded DNA. Another common approach is quantifying cDNA libraries via qPCR with primers designed to adaptor sequences. Since only functional molecules are captured in the analysis, qPCR and its derivatives seem to precisely predict actual cluster densities (121). Increasingly sensitive methods of library quantification allow for both less input material and fewer PCR cycles, which in turn facilitates sequencing of limited samples and reduces distortion of the initial sequence distribution. Although more costly than other methods, calibration-free absolute quantification of cDNA libraries by digital PCR was found to be a highly accurate tool for quantification of amplifiable molecules in sequencing libraries (122,123).

Loading precision can also be increased by using artificial or exogenous spike-ins. Adding known quantities of a synthetic sequence to samples and quantifying their read count allows for additional control of sequencing parameters. Additionally, technical biases and sequencing errors can be assessed by correlating the amount of spiked-in RNA to read counts mapping to those standards. Fahlgren *et al*. spiked sequencing libraries with three synthetic 21-nt sequences, and found a linear correlation between spike-in concentration and mapped spike-in reads that reached saturation at 10 pmol spike-in per 100 μg of total RNA (124). Another publication using poly-A-tailed mRNA-mimetic standards reported a linear correlation spanning six orders of magnitude while suggesting that the detection of standards is robust to the endogenous complexity of RNA samples (125). As for the analysis of target transcripts, the recovery of standard reads was limited by sequence abundance and sequencing depth, both of which increased spike-in detection.

Critical steps in small RNA-Seq experimental design, sampling and library preparation as well as recommendations by the authors are summarized in Table 1.

SEQUENCING - THE SEQUENCING BIAS

Introduction to sequencing bias

While researchers used to increase sequencing depth rather than introduce additional biological replicates, the ever-subsiding costs of sequencing assays nowadays allow for more replication (126). This, in turn, increases specificity and sensitivity of NGS experiments, and helps correct for biases that cannot be mitigated by bioinformatics methods, such as batch or library preparation effects. Merely increasing sequencing depth in order to improve the specificity of experiments might seem a straightforward strategy, but in reality does not help alleviate sequencing-specific errors (126). Even though a major cause of bias lies in the library preparation of small RNA samples, the sequencing reaction itself can also lead to substantial errors in NGS data. A great number of factors pertaining to the sequencing reaction have to be considered when conceptualizing RNA-Seq experiments. Regardless of the particular experimental question, fundamental aspects such as randomization, replication and blocking need to be properly addressed (127). The most basic decisions relate to choosing a particular sequencing platform and type of flow cell, and designing an experiment that tailors the sequencing chemistry specifically to the question at hand. Additionally, insufficient replication, unsatisfactory sequencing depth and PCR errors are known to increase bias in sequencing data. It is also important to notice that batch effects may result from different kits, reagents, chips, platforms, instruments, handling by different technicians, and day-to-day variations. Batch effects may even occur between different lanes on an Illumina flow cell, or between sequencing runs (110,128). In light of Illumina's dominance in the NGS market, most types of bias discussed in this review are focused on this particular sequencing chemistry.

Batch, lane- and flow cell effects

A major concern in all experiments is detaching biological from technical variation since confounding both makes it impossible to interpret changes in data. For RNA-Seq experiments, it was shown that library preparation introduces the largest bias. This so-called batch effect is an often underestimated problem in high-throughput techniques. As shown above, variations in cDNA preparation from a singular biological source can arise from laboratory conditions, varying quality and reagent lots, skills of the particular operator, changes in personnel, or more subtle factors such as laboratory temperature or ozon levels (128). Quality and quantity of input material, primer concentration, size selection and number of PCR cycles are only a few of many critical parameters of an RNA-Seq protocol that can lead to profound batch effects. A recent article even reported that the composition of small RNA sequencing libraries is more heavily influenced by RNA extraction than by library preparation itself (129). Confounding batch effects with the question of interest, e.g. preparing sequenc-

ing libraries of all treated and control samples on different days or by different operators, can skew the data and directly lead to false biological conclusions. While shown to be of less impact than batch effects, there are also lane and flow cell effects that need to be taken into consideration when designing RNA-Seq experiments (29). These effects pertain to technical variations arising after the cDNA library is loaded onto the sequencer. Marioni *et al*. reported a high replicability in Illumina sequencing data with only a small percentage of genes featuring a systematic difference between different lanes of a flow cell (30). Ross *et al*., on the other hand, found substantial variation between separate flow cells, but not between lanes within a flow cell (130). It should be noted, though, that intra- and inter-assay variation was shown to be less prominent than variation between sequencing platforms.

Multiplexing

The ability to multiplex—adding specific barcodes to separate samples and sequencing them on the same lane of a flow cell—nowadays allows researchers to mitigate lane effects and create more effective experimental designs. Auer *et al*. proposed creating 'balanced blocks' by subjecting all samples to the same experimental conditions, including library preparation and sequencing (i.e. equal proportions of all samples are loaded onto all lanes of the flow cell) (131). For more sophisticated and larger experiments, it is advisable to spread library preparation batches, sequencing lanes and flow cells across all biological groups and replicates to minimize technical variability. Multiplexing and pooling samples as early as possible is advantageous since they can then be processed through the library preparation workflow together, which further alleviates batch effects. Multiplexing also helps to reduce sampling bias when loading the cDNA library onto the flow cell. Loading entails a large dilution step since only a fraction of the cDNA pool is used for cluster generation. An uneven distribution of molecules results in skewed library representation on the flow cell, and thus profoundly alters data output (132). Multiplexing and pooling all samples tones down sequencing errors by reducing sampling bias to only one dilution step.

Paired-end versus single-end sequencing

Paired-end sequencing is a powerful innovation in transcriptomics, yielding more information on transcripts at the same sequencing depth (29). While useful for detection of alternative splice variants and chimeric transcripts, paired-end sequencing usually offers no advantage in small RNA-Seq. Since inserts are short, most experiments do not exceed 50 cycles of sequencing even for small RNA discovery applications. Illumina in fact suggests lowering cycle numbers to 18–36 for miRNA expression profiling studies. Even for profiling of protein-coding genes, 50-bp single-end reads were previously recommended in the literature (26).

Sequencing depth

Since the amount of binding sites on a flow cell is a finite resource, the number of samples in a sequencing run and the sequencing depth are intimately connected. While depth usually refers to the number of reads contributing to an assembly, the respective coverage depends on the abundance of the transcript of interest. For high-level transcripts, even a lower depth might be sufficient to analyze differential gene expression, whereas low-level transcripts require much higher sequencing depths to yield sufficient coverage. Since small RNA copy numbers span a wide range of expression, higher depth is usually required to accurately capture less abundant transcripts. When designing RNA-Seq experiments, sequencing depth has to be tailored to the outcome of interest: a rough snapshot of gene expression requires far lower coverage than the analysis of rare transcripts. For miRNA discovery, Illumina nowadays recommends at least 10M mapped reads. Metpally *et al*. found that while increasing sequencing depth facilitates the detection of new miRNAs, even a moderate depth of only 1.5M mapped reads reliably represents the miRNA distribution in the sample (133). For a given sample type, increasing sequencing depth seems to positively correlate with increasing the proportion of mapped reads. Previous RNA-Seq studies stated that increasing sequencing depth reduces errors in differential gene expression experiments with the caveat of diminishing returns at a certain level of coverage (134). For mRNA-Seq experiments, a stable detection of transcripts seems to be reached at coverage of about 30× with greater coverage only yielding marginal error reduction rates (23). These guidelines could also be applied to small RNA-Seq studies. Since the percentage of initial reads mapping to known miRNAs varies across sample types and library preparation batches, it might be advisable to run a small pilot study in order to determine how many mapped reads are appropriate for the particular biological problem, and how much coverage is needed to generate those reads (133). This ultimately also determines how many samples can be multiplexed on each flow cell of the main experiment. The decision as to whether increase sequencing depth or include more samples depends on the outcome of interest, and is oftentimes limited by the given research budget.

Systematic PCR error

While careful experimental design, library preparation, and loading of the flow cell support bias reduction, the sequencing reaction itself bears additional risk for skewing NGS data. PCR errors induce bias not only during library preparation, but also affect cluster generation and sequencing by synthesis chemistry. Even in the days of high-fidelity DNA polymerases, false incorporation of nucleotides cannot be prevented completely, resulting in DNA strands deviating from the original template. Sequence errors during cluster generation are particularly detrimental since erroneous molecules are exponentially amplified and impair base calling during the subsequent sequencing reaction, ultimately resulting in poor read quality. Growing mixed clusters from more than one template molecule results in a heterogeneous colony of PCR products, and thus an inconclusive fluorescence signal during imaging (135). While amplification efficiency is a significant cause of bias in library preparation, differences in template-specific amplification during cluster generation do not majorly skew read count results since

only the fluorescence intensity of the respective cluster is affected.

Polymerase errors also occur during the sequencing reaction itself. Phasing, the lagging behind of a strand that failed to incorporate a base, hampers base-calling since a more heterogeneous fluorescence signal of the cluster is recorded in each imaging cycle. The enzyme can also erroneously insert multiple bases, which is referred to as prephasing (136). Both of these problems are independent of the template DNA sequence, and lead to an increased frequency of base-calling errors toward the end of a read since more and more noise from preceding and ensuing cycles is introduced. Imaging is further impeded by cross-talk, the partial overlap of emission spectra of the four dyes used in Illumina sequencing technology. This additional noise factor seems to be cycle-dependent and also increases error rates in later cycles (137). Further factors contributing to sequence-independent base-calling errors are dead fluorophores and uneven signal intensities across each tile of the flow cell (138,139). Base-calling algorithms need to be aware of and account for these biases. After signal detection and error correction, the base with the highest intensity is chosen. Remaining uncertainties about called bases are then expressed in quality metrics such as the widely adopted Phred score (140). Originally published in 1998, Phred employs log-transformed error probabilities to generate ASCII-encoded quality scores for each nucleobase. According to the algorithm $q = -10 \times \log_{10}(p)$ where p is the probability of an incorrect base-call, high quality scores equal low error probabilities and the ubiquitous benchmark of Q30 reads corresponds to an error probability of 0.001. The better a base-caller works, the higher the accuracy of sequencing, which ultimately reduces coverage requirements.

Sequence-specific PCR errors in Illumina sequencing

In addition to the abovementioned systematic errors, there are also several sequence-dependent biases in sequencing by synthesis. It is well-known that miscalls on the Illumina platform occur more frequently in GC-rich regions and increase in later cycles (141). Sequence-related biases resulting in failed single-nucleotide elongation might be induced by altered substrate preference of the DNA polymerase or specific inhibition of the enzyme. Indeed, Nakamura *et al.* identified sequence-specific dephasing triggered by GGC sequences to be a consistent bias in Illumina datasets (142). Another cause of sequence-specific errors in Illumina sequencing, albeit potentially of less relevance for small RNA-Seq, are secondary structures of the flow cell-bound single-stranded DNA (ssDNA). According to Nakamura *et al.*, ssDNA folding induced by inverted repeats contributes to polymerase inhibition, while Stein *et al.* illustrated how secondary structures can facilitate or hinder priming during Illumina bridge amplification (142,143). Sequence-induced errors are not only detrimental for applications such as SNP detection or transcriptome assembly, but can also interfere with small RNA-Seq due to the close homology of miRNAs.

Platform-specific error profiles

Previous publications about NGS error rates reported that a majority of miscalled bases is not associated with insufficient coverage, but rather stems from systematic biases in the respective sequencing chemistry (144). It is well known that single base substitutions are the dominant error in Illumina data, while pyrosequencing and ion semiconductor sequencing are more prone to insertions and deletions (indels) (145). In a recent comparison of common platforms, Illumina MiSeq sequencing was shown to produce the highest quality data with a substitution rate of 0.1/100 bases and an indel rate of <0.001/100 bases (146). The frequency of indels was markedly higher when using the Life Technologies Ion Torrent Personal Genome Machine (PGM) and Roche 454 GS Junior systems, featuring 1.5/100 bases and 0.38/100 bases, respectively. Another publication on Illumina sequencing reported error rates as low as 0.3% and an increased frequency of A>C conversion (141). Since the early days of high-throughput sequencing, significant improvements in sequencing chemistry and software have markedly lowered error rates in Illumina data and led to more robust performance. Still, certain error patters characteristic for the technology and independent of the input sequence still pertain to newer generations of sequencers (147). Error rates were shown to be reproducible and predictable across multiple samples in a recent publication on cellular barcoding (148). While indels are fairly rare in Illumina data, they can account for up to two thirds of all errors in 454 pyrosequencing (149). Both Ion Torrent and 454 are known to struggle with homopolymer stretches that oftentimes induce frameshifts. In 454 sequencing, homopolymer errors are more frequent in A and T rich regions and increase with longer sequences of identical bases, while Illumina errors are more randomly distributed (150).

DATA ANALYSIS - THE DATA ANALYSIS BIAS

Small RNA data analysis

Having successfully avoided any pitfalls and biases during experimental setup, library preparation and sequencing, scientists are challenged by processing the frequently huge amounts of sequence data, and extracting meaningful and reliable information from millions and millions of reads. Although digital datasets provide the opportunity to test and validate a seemingly endless array of analyses without spending more than time and computational resources, beginners in the field are often overwhelmed and deterred by the multitude of offered software tools and pipelines. Since a complete discussion of all possible analyses would go beyond the scope of this review, the following part will be centered on the currently prevalent aim of most small RNA-Seq experiments: the detection and comparison of small RNA (mainly miRNA) expression profiles in differently treated samples. In addition, we will focus on 'free to use' software tools or R packages (151) that, while sometimes lacking in user friendliness, are readily available to anyone. Even though most of the software provides comprehensive manuals and tutorials, scientists not already familiar with command line tools may want to try a more intuitively usable software suite, in particular Galaxy (152–

154) or eRNA (155), which implement many of the tools discussed here in a user-friendly graphical interface or invest in commercially distributed programs such as CLC Genomics Workbench (Qiagen), Ingenuity Pathway Analysis (Qiagen) or Genomatix Genome Analyzer (Genomatix). Unfortunately, due to the complexity of varying genomes, small RNA species, data bases and constant updates and improvements of existing software tools, a uniformly valid and standardized analysis approach for all datasets has yet to be established. The fact that most extensive evaluations of methods are carried out on sequencing runs of longer RNAs, and do not take into account the special nature of small RNA datasets further complicates this. The following chapter will highlight all major sources of bias or unwanted variation that need to be addressed and reported to nonetheless guarantee reproducibility and comparability between experimental setups or computational pipelines.

The starting point for all explorations is a fastq file comprising all read sequences with their associated quality scores, indicating the probability of a wrong base call for any given nucleotide. Small RNA data analysis can be generally divided into four individual parts of equal importance: **data preprocessing,** including quality control and adapter trimming, the **alignment** of reads to the respective reference genome or small RNA database, **normalization** of mapped reads, and **differential expression analysis** between samples. A summarizing overview of critical steps and recommended tools for small RNA-Seq data analysis is provided in Table 2.

Data preprocessing

As discussed previously, sequencing errors accumulate with read length, and quality of sequencing data drastically affects downstream analysis (141). Furthermore, sizes of many small RNA transcripts such as miRNAs (∼22 nt) and piRNAs (∼31 nt) (156) fall short of usual sequencing lengths (∼36–50 nt), and resulting reads inevitably incorporate 3′-end adapter sequences from library preparation. To facilitate correct alignments, small RNA read data must therefore be trimmed of adapter artifacts. Complementarily, a significant reduction in false positive alignments to multiple genomic locations can be achieved by filtering for sequences with inadequate lengths (157,158). Removal of these reads with less than 16–18 nt, representing almost exclusively degraded RNA or adapter dimers from library preparation, can also crucially save computational time and associated costs. With the adapter sequences supplied by library preparation kit manufacturers, this can be achieved by a number of programs including Btrim (159), the fastx_clipper tool from the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), cutadapt (160) or FaQCs (161). Although current library preparation and sequencing protocols, in conjunction with small read lengths after adapter trimming, do a good job of minimizing sequencing errors, low quality datasets can still occur and will struggle finding accurate alignments. While there are algorithms such as Quake (162) or ALLPATHS-LG (163) that try to correct unreliable base callings by superimposing the most frequent, similar patterns on them, the intrinsically non-uniform sequence abundances found in small RNA-Seq (164) prohibit their application. Low quality reads can nonetheless be mitigated in part by removing bases with low Phred scores from reads up to a minimum length (∼18 nt) or, less preferably, by filtering them out completely (165). Popular quality trimming algorithms implement either some variation of a running sum of the quality scores from 3′- to 5′-end looking for a minimal (Cutadapt), or a moving window that determines the longest continuous stretch of nucleotides above the threshold and trims the rest (Btrim, fastq_quality_trimmer from FASTX Toolkit, FaQCs, SolexaQA (166)). Prior to alignment, filtered and adapter- as well as quality-trimmed reads should then be evaluated in terms of quality scores and typical length distribution of reads. Remaining reads should be free of low quality sequences indicating sequencing errors (quality score <20), and read lengths should show a distinct peak for the targeted small RNA species (e.g. 21–23 nt for miRNA, 30–32 nt for piRNA). An absence of these typical read lengths can originate from a multitude of causes, including incorrect small RNA isolation, inaccurate size selection during library preparation, as well as degradation during, for instance, storage of samples. A fairly uniform increase in read numbers from longer to shorter reads is further proof of low RNA integrity. Additionally, read data can be examined for over-represented sequences potentially deriving from amplification bias during library preparation or contamination with longer RNAs, especially rRNA. k-mer distribution can be assessed by, inter alia, FAQCs or FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Readers interested in benchmarking performances (computation time, memory consumption, possibility of multi-threading etc.) or further quality control checks can find short overviews of existing software tools in (165), (167) or (161).

Small RNA read alignment

To extract meaning from the carefully preprocessed data, reads must be mapped to their respective reference and matched with an appropriate annotation. Almost all existing tools start this process by creating an index for either the reads or the reference, which can then be used to find the corresponding sequence or genomic position. Using these indices allows alignment tools to quickly reduce the number of potential locations on the reference by a first heuristic match of reads, followed by a thorough local alignment for each possible match to evaluate the correct alignment. Without this inexact first pass, alignment of millions of nucleotides would take prohibitively long and overtax all but the most sophisticated computational clusters. Common indexing algorithms include hash tables based on principles used by the well-known BLAST aligner (168), or suffix/prefix tries based on Burrows-Wheeler Transform (169). While hash table based aligners have fewer problems identifying even complicated mismatches between read and reference, the computational requirements to do so escalate quickly. Burrows-Wheeler Transform aligners, on the other hand, are extremely fast and efficient in mapping closely matching read-reference pairs, but slow down significantly when challenged with complex misalignments. In general, there is no single 'best' software tool, and the individual per-

formance varies, among other things, with the error rate or genome type of the particular dataset, as well as the allowed mismatch rate (158), although reference indexing tends to outperform read indexing. Frequently used aligning software for small RNA-Seq include Bowtie (170), BWA (171), or SOAP2 (172), but an evaluation of mapping sensitivity and specificity based on an actual dataset is strongly recommended. Readers interested in benchmarking performances (indexing time, mapping throughput, mapping sensitivity etc.), as well as software-specific algorithm variations such as spaced seeding, q-gram filters, and FM-indices can find short overviews of existing software tools in (173), (174) or (158). Researchers with exceptionally large datasets or facing limiting time constraints could benefit from exploring the possibilities offered by multiple processors in high-end graphic cards (e.g. BarraCUDA (175) or SOAP3-dp (176)), or high-performance computing clusters (e.g. MICA (177)).

Classic read alignment strategies include mapping to a reference genome or a specific small RNA database such as mirBase (178,179) or Rfam (180). While reference genomes enable researchers to get the most comprehensive view of their data, allocating reads to all small RNA classes, as well as potential degraded mRNAs and rRNAs, their annotations often lack the extensiveness found in specific small RNA databases, especially in the case of less researched organisms. Additionally, alignment to a genome can lead to problems with reads that map to multiple genomic locations (multireads). Reads without unique genomic locations are mostly caused by sequencing errors or repetitive sequences, but can also originate from genes with multiple genuine copies in the genome (e.g. hsa-let-7a), and incorrect handling of them can lead to a severe bias (181–183). On the other hand, mapping to a reference genome allows for further characterization of unannotated sequences on the basis of their location or accumulation (e.g. novel miRNA prediction). Alignment to a specific small RNA database, however, has its own pros and cons, mostly stemming from a vastly downscaled mapping reference. Most noticeably, alignment is significantly faster and has a considerably reduced memory footprint. Even though multireads are extremely improbable to occur, the likelihood of false positive mappings of reads from non-targeted small RNAs is increased manifold due to the absence of their sequences in the reference. A more conservative mapping with less mismatches is as crucial in avoiding false positive mappings as is filtering for non-targeted small RNA classes (184). Further complicating this is the existence of functionally relevant isoforms such as isoMirs that often differ substantially from their canonical sequence, but have to be taken into account when determining mismatch thresholds and, ultimately, differential expression (185–187). By comparing reads directly to specific sequences, researchers can also take advantage of homologous datasets from well-explored organisms due to the strong conservation of seed sequences between most small RNA classes in different species (e.g miRNAs or piRNAs (188)). After deciding on a mapping strategy, the final step in alignment is matching the database sequence or genomic position to its corresponding small RNA and counting all reads related to the same feature. With annotations available for all major sequenced genomes, these countlists can be easily generated using HTSEQ (189) or R packages such as IRanges, GenomicFeatures (190) or, in the case of an alignment against a specific sequence database, with e.g. SAMtools (191).

Normalization strategies

Although small RNA-Seq features distinctively less noise and technical bias compared to former holistic screening methods such as microarrays (192), it still generates systematic variation that needs to be addressed prior to differential expression analysis. Unwanted differences between libraries commonly occur in size (sequencing depth) (193) as well as within libraries in GC-content (194) or as batch effects (128). Variation introduced by different gene lengths (195), as is frequently encountered in sequencings of longer RNAs, has a negligible effect. Since usual sequencing lengths cover the whole transcript and fragmentation is not necessary during library prep, the still popular Reads-per-Kilobase-per-Million-mapped-reads (193) is therefore not suited for small RNA-Seq. Overall, the general importance of normalization and its impact on differential expression was clearly shown by Bullard *et al*. in 2010 (196). Special attention has to be paid to experimental setups such as degradation studies, where read distributions differ fundamentally from the underlying assumptions of most methods. Most of the currently established and preferred normalization strategies evolve around a global scaling factor per sample to adjust read counts with. Widespread normalization methods include: (i) library size or total mapped reads, where individual read counts are first divided by their respective library size and then multiplied by the arithmetic mean of all library sizes or counts of total mapped reads, respectively. Since individual read counts are not only directly related to sequencing depth, but also dependent on their relative expression compared to all other small RNA expression levels in a sample, this normalization should be avoided (196). (ii) Upper quartile of reads, where transcripts with zero counts across all samples are filtered from the dataset, and a scaling factor is derived for each sample from the 75th quartile of the remaining reads (196). (iii) Quantile, where the distribution of each gene is assumed to be identical, and read counts are adjusted according to a reference obtained from the median of each quantile across all samples (197). (iv) Trimmed Mean of M-values, where a weighted trimmed mean of log expression ratios is calculated for each sample compared to a reference sample. Working under the assumption that expression of most genes will not be significantly altered in the experiment, these means should be close to 1, and a scaling factor is derived from this difference, and finally adjusted by the mean of the normalized libraries (198). (v) Median of expression ratios from geometric means, where a pseudoreference is first created by computing the geometric mean of all genes across samples, and then the ratio for each count to its respective mean is determined. The scaling factor is finally obtained from the median of all ratios for each sample. Similar to (iv), median normalization also assumes most genes to not be differentially expressed (199,200). (vi) Artificial spike-in standards, where reads are quantified using a standard curve derived from a set of pre-determined small RNAs independent of the samples (125). (vii) Surrogate variable analysis, which is

specifically targeted on batch effects, and helps identifying genomic data affected by artifacts. It adjusts read counts by estimating these artifacts with the help of singular vectors of the specific subset of the data (201).

Although the variety in experimental and genomic set ups so far makes it impossible to universally recommend a single normalization strategy, recent evaluations of these methods have found Median normalizing of expression ratios from geometric means to work favorably with various kinds of datasets (202,203). Additionally, Zyprich-Walczak *et al*. proposed a step-by-step workflow to determine the most appropriate normalization method for a specific dataset in terms of bias, variance, sensitivity, specificity and prediction errors to avoid data distortion by using the wrong normalization (203).

Differential expression analysis

In comparison to normalization strategies that were mostly extending existing methods for microarrays, the distinctly different data type of NGS made the development of new algorithms for differential expression analysis imperative. While microarray data consists of continuous intensities coupled with a high background, NGS read counts give discrete measurements for each gene, and should not, unlike microarray intensities, be modeled on a normal distribution. Although early RNA-Seq reported a good fit to a Poisson distribution for single sample sequencings and technical replicates (30,196), studies with biological replicates are extremely likely to show variances greater than the mean for many genes (204). This so-called overdispersion makes analyses working under the Poisson assumption prone to high false-positive rates due to an underestimation of sampling error. One way to overcome this is an extension of the Poisson model with a quasi-likelihood approach, where each gene is tested individually for overdispersion (Two-Stage-Poisson-Model (205)). Another way to account for biological variability is the negative binomial distribution, which adds the dispersion to the mean as a second parameter (206). Correct estimation of gene-wise dispersion factors is crucial, but unfortunately also hampered by the still prevalent low number of sample in most RNA-Seq studies. To obtain more accurate dispersion factors, analysis tools share information across all genes in the dataset by, among other things, a weighted likelihood approach toward the common dispersion (edgeR (207)) or by modeling the observed mean-variance relationship for all genes via regression (DESeq (199,200)). Differential expression can then be tested by either exact tests (edgeR, DESeq) or empirical Bayesian frameworks (EBSeq (208), baySeq (209)). Apart from these distribution assumptions, differential expression can also be assessed by non-parametric approaches based for instance on Wilcoxon rank statistics and resampling strategies (SAMSeq (210)), or by comparing the absolute and relative expression differences between and within experimental conditions (NOISeq (211)). A major drawback of these methods is their relatively low power and specificity in experiments with low sample numbers. In addition, robust methods established for microarrays (limma (212,213)) can be made applicable through transformation of discrete read count data (voom (214)). Irrelevant of the employed algorithm, all tools will produce a list of significantly regulated genes that should be treated with caution. Due to the large number of tests, the false discovery rate should be controlled for all results to avoid accumulation of type-1-errors (215). Additionally, the ratio of expression signal to experimental noise should be monitored for lowly expressed genes by assessing the biological relevance of the fold change, as well as absolute read count values.

More so than any other tools, software for differential expression is subject to frequent updates, which can alter their behavior dramatically and new algorithms are published continually. Even though comparisons of software performances on small RNA-Seq data are scarce, a number of independent and extensive evaluations for mRNAs based on either synthetic data with clearly defined properties (216), or on biological datasets with validated gene expressions (217,218) have been made recently. While it was shown that statistical power of almost all methods is heavily dependent on the number of samples per condition and less on sequencing depth, the variability of expression changes in biological datasets affects each analysis tool differently. Outliers, 'ON/OFF' expression changes, where a gene is detected in only one condition, and lopsided expression patterns, where upregulations drastically outweigh downregulations or *vice versa*, influence specificity (false positive rate) and sensitivity (false negative rate) of each method unequally. Nonetheless, some methods appear to capture the true expression status of small RNAs better than others. Most independent evaluations seem to agree that calling differential expression with SAMSeq works well for datasets with sufficient sample sizes of 10 or more. For smaller datasets, edgeR and especially the more conservative DESeq (or DESeq2) are found to be the methods of choice. On top of that, the voom + limma method was reported to generally perform well for different datasets (216). Additionally, a recent publication on RNA-Seq showed that most of the frequently used tools correctly assess differential gene expression when sufficient biological replication is employed (28). For a low number of replicates, edgeR outperformed its competitors, while DESeq exceled in experiments with more than 12 replicates, suggesting that data analysis tools need to fit the respective experimental setup. Efforts with mixed results have also been made to weigh differential expression results of various methods and combine them to an optimized consensus bypassing the individual flaws of each algorithm (219). Considering all this, choosing the optimal tool for differential expression analysis is still strongly dependent on the individual dataset, highlighting once again the fact that researchers need to thoroughly acquaint themselves with the details and specifics of their individual setup and data distribution before starting any analyses.

BIOMARKER IDENTIFICATION AND VALIDATION

After biomarker candidates have been identified in the differential expression analysis, these markers have to be statistically validated. Since univariate analyses, like most differential expression tests, treat each biomarker (i.e. small RNA) as independent, they are unable to capture the complete reality of highly multivariate (variables >> observa-

tions) and correlated datasets such as NGS read counts. By taking the synergies, antagonisms and redundancy inherent in each NGS dataset into consideration, multivariate analyses can reach much higher discriminative power and separate noise from signal (19,220). In reality, there will most likely be no single valid transcriptional biomarker for the physiological situation of interest. In most cases, only a set of multiple biomarkers can ensure the high sensitivity, specificity and reliability needed for diagnostic and prognostic analyses. Appropriately dealing with these data to retrieve the desired outcome of a stable and valid biomarker signature is, however, not trivial.

The most promising approach is to first screen read counts for general trends or potential outliers in an unsupervised manner (no classification information is given to the algorithm), and subsequently assess the discriminative power of potential biomarker candidates (221). These analyses generate clusters of similarities, specifically similar gene expression patterns in the case of RNA-Seq, by using methods for dimension reduction combined with pattern recognition technologies and visualize them in two- or three-dimensional graphs (222). Similar to differential expression profiling, read count lists need to be preprocessed. Input data for any cluster or classification analysis can either be normalized read counts, as described previously, or ratios thereof, and in addition should be transformed to address their skewed distribution. A simple shifted log transformation ($\log_2 (n + 0.5)$) to make the data conform to normality is most commonly used, but more sophisticated alternatives such as regularized log transformation (rlog, (200)) and variance stabilizing transformation (vst, (223)) might be better suited for small RNA-Seq data (both algorithms are implemented in DESeq2, (200)). Cluster algorithms are implemented, for instance, in the base distribution of R, as well as more comprehensive packages such as pcaMethods (224) and the excellent mixOmics (225), or the commercially available Simca-Q software (Umetrics).

Widely accepted unsupervised multivariate analyses include clustering analyses such as hierarchical clustering (HCA), partitioning methods such as k-means and self-organizing maps (SOM), as well as projections on latent variables such as the powerful principal component analysis (PCA). In agglomerative HCA, samples (or genes) start as single entity clusters and are then joined step-by-step based on a similarity measure and a linkage function, defining inter-cluster distances. For log-transformed data, it was shown that Euclidian distances and Pearson correlation perform well as distance measures, while complete linkage (or Ward's method) strictly surpass single or average linkage functions (226). The result and graphical output of HCA is a tree dendrogram emphasizing the distances between the individual samples (or genes) with rising node lengths and clusters can be obtained by, among others things, cutting at fixed heights (227,228). Combining HCA of samples and genes with a two-dimensional color-coded description of the whole experimental matrix creates a heatmap, which allows for easy detection of similarities and dissimilarities in a read count list. Although HCA is the still the most common clustering algorithm, it is in most cases outperformed by partitioning methods such as k-means and SOM (226,229). Both work by subdividing the dataset into a predetermined number of unhierarchical subsets based on randomly chosen centroids. In k-means, samples (or genes) are iteratively assigned to the closest centroid with each iteration replacing the former centroid by the average of each entity in its cluster until all samples (or genes) are set. In SOM, the centroids are linked by a grid structure, and with each iteration the closest centroid, as well as its neighbors, is moved toward a randomly chosen sample (or gene). By gradually shrinking the radius of each adjacent centroid, this will result in a grid of clusters comprising all samples (or genes) with related expression patterns. Since both k-means and SOM start with randomly placed centroids and the optimal number of cluster is usually not apparent, these algorithms should be rerun with random seeds and different numbers of clusters to obtain a stable classification.

Even more information on potential biomarkers can be obtained by PCA, which converts a multidimensional dataset into a lower number of variables called principal components (PCs) (228,230). The read count data is thus decomposed in a score matrix describing small RNA genes, a loadings matrix describing the samples, and a residual matrix expressing deviations between the original variables and the projections. PCs are calculated ranked with the first PC accounting for the greatest variance in the dataset and subsequent PCs comprising the respective maximum residual variance. Since PCs are computed orthogonally to each other, they each describe independent sources of information, and with decreasing variance explained by later PCs, they can be used to separate systematic effects, explained by the molecular biomarker set, from random expression noise (227). Variance derived from experimental study design is expected to be systematic, while confounding variance is expected to be small and random and can therefore be found in later PCs. The advantage of PCA in comparison to clustering and partitioning methods is obvious, since it allows a much clearer recognition and more precise differentiation of the experimental groups. In PCA, the commonalities (or differences) in gene expression pattern are clearly visualized by the symbol interspaces in at least two dimensions (228,231). By plotting scores and loadings plots side by side and looking at their corresponding positioning, it is also possible to identify which small RNA genes are responsible for the separations of samples. Potential biomarkers can be assessed by their contribution plots, and outliers can be detected by either Hotelling's T^2 or by their residual standard deviation (distance to model, DModX) (232).

All of the unsupervised methods mentioned above generate groupings of samples (or genes) with similar expression patterns. While this allows for easy detection of outliers and inconsistencies in experimental setup, it does not necessarily mean that resulting clusters will reflect the desired classification of samples or genes. An underlying treatment effect can sometimes be veiled by other dominating effects, be they intentional (different cell types, time points etc.) or not (batch effects). By incorporating information on experimental setup, researchers are able to filter out genes inducing the greatest separation between treatment groups or, in other words, potential biomarkers. Although a number of supervised classifications algorithms exist, it was shown that the widely used partial least squares projection to latent structures (PLS) and its modifications such as PLS discrim-

inant analysis (PLS-DA, (233)), sparse PLS-DA (sPLS-DA, (234)) or orthogonal PLS (OPLS, (235)) are well suited for dimension reduction and discrimination (233,236).

PLS is related to linear discriminant analyses (LDA), and is a regression extension of PCA that shares many characteristics with it. By adding a second matrix containing the responses or dependent variables to the read count matrix, PLS attempts to find latent variables (LV) that predict the responses from gene expression profiles and describe the common structure of both matrices. LVs are calculated hierarchically similar to PCs, but LVs maximize covariance instead of variance. In PLS-DA, the response matrix is replaced by an optimized dummy matrix containing only 0 and 1 for every respective class, and the resulting projection model therefore focuses on maximum discrimination between classes in the responses rather than 'optimal class modeling' (221). Biomarkers can then be evaluated by a number of variable selection methods including variable importance in projection (equivalent to a contribution plot in PCA) or target projection with selectivity ratio test (237), and by drawing a consensus between differentially expressed genes and multivariate analyses.

The biological functionality of detected small RNA biomarkers, mainly based on miRNAs, can be further verified in functional experimental tests using miRNA overexpression, knockdown or even knockout experiments. Various tools and software packages are available for the *in silico* functional analysis of miRNAs. For *in silico* target prediction, we recommend the TargetScan package (http://www.targetscan.org/) (238,239) or miRanda (http://www.microrna.org/) (240,241). For analyzing the inverse relation of expressed miRNAs and mRNAs in conjunction with target predictions, we recommend using a Lasso regression model (242,243). If an integrative analysis of miRNAs and their target genes is of interest, the miRNA–mRNA relations can be tested on the basis of regression analysis, and further processed by testing for enrichment in gene ontology terms or KEGG pathways (http://www.genome.jp/kegg/pathway.html), amongst others (244,245). In addition, several all-in-one software packages such as CLC Genomics Workbench (Qiagen), Ingenuity Pathway Analysis (Qiagen) or Genomatix Genome Analyzer (Genomatix) are available to allow a relatively easy, graphic user interface (GUI)-based *in silico* functional analysis of miRNAs. Applying Genomatix Pathway System (GEPS) or Ingenuity Pathway Analysis facilitates the creation and extension of miRNA networks based on information extracted from public and proprietary databases and co-citations in the literature.

Conclusion - where are the real bottlenecks?

Today, liquid biopsies and the small RNA biomarker signatures they may inclose are considered the promising new generation of transcriptional biomarkers. The RNA is easily accessible, often by non-invasive procedures, physiologically stable and protected by microvesicles or associated proteins. Due to its chemical nature, it can be rapidly amplified and quantified using RT and PCR-related methods. Small RNA-based biomarker signatures can therefore be detected at low concentrations and early disease stages, and the discovery workflow can be further optimized and standardized. This sustains the idea of the MIQE and dMIQE guidelines previously published by an international consortium (headed by SA Bustin and JF Huggett) in the field of qPCR and dPCR (21,22).

Thoroughly and accurately following our recommendations by optimizing and standardizing the small RNA-Seq workflow will result in reproducible data and, subsequently, reliable hypotheses. The digital and holistic nature of the small RNA-Seq approach provides vast transcriptional data that is highly informative in terms of both quality and quantity (246). The subsequent complex, comparative and multivariate data analysis can result in valid biomarker signatures. The technological developments in the entire workflow (from sampling to multivariate data analysis) are very dynamic, and will continue to improve in the future. While proven standards and optimized methodologies to identify promising biomarkers in liquid biopsies are still lacking, the optimization and validation process will continue to develop.

Where are the real bottlenecks in small RNA-Seq analysis of liquid biopsies? The most significant factor leading to success is probably the number of variables and conditions being tested, and the number of real biological replicates used for sequencing. What appears to be specific in the particular biological samples analyzed by small RNA-Seq may not necessarily be reflected in a larger group, or even in the entire population. Therefore, the more individuals tested, and the more conditions or variables being evaluated, the better the outcome of the prediction and the validity of the discovered biomarker signature will be (247,248).

No step in the workflow is free of bias, but some are more prone to produce noise in the resulting data. Due to financial reasons, researchers still employ too few biological replicates. Only biological replicates can explain any biological difference, while technical replicates are limited to only report the technical noise researchers introduce. In our opinion, the largest noise impact is introduced by RNA extraction and the complex library preparation, which can be performed in various ways, but always highly depends on enzyme efficiency. Depending on the respective library preparation chemistry, numerous individual barcodes are used. These not only cause technical bias, but also affect RT efficiency and PCR amplification.

In general, it is recommended to perform as few PCR cycles as possible for pre-amplification, and to only compare replicates with the identical number of cycles. The sequencing or clonal amplification as such is not a major source of variation, since error rates of polymerases are acceptably low, sequencing chemistry exhibits high purity and the hardware operates very precisely and reproducibly. A further big challenge is the off-instrument data analysis, which requires the majority of manpower and time in the quantification workflow. We should put major focus on alignment, normalization and differential expression analysis, since these are the most critical steps. Biases introduced at earlier stages can in part be corrected and compensated by an appropriate normalization strategy.

As a final and essential step after small RNA-Seq, we recommend additional validation of the identified transcriptional biomarker signatures. This confirmation should be

carried out using established and highly standardized methods such as RT in combination with real-time PCR or digital PCR. The consistency and correctness of the discovered transcriptional biomarker signature in the liquid biopsy can only be assumed after data verification and demonstration of a statistically validated correlation between small RNA-Seq and RT-qPCR or dPCR.

ACKNOWLEDGEMENT

This work was supported by the Vereinigung zur Förderung der Milchwissenschaftlichen Forschung an der Technischen Universität München e.V.

FUNDING

Funding for open access charge: Internal institutional budget.
Conflict of interest statement. None declared.

REFERENCES

1. Riedmaier,I., Pfaffl,M.W. and Meyer,H.H. (2012) The physiological way: monitoring RNA expression changes as new approach to combat illegal growth promoter application. *Drug Test Anal.*, **4**(Suppl. 1), 70–74.
2. Pfaffl,M.W. (2013) Transcriptional biomarkers. *Methods*, **59**, 1–2.
3. Sewall,C.H., Bell,D.A., Clark,G.C., Tritscher,A.M., Tully,D.B., Vanden Heuvel,J. and Lucier,G.W. (1995) Induced gene transcription: implications for biomarkers. *Clin. Chem.*, **41**, 1829–1834.
4. Pfaffl,M.W. (2015) Guest editor's introduction for BDQ special issue: 'Advanced Molecular Diagnostics for Biomarker Discovery'. *Biomol. Detect. Quantif.*, **5**, 1–2.
5. Karachaliou,N., Mayo-de-Las-Casas,C., Molina-Vila,M.A. and Rosell,R. (2015) Real-time liquid biopsies become a reality in cancer treatment. *Ann. Transl. Med.*, **3**, doi:10.3978/j.issn.2305-5839.2015.01.16.
6. Buder,A., Tomuta,C. and Filipits,M. (2016) The potential of liquid biopsies. *Curr. Opin. Oncol.*, **28**, 130–134.
7. Properzi,F., Logozzi,M. and Fais,S. (2013) Exosomes: the future of biomarkers in medicine. *Biomark. Med.*, **7**, 769–778.
8. Fleischhacker,M. and Schmidt,B. (2007) Circulating nucleic acids (CNAs) and cancer–a survey. *Biochim. Biophys. Acta*, **1775**, 181–232.
9. Pinzani,P., Salvianti,F., Pazzagli,M. and Orlando,C. (2010) Circulating nucleic acids in cancer and pregnancy. *Methods*, **50**, 302–307.
10. Calin,G.A., Ferracin,M., Cimmino,A., Di Leva,G., Shimizu,M., Wojcik,S.E., Iorio,M.V., Visone,R., Sever,N.I., Fabbri,M. *et al.* (2005) A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *N. Engl. J. Med.*, **353**, 1793–1801.
11. Reid,G., Kirschner,M.B. and van Zandwijk,N. (2011) Circulating microRNAs: association with disease and potential use as biomarkers. *Crit. Rev. Oncol. Hematol.*, **80**, 193–208.
12. Kirschner,M.B., van Zandwijk,N. and Reid,G. (2013) Cell-free microRNAs: potential biomarkers in need of standardized reporting. *Front. Genet.*, **4**, doi:10.3389/fgene.2013.00056.
13. Ono,S., Lam,S., Nagahara,M. and Hoon,D.S. (2015) Circulating microRNA biomarkers as liquid biopsy for cancer patients: pros and cons of current assays. *J. Clin. Med.*, **4**, 1890–1907.
14. Witwer,K.W. (2015) Circulating microRNA biomarker studies: pitfalls and potential solutions. *Clin. Chem.*, **61**, 56–63.
15. Hayes,C.J. and Dalton,T.M. (2015) Microfluidic droplet-based PCR instrumentation for high-throughput gene expression profiling and biomarker discovery. *Biomol. Detect. Quantif.*, **4**, 22–32.
16. Sanders,R., Mason,D.J., Foy,C.A. and Huggett,J.F. (2013) Evaluation of digital PCR for absolute RNA quantification. *PLoS One*, **8**, e75296.
17. Meyer,S.U., Kaiser,S., Wagner,C., Thirion,C. and Pfaffl,M.W. (2012) Profound effect of profiling platform and normalization strategy on detection of differentially expressed microRNAs–a comparative study. *PLoS One*, **7**, e38946.
18. Meyer,S.U., Pfaffl,M.W. and Ulbrich,S.E. (2010) Normalization strategies for microRNA profiling experiments: a 'normal' way to a hidden layer of complexity? *Biotechnol. Lett.*, **32**, 1777–1788.
19. Spornraft,M., Kirchner,B., Pfaffl,M.W. and Irmgard,R. (2015) The potential of circulating extracellular small RNAs (smexRNA) in veterinary diagnostics—Identifying biomarker signatures by multivariate data analysis. *Biomol. Detect. Quantif.*, **5**, 15–22.
20. Bustin,S.A. (2014) The reproducibility of biomedical research: sleepers awake! *Biomol. Detect. Quantif.*, **2**, 35–42.
21. Bustin,S.A., Benes,V., Garson,J.A., Hellemans,J., Huggett,J., Kubista,M., Mueller,R., Nolan,T., Pfaffl,M.W., Shipley,G.L. *et al.* (2009) The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin. Chem.*, **55**, 611–622.
22. Huggett,J.F., Foy,C.A., Benes,V., Emslie,K., Garson,J.A., Haynes,R., Hellemans,J., Kubista,M., Mueller,R.D., Nolan,T. *et al.* (2013) The digital MIQE guidelines: minimum information for publication of quantitative digital PCR experiments. *Clin Chem*, **59**, 892–902.
23. Tarazona,S., Garcia-Alcalde,F., Dopazo,J., Ferrer,A. and Conesa,A. (2011) Differential expression in RNA-seq: a matter of depth. *Genome Res.*, **21**, 2213–2223.
24. Hart,S.N., Therneau,T.M., Zhang,Y., Poland,G.A. and Kocher,J.-P. (2013) Calculating sample size estimates for RNA sequencing data. *J. Comput. Biol.*, **20**, 970–978.
25. Liu,Y., Zhou,J. and White,K.P. (2013) RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics*, **30**, 301–304.
26. Williams,A.G., Thomas,S., Wyman,S.K. and Holloway,A.K. (2014) RNA-seq data: challenges in and recommendations for experimental design and analysis. *Curr. Protoc. Hum. Genet.*, **83**, 11–20.
27. Kvam,V.M., Liu,P. and Si,Y. (2012) A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *Am. J. Bot.*, **99**, 248–256.
28. Schurch,N.J., Schofield,P., Gierliński,M., Cole,C., Sherstnev,A., Singh,V., Wrobel,N., Gharbi,K., Simpson,G.G., Owen-Hughes,T. *et al.* (2016) How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, **22**, 839–851.
29. Fang,Z. and Cui,X. (2011) Design and validation issues in RNA-seq experiments. *Brief. Bioinform.*, **12**, 280–287.
30. Marioni,J.C., Mason,C.E., Mane,S.M., Stephens,M. and Gilad,Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.
31. Tichopad,A., Kitchen,R., Riedmaier,I., Becker,C., Stahlberg,A. and Kubista,M. (2009) Design and optimization of reverse-transcription quantitative PCR experiments. *Clin. Chem.*, **55**, 1816–1823.
32. Stahlberg,A., Kubista,M. and Pfaffl,M. (2004) Comparison of reverse transcriptases in gene expression analysis. *Clin. Chem.*, **50**, 1678–1680.
33. Scicchitano,M.S., Dalmas,D.A., Bertiaux,M.A., Anderson,S.M., Turner,L.R., Thomas,R.A., Mirable,R. and Boyce,R.W. (2006) Preliminary comparison of quantity, quality, and microarray performance of RNA extracted from formalin-fixed, paraffin-embedded, and unfixed frozen tissue samples. *J. Histochem. Cytochem.*, **54**, 1229–1237.
34. Li,J., Smyth,P., Cahill,S., Denning,K., Flavin,R., Aherne,S., Pirotta,M., Guenther,S.M., O'Leary,J.J. and Sheils,O. (2008) Improved RNA quality and TaqMan® Pre-amplification method (PreAmp) to enhance expression analysis from formalin fixed paraffin embedded (FFPE) materials. *BMC Biotechnol.*, **8**, doi:10.1186/1472-6750-8-10.
35. Micke,P., Ohshima,M., Tahmasebpoor,S., Ren,Z.P., Ostman,A., Ponten,F. and Botling,J. (2006) Biobanking of fresh frozen tissue: RNA is stable in nonfixed surgical specimens. *Lab. Invest.*, **86**, 202–211.
36. Dekairelle,A.F., Van der Vorst,S., Tombal,B. and Gala,J.L. (2007) Preservation of RNA for functional analysis of separated alleles in

yeast: comparison of snap-frozen and RNALater solid tissue storage methods. *Clin. Chem. Lab. Med.*, **45**, 1283–1287.
37. Schroeder,A., Mueller,O., Stocker,S., Salowsky,R., Leiber,M., Gassmann,M., Lightfoot,S., Menzel,W., Granzow,M. and Ragg,T. (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.*, **7**, doi:10.1186/1471-2199-7-3.
38. Jahn,C.E., Charkowski,A.O. and Willis,D.K. (2008) Evaluation of isolation methods and RNA integrity for bacterial RNA quantitation. *J. Microbiol. Methods*, **75**, 318–324.
39. Die,J.V. and Roman,B. (2012) RNA quality assessment: a view from plant qPCR studies. *J. Exp. Bot.*, **63**, 6069–6077.
40. Heera,R., Sivachandran,P., Chinni,S.V., Mason,J., Croft,L., Ravichandran,M. and Yin,L.S. (2015) Efficient extraction of small and large RNAs in bacteria for excellent total RNA sequencing and comprehensive transcriptome analysis. *BMC Res. Notes*, **8**, doi:10.1186/s13104-015-1726-3.
41. Die,J.V., Obrero,Á., González-Verdejo,C.I. and Román,B. (2011) Characterization of the 3′:5′ ratio for reliable determination of RNA quality. *Anal. Biochem.*, **419**, 336–338.
42. Copois,V., Bibeau,F., Bascoul-Mollevi,C., Salvetat,N., Chalbos,P., Bareil,C., Candeil,L., Fraslon,C., Conseiller,E., Granci,V. *et al.* (2007) Impact of RNA degradation on gene expression profiles: assessment of different methods to reliably determine RNA quality. *J. Biotechnol.*, **127**, 549–559.
43. Fleige,S. and Pfaffl,M.W. (2006) RNA integrity and the effect on the real-time qRT-PCR performance. *Mol. Aspects Med.*, **27**, 126–139.
44. Fleige,S., Walf,V., Huch,S., Prgomet,C., Sehm,J. and Pfaffl,M.W. (2006) Comparison of relative mRNA quantification models and the impact of RNA integrity in quantitative real-time RT-PCR. *Biotechnol. Lett.*, **28**, 1601–1613.
45. Gallego Romero,I., Pai,A.A., Tung,J. and Gilad,Y. (2014) RNA-seq: impact of RNA degradation on transcript quantification. *BMC Biol.*, **12**, doi:10.1186/1741-7007-12-42.
46. Feng,H., Zhang,X. and Zhang,C. (2015) mRIN for direct assessment of genome-wide and gene-specific mRNA integrity from large-scale RNA-sequencing data. *Nat. Commun.*, **6**, 7816–7826.
47. Becker,C., Hammerle-Fickinger,A., Riedmaier,I. and Pfaffl,M.W. (2010) mRNA and microRNA quality control for RT-qPCR analysis. *Methods*, **50**, 237–243.
48. Etheridge,A., Gomes,C.P.C., Pereira,R.W., Galas,D. and Wang,K. (2013) The complexity, function and applications of RNA in circulation. *Front. Genet.*, **4**, doi:10.3389/fgene.2013.00115.
49. Vickers,K.C., Palmisano,B.T., Shoucri,B.M., Shamburek,R.D. and Remaley,A.T. (2011) MicroRNAs are transported in plasma and delivered to recipient cells by high-density lipoproteins. *Nat. Cell Biol.*, **13**, 423–433.
50. Arroyo,J.D., Chevillet,J.R., Kroh,E.M., Ruf,I.K., Pritchard,C.C., Gibson,D.F., Mitchell,P.S., Bennett,C.F., Pogosova-Agadjanyan,E.L., Stirewalt,D.L. *et al.* (2011) Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 5003–5008.
51. Valadi,H., Ekstrom,K., Bossios,A., Sjostrand,M., Lee,J.J. and Lotvall,J.O. (2007) Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. *Nat. Cell Biol.*, **9**, 654–659.
52. Taylor,D.D. and Gercel-Taylor,C. (2008) MicroRNA signatures of tumor-derived exosomes as diagnostic biomarkers of ovarian cancer. *Gynecol. Oncol.*, **110**, 13–21.
53. Chim,S.S., Shing,T.K., Hung,E.C., Leung,T.Y., Lau,T.K., Chiu,R.W. and Lo,Y.M. (2008) Detection and characterization of placental microRNAs in maternal plasma. *Clin. Chem.*, **54**, 482–490.
54. Spornraft,M., Kirchner,B., Haase,B., Benes,V., Pfaffl,M.W. and Riedmaier,I. (2014) Optimization of extraction of circulating RNAs from plasma—enabling small RNA sequencing. *PLoS One*, **9**, e107259.
55. Yanez-Mo,M., Siljander,P.R., Andreu,Z., Zavec,A.B., Borras,F.E., Buzas,E.I., Buzas,K., Casal,E., Cappello,F., Carvalho,J. *et al.* (2015) Biological properties of extracellular vesicles and their physiological functions. *J. Extracell. Vesicles*, **4**, doi:10.3402/jev.v4.27066.
56. Yokoi,A., Yoshioka,Y. and Ochiya,T. (2015) Towards the realization of clinical extracellular vesicle diagnostics: challenges and opportunities. *Expert Rev. Mol. Diagn.*, **12**, 1555–1566.
57. Gould,S.J. and Raposo,G. (2013) As we wait: coping with an imperfect nomenclature for extracellular vesicles. *J. Extracell. Vesicles*, **2**, doi:10.3402/jev.v2i0.20389.
58. Lobb,R.J., Becker,M., Wen,S.W., Wong,C.S., Wiegmans,A.P., Leimgruber,A. and Moller,A. (2015) Optimized exosome isolation protocol for cell culture supernatant and human plasma. *J. Extracell. Vesicles*, **4**, doi:10.3402/jev.v4.27031.
59. He,M., Crow,J., Roth,M., Zeng,Y. and Godwin,A.K. (2014) Integrated immunoisolation and protein analysis of circulating exosomes using microfluidic technology. *Lab. Chip*, **14**, 3773–3780.
60. Welton,J.L., Webber,J.P., Botos,L.A., Jones,M. and Clayton,A. (2015) Ready-made chromatography columns for extracellular vesicle isolation from plasma. *J. Extracell. Vesicles*, **4**, doi:10.3402/jev.v4.27269.
61. Baranyai,T., Herczeg,K., Onodi,Z., Voszka,I., Modos,K., Marton,N., Nagy,G., Mager,I., Wood,M.J., El Andaloussi,S. *et al.* (2015) Isolation of exosomes from blood plasma: qualitative and quantitative comparison of ultracentrifugation and size exclusion chromatography methods. *PLoS One*, **10**, e0145686.
62. Greening,D.W., Xu,R., Ji,H., Tauro,B.J. and Simpson,R.J. (2015) A protocol for exosome isolation and characterization: evaluation of ultracentrifugation, density-gradient separation, and immunoaffinity capture methods. *Methods Mol. Biol.*, **1295**, 179–209.
63. Kalra,H., Drummen,G.P. and Mathivanan,S. (2016) Focus on extracellular vesicles: introducing the next small big thing. *Int. J. Mol. Sci.*, **17**, doi:10.3390/ijms17020170.
64. Szatanek,R., Baran,J., Siedlar,M. and Baj-Krzyworzeka,M. (2015) Isolation of extracellular vesicles: determining the correct approach (Review). *Int. J. Mol. Med.*, **36**, 11–17.
65. Zeringer,E., Barta,T., Li,M. and Vlassov,A.V. (2015) Strategies for isolation of exosomes. *Cold Spring Harb. Protoc.*, **2015**, 319–323.
66. Van Deun,J., Mestdagh,P., Sormunen,R., Cocquyt,V., Vermaelen,K., Vandesompele,J., Bracke,M., De Wever,O. and Hendrix,A. (2014) The impact of disparate isolation methods for extracellular vesicles on downstream RNA profiling. *J. Extracell. Vesicles*, **3**, doi:10.3402/jev.v3.24858.
67. Rekker,K., Saare,M., Roost,A.M., Kubo,A.L., Zarovni,N., Chiesi,A., Salumets,A. and Peters,M. (2014) Comparison of serum exosome isolation methods for microRNA profiling. *Clin. Biochem.*, **47**, 135–138.
68. Royo,F., Zuniga-Garcia,P., Sanchez-Mosquera,P., Egia,A., Perez,A., Loizaga,A., Arceo,R., Lacasa,I., Rabade,A., Arrieta,E. *et al.* (2016) Different EV enrichment methods suitable for clinical settings yield different subpopulations of urinary extracellular vesicles from human samples. *J. Extracell. Vesicles*, **5**, doi:10.3402/jev.v5.29497.
69. Kowal,J., Arras,G., Colombo,M., Jouve,M., Morath,J.P., Primdal-Bengtson,B., Dingli,F., Loew,D., Tkach,M. and Thery,C. (2016) Proteomic comparison defines novel markers to characterize heterogeneous populations of extracellular vesicle subtypes. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E968–E977.
70. Willms,E., Johansson,H.J., Mager,I., Lee,Y., Blomberg,K.E., Sadik,M., Alaarg,A., Smith,C.I., Lehtio,J., El Andaloussi,S. *et al.* (2016) Cells release subpopulations of exosomes with distinct molecular and biological properties. *Sci. Rep.*, **6**, doi:10.1038/srep22519.
71. Paolini,L., Zendrini,A., Noto,G.D., Busatto,S., Lottini,E., Radeghieri,A., Dossi,A., Caneschi,A., Ricotta,D. and Bergese,P. (2016) Residual matrix from different separation techniques impacts exosome biological activity. *Sci. Rep.*, **6**, doi:10.1038/srep23550.
72. Abramowicz,A., Widlak,P. and Pietrowska,M. (2016) Proteomic analysis of exosomal cargo: the challenge of high purity vesicle isolation. *Mol. Biosyst.*, **12**, 1407–1419.
73. Müller,R., Häntzsch,M., Tolios,A., Beutner,F., Nagel,D., Thiery,J., Teupser,D. and Holdt,L.M. (2014) Comparison of whole blood RNA preservation tubes and novel generation RNA extraction kits for analysis of mRNA and MiRNA profiles. *PLoS One*, **9**, e113298.
74. Nikula,T., Mykkänen,J., Simell,O. and Lahesmaa,R. (2013) Genome-wide comparison of two RNA-stabilizing reagents for transcriptional profiling of peripheral blood. *Transl. Res.*, **161**, 181–188.
75. Hammerle-Fickinger,A., Riedmaier,I., Becker,C., Meyer,H.H., Pfaffl,M.W. and Ulbrich,S.E. (2010) Validation of extraction

methods for total RNA and miRNA from bovine blood prior to quantitative gene expression analyses. *Biotechnol. Lett.*, **32**, 35–44.
76. Jiang,Z., Uboh,C.E., Chen,J. and Soma,L.R. (2013) Isolation of RNA from equine peripheral blood cells: comparison of methods. *Springerplus*, **2**, 478–484.
77. Bayatti,N., Cooper-Knock,J., Bury,J.J., Wyles,M., Heath,P.R., Kirby,J. and Shaw,P.J. (2014) Comparison of blood RNA extraction methods used for gene expression profiling in amyotrophic lateral sclerosis. *PLoS One*, **9**, e87508.
78. Wieczorek,D., Delauriere,L., Schagat,T. and Promega Corporation. (2012) Methods of RNA Quality Assessment. http://www.promega.de/resources/pubhub/methods-of-rna-quality-assessment/.
79. McAlexander,M.A., Phillips,M.J. and Witwer,K.W. (2013) Comparison of methods for miRNA extraction from plasma and quantitative recovery of RNA from cerebrospinal fluid. *Front. Genet.*, **4**, 1–8.
80. Li,Y. and Kowdley,K.V. (2012) Method for microRNA isolation from clinical serum samples. *Anal. Biochem.*, **431**, 69–75.
81. Burgos,K.L., Javaherian,A., Bomprezzi,R., Ghaffari,L., Rhodes,S., Courtright,A., Tembe,W., Kim,S., Metpally,R. and Van Keuren-Jensen,K. (2013) Identification of extracellular miRNA in human cerebrospinal fluid by next-generation sequencing. *RNA*, **19**, 712–722.
82. Raabe,C.A., Tang,T.H., Brosius,J. and Rozhdestvensky,T.S. (2014) Biases in small RNA deep sequencing data. *Nucleic Acids Res.*, **42**, 1414–1426.
83. Linsen,S.E., de Wit,E., Janssens,G., Heater,S., Chapman,L., Parkin,R.K., Fritz,B., Wyman,S.K., de Bruijn,E., Voest,E.E. *et al.* (2009) Limitations and possibilities of small RNA digital gene expression profiling. *Nat. Methods*, **6**, 474–476.
84. Sterling,C.H., Veksler-Lublinsky,I. and Ambros,V. (2015) An efficient and sensitive method for preparing cDNA libraries from scarce biological samples. *Nucleic Acids Res.*, **43**, e1.
85. Blomquist,T., Crawford,E.L., Yeo,J., Zhang,X. and Willey,J.C. (2015) Control for stochastic sampling variation and qualitative sequencing error in next generation sequencing. *Biomol. Detect. Quantif.*, **5**, 30–37.
86. van Dijk,E.L., Jaszczyszyn,Y. and Thermes,C. (2014) Library preparation methods for next-generation sequencing: tone down the bias. *Exp. Cell Res.*, **322**, 12–20.
87. Kim,Y.K., Yeo,J., Kim,B., Ha,M. and Kim,V.N. (2012) Short structured RNAs with low GC content are selectively lost during extraction from a small number of cells. *Mol. Cell*, **46**, 893–895.
88. Jackson,T.J., Spriggs,R.V., Burgoyne,N.J., Jones,C. and Willis,A.E. (2014) Evaluating bias-reducing protocols for RNA sequencing library preparation. *BMC Genomics*, **15**, doi:10.1186/1471-2164-15-569.
89. Zhuang,F., Fuchs,R.T., Sun,Z., Zheng,Y. and Robb,G.B. (2012) Structural bias in T4 RNA ligase-mediated 3′-adapter ligation. *Nucleic Acids Res.*, **40**, e54.
90. Zhuang,F., Fuchs,R.T. and Robb,G.B. (2012) Small RNA expression profiling by high-throughput sequencing: implications of enzymatic manipulation. *J. Nucleic Acids*, **2012**, doi:10.1155/2012/360358.
91. Sun,G., Wu,X., Wang,J., Li,H., Li,X., Gao,H., Rossi,J. and Yen,Y. (2011) A bias-reducing strategy in profiling small RNAs using Solexa. *RNA*, **17**, 2256–2262.
92. Alon,S., Vigneault,F., Eminaga,S., Christodoulou,D.C., Seidman,J.G., Church,G.M. and Eisenberg,E. (2011) Barcoding bias in high-throughput multiplex sequencing of miRNA. *Genome Res.*, **21**, 1506–1511.
93. Jayaprakash,A.D., Jabado,O., Brown,B.D. and Sachidanandam,R. (2011) Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic Acids Res.*, **39**, e141.
94. Hafner,M., Renwick,N., Brown,M., Mihailovic,A., Holoch,D., Lin,C., Pena,J.T., Nusbaum,J.D., Morozov,P., Ludwig,J. *et al.* (2011) RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA*, **17**, 1697–1712.
95. Sorefan,K., Pais,H., Hall,A.E., Kozomara,A., Griffiths-Jones,S., Moulton,V. and Dalmay,T. (2012) Reducing ligation bias of small RNAs in libraries for next generation sequencing. *Silence*, **3**, doi:10.1186/1758-907X-3-4.
96. Fuchs,R.T., Sun,Z., Zhuang,F. and Robb,G.B. (2015) Bias in ligation-based small RNA sequencing library construction is determined by adaptor and RNA structure. *PLoS One*, **10**, e0126049.
97. Baran-Gale,J., Kurtz,C.L., Erdos,M.R., Sison,C., Young,A., Fannin,E.E., Chines,P.S. and Sethupathy,P. (2015) Addressing bias in small RNA library preparation for sequencing: a new protocol recovers microRNAs that evade capture by current methods. *Front. Genet.*, **6**, doi:10.3389/fgene.2015.00352.
98. Van Nieuwerburgh,F., Soetaert,S., Podshivalova,K., Ay-Lin Wang,E., Schaffer,L., Deforce,D., Salomon,D.R., Head,S.R. and Ordoukhanian,P. (2011) Quantitative bias in Illumina TruSeq and a novel post amplification barcoding strategy for multiplexed DNA and small RNA deep sequencing. *PLoS One*, **6**, e26969.
99. Lopez-Barragan,M.J., Quinones,M., Cui,K., Lemieux,J., Zhao,K. and Su,X.Z. (2011) Effect of PCR extension temperature on high-throughput sequencing. *Mol. Biochem. Parasitol.*, **176**, 64–67.
100. Kirino,Y. and Mourelatos,Z. (2007) Mouse Piwi-interacting RNAs are 2′-O-methylated at their 3′ termini. *Nat. Struct. Mol. Biol.*, **14**, 347–348.
101. Ohara,T., Sakaguchi,Y., Suzuki,T., Ueda,H. and Miyauchi,K. (2007) The 3′ termini of mouse Piwi-interacting RNAs are 2′-O-methylated. *Nat. Struct. Mol. Biol.*, **14**, 349–350.
102. Viollet,S., Fuchs,R.T., Munafo,D.B., Zhuang,F. and Robb,G.B. (2011) T4 RNA ligase 2 truncated active site mutants: improved tools for RNA analysis. *BMC Biotechnol.*, **11**, doi:10.1186/1472-6750-11-72.
103. Munafo,D.B. and Robb,G.B. (2010) Optimization of enzymatic reaction conditions for generating representative pools of cDNA from small RNA. *RNA*, **16**, 2537–2552.
104. Svec,D., Tichopad,A., Novosadova,V., Pfaffl,M.W. and Kubista,M. (2015) How good is a PCR efficiency estimate: recommendations for precise and robust qPCR efficiency assessments. *Biomol. Detect. Quantif.*, **3**, 9–16.
105. Orpana,A.K., Ho,T.H. and Stenman,J. (2012) Multiple heat pulses during PCR extension enabling amplification of GC-rich sequences and reducing amplification bias. *Anal. Chem.*, **84**, 2081–2087.
106. Sendler,E., Johnson,G.D. and Krawetz,S.A. (2011) Local and global factors affecting RNA sequencing analysis. *Anal. Biochem.*, **419**, 317–322.
107. Dabney,J. and Meyer,M. (2012) Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques*, **52**, 87–94.
108. Mamedov,T.G., Pienaar,E., Whitney,S.E., TerMaat,J.R., Carvill,G., Goliath,R., Subramanian,A. and Viljoen,H.J. (2008) A fundamental study of the PCR amplification of GC-rich DNA templates. *Comput. Biol. Chem.*, **32**, 452–457.
109. Aird,D., Ross,M.G., Chen,W.S., Danielsson,M., Fennell,T., Russ,C., Jaffe,D.B., Nusbaum,C. and Gnirke,A. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, **12**, doi:10.1186/gb-2011-12-2-r18.
110. Head,S.R., Komori,H.K., LaMere,S.A., Whisenant,T., Van Nieuwerburgh,F., Salomon,D.R. and Ordoukhanian,P. (2014) Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*, **56**, 61–68.
111. Okino,S.T., Kong,M., Sarras,H. and Wang,Y. (2016) Evaluation of bias associated with high-multiplex, target-specific pre-amplification. *Biomol. Detect. Quantif.*, **6**, 13–21.
112. Kivioja,T., Vaharautio,A., Karlsson,K., Bonke,M., Enge,M., Linnarsson,S. and Taipale,J. (2012) Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods*, **9**, 72–74.
113. Islam,S., Zeisel,A., Joost,S., La Manno,G., Zajac,P., Kasper,M., Lonnerberg,P. and Linnarsson,S. (2014) Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods*, **11**, 163–166.
114. Locati,M.D., Terpstra,I., de Leeuw,W.C., Kuzak,M., Rauwerda,H., Ensink,W.A., van Leeuwen,S., Nehrdich,U., Spaink,H.P., Jonker,M.J. *et al.* (2015) Improving small RNA-seq by using a synthetic spike-in set for size-range quality control together with a set for data normalization. *Nucleic Acids Res.*, **43**, e89.

115. Grundemann,D. and Schomig,E. (1996) Protection of DNA during preparative agarose gel electrophoresis against damage induced by ultraviolet light. *Biotechniques*, **21**, 898–903.
116. Cariello,N.F., Keohavong,P., Sanderson,B.J. and Thilly,W.G. (1988) DNA damage produced by ethidium bromide staining and exposure to ultraviolet light. *Nucleic Acids Res.*, **16**, 4157–4157.
117. Alba,F.J., Bermudez,A. and Daban,J.R. (2001) Green-light transilluminator for the detection without photodamage of proteins and DNA labeled with different fluorescent dyes. *Electrophoresis*, **22**, 399–403.
118. Seville,M. (2001) A whole new way of looking at things: the use of Dark Reader technology to detect fluorophors. *Electrophoresis*, **22**, 814–828.
119. Quail,M.A., Kozarewa,I., Smith,F., Scally,A., Stephens,P.J., Durbin,R., Swerdlow,H. and Turner,D.J. (2008) A large genome center's improvements to the Illumina sequencing system. *Nat. Methods*, **5**, 1005–1010.
120. Buehler,B., Hogrefe,H.H., Scott,G., Ravi,H., Pabon-Pena,C., O'Brien,S., Formosa,R. and Happe,S. (2010) Rapid quantification of DNA libraries for next-generation sequencing. *Methods*, **50**, S15–S18.
121. Laurie,M.T., Bertout,J.A., Taylor,S.D., Burton,J.N., Shendure,J.A. and Bielas,J.H. (2013) Simultaneous digital quantification and fluorescence-based size characterization of massively parallel sequencing libraries. *Biotechniques*, **55**, 61–67.
122. Robin,J.D., Ludlow,A.T., LaRanger,R., Wright,W.E. and Shay,J.W. (2016) Comparison of DNA quantification methods for next generation sequencing. *Sci. Rep.*, **6**, doi:10.1038/srep24067.
123. White,R.A. 3rd, Blainey,P.C., Fan,H.C. and Quake,S.R. (2009) Digital PCR provides sensitive and absolute calibration for high throughput sequencing. *BMC Genomics*, **10**, doi:10.1186/1471-2164-10-116.
124. Fahlgren,N., Sullivan,C.M., Kasschau,K.D., Chapman,E.J., Cumbie,J.S., Montgomery,T.A., Gilbert,S.D., Dasenko,M., Backman,T.W., Givan,S.A. *et al.* (2009) Computational and analytical framework for small RNA profiling by high-throughput sequencing. *RNA*, **15**, 992–1002.
125. Jiang,L., Schlesinger,F., Davis,C.A., Zhang,Y., Li,R., Salit,M., Gingeras,T.R. and Oliver,B. (2011) Synthetic spike-in standards for RNA-seq experiments. *Genome Res.*, **21**, 1543–1551.
126. Robasky,K., Lewis,N.E. and Church,G.M. (2014) The role of replicates for error mitigation in next-generation sequencing. *Nat. Rev. Genet.*, **15**, 56–62.
127. Fisher,R.A. (1935) *The Design of Experiments*. Oliver and Boyde, London.
128. Leek,J.T., Scharpf,R.B., Bravo,H.C., Simcha,D., Langmead,B., Johnson,W.E., Geman,D., Baggerly,K. and Irizarry,R.A. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.*, **11**, 733–739.
129. t Hoen,P.A., Friedlander,M.R., Almlof,J., Sammeth,M., Pulyakhina,I., Anvar,S.Y., Laros,J.F., Buermans,H.P., Karlberg,O., Brannvall,M. *et al.* (2013) Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat. Biotechnol.*, **31**, 1015–1022.
130. Ross,M.G., Russ,C., Costello,M., Hollinger,A., Lennon,N.J., Hegarty,R., Nusbaum,C. and Jaffe,D.B. (2013) Characterizing and measuring bias in sequence data. *Genome Biol.*, **14**, doi:10.1186/gb-2013-14-5-r51.
131. Auer,P.L. and Doerge,R.W. (2010) Statistical design and analysis of RNA sequencing data. *Genetics*, **185**, 405–416.
132. McIntyre,L.M., Lopiano,K.K., Morse,A.M., Amin,V., Oberg,A.L., Young,L.J. and Nuzhdin,S.V. (2011) RNA-seq: technical variability and sampling. *BMC Genomics*, **12**, doi:10.1186/1471-2164-12-293.
133. Metpally,R.P., Nasser,S., Malenica,I., Courtright,A., Carlson,E., Ghaffari,L., Villa,S., Tembe,W. and Van Keuren-Jensen,K. (2013) Comparison of analysis tools for miRNA high throughput sequencing using nerve crush as a model. *Front. Genet.*, **4**, doi:10.3389/fgene.2013.00020.
134. Fonseca,N.A., Marioni,J. and Brazma,A. (2014) RNA-Seq gene profiling–a systematic empirical comparison. *PLoS One*, **9**, e107026.
135. Kircher,M., Sawyer,S. and Meyer,M. (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.*, **40**, e3.
136. Ledergerber,C. and Dessimoz,C. (2011) Base-calling for next-generation sequencing platforms. *Brief. Bioinform.*, **12**, 489–497.
137. Erlich,Y., Mitra,P.P., delaBastide,M., McCombie,W.R. and Hannon,G.J. (2008) Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nat. Methods*, **5**, 679–682.
138. Rougemont,J., Amzallag,A., Iseli,C., Farinelli,L., Xenarios,I. and Naef,F. (2008) Probabilistic base calling of Solexa sequencing data. *BMC Bioinformatics*, **9**, doi:10.1186/1471-2105-9-431.
139. Fuller,C.W., Middendorf,L.R., Benner,S.A., Church,G.M., Harris,T., Huang,X., Jovanovich,S.B., Nelson,J.R., Schloss,J.A., Schwartz,D.C. *et al.* (2009) The challenges of sequencing by synthesis. *Nat. Biotechnol.*, **27**, 1013–1023.
140. Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.
141. Dohm,J.C., Lottaz,C., Borodina,T. and Himmelbauer,H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, **36**, e105.
142. Nakamura,K., Oshima,T., Morimoto,T., Ikeda,S., Yoshikawa,H., Shiwa,Y., Ishikawa,S., Linak,M.C., Hirai,A., Takahashi,H. *et al.* (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.*, **39**, e90.
143. Stein,A., Takasuka,T.E. and Collings,C.K. (2010) Are nucleosome positions in vivo primarily determined by histone-DNA sequence preferences? *Nucleic Acids Res.*, **38**, 709–719.
144. Harismendy,O., Ng,P.C., Strausberg,R.L., Wang,X., Stockwell,T.B., Beeson,K.Y., Schork,N.J., Murray,S.S., Topol,E.J., Levy,S. *et al.* (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.*, **10**, doi:10.1186/gb-2009-10-3-r32.
145. Yang,X., Chockalingam,S.P. and Aluru,S. (2013) A survey of error-correction methods for next-generation sequencing. *Brief. Bioinform.*, **14**, 56–66.
146. Loman,N.J., Misra,R.V., Dallman,T.J., Constantinidou,C., Gharbia,S.E., Wain,J. and Pallen,M.J. (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.*, **30**, 434–439.
147. Schirmer,M., D'Amore,R., Ijaz,U.Z., Hall,N. and Quince,C. (2016) Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*, **17**, doi:10.1186/s12859-016-0976-y.
148. Beltman,J.B., Urbanus,J., Velds,A., van Rooij,N., Rohr,J.C., Naik,S.H. and Schumacher,T.N. (2016) Reproducibility of Illumina platform deep sequencing errors allows accurate determination of DNA barcodes in cells. *BMC Bioinformatics*, **17**, doi:10.1186/s12859-016-0999-4.
149. Huse,S.M., Huber,J.A., Morrison,H.G., Sogin,M.L. and Welch,D.M. (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.*, **8**, doi:10.1186/gb-2007-8-7-r143.
150. Luo,C., Tsementzi,D., Kyrpides,N., Read,T. and Konstantinidis,K.T. (2012) Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS One*, **7**, e30087.
151. Team, R.C. (2014) *R Foundation for Statistical Computing*. Vienna.
152. Blankenberg,D., Kuster,G.V., Coraor,N., Ananda,G., Lazarus,R., Mangan,M., Nekrutenko,A. and Taylor,J. (2010) Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol.*, doi:10.1002/0471142727.mb1910s89.
153. Giardine,B., Riemer,C., Hardison,R.C., Burhans,R., Elnitski,L., Shah,P., Zhang,Y., Blankenberg,D., Albert,I., Taylor,J. *et al.* (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, **15**, 1451–1455.
154. Goecks,J., Nekrutenko,A. and Taylor,J. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, doi:10.1101/gr.4086505.
155. Yuan,T., Huang,X., Dittmar,R.L., Du,M., Kohli,M., Boardman,L., Thibodeau,S.N. and Wang,L. (2014) eRNA: a graphic user interface-based tool optimized for large data analysis from high-throughput RNA sequencing. *BMC Genomics*, **15**, doi:10.1186/1471-2164-15-176.
156. Kim,V.N., Han,J. and Siomi,M.C. (2009) Biogenesis of small RNAs in animals. *Nat. Rev. Mol. Cell Biol.*, **10**, 126–139.

157. Yang,J., Ding,X., Sun,X., Tsang,S.-Y. and Xue,H. (2015) SAMSVM: A tool for misalignment filtration of SAM-format sequences with support vector machine. *J. Bioinform. Comput. Biol.*, **13**, doi:10.1142/s0219720015500250.
158. Hatem,A., Bozdağ,D., Toland,A.E. and Çatalyürek,Ü.V. (2013) Benchmarking short sequence mapping tools. *BMC Bioinformatics*, **14**, doi:10.1186/1471-2105-14-184.
159. Kong,Y. (2011) Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. *Genomics*, **98**, 152–153.
160. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10–12.
161. Lo,C.-C. and Chain,P.S.G. (2014) Rapid evaluation and quality control of next generation sequencing data with FaQCs. *BMC Bioinformatics*, **15**, doi:10.1186/s12859-014-0366-2.
162. Kelley,D.R., Schatz,M.C. and Salzberg,S.L. (2010) Quake: quality-aware detection and correction of sequencing errors. *Genome Biol.*, **11**, doi:10.1186/gb-2010-11-11-r116.
163. Gnerre,S., Maccallum,I., Przybylski,D., Ribeiro,F.J., Burton,J.N., Walker,B.J., Sharpe,T., Hall,G., Shea,T.P., Sykes,S. *et al.* (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 1513–1518.
164. Robertson,G., Schein,J., Chiu,R., Corbett,R., Field,M., Jackman,S.D., Mungall,K., Lee,S., Okada,H.M., Qian,J.Q. *et al.* (2010) De novo assembly and analysis of RNA-seq data. *Nat. Methods*, **7**, 909–912.
165. Del Fabbro,C., Scalabrin,S., Morgante,M. and Giorgi,F.M. (2013) An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS One*, **8**, e85024.
166. Cox,M.P., Peterson,D.A. and Biggs,P.J. (2010) SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*, **11**, doi:10.1186/1471-2105-11-485.
167. Chen,C., Khaleel,S.S., Huang,H. and Wu,C.H. (2014) Software for pre-processing Illumina next-generation sequencing short read sequences. *Source Code Biol. Med.*, **9**, doi:10.1186/1751-0473-9-8.
168. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
169. Burrows,M. and Wheeler,D.J. (1994) *Technical Report*. Systems Research Center, Paöp Alto.
170. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, doi:10.1186/gb-2009-10-3-r25.
171. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
172. Li,R., Yu,C., Li,Y., Lam,T.-W., Yiu,S.-M., Kristiansen,K. and Wang,J. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966–1967.
173. Li,H. and Homer,N. (2010) A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform.*, **11**, 473–483.
174. Bao,S., Jiang,R., Kwan,W., Wang,B., Ma,X. and Song,Y.-Q. (2011) Evaluation of next-generation sequencing software in mapping and assembly. *J. Hum. Genet.*, **56**, 406–414.
175. Klus,P., Lam,S., Lyberg,D., Cheung,M., Pullan,G., McFarlane,I., Yeo,G.S.H. and Lam,B.Y.H. (2012) BarraCUDA - a fast short read sequence aligner using graphics processing units. *BMC Res. Notes*, **5**, doi:10.1186/1756-0500-5-27.
176. Luo,R., Wong,T., Zhu,J., Liu,C.-M., Zhu,X., Wu,E., Lee,L.-K., Lin,H., Zhu,W., Cheung,D.W. *et al.* (2013) SOAP3-dp: fast, accurate and sensitive GPU-based short read aligner. *PLoS One*, **8**, e65632.
177. Luo,R., Cheung,J., Wu,E., Wang,H., Chan,S.-H., Law,W.-C., He,G., Yu,C., Liu,C.-M., Zhou,D. *et al.* (2015) MICA: a fast short-read aligner that takes full advantage of many integrated core architecture (MIC). *BMC Bioinformatics*, **16**, doi:10.1186/1471-2105-16-s7-s10.
178. Griffiths-Jones,S. (2004) The microRNA registry. *Nucleic Acids Res.*, **32**, D109–D111.
179. Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68–D73.
180. Nawrocki,E.P., Burge,S.W., Bateman,A., Daub,J., Eberhardt,R.Y., Eddy,S.R., Floden,E.W., Gardner,P.P., Jones,T.A., Tate,J. *et al.* (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, **43**, D130–D137.
181. Lipson,D., Speed,T.P. and Taub,M. (2010) Methods for allocating ambiguous short-reads. *Commun. Inform. Syst.*, **10**, 69–82.
182. Li,B., Ruotti,V., Stewart,R.M., Thomson,J.A. and Dewey,C.N. (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**, 493–500.
183. Ji,Y., Xu,Y., Zhang,Q., Tsui,K.-W., Yuan,Y., Norris,C., Liang,S. and Liang,H. (2011) BM-map: Bayesian mapping of multireads for next-generation sequencing data. *Biometrics*, **67**, 1215–1224.
184. Motameny,S., Wolters,S., Nürnberg,P. and Schumacher,B. (2010) Next generation sequencing of miRNAs—strategies, resources and methods. *Genes*, **1**, 70–84.
185. Guo,L. and Chen,F. (2014) A challenge for miRNA: multiple isomiRs in miRNAomics. *Gene*, **544**, 1–7.
186. Cloonan,N., Wani,S., Xu,Q., Gu,J., Lea,K., Heater,S., Barbacioru,C., Steptoe,A.L., Martin,H.C., Nourbakhsh,E. *et al.* (2011) MicroRNAs and their isomiRs function cooperatively to target common biological pathways. *Genome Biol.*, **12**, doi:10.1186/gb-2011-12-12-r126.
187. Neilsen,C.T., Goodall,G.J. and Bracken,C.P. (2012) IsomiRs–the overlooked repertoire in the dynamic microRNAome. *Trends Genet.*, **28**, 544–549.
188. Sai Lakshmi,S. and Agrawal,S. (2008) piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Res.*, **36**, D173–D177.
189. Anders,S., Pyl,P.T. and Huber,W. (2015) HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.
190. Lawrence,M., Huber,W., Pagès,H., Aboyoun,P., Carlson,M., Gentleman,R., Morgan,M.T. and Carey,V.J. (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**, e1003118.
191. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
192. Zhao,S., Fung-Leung,W.-P., Bittner,A., Ngo,K. and Liu,X. (2014) Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One*, **9**, e78644.
193. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
194. Risso,D., Schwartz,K., Sherlock,G. and Dudoit,S. (2011) GC-content normalization for RNA-Seq data. *BMC Bioinformatics*, **12**, doi:10.1186/1471-2105-12-480.
195. Oshlack,A. and Wakefield,M.J. (2009) Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct*, **4**, doi:10.1186/1745-6150-4-14.
196. Bullard,J.H., Purdom,E., Hansen,K.D. and Dudoit,S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, doi:10.1186/1471-2105-11-94.
197. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
198. Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, **11**, doi:10.1186/gb-2010-11-3-r25.
199. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, doi:10.1186/gb-2010-11-10-r106.
200. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, doi:10.1186/s13059-014-0550-8.
201. Leek,J.T. (2014) svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.*, **42**, e161.
202. Dillies,M.-A., Rau,A., Aubert,J., Hennequet-Antier,C., Jeanmougin,M., Servant,N., Keime,C., Marot,G., Castel,D., Estelle,J. *et al.* (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.*, **14**, 671–683.

203. Zyprych-Walczak,J., Szabelska,A., Handschuh,L., Górczak,K., Klamecka,K., Figlerowicz,M. and Siatkowski,I. (2015) The impact of normalization methods on RNA-Seq data analysis. *Biomed Res. Int.*, **2015**, doi:10.1155/2015/621690.
204. Oberg,A.L., Bot,B.M., Grill,D.E., Poland,G.A. and Therneau,T.M. (2012) Technical and biological variance structure in mRNA-Seq data: life in the real world. *BMC Genomics*, **13**, doi:10.1186/1471-2164-13-304.
205. Auer,P.L. and Doerge,R.W. (2011) A two-stage poisson model for testing RNA-Seq data. *Stat. Appl. Genet. Mol. Biol.*, **10**, doi:10.2202/1544-6115.1627.
206. Di,Y., Schafer,D.W., Cumbie,J.S. and Chang,J.H. (2011) The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Stat. Appl. Genet. Mol. Biol.*, **10**, doi:10.2202/1544-6115.1637.
207. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
208. Leng,N., Dawson,J.A., Thomson,J.A., Ruotti,V., Rissman,A.I., Smits,B.M.G., Haag,J.D., Gould,M.N., Stewart,R.M. and Kendziorski,C. (2013) EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, **29**, 1035–1043.
209. Hardcastle,T.J. and Kelly,K.A. (2010) baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**, doi:10.1186/1471-2105-11-422.
210. Li,J. and Tibshirani,R. (2013) Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat. Methods Med. Res.*, **22**, 519–536.
211. Tarazona,S., Furió-Tarí,P., Turrà,D., Di Pietro,A., Nueda,M.J., Ferrer,A. and Conesa,A. (2015) Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res.*, **43**, e140.
212. Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, doi:10.2202/1544-6115.1027.
213. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.
214. Law,C.W., Chen,Y., Shi,W. and Smyth,G.K. (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, **15**, doi:10.1186/gb-2014-15-2-r29.
215. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.
216. Soneson,C. and Delorenzi,M. (2013) A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, **14**, doi:10.1186/1471-2105-14-91.
217. Rapaport,F., Khanin,R., Liang,Y., Pirun,M., Krek,A., Zumbo,P., Mason,C.E., Socci,N.D. and Betel,D. (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.*, **14**, doi:10.1186/gb-2013-14-9-r95.
218. Seyednasrollah,F., Laiho,A. and Elo,L.L. (2015) Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief. Bioinformatics*, **16**, 59–70.
219. Moulos,P. and Hatzis,P. (2015) Systematic integration of RNA-Seq statistical algorithms for accurate detection of differential gene expression patterns. *Nucleic Acids Res.*, **43**, e25.
220. Marcello Manfredi,E.R. (2013) Biomarkers discovery through multivariate statistical methods: a review of recently developed methods and applications in proteomics. *J. Proteomics Bioinform.*, **s3**, doi:10.4172/jpb.S3-003.
221. Eriksson,L., Antti,H., Gottfries,J., Holmes,E., Johansson,E., Lindgren,F., Long,I., Lundstedt,T., Trygg,J. and Wold,S. (2004) Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm). *Anal. Bioanal. Chem.*, **380**, 419–429.
222. Herrero,J., Valencia,A. and Dopazo,J. (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, **17**, 126–136.
223. Tibshirani,R. (1988) Estimating transformations for regression via additivity and variance stabilization. *J. Am. Stat. Assoc.*, **83**, 394–405.
224. Stacklies,W., Redestig,H., Scholz,M., Walther,D. and Selbig,J. (2007) pcaMethods–a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, **23**, 1164–1167.
225. Le Cao,K.A., Gonzalez,I. and Dejean,S. (2009) integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics*, **25**, 2855–2856.
226. Gibbons,F.D. and Roth,F.P. (2002) Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res.*, **12**, 1574–1581.
227. Bergkvist,A., Rusnakova,V., Sindelka,R., Garda,J.M., Sjogreen,B., Lindh,D., Forootan,A. and Kubista,M. (2010) Gene expression profiling–clusters of possibilities. *Methods*, **50**, 323–335.
228. Beyene,J., Tritchler,D., Bull,S.B., Cartier,K.C., Jonasdottir,G., Kraja,A.T., Li,N., Nock,N.L., Parkhomenko,E., Rao,J.S. *et al.* (2007) Multivariate analysis of complex gene expression and clinical phenotypes with genetic marker data. *Genet. Epidemiol.*, **31**(Suppl. 1), S103–S109.
229. Costa,I.G., Carvalho,F.d.A.T.d. and Souto,M.C.P.d. (2004) Comparative analysis of clustering methods for gene expression time course data. *Genet. Mol. Biol.*, **27**, 623–631.
230. Kubista,M., Andrade,J.M., Bengtsson,M., Forootan,A., Jonak,J., Lind,K., Sindelka,R., Sjoback,R., Sjogreen,B., Strombom,L. *et al.* (2006) The real-time polymerase chain reaction. *Mol. Aspects Med.*, **27**, 95–125.
231. Lee,G., Rodriguez,C. and Madabhushi,A. (2008) Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **5**, 368–384.
232. Shaffer,R.E. (2002) Multi- and megavariate data analysis. Principles and applications, I. Eriksson, E. Johansson, N. Kettaneh-Wold and S. Wold, Umetrics Academy, Umeå, 2001, ISBN 91-973730-1-X, 533pp. *J. Chemom.*, **16**, 261–262.
233. Barker,M. and Rayens,W. (2003) Partial least squares for discrimination. *J. Chemom.*, **17**, 166–173.
234. Le Cao,K.A., Boitard,S. and Besse,P. (2011) Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, **12**, doi:10.1186/1471-2105-12-253.
235. Trygg,J. and Wold,S. (2002) Orthogonal projections to latent structures (O-PLS). *J. Chemom.*, **16**, 119–128.
236. Liu,Y. and Rayens,W. (2007) PLS and dimension reduction for classification. *Comput. Stat.*, **22**, 189–208.
237. Rajalahti,T., Arneberg,R., Kroksveen,A.C., Berle,M., Myhr,K.-M. and Kvalheim,O.M. (2009) Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles. *Anal. Chem.*, **81**, 2581–2590.
238. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
239. Friedman,R.C., Farh,K.K., Burge,C.B. and Bartel,D.P. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.
240. Betel,D., Wilson,M., Gabow,A., Marks,D.S. and Sander,C. (2008) The microRNA.org resource: targets and expression. *Nucleic Acids Res.*, **36**, D149–D153.
241. Enright,A.J., John,B., Gaul,U., Tuschl,T., Sander,C. and Marks,D.S. (2003) MicroRNA targets in Drosophila. *Genome Biol.*, **5**, doi:10.1186/gb-2003-5-1-r1.
242. Lu,Y., Zhou,Y., Qu,W., Deng,M. and Zhang,C. (2011) A Lasso regression model for the construction of microRNA-target regulatory networks. *Bioinformatics*, **27**, 2406–2413.
243. Meyer,S.U., Stoecker,K., Sass,S., Theis,F.J. and Pfaffl,M.W. (2014) Posttranscriptional regulatory networks: from expression profiling to integrative analysis of mRNA and microRNA data. *Methods Mol. Biol.*, **1160**, 165–188.
244. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
245. Meyer,S.U., Sass,S., Mueller,N.S., Krebs,S., Bauersachs,S., Kaiser,S., Blum,H., Thirion,C., Krause,S., Theis,F.J. *et al.* (2015) Integrative analysis of microRNA and mRNA data reveals an orchestrated function of microRNAs in skeletal myocyte

differentiation in response to TNF-alpha or IGF1. *PLoS One*, **10**, e0135284.
246. McCormick,K.P., Willmann,M.R. and Meyers,B.C. (2011) Experimental design, preprocessing, normalization and differential expression analysis of small RNA sequencing experiments. *Silence*, **2**, doi:10.1186/1758-907X-2-2.
247. Hruz,T., Laule,O., Szabo,G., Wessendorp,F., Bleuler,S., Oertle,L., Widmayer,P., Gruissem,W. and Zimmermann,P. (2008) Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. *Adv. Bioinformatics*, **2008**, doi:10.1155/2008/420747.
248. Zimmermann,P., Laule,O., Schmitz,J., Hruz,T., Bleuler,S. and Gruissem,W. (2008) Genevestigator transcriptome meta-analysis and biomarker search using rice and barley gene expression databases. *Mol. Plant*, **1**, 851–857.