MOLECULAR ECOLOGY

Molecular Ecology (2016) doi: 10.1111/mec.13526

INVITED REVIEWS AND SYNTHESES

The power and promise of RNA-seq in ecology and evolution

ERICA V. TODD,* MICHAEL A. BLACK† and NEIL J. GEMMELL*

*Department of Anatomy, University of Otago, PO Box 913, Dunedin 9054, New Zealand, †Department of Biochemistry, University of Otago, PO Box 56, Dunedin 9054, New Zealand

Abstract

Reference is regularly made to the power of new genomic sequencing approaches. Using powerful technology, however, is not the same as having the necessary power to address a research question with statistical robustness. In the rush to adopt new and improved genomic research methods, limitations of technology and experimental design may be initially neglected. Here, we review these issues with regard to RNA sequencing (RNA-seq). RNA-seq adds large-scale transcriptomics to the toolkit of ecological and evolutionary biologists, enabling differential gene expression (DE) studies in nonmodel species without the need for prior genomic resources. High biological variance is typical of field-based gene expression studies and means that larger sample sizes are often needed to achieve the same degree of statistical power as clinical studies based on data from cell lines or inbred animal models. Sequencing costs have plummeted, yet RNA-seq studies still underutilize biological replication. Finite research budgets force a trade-off between sequencing effort and replication in RNA-seq experimental design. However, clear guidelines for negotiating this trade-off, while taking into account study-specific factors affecting power, are currently lacking. Study designs that prioritize sequencing depth over replication fail to capitalize on the power of RNA-seq technology for DE inference. Significant recent research effort has gone into developing statistical frameworks and software tools for power analysis and sample size calculation in the context of RNA-seq DE analysis. We synthesize progress in this area and derive an accessible rule-of-thumb guide for designing powerful RNA-seq experiments relevant in eco-evolutionary and clinical settings alike.

Keywords: biological replication, differential expression, experimental design, power analysis, RNA sequencing

Received 2 October 2015; revision received 5 December 2015; accepted 27 December 2015

The rise of RNA-seq

RNA sequencing (RNA-seq) (Wang *et al.* 2009) has driven the rapid expansion of transcriptomics beyond clinical biology and into the fields of ecology and evolution (Ekblom & Galindo 2011; Alvarez *et al.* 2015) (Box 1). Quickly surpassing microarrays as the high-throughput method of choice to study differential expression (DE) in nonmodel species, RNA-seq promised unprecedented sensitivity for detecting expression differences among rare transcripts, splice variants and microRNAs (Ozsolak & Milos 2011). However, in the slipstream of significant technological advancement, careful experimental design is frequently overlooked and the power of new technologies is sometimes overstated. The true sensitivity of RNA-seq for detecting subtle expression differences has been questioned as its wide dynamic range also makes RNA-seq data potentially very noisy (McIntyre *et al.* 2011; Tarazona *et al.* 2011).

With whole-transcriptome analysis now possible in almost any organism, biologists have quickly adopted RNA-seq to address big questions in ecology and evolution. Where microarrays suffer strain- or species-specific probe biases, RNA-seq permits studying the evolutionary forces shaping gene expression at the

Correspondence: Erica V. Todd, Fax: +64 3 479 7254; E-mail: ericavtodd@gmail.com

Box 1. RNA-seq in ecology and evolution: the state of play

RNA-seq technology has rapidly become an important tool in ecological and evolutionary research (Box Fig. 1A), mirroring an exponential rise in the number of studies using RNA-seq in empirical research more broadly, albeit on a smaller scale. Searching the Web of Science (Thomson Reuters) database for the terms 'RNA-seq OR RNAseq' returns 5630 articles published 2008–2014, with 2475 published in 2014. Refining this search to include only those studies from the Web of Science research areas 'Evolutionary Biology' and 'Environmental Sciences Ecology', and including five early studies from 2008/2009 published prior to widespread use of the term 'RNA-seq', returns 430 articles, with 190 published in 2014 (Box Fig. 1A). In a recent decadal review of the development of transcriptomics in ecological and evolutionary research (2004–2013), Alvarez *et al.* (2015) report 45% of transcriptomic studies in these fields used RNA-seq over microarrays (256 of 575 studies).

We reviewed biological replicate usage by eco-evolutionary studies using RNA-seq data for DE analysis, building on the review by Alvarez *et al.* (2015). We considered 256 RNA-seq studies from their review, published 2008–2013, and used the same search criteria to include 190 additional articles published in 2014. We searched the Web of Science database for the wildcard 'transcriptom*' and filtered articles within the research areas 'Evolutionary Biology' and 'Environmental Sciences Ecology', before retaining records sharing the term 'RNA-seq OR RNAseq'. References referring primarily to toxicology or agriculture without an obvious ecological context were excluded. We defined true biological replication as requiring independent library preparations. Pooling multiple independent biological samples into a single RNA-seq library for sequencing is a common cost-saving strategy, but provides only an average of the expression states across samples such that a single pooled library cannot be considered true biological replication. Therefore, pooled libraries are counted as one biological replicate in Box Fig. 1B. However, as a single pooled library captures greater biological variation than a single sample, pooled designs are also considered separately. Sample pooling for RNA-seq DE analysis is discussed as a separate experimental design issue in Box 2. An aliquot from a culture of microorganisms was not considered a pooled sample, but a single true biological replicate. Such samples may comprise many (genetically identical/similar) individual organisms, but provide the most relevant approximation of the within-group variance when comparing differences between strains (not individual cells). This can be considered analogous to comparing expression states among samples of a given tissue type (comprising many individual cells) taken from a large multicellular organism.

Biological replicate usage was low, regardless of whether pooled libraries were counted as one biological replicate. Of 158 eco-evolutionary studies reporting statistical DE analysis from RNA-seq data (Appendix S1, Supporting information), 89 (56%) sequenced a single library per treatment. Therefore, true biological replication (counting pooled libraries as one replicate) was absent in the majority of cases (Box Fig. 1B). Most single-replicate studies did use libraries constructed from pooled biological samples, and so did include some level of biological replication (Box Fig. 1C). However, 20 single-replicate studies appeared to sequence a single biological sample per treatment, meaning that 13% of all studies surveyed lack any form of biological replication. Only 23 studies (15%) report using more than three replicate libraries per treatment, and there was only a weak trend suggesting greater replicate usage in more recent articles: average number of replicates used was lowest in 2011 (1.2, 15 studies) and highest in 2014 (2.4, 67 studies).

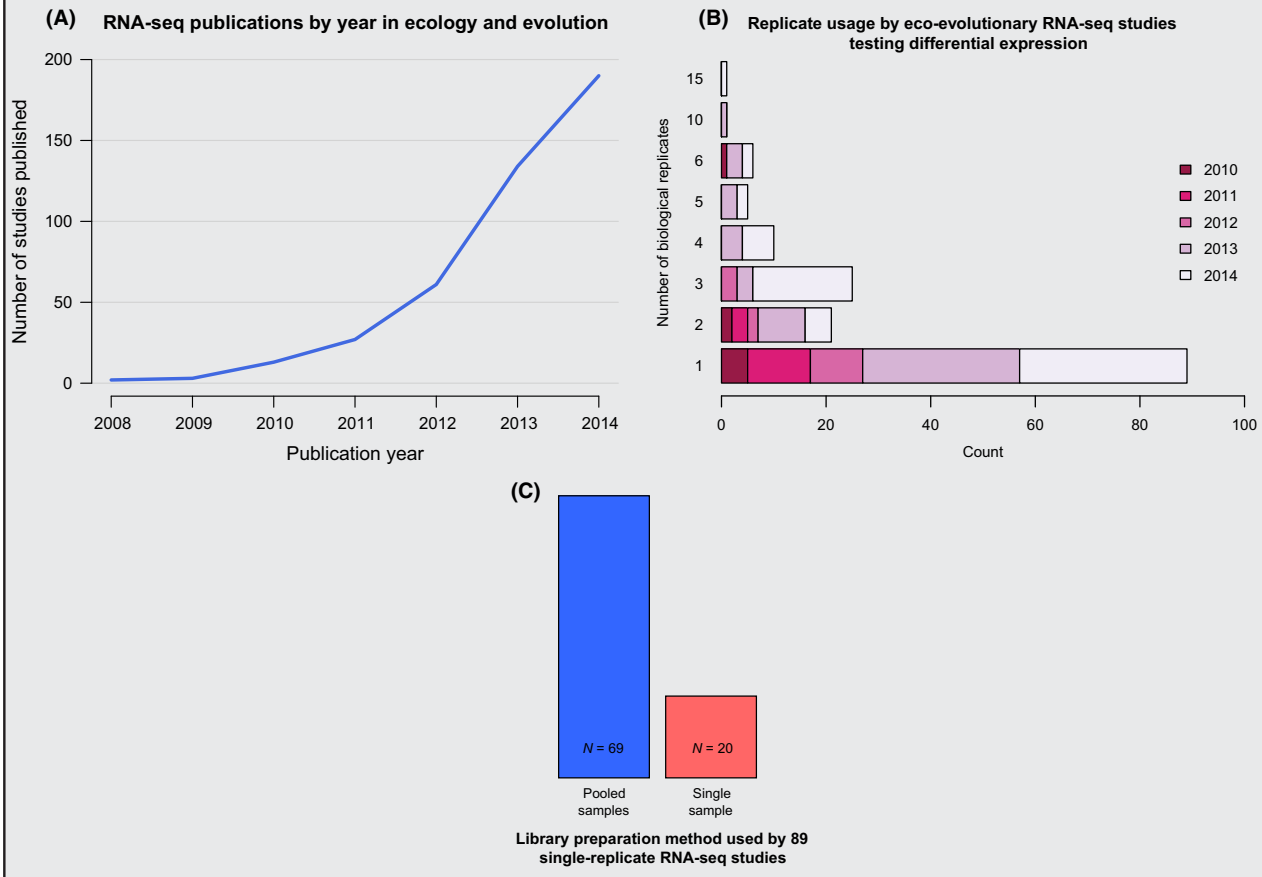
Pooled study designs were common overall, with 100 studies (63%) sequencing libraries representing pools of biological samples. However, the vast majority of these sequenced a single replicate pool per treatment (69 articles). Only 31 studies employing a pooled design included replicate libraries per treatment.

Much of the current RNA-seq literature appears under-replicated. While DE analysis was often used primarily for hypothesis discovery (e.g. to identify potential candidate genes for further detailed study), many studies also derived broader biological conclusions from data with little or no true biological replication. Interpreting the results from RNA-seq studies based on limited replication and/or subtle fold changes requires caution. With limited biological replication, not only will power and precision to detect true differences be low (see main text), but significant results may reflect biological (or uncontrolled technical) variation and may not be reproducible or biologically relevant when generalized to the study populations (Hansen *et al.* 2011).

It was often difficult to discern key details of the experimental design for the studies we reviewed. This includes the number of true biological replicates per treatment, whether or not library preparation involved pooled samples, and if so, how many samples were pooled per library. This raises general concerns regarding adequate reporting of methods and the reproducibility of genomics research.

Box 1. Continued

Although we did not review sequencing effort, the number of mapped reads used per sample will impact power of DE tests in the studies we reviewed. Comparing sequencing efforts across studies is difficult due to substantial differences (and continuous improvements) in sequencing efficiency and read length of the different technologies, and the many ways that sequencing effort is reported (e.g. raw reads, trimmed reads, mapped reads, either in total or by sample).



Box Fig. 1 (A) Number of studies published each year using RNA-seq data in ecological and evolutionary research. (B) Biological replicate usage by published eco-evolutionary studies using RNA-seq data for differential expression analysis. True biological replicates were considered to require independent library preparations, where pooling of multiple biological samples into a single library for sequencing was counted as one biological replicate. Where the number of biological replicates used differed across conditions, the smaller number is represented. (C) Proportion of 89 single-replicate studies using libraries representing single or pools of biological samples.

whole-transcriptome level (Busby *et al.* 2011; Romero *et al.* 2012). In ecological contexts, RNA-seq enables the examination of expression differences underlying interindividual or interpopulation variation in ecologically important traits such as disease resistance (Bonneaud *et al.* 2011) and mating behaviour (Fraser *et al.* 2014; Schunter *et al.* 2014), and the identification of genes of potential adaptive significance in changing environments (Meyer *et al.* 2011; Smith *et al.* 2013; Veilleux *et al.* 2015). RNA-seq is a key technology facilitating the recent push towards using integrative biology to understand molecular mechanisms of phenotypic and behavioural plasticity in wild populations (Aubin-Horth & Renn 2009; Harris & Hofmann 2014). For example, integrating RNA-seq DE analysis with quantitative PCR (qPCR) and information on transcription

factor binding sites permitted identification of gene regulatory networks underlying development of alternative jaw phenotypes in a cichlid fish (Gunter *et al.* 2013; Schneider *et al.* 2014).

Measuring gene expression in wild populations and natural settings still presents big challenges. Not least is that gene expression measurements made for nonmodel species in natural settings are subject to high biological and technical variance (Box 3). Regardless of the technology applied, transcription is by nature an inherently stochastic process and biological variability must be taken into account during experimental design. The power of any test of statistical significance is the probability of correctly rejecting the null hypothesis, which in the context of RNA-seq, is the likelihood of correctly identifying a gene or transcript differentially expressed between conditions. When it comes to establishing the statistical significance of DE tests, biological replication is as necessary for RNA-seq as it ever was for microarrays (Hansen *et al.* 2011). Despite more restrictive budgets, ecological and evolutionary studies will often require larger sample sizes to achieve the same power as their clinical counterparts for data sets representing cell lines or inbred strains (Fig. 1). Given finite financial resources, the acceptable minimum number of biological replicates in a given experimental situation is a key experimental design question.

When designing RNA-seq DE experiments, research budgets force a trade-off between increasing the sequencing depth and increasing the sample size. How to distribute sequencing effort between sequencing depth and sample size is a major decision affording considerable flexibility to RNA-seq experimental design, but one that lacks clear guidelines in the literature. Sequencing depths in published RNA-seq studies vary over several orders of magnitude compared to other NGS approaches that have more standardized read depths, such as whole-exome and whole-genome sequencing (Sims *et al.* 2014). Despite precipitous drops in the cost of sequencing, the cost of individual library preparation makes biological replicates expensive (Fig. 2). For complex multifactorial designs (e.g. time-series experiments), increasing the sample size quickly escalates total project costs. We find that biological replicate usage by eco-evolutionary studies using RNA-seq data for DE analysis is low (Box 1). Sequencing fewer replicates more deeply is the more popular (cheaper) strategy, but is unlikely to provide optimal power for DE tests.

Here, we synthesize recent research effort addressing the depth–replication trade-off in RNA-seq experimental design and derive a rule-of-thumb guide for optimizing statistical power of RNA-seq DE experiments. While we focus on aspects relevant to studying gene

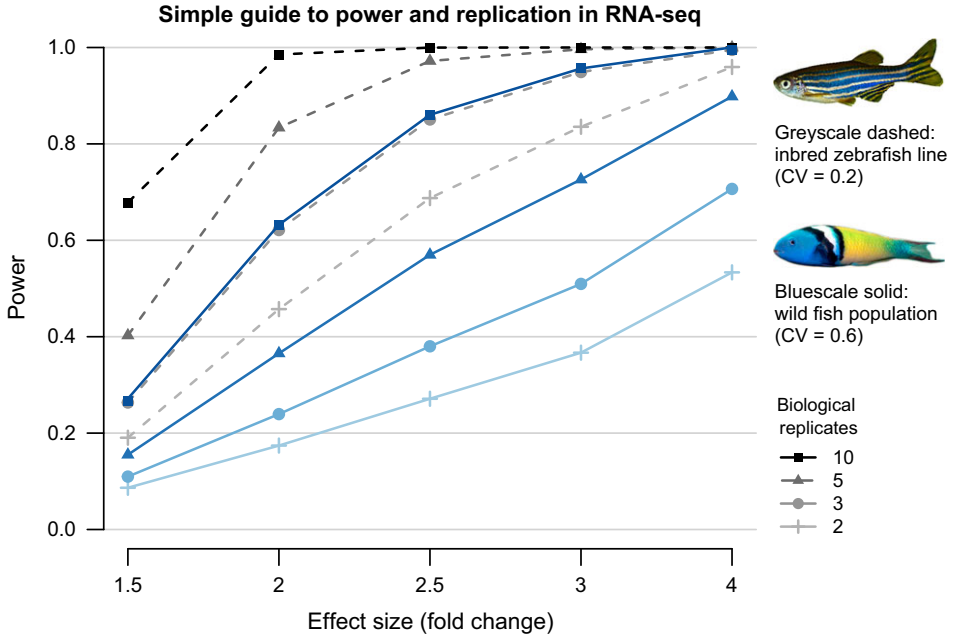


Fig. 1 Power in RNA-seq differential expression analysis depends on replication, biological variance and effect size. Expected statistical power is plotted for detecting different effect sizes of expression difference (as fold change) given different sample sizes in hypothetical cases of low biological variance (e.g. an inbred zebrafish line, CV = 0.2) and high biological variance (e.g. a wild reef fish population, CV = 0.6). Calculations were performed in the RNASEQPOWER package (Hart *et al.* 2013) in R, assuming 10 reads average sequencing depth and a 5% false positive rate. CV: coefficient of variation. A fold change of 2 is equivalent to a $\log_2$ fold change of 1.

Box 2. Pooling biological samples in RNA-seq experiments – when is it a good idea for DE analysis?

Sample pooling is widespread in transcriptomic analysis. Here, several independent biological samples (e.g. from separate individuals or cultures) are combined into single libraries for sequencing to reduce costs while still attempting to represent biological variation within the data. The impact of this practice on microarray studies is well researched, but is not yet well understood for RNA-seq.

For microarrays, the estimation of gene expression levels does not appear to be overly affected by pooling, and pooling is often recommended when 'fewer than three arrays are used in each condition' (Kendziorski *et al.* 2005). However, probe-based microarray technology does not force low-abundance genes to compete with more highly expressed genes for detection. In the case of RNA-seq, pooling may make the detection of low-abundance reads more difficult. Limited research into the utility of sample pooling for RNA-seq DE analysis suggests that while differentially expressed genes can be identified from pooled designs, sample pooling results in lower precision and higher false positive rates relative to analyses performed on samples sequenced separately (Biswas *et al.* 2013; Rajkumar *et al.* 2015). Rajkumar *et al.* (2015) found poor agreement between DE results obtained for sample pools compared to corresponding samples sequenced separately, but this experiment was compromised by the use of technical rather than biological replicates for each pool. The utility of using pooled data for RNA-seq DE analysis warrants further research. Our survey of the literature suggests this issue needs urgent attention, with 100 of 158 studies using pooled designs (Box 1).

For now, where only a few RNA-seq libraries are affordable, sample pooling offers the clear advantage of incorporating information from more individuals into the analysis. From a statistical perspective, each pool is simply treated as a single sample. When working with samples that may exhibit a large amount of biological variation, analysing several sample pools, rather than a few single samples per condition, will also lessen the impact of single aberrant samples (Kendziorski *et al.* 2005). Including larger numbers of samples per pool should help to prevent bias caused by a nonrepresentative sample. In analysing pooled data, a more stringent false discovery correction may be necessary to counter the potentially lower precision of pooled designs.

It is important to separate the issue of pooling from that of replication – creating multiple pools of biologically distinct samples per experimental condition still allows for the estimation of biological variability, whereas an absence of replication (regardless of whether pools or individual samples are being used) does not. If no replicates are available, the count-based nature of RNA-seq data means that the negative-binomial distribution can still be used to model biological variation, using the relationship between the mean and variance to estimate the likelihood that a gene is differentially expressed (Anders & Huber 2010; Robinson *et al.* 2010). In this setting, although results are likely to 'lead to conclusions of limited reliability' (DESEQ BIOCONDUCTOR vignette, http://bioconductor.org/packages/release/bioc/vignettes/DESeq/inst/doc/DESeq.pdf), analysing nonreplicated pooled samples is still more appropriate than comparing a single sample from each condition. However, without true biological replication (i.e. multiple libraries representing biologically distinct samples from each condition), there is still no way to accurately determine the amount of variation inherent within an experimental group, and thus reliably identify changes in gene expression between conditions. Therefore, sufficient biological replicates sequenced independently provide the most statistically robust data for variance estimation and DE inference, and should be prioritized in RNA-seq experimental designs. However, when budgets restrict the number of affordable library preparations much more than the number of individual samples that can be collected, pooled designs may be beneficial.

expression in wild populations, our discussion is relevant to anyone using RNA-seq data for DE inference.

Complexities of RNA-seq power analysis

The unique count-based nature of RNA-seq data (Fig. 2; Box 3) introduces complexities into DE analysis and power assessment not encountered for microarrays, which measure expression by fluorescence intensity (a continuous measurement). The tens of thousands of transcripts measured in a typical RNA-seq data set will be represented by a highly skewed distribution of read counts, such that detection probability and power to infer DE will vary considerably across transcripts. Transcript detection and the relative importance of technical sampling error depends on read coverage, which itself depends on expression level, transcript length and sequencing effort (McIntyre *et al.* 2011). This results in an inherent power bias in RNA-seq towards longer transcripts and transcripts with higher expression (Anders & Huber 2010; Tarazona *et al.* 2011). Therefore, DE tests for lowly expressed and short transcripts suffer

Box 3. Sources of noise in RNA-seq experiments

Statistical power to detect meaningful expression differences reflects our ability to distinguish true differential expression (i.e. due to treatment effect) from background noise. Three sources of noise contribute different degrees of uncertainty to gene expression measurements derived from RNA-seq data: (i) Poisson counting error, (ii) non-Poisson technical variance and (iii) biological variance (Busby *et al.* 2011, 2013 Supplementary data). How effectively these sources of noise are controlled and accounted for during RNA-seq experimental design and data analysis will significantly impact the accuracy of DE calls, the reliability of conclusions and the reproducibility of results (Busby *et al.* 2011; McIntyre *et al.* 2011; Soneson & Delorenzi 2013). Avoiding additional sources of variance through careful design will improve the power of the final experiment.

Poisson counting error is the uncertainty inherent in any count-based measurement. Poisson noise is disproportionately large for low count data and dominates the variance for counts below 10 (e.g. see Fig. S7 of Busby *et al.* 2013). This is because the variance of the Poisson distribution is equal to its mean, which increases the impact of the error for low counts. For RNA-seq data, where the number of reads mapping to transcripts serves as a proxy for relative expression level, there will be high uncertainty that low counts accurately reflect the true expression level. As a simple example: we can be less certain about a twofold expression difference represented by 1 vs. 2 counts, than the same difference represented by 100 vs. 200 counts. An effective minimum sequencing depth will be one that minimizes bias caused by many genes being measured with low read counts and high Poisson noise (see main text).

Non-Poisson technical variance is the imprecision observed between repeat measurements of the same sample (e.g. duplicate samples prepared and sequenced as separate RNA-seq libraries, or aliquots of the same library preparation sequenced on separate lanes). RNA-seq technology has been shown to be highly replicable (Marioni *et al.* 2008) and technical replicates are no longer considered necessary for standard RNA-seq experiments. However, technical variance can arise from multiple sources in RNA-seq experiments, and is not so low that it can be ignored (McIntyre *et al.* 2011).

Shotgun sequencing captures only a small fraction of the RNA molecules in a typical sample (e.g. <0.01% for 30 M reads, McIntyre *et al.* 2011) and random sampling noise, like Poisson noise, is a source of measurement error common to all RNA-seq data. Random noise due to the very small sampling fraction in RNA-seq data causes variability in expression estimates at all levels of coverage, but has been shown to be especially dramatic below average 5 mapped reads per nucleotide, where exon detection is highly inconsistent (McIntyre *et al.* 2011).

Sample collection, storage and processing are all sources of potentially confounding technical variance. RNA degrades quickly and RNA-seq expression profiling is extremely sensitive to inconsistencies in RNA quality (e.g. through reduced library complexity) (Romero *et al.* 2014). This presents a particular challenge for archived tissues, and for field-based studies where tissue can neither be processed immediately or stored under ideal conditions for RNA stability (i.e. cryopreservation). RNA-stabilizing reagents such as RNAlater help circumvent the problem for field-collected tissues, but are not always ideal (Camacho-Sanchez *et al.* 2013). The RNA integrity number (RIN) is a useful standardized metric of RNA quality (Schroeder *et al.* 2006; Romero *et al.* 2014), but is not valid for all samples (e.g. those with nontypical RNA profiles, Diaz de Cerio *et al.* 2012), and there is no threshold for deciding a sample is too degraded for whole-transcriptome analysis. Specific library preparation techniques and analytical corrections are available to help optimize the biological signal from degraded but valuable samples whose exclusion would otherwise compromise overall power (Adiconis *et al.* 2013; Romero *et al.* 2014; Cieslik *et al.* 2015).

Other well-known sources of technical variance in RNA-seq data are library preparation (e.g. sample handling and PCR biases; Bullard *et al.* 2010), and flow cell and lane effects during Illumina sequencing (Bullard *et al.* 2010; McIntyre *et al.* 2011). Sample barcoding facilitates multiplexed sequencing of many libraries over one or multiple lanes of the sequencer and is an effective strategy to reduce cost and avoid confounding lane effects in RNA-seq experiments.

Biological variance is the natural variation in gene expression measurements observed among samples of the same condition due to environmental or genetic differences. Biological variance will represent the greatest source of within-group variance for most RNA-seq data sets; it is usually low for data sets from cell lines and inbred animal strains (biological coefficient of variation, CV, typically ≤0.2), but can be appreciably larger for data sets representing unrelated individuals (CV >0.3). Therefore, as eco-evolutionary studies are typically based on wild populations and nonmodel species, high biological variance can be expected to add considerable uncertainty to gene expression measurements regardless of the technology applied (qPCR, microarray or RNA-seq).

Box 3. Continued

It is important to note that with larger sample sizes, technical and biological variation can become harder to control. The experimental design should include measures for constraining potential sources of variance as much as possible. Overall, the fundamental principles of randomization, replication and blocking (Fisher 1935) still apply and should be followed during RNA-seq study design (Auer & Doerge 2010).

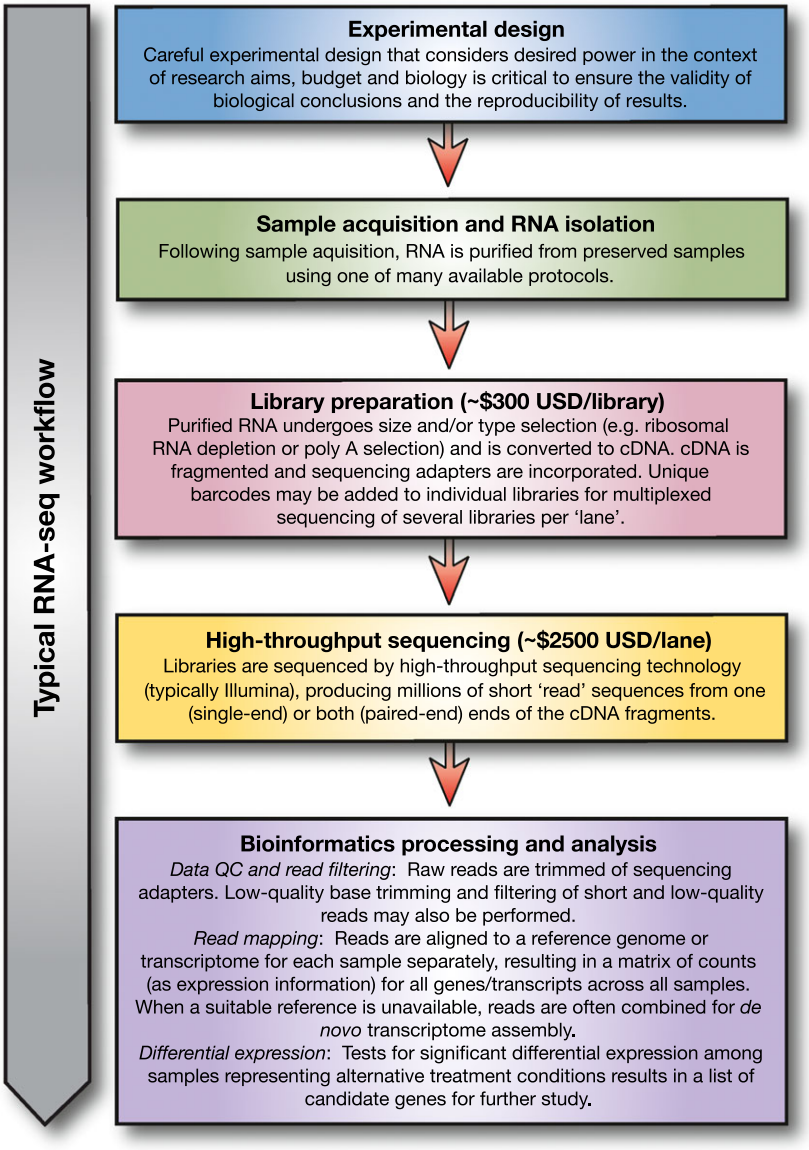


Fig. 2 A typical experimental workflow for differential expression inference from RNA-seq data, including approximate costs of library preparation and sequencing.

low power due to lower detection probability and higher measurement uncertainty (McIntyre *et al.* 2011; Robinson *et al.* 2015). For example, noncoding RNAs typically have lower expression levels relative to coding sequences, and a given RNA-seq data set will have lower power to identify significant DE in the former (Busby *et al.* 2011; Tarazona *et al.* 2011; Ching *et al.* 2014).

The highly dynamic nature of gene expression means that transcriptional landscapes captured by RNA-seq data will vary widely depending on the sampling context. Because power is affected by baseline expression level, expression landscape (i.e. the magnitude and proportion of expression differences among samples) will influence statistical power in a highly study-specific manner (Box 4). For example, if a few highly expressed

Box 4. Pilot sequencing – an example in a wild population: the bluehead wrasse

In the experimental design phase of an RNA-seq study, it is important to know what power is realistically achievable for detecting different degrees of DE under alternative experimental configurations (of sample size and sequencing depth). Because the power-sample size equation in RNA-seq is influenced by study-specific factors that cannot be reliably known beforehand (such as the amount of biological variance present within groups, the landscape of expression differences among groups, and transcriptome size and complexity), pilot sequencing and preliminary power analysis can be invaluable for optimizing experimental designs that maximize power while limiting cost.

The following example uses pilot RNA-seq data and newly available software to evaluate power to detect sex-biased gene expression in the bluehead wrasse (*Thalassoma bifasciatum*). Bluehead wrasse are common on tropical reefs of the Caribbean and, as protogynous hermaphrodites, undergo female-to-male sex change as adults in response to social cues (Godwin 2009). Here, pilot data represents gonad and forebrain transcriptomes for three female and three sex-reversed male fish, sequenced separately to a depth of 6.7 to 8.0 million mapped reads per replicate. These data form part of a larger experiment investigating gene expression changes underlying protogynous sex change in this species. Power analyses were performed using the R/BIOCONDUCTOR package PROPER (Wu *et al.* 2015). PROPER simulates count data from a negative-binomial model, using biological variance and baseline expression estimates from pilot data (estimated via the DSS package, Wu *et al.* 2013) (see Appendix S2, Supporting information for R code and simulation parameters). DE analysis was then performed on the simulated data using inbuilt DE software (e.g. DESEQ) and a range of power metrics calculated.

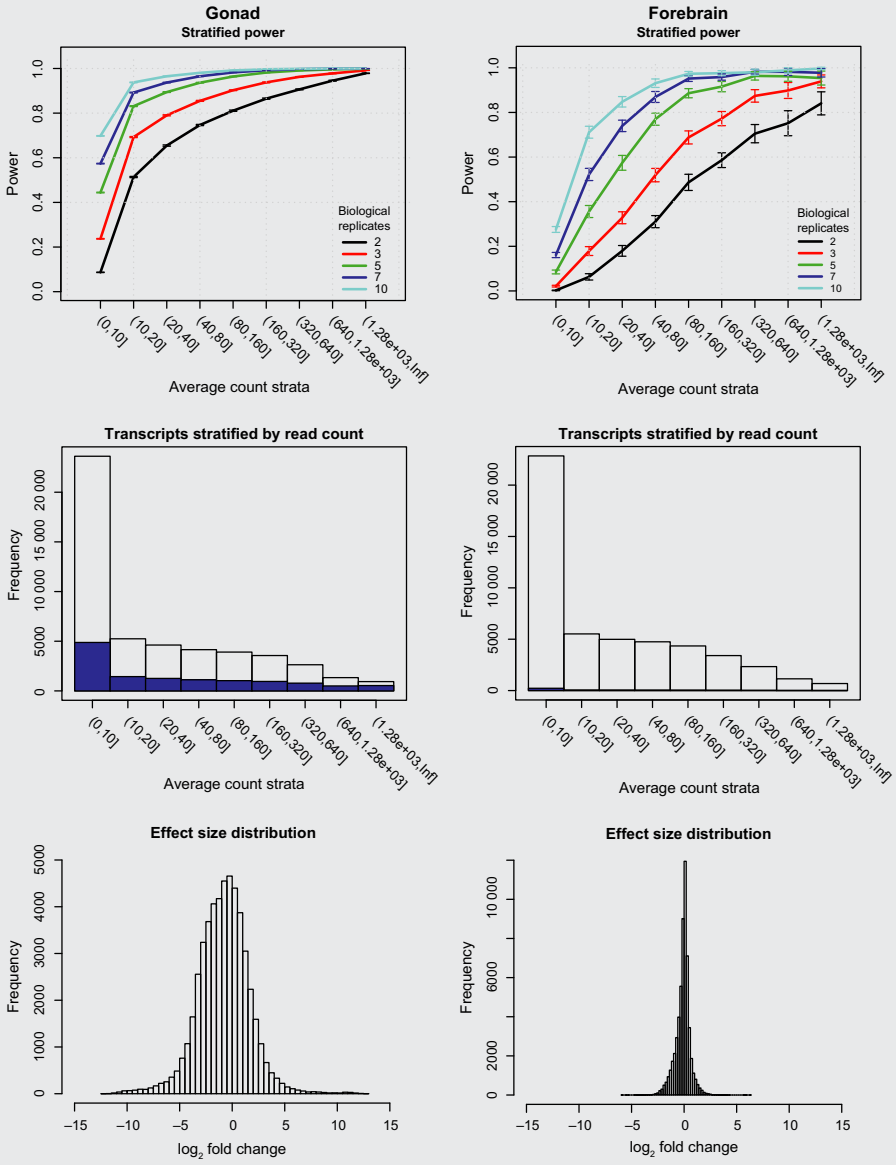
Box Fig. 2 (top) contrasts power to detect sex-biased gene expression in the gonad (left) and forebrain (right) (assuming a 1.5-fold change in expression, equivalent to a difference of 0.58 on the $\log_2$ scale). Stratifying transcripts by average read count reveals how relative expression level impacts power for different groups of transcripts in the data set. For gonad, five replicates per condition are predicted to achieve reasonable power (>80%) to detect differentially expressed transcripts with average counts above 10. In the forebrain analysis, five replicates would achieve sufficient power only for transcripts with average counts above 80 mapped reads. For most RNA-seq data sets, the majority of transcripts fall in the lowest count strata (counts <10).

The observed difference in power between the two data sets is due largely to their vastly different expression landscapes (Box Fig. 2, bottom). Both data sets have high biological variance (mean CV 0.5 and 0.3 for gonad and forebrain, respectively) and similar baseline expression (mean $\log_2$ of counts 3.9 and 4.2 for gonad and forebrain, respectively). However, a much broader distribution of effect sizes describes the much more varied expression landscape between male and female gonad, compared to the shallow expression landscape of the brain, where very few genes show sex-specific expression differences. In bluehead wrasse gonad, tens of thousands of transcripts show significant sex-biased expression, while fewer than 10 transcripts are significantly differentially expressed between male and female forebrain (Liu *et al.* 2015). Where expression landscapes are shallow and most meaningful expression differences are subtle, DE tests will have lower power and study designs will require more biological replicates. Taking a stratified view of power can be useful in defining fold change and expression thresholds above which DE calls have acceptable confidence.

transcripts contribute the majority of reads, remaining transcripts may be left with low coverage and low power. Similarly, where expression landscapes are shallow and most meaningful expression differences are subtle, low signal-to-noise ratio will limit the resolution of DE tests (Box 4). Low signal-to-noise ratio may also limit power in cross-species comparisons of gene expression (Busby *et al.* 2011; Romero *et al.* 2012).

The complex nature of RNA-seq data, including transcript- and sample-specific factors influencing power, plus the need to control for multiple hypothesis testing [i.e. false discovery rate (FDR)], makes power assessment in RNA-seq DE studies especially challenging. The power equation in RNA-seq includes factors both under the control of the investigator (sample size, sequencing depth, choice of Type 1 error rate), as well as those determined largely by the data itself (expression landscape, degree of biological and technical variation). A further consideration is the minimum effect size (i.e. fold change) of expression difference that is deemed to be biologically interesting. There is no consensus on what defines a biologically meaningful expression difference, and the chosen cut-off will depend on both the transcriptome being assayed and the study question. It is only in the past 1–2 years that statistical frameworks have been formalized into soft-

Box 4. Continued



Box Fig. 2 Power and transcription landscape differ between gonad (left) and forebrain (right) pilot RNA-seq experiments examining sex-biased expression in bluehead wrasse. Top: Power to detect DE for transcripts stratified by average count, with different sample sizes (number of biological replicates per condition, i.e. sex). DE is defined as a 1.5-fold change difference in expression. Results are averaged over 100 simulations, with proportion of DE simulated at 0.25 for gonad and 0.01 for forebrain, based on observations from real data.
Middle: Histogram of transcripts stratified by average count, based on the simulated data. Open bars are for the total number of transcripts, with blue bars representing the number of transcripts differentially expressed.
Bottom: Distribution of effect sizes ($\log_2$ fold change) for sex-biased gene expression describes the deep vs. shallow transcription landscapes of the gonad (left) and forebrain (right).

ware packages for evaluating the complex relationship among these variables on a study-specific basis (Table 1). The considerable recent research effort addressing the interaction between read depth, sample size and statistical power in RNA-seq DE analysis represents an important step forward in the field (Busby *et al.* 2013; Hart *et al.* 2013; Li *et al.* 2013; Ching *et al.* 2014; Liu *et al.* 2014; Robinson & Storey 2014).

Table 1 A summary of available software tools for designing and evaluating statistical power of RNA-seq experiments examining differential expression. Power calculators range from simple formulae examining the interaction between fixed experimental design variables and power (i.e. sample size, read depth, biological variation, fold change of expression difference, Type 1 error rate), to more detailed programs written in R that use pilot data to comprehensively assess power in a study-specific manner.

Name (Reference)	Brief description	Input variables	Output	Pilot data?	Multifactor designs?	Implementation and availability
RNASEQPOWER (Hart *et al.* 2013)	Simple sample size formula based on the negative-binomial distribution. Key experimental design variables can be fixed or estimated to evaluate their interactive effects on power.	Sample size, average sequencing depth, biological variance, effect size, desired power, type 1 error rate.	Solves for any variable given fixed estimates of remaining variables.	No	No	Excel worksheet with basic formula provided in Supplementary Data. R package also via BIOCONDUCTOR: http://bioconductor.org/packages/release/bioc/html/RNA-seqPower.html
Unnamed metric (Liu *et al.* 2014)	Simple metric for estimating cost per 1% power for a given experimental design. Considers the trade-off between sequencing depth, sample size and cost.	Sample size, cost per sample (sample collection and library preparation), sequencing costs (per lane), desired power.	Cost per 1% power	No	No	Cost per 1% power = (fixed costs per sample * no. samples + sequencing costs)/power
Unnamed formula (Li *et al.* 2013)	Sample size formula based on the exact test under the negative-binomial distribution, and including FDR control.	Minimum fold change, average read count, maximum biological variance.	Sample size estimate	Yes	No	R code available from corresponding author.
SCOTTY (Busby *et al.* 2013)	Web-based tool for calculating power, cost and measurement bias for a matrix of experimental configurations (read depths and sample sizes) within user-defined experimental and budgetary constraints. Measures power as a percentage of true DE genes detected at a given fold change, based on a t-test.	Pilot data in the form of a read count matrix. Maximum sample size, sequencing depth range, minimum % of genes detectable, desired minimum fold change, maximum allowable measurement bias, type 1 error rate, cost per replicate, cost per million reads sequenced, total budget.	Matrices of experimental configurations permitted within user-defined budgetary, power and bias constraints. Also, graphical depiction of dispersion within treatment groups, genes detected as a function of sequencing depth, and predicted statistical power at different expression fold changes.	Yes	No	Web-based interface: http://bioinformatics.bc.edu/marthlab/scotty/scotty.php Matlab code available from GitHub.

Table 1 *Continued*

Name (Reference)	Brief description	Input variables	Output	Pilot data?	Multifactor designs?	Implementation and availability
SUBSEQ (Robinson & Storey 2014)	Determines optimal sequencing depth by subsampling reads from a pilot sequencing experiment to establish where saturation of power and accuracy occurs. Uses binomial sampling from count data and built-in DE software to analyse each subsample.	Pilot data in the form of a read count matrix and subsampling proportions.	Results include several graphs depicting the number of genes detected DE, estimated false discovery proportion, Spearman correlation and mean squared error as a function of subsampling depth.	Yes	No	R package available via http://github.com/StoreyLab/subSeq/.
RNASEQPOWERCALCULATOR (Ching *et al.* 2014)	Calculates optimal sample size and expected power given specified budgetary constraints and data set parameters estimated from pilot data, using a range of inbuilt DE software. Based on the negative-binomial distribution within a generalized linear model framework.	Pilot data in the form of a read count matrix. Total budget and sample size range.	Returns a matrix of results for each simulation in the conditions tested.	Yes	Yes (paired-sample designs)	R package available via http://www2.hawaii.edu/~lgarmire/RNA-seqPowerCalculator.htm
PROspective Power Evaluation for RNAseq: PROPER (Wu *et al.* 2015)	Prospective power evaluation (without assuming single values for fold change, variance, read count, etc.) using simulated count data given dispersion and expression parameters estimated from real data. Provides a stratified view of power and enables calculating true error rates. Incorporates FDR control.	Biological variance and expression level estimated from pilot data, plus values for sample size, type 1 error rate and minimum detectable log fold change.	A table summarizing power-related metrics under each sample size, including overall power, true discovery rate, false discovery rate and false discovery cost. Graphs can be plotted for each metric to visualize, for example, power stratified by sample size and average read count.	Yes	No	R package available in BIOCONDUCTOR: https://www.bioconductor.org/packages/release/bioc/html/PROPER.html

More sequence or more replication? The depth–replication trade-off in RNA-seq experimental design

Research budgets impose limits on how much sequencing can be performed: forcing a trade-off between sequencing fewer samples more deeply, or including more samples at the cost of per-sample read depth. Current generic guidelines for RNA-seq experimental design ignore study-specific factors influencing power, like individual research goals and underlying biology. The Encyclopedia of DNA Elements Consortium (ENCODE) guidelines recommend at least two biological replicates and 30 million (M) paired-end reads for gene expression estimation from human RNA-seq data. For DE inference, these guidelines are likely inadequate in most cases, as two replicates would provide <20% power to detect a twofold expression difference (Fig. 1). The following sections summarize recent statistical research that shows investment in sample size, rather than sequencing depth, provides the greatest power for differential expression analysis from RNA-seq data.

More sequence is not necessarily better

With deeper sequencing, Poisson noise and random sampling error are reduced across the data set and transcripts with lower expression, lower fold changes and higher variance become more detectable (Tarazona *et al.* 2011). However, recent work demonstrates diminishing returns on power for DE detection with deeper sequencing and reaches a clear consensus on two important points regarding RNA-seq experimental design. (i) Power gains quickly plateau once average read depth reaches ~10 mapped reads per transcript (Busby *et al.* 2011; Hart *et al.* 2013; Wu *et al.* 2015). At this depth, bias arising from random sampling and Poisson counting error is overcome (see Box 3). As these sources of variance derive from the count itself and are thus not experiment-specific, 10 reads average depth serves as a generalizable rule across studies regarding a suitable minimum sequencing depth for DE analysis. (ii) Sequencing efforts in the range of 5–20 M mapped reads per sample provide sufficient depth to accurately quantify gene expression across a broad range of expression levels in diverse eukaryotic transcriptomes (Tarazona *et al.* 2011; Wang *et al.* 2011; Hart *et al.* 2013; Vijay *et al.* 2013; Ching *et al.* 2014; Liu *et al.* 2014; Williams *et al.* 2014). For example, Hart *et al.* (2013) examined expression distributions for 127 RNA-seq experiments (six replicated studies; human and zebrafish), finding that 10 M mapped reads were sufficient to cover approximately 90% of transcripts with >10 reads in a range of biosamples (cell lines, tissue/organ and population comparisons). Larger, more complex transcriptomes, and data sets with higher dispersion or lower fold changes, will require more reads to reach power saturation (Ching *et al.* 2014).

We focus our discussion on the analysis of large, complex eukaryotic transcriptomes (>17 000 genes) most often studied by molecular ecologists. Obviously, less sequencing effort will be required for RNA-seq analysis of prokaryote transcriptomes. For bacteria, Haas *et al.* (2012) suggest 2–3 M mapped reads per replicate enable statistically robust DE inference above a twofold expression difference. As described for eukaryotes above, further substantial increases in read depth gave diminishing returns on gene detection.

Power derives from sample size

Biological replication improves estimates for all sources of variance and is the only way of quantifying biological variation; thus, increasing the sample size has a more potent effect on power than increasing the sequencing depth. For example, Liu *et al.* (2014) examined the impact of increasing the sample size or sequencing depth on the number of genes found significantly differentially expressed in a human MCF7 cell line comparison (17β-estradiol treated vs. control). Increasing read depth from 10 M to 15 M reads for each of two replicates (i.e. from 20 M to 30 M total reads) resulted in only a 6% increase in the number of genes detected as differentially expressed (for a 50% increase in reads). Using the same number of total reads to sequence an additional replicate per group (three replicates, 30 M total reads) increased DE detection by 35%. Going from six to seven replicates resulted in a further 26% increase in DE detection.

Reducing per-sample read depth in favour of larger sample sizes may sacrifice some technical precision (although mainly for low-abundance RNAs), but achieves overall higher power through improved biological variance estimation (Soneson & Delorenzi 2013; Sims *et al.* 2014). Parametric methods typically used for DE analysis, for example DESEQ (Anders & Huber 2010) and EDGER (Robinson *et al.* 2010), depend on accurately modelling biological variance. When sample sizes are small (≤3 replicates per condition), these methods suffer high false positive rates that can widely exceed the desired FDR threshold (Tarazona *et al.* 2011; Soneson & Delorenzi 2013). Therefore, when budgets are limiting and sample sizes are necessarily small, even a small increase in sample size (e.g. from 2 to 3, or 4 to 5 replicates per condition) can significantly improve experimental power and the accuracy of DE calls from RNA-seq data (Soneson & Delorenzi 2013; Liu *et al.*

Box 5. Rules-of-thumb for designing powerful RNA-seq experiments

Unguided decisions regarding key aspects of RNA-seq experimental design can lead to underpowered experiments, wasted resources and an inability to address primary research goals. Based on recent research effort addressing the trade-off between sequencing depth and sample size in RNA-seq, we derive the following 'rules-of-thumb' as general experimental design guidelines for optimizing power for DE inference from RNA-seq data.

1 *Sequence more replicates rather than increasing read depth*. The most efficient approach to the depth–replication trade-off in RNA-seq experimental design is to sequence more replicates rather than obtaining high read depth on a small number of samples. As further depth is added, the majority of additional reads will map to transcripts that are already relatively well covered (i.e. >10 reads), whereas using additional replication achieves overall higher power through improved biological variance estimation.
2 *Sequence each sample to a depth that ensures the majority of transcripts are covered by >10 reads*. For eukaryotic transcriptomes, aim for ~10 M mapped reads/sample. Beyond 10 reads average depth, bias caused by sampling noise and Poisson counting error is minimized and resources are better spent on increasing the sample size. Sequencing efforts producing 10–20 M mapped reads per sample should achieve this for most eukaryotic transcriptomes. Accurately estimating expression for rare transcripts, and robust DE analysis at the isoform level will require greater sequencing effort.
3 *Sequence at least three biological replicates per condition, more when biological variance is high and/or when the research question includes small expression differences*. We advocate three biological replicates per condition as an absolute minimum for DE analysis from RNA-seq data because this provides some ability to identify outliers. However, three replicates will often only be sufficient to detect large expression differences (≥fourfold). A minimum number of replicates required to achieve acceptable power (e.g. >80%) for DE inference in a given study will depend critically on the within-group biological variance and fold change of expression differences (Fig. 1 main text). A well-defined research question must drive the entire experimental design process to establish a minimum biologically meaningful level of expression difference and, therefore, the required power.
4 *Conduct a pilot sequencing experiment*. Large-scale RNA-seq experiments, with full replication, remain costly. It is therefore imperative to know, for a particular study system, what sample size and read depth is sufficient to achieve a desired level of power for detecting the expression differences of interest. A pilot sequencing experiment is the only sure way to evaluate the feasibility of larger experiments and assess their likely benefits vs. costs. Specifically, a good pilot experiment should answer two key questions: 'What is the best (most powerful) experiment that I can afford to do?' and 'What is the smallest fold change I can reliably detect?' A pilot sequencing experiment might consist of a few biological replicates representing each of the main conditions of interest, multiplexed over one lane of sequencing. Such an experiment may cost a few thousand dollars, but will provide the necessary data for estimating biological variance and baseline expression levels and, using newly available tools (Table 1; Box 4), for calculating sample sizes and sequencing efforts needed in a larger experiment. Sequencing pilot samples more deeply than in a larger DE experiment also provides the additional depth necessary for *de novo* transcriptome assembly.

2014) (Fig. 1). By contrast, incorporating more reads per sample has been shown to progressively introduce more false positives (primarily genes of shorter length, lower expression level and smaller fold changes, as well as off-target RNA species) (Tarazona *et al.* 2011). Therefore, in the context of RNA-seq DE analysis, deep sequencing equates to wasted effort and can in fact prove counterproductive.

For a given RNA-seq DE study, an effective minimum sample size that achieves acceptable power (e.g. >80%) will depend most critically on the magnitude of biological variance and the scale of expression differences of interest (Hart *et al.* 2013; Ching *et al.* 2014; Wu *et al.* 2015). A well designed RNA-seq study can detect expression levels that are difficult to assess using qPCR or microarrays, due to its superior dynamic range. Figure 1 provides a general guide to statistical power for detecting different effect sizes of expression difference (i.e. fold change) given different sample sizes, for hypothetical RNA-seq data sets with low vs. high biological variance. Overall, large expression differences (≥fourfold) can be reliably detected with modest biological replication (i.e. 3–5 replicates per condition) in most data sets (Fig. 1). However, data sets with high variance or small fold changes require much larger sample sizes to achieve suitable power. For example, for low-variance data (e.g. an inbred zebrafish strain), detecting at least a twofold expression difference with at least 80% power is achievable with three or more biological replicates. For high-variance data (e.g. a wild reef fish population), detecting a twofold change with 80% power requires at least 10 replicates per condition. In either example, detecting subtle fold changes (<twofold) with confidence requires very large sample sizes (>10)

because of the difficulty in differentiating true expression differences from background noise. Box 4 describes a pilot sequencing experiment that evaluates power to infer DE in two real RNA-seq data sets from a wild reef fish population, and demonstrates how the expression landscape can vary by tissue type and impact the power of DE tests.

Designing better RNA-seq experiments

Designing RNA-seq DE experiments that optimize power while limiting cost is now easier. Recent work, summarized above, convincingly demonstrates that prioritizing replication over sequencing depth achieves the greatest power for DE inference from RNA-seq data. This work is synthesized into four rules-of-thumb for designing powerful RNA-seq experiments (Box 5), which are complimented by a summary of available software tools for evaluating statistical power and calculating appropriate sample sizes and sequencing depths for RNA-seq DE experiments within definable budgetary limits (Table 1). Currently, most software tools only consider single-factor (two-condition) designs; power assessment for more complex multifactorial RNA-seq experiments should be a focus of future work.

Generalizable rules-of-thumb are helpful, but striking the right balance between read depth and sample size in RNA-seq experimental design must be carefully considered on a study-specific basis, taking into account the research question as much as budget and biology. If the scientific aims of the experiment include rare transcripts, subtle fold changes and/or an isoform-level analysis, both deeper sequencing and greater replication will be necessary. Robust analysis of rare transcripts (e.g. noncoding RNAs or rare splice isoforms) may benefit from using RNA-capture techniques enriching for low-abundance RNAs (Halvardson *et al.* 2013), as well as technical replication in addition to more extensive biological replication (McIntyre *et al.* 2011). By contrast, fewer reads may be necessary to characterize gene expression for low-complexity libraries, as is often the case for degraded samples. Because expression landscape, library complexity and the distribution of read counts in RNA-seq data will also be tissue-specific, experimental designs will potentially vary even among tissues of the same organism (Attolini *et al.* 2015) (Box 4). We advocate a pilot sequencing approach (Box 5), for evaluating suitable sequencing depths and sample sizes to achieve optimal power to address the research question of interest given the data (and budget) at hand. In addition, greater sequencing depths are necessary in instances where *de novo* transcriptome assembly is required to provide a mapping reference for transcript quantification (Fig. 2). Here, given the relatively low cost of sequencing, pilot work can be expanded to obtain ~100 M paired-end reads (>100 bp), recommended in the current literature as sufficient to capture the majority of RNAs expressed in eukaryotic samples (Wang *et al.* 2011; Francis *et al.* 2013; Vijay *et al.* 2013; Wolf 2013).

Choice of mapping reference matters

Success of downstream DE inference also hinges on the quality and completeness of the mapping reference used for transcript quantification (Busby *et al.* 2011; Vijay *et al.* 2013). As annotated genome sequences become available for an ever-widening diversity of taxa (Koepfli *et al.* 2015), the option of mapping to a genome from a related species will increasingly become an option for many nonmodel species. Using a genomic reference achieves more accurate isoform counts, because isoform recovery by *de novo* transcriptome assembly is more error prone (Vijay *et al.* 2013). However, genomic divergence together with the accuracy and completeness of the genome assembly is crucial to the success of this strategy. Simulations show that a genomic mapping approach produces more accurate gene expression estimates for DE inference, at up to 15% sequence divergence from the study species, but that at 30% divergence, incorrect mapping makes a *de novo* assembled transcriptome the better choice of reference (Vijay *et al.* 2013). Unfortunately, many current genome assemblies are incomplete or inaccurately assembled. As a consequence, genes expressed in RNA-seq data but missing (or misassembled) in the reference genome will not be counted. Assembling a *de novo* transcriptome from the same data used for DE analysis largely circumvents this issue (Haas *et al.* 2013) and the best current rule-of-thumb we can provide is to encourage investigators to pursue both approaches if available to them.

Transcriptome assembly and quality assessment are active areas of research (Haas *et al.* 2013; O'Neil & Emrich 2013; Vijay *et al.* 2013; Yang & Smith 2013; Li *et al.* 2014), but how transcriptome assembly and the choice of mapping reference influences success of downstream DE inference is an aspect of RNA-seq experimental design that warrants further study.

Ongoing challenges and prospects for RNA-seq in ecology and evolution

Ultimately, gleaning meaningful biological insights from lists of genes found differentially expressed in RNA-seq data depends crucially on the availability of accurate gene annotation information. Reliably inferring functionally relevant information is a major challenge in ecological and evolutionary genomics research generally

(Primmer *et al.* 2013). Gene name assignments come for free when a well-annotated genomic reference is available for transcript quantification. But this is still rarely the case in eco-evolutionary research, and contigs in a *de novo* transcriptome assembly come with no information on their potential biological function. The Gene Ontology (GO) database (Harris *et al.* 2004) provides the largest organized resource for transferring functional gene annotations from genetic model organisms to nonmodel species based on inferred sequence homology (Primmer *et al.* 2013). Unfortunately, often only a small proportion of DE contigs can be matched to known proteins [e.g. <20% in the bluehead wrasse (Liu *et al.* 2015), 31% in *Acropora* coral (Meyer *et al.* 2011)]. There will be many (e.g. taxon-specific) transcripts for which a physiologically relevant context cannot be determined. Many factors can also lead to erroneous orthology assignment, and the assumption that orthologous genes retain the same function across species will not always hold true. As gene annotation databases grow to include yet greater molecular and taxonomic diversity, so will the depth of biological insight possible through RNA-seq analysis in nonmodel species.

Caveats discussed herein do not undermine RNA-seq technology as a practical tool in eco-evolutionary research when used within the bounds of current experimental design limitations. The strength of RNA-seq is that it provides unprecedented access to transcriptome-wide gene expression data in any species. For nonmodel species, which typically lack extensive existing genomic resources (e.g. a suitable microarray platform), or when novel transcripts are of interest, RNA-seq is an extremely efficient and cost-effective tool for exploring transcript diversity. When extensive replication is unaffordable, integrating RNA-seq with established technologies, such as qPCR and pre-existing microarrays, remains a robust strategy to study gene expression in eco-evolutionary contexts and nonmodel species. In a recent review of ecological transcriptomics research over the last decade, Alvarez *et al.* (2015) suggest leveraging RNA-seq as an exploratory tool to first establish whether genes of interest are active under given experimental conditions and to discover novel candidate genes, both of which can be followed up with further research using targeted technologies such as qPCR.

The nascent use of biological replication in RNA-seq is typical of a maturing research field. The major limitation currently preventing larger sample sizes and more powerful RNA-seq study designs is the expense of per-sample library preparation. High-throughput library preparation protocols are being developed, as too are library protocols that work with increasingly small quantities of RNA, such as that obtained from a single cell (scRNA-seq, reviewed by Kolodziejczyk *et al.* 2015; Hicks *et al.* 2015). Such approaches further empower RNA-seq. Shishkin *et al.* (2015) recently reported RNA-tag-Seq, a method for generating single RNA-seq libraries containing many RNA samples, which are individually barcoded and pooled before library construction and should significantly reduce costs. Increased use of robotics and other forms of automation should also reduce overall costs as well as technical variance introduced during high-throughput library preparation.

Looking to the future, integrating RNA-seq data with other omics technologies will soon bring a truly multidimensional approach to systems biology and to quantifying the link between genotype and phenotype. Whole-genome scans of DNA methylation via bisulphite sequencing and protein–DNA interaction sites via chromatin immunoprecipitation sequencing (ChIP-seq), are bridging the gap between observed mRNA abundances and epigenetic regulation of gene expression. High-throughput proteomics data can likewise place gene expression into a direct functional context (Diz *et al.* 2012). If cost barriers can be overcome, and with appropriate experimental designs, the combination of new RNA-seq approaches and integrated omics seems set to usher in a new era of science that can investigate patterns of change from the most fundamental building block, the single cell, through to higher levels of complexity.

Conclusions

RNA-seq holds great promise for whole-transcriptome gene expression analysis in ecology and evolution, but considerable challenges remain. The rapid rise in popularity of RNA-seq in published research follows dramatic falls in the cost of sequencing and significant improvements in the efficiency of sequencing technologies. However, this has not yet translated into clear improvements in experimental design when it comes to biological replication. Hypothesis-driven RNA-seq requires careful experimental design that considers desired power in the context of study aims. The inherent sensitivity of gene expression to environmental stimulus means that while field-based RNA-seq experiments on wild populations can be valuable for generating hypotheses, drawing robust conclusions regarding processes underlying observed expression differences may be challenging without high levels of replication. The guidelines and software tools presented here in should assist investigators in designing RNA-seq experiments that optimize statistical power while limiting cost. Adoption of these approaches should lead to further strengthening of the experimental basis of a field still in its infancy. We have no doubt that as RNA-seq data are integrated with other omics technologies and as available genomic resources for nonmodel

species expand, so too will the range of biological insights possible from RNA-seq analysis in the ecological and evolutionary sciences.

Acknowledgements

We are grateful to Helen Taylor, Kim Rutherford and four anonymous reviewers for their constructive feedback that helped to improve the final manuscript. This research was supported by Grant UOO1308 awarded to NJG by the Marsden Fund, Royal Society of New Zealand.

References

Adiconis X, Borges-Rivera D, Satija R *et al.* (2013) Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nature Methods*, **10**, 623–629.

Alvarez M, Schrey AW, Richards CL (2015) Ten years of transcriptomics in wild populations: what have we learned about their ecology and evolution? *Molecular Ecology*, **24**, 710–725.

Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biology*, **11**, 12.

Attolini CSO, Pena V, Rossell D (2015) Designing alternative splicing RNA-seq studies. Beyond generic guidelines. *Bioinformatics*, **31**, 3631–3637.

Aubin-Horth N, Renn SCP (2009) Genomic reaction norms: using integrative biology to understand molecular mechanisms of phenotypic plasticity. *Molecular Ecology*, **18**, 3763–3780.

Auer PL, Doerge RW (2010) Statistical design and analysis of RNA sequencing data. *Genetics*, **185**, 405–416.

Biswas S, Agrawal YN, Mucyn TS, Dangl JL, Jones CD (2013) Biological averaging in RNA-seq. arXiv 1309.0670 [q-bio. QM].

Bonneaud C, Balenger SL, Russell AF *et al.* (2011) Rapid evolution of disease resistance is accompanied by functional changes in gene expression in a wild bird. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 7866–7871.

Bullard JH, Purdom E, Hansen KD, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94.

Busby MA, Gray JM, Costa AM *et al.* (2011) Expression divergence measured by transcriptome sequencing of four yeast species. *BMC Genomics*, **12**, 635.

Busby MA, Stewart C, Miller CA, Grzeda KR, Marth GT (2013) Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics*, **29**, 656–657.

Camacho-Sanchez M, Burraco P, Gomez-Mestre I, Leonard JA (2013) Preservation of RNA and DNA from mammal samples under field conditions. *Molecular Ecology Resources*, **13**, 663–673.

Ching T, Huang S, Garmire LX (2014) Power analysis and sample size estimation for RNA-Seq differential expression. *RNA*, **20**, 1684–1696.

Cieslik M, Chugh R, Wu Y-M *et al.* (2015) The use of exome capture RNA-seq for highly degraded RNA with application to clinical cancer sequencing. *Genome Research*, **9**, 1372–1381.

Diaz de Cerio O, Rojo-Bartolomé I, Bizarro C, Ortiz-Zarragoitia M, Cancio I (2012) 5S rRNA and accompanying proteins in gonads: powerful markers to identify sex and reproductive endocrine disruption in fish. *Environmental Science & Technology*, **46**, 7763–7771.

Diz AP, Martinez-Fernandez M, Rolan-Alvarez E (2012) Proteomics in evolutionary ecology: linking the genotype with the phenotype. *Molecular Ecology*, **21**, 1060–1080.

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.

Fisher RA (1935) *The Design of Experiments*, 2nd edn. Oliver & Boyd, Edinburgh.

Francis WR, Christianson LM, Kiko R *et al.* (2013) A comparison across non-model animals suggests an optimal sequencing depth for *de novo* transcriptome assembly. *BMC Genomics*, **14**, 167.

Fraser BA, Janowitz I, Thairu M, Travis J, Hughes KA (2014) Phenotypic and genomic plasticity of alternative male reproductive tactics in sailfin mollies. *Proceedings of the Royal Society B-Biological Sciences*, **281**, 20132310.

Godwin J (2009) Social determination of sex in reef fishes. *Seminars in Cell & Developmental Biology*, **20**, 264–270.

Gunter HM, Fan SH, Xiong F *et al.* (2013) Shaping development through mechanical strain: the transcriptional basis of diet-induced phenotypic plasticity in a cichlid fish. *Molecular Ecology*, **22**, 4516–4531.

Haas BJ, Chin M, Nusbaum C, Birren BW, Livny J (2012) How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? *BMC Genomics*, **13**, 734.

Haas BJ, Papanicolaou A, Yassour M *et al.* (2013) *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, **8**, 1494–1512.

Halvardson J, Zaghlool A, Feuk L (2013) Exome RNA sequencing reveals rare and novel alternative transcripts. *Nucleic Acids Research*, **41**, e6.

Hansen KD, Wu Z, Irizarry RA, Leek JT (2011) Sequencing technology does not eliminate biological variability. *Nature Biotechnology*, **29**, 572–573.

Harris R, Hofmann H (2014) Neurogenomics of Behavioral Plasticity. In: *Ecological Genomics: Ecology and the Evolution of Genes and Genomes* (eds Landry CR, Aubin-Horth N), pp. 149–168. Springer Netherlands, Dordrecht.

Harris MA, Clark J, Ireland A *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, **32**, D258–D261.

Hart SN, Therneau TM, Zhang Y, Poland GA, Kocher J-P (2013) Calculating sample size estimates for RNA sequencing data. *Journal of Computational Biology*, **20**, 970–978.

Hicks SC, Teng M, Irizarry RA (2015) On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. *bioRxiv*, doi: http://dx.doi.org/10.1101/025528.

Kendziorski C, Irizarry RA, Chen K-S, Haag JD, Gould MN (2005) On the utility of pooling biological samples in microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 4252–4257.

Koepfli K-P, Benedict P, Genome 10K Community of Scientists, O'Brien SJ (2015) The Genome 10K Project: A

Way Forward. *Annual Review of Animal Biosciences*, **3**, 57–111.

Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA (2015) The technology and biology of single-cell RNA sequencing. *Molecular Cell*, **58**, 610–620.

Li CI, Su PF, Shyr Y (2013) Sample size calculation based on exact test for assessing differential expression analysis in RNA-seq data. *BMC Bioinformatics*, **14**, 7.

Li B, Fillmore N, Bai Y *et al.* (2014) Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biology*, **15**, 553.

Liu YW, Zhou J, White KP (2014) RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics*, **30**, 301–304.

Liu H, Lamm MS, Ruthorford K, Black MA, Godwin JR, Gemmell NJ (2015) Large-scale transcriptome sequencing reveals novel expression patterns for key sex-related genes in a sex-changing fish. *Biology of Sex Differences*, **6**, 26.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, **18**, 1509–1517.

McIntyre LM, Lopiano KK, Morse AM *et al.* (2011) RNA-seq: technical variability and sampling. *BMC Genomics*, **12**, 293.

Meyer E, Aglyamova GV, Matz MV (2011) Profiling gene expression responses of coral larvae (*Acropora millepora*) to elevated temperature and settlement inducers using a novel RNA-Seq procedure. *Molecular Ecology*, **20**, 3599–3616.

O'Neil ST, Emrich SJ (2013) Assessing *de novo* transcriptome assembly metrics for consistency and utility. *BMC Genomics*, **14**, 12.

Ozsolak F, Milos PM (2011) RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, **12**, 87–98.

Primmer CR, Papakostas S, Leder EH, Davis MJ, Ragan MA (2013) Annotated genes and nonannotated genomes: cross-species use of Gene Ontology in ecology and evolution research. *Molecular Ecology*, **22**, 3216–3241.

Rajkumar AP, Qvist P, Lazarus R *et al.* (2015) Experimental validation of methods for differential gene expression analysis and sample pooling in RNA-seq. *BMC Genomics*, **16**, 548.

Robinson DG, Storey JD (2014) subSeq: determining appropriate sequencing depth through efficient read subsampling. *Bioinformatics*, **30**, 3424–3426.

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Robinson DG, Wang JY, Storey JD (2015) A nested parallel experiment demonstrates differences in intensity-dependence between RNA-seq and microarrays. *Nucleic Acids Research*, **20**, e131.

Romero IG, Ruvinsky I, Gilad Y (2012) Comparative studies of gene expression and the evolution of gene regulation. *Nature Reviews Genetics*, **13**, 505–516.

Romero IG, Pai AA, Tung J, Gilad Y (2014) RNA-seq: impact of RNA degradation on transcript quantification. *BMC Biology*, **12**, 13.

Schneider RF, Li Y, Meyer A, Gunter HM (2014) Regulatory gene networks that shape the development of adaptive phenotypic plasticity in a cichlid fish. *Molecular Ecology*, **23**, 4511–4526.

Schroeder A, Mueller O, Stocker S *et al.* (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Molecular Biology*, **7**, 3.

Schunter C, Vollmer SV, Macpherson E, Pascual M (2014) Transcriptome analyses and differential gene expression in a non-model fish species with alternative mating tactics. *BMC Genomics*, **15**, 13.

Shishkin AA, Giannoukos G, Kucukural A *et al.* (2015) Simultaneous generation of many RNA-seq libraries in a single reaction. *Nature Methods*, **12**, 323–325.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, **15**, 121–132.

Smith S, Bernatchez L, Beheregaray LB (2013) RNA-seq analysis reveals extensive transcriptional plasticity to temperature stress in a freshwater fish species. *BMC Genomics*, **14**, 375.

Soneson C, Delorenzi M (2013) A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, **14**, 91.

Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, Conesa A (2011) Differential expression in RNA-seq: a matter of depth. *Genome Research*, **21**, 2213–2223.

Veilleux HD, Ryu T, Donelson JM *et al.* (2015) Molecular processes of transgenerational acclimation to a warming ocean. *Nature Climate Change*, **5**, 1074–1078.

Vijay N, Poelstra JW, Kunstner A, Wolf JBW (2013) Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Molecular Ecology*, **22**, 620–634.

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**, 57–63.

Wang Y, Ghaffari N, Johnson CD *et al.* (2011) Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens. *BMC Bioinformatics*, **12**, 7.

Williams AG, Thomas S, Wyman SK, Holloway AK (2014) RNA-seq Data: Challenges in and Recommendations for Experimental Design and Analysis. *Current Protocols in Human Genetics*, **83**, 11.13.1–11.13.20.

Wolf JBW (2013) Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. *Molecular Ecology Resources*, **13**, 559–572.

Wu H, Wang C, Wu Z (2013) A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*, **14**, 232–243.

Wu H, Wang C, Wu Z (2015) PROPER: comprehensive power evaluation for differential expression using RNA-seq. *Bioinformatics*, **31**, 233–241.

Yang Y, Smith SA (2013) Optimizing *de novo* assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics*, **14**, 11.

E.V.T. and N.J.G. conceived the ideas, E.V.T. and M.A.B. performed the analyses, and E.V.T. wrote the manuscript. All authors contributed editorially and approved the final version of the manuscript.

Data accessibility

References for studies reviewed in Box 1, and the R code and parameters for power analyses performed in Box 4, are supplied in the online Supporting information. Data files containing raw counts for samples analysed in Box 4 are available on Dryad DOI:10.5061/dryad.vp42s. Raw sequences used in assembling the reference transcriptome for mapping these data are available through NCBI's Sequence Read Archive under accession SRP06302n.

Supporting information

Additional supporting information may be found in the online version of this article.

Appendix S1 References reviewed.

Appendix S2 R code used for power analysis.