



Transcriptomic Analysis in the Sea Anemone *Nematostella vectensis*

Jacob Warner, Eric Röttinger

► To cite this version:

Jacob Warner, Eric Röttinger. Transcriptomic Analysis in the Sea Anemone *Nematostella vectensis*.
Methods in Molecular Biology, 2020, 10.1007/978-1-0716-0974-3_14 . hal-03021243

HAL Id: hal-03021243

<https://hal.science/hal-03021243>

Submitted on 24 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Chapter 14

Transcriptomic Analysis in the Sea Anemone *Nematostella vectensis*

Jacob F. Warner and Eric Röttinger

Abstract

The sea anemone *Nematostella vectensis* is an emerging research model to study embryonic development and regeneration at the molecular and global transcriptomic level. Transcriptomics analysis is now routinely used to detect differential expression at the genome level. Here we present the latest procedures for isolating high-quality RNA required for next generation sequencing, as well as methods and resources for quantifying transcriptomic data.

Key words RNA isolation, transcriptome, De novo assembly, Cnidarian, *Nematostella vectensis*

1 Introduction

The anthozoan cnidarian *Nematostella vectensis* has long been used as a model for evolutionary developmental (EvoDevo) studies as it is readily cultivable under laboratory conditions [1]. More recently, *Nematostella* emerged as a powerful and complementary whole-body regeneration model, as it is able to regrow missing body parts within days after amputation [2–5]. The genome of *Nematostella* was the first cnidarian genome to be sequenced and has revealed surprising similarities with mammalian genomes [6, 7]. Since then, a variety of functional genomics approaches have been adapted for use with *Nematostella* including gene knockdowns [8–11], gene knockouts [11, 12] and CRISPR-mediated knockins [13]. As high throughput transcriptomic approaches have advanced, several studies have assessed global gene expression during *Nematostella* early embryonic development [14–17] and regeneration [17]. Thus, this cnidarian model has been demonstrated to be a powerful system for transcriptomic studies addressing a variety of developmental questions. In this chapter, we present methods for performing such transcriptomic studies. We describe the procedure for isolating

high-quality RNA from embryonic and juvenile tissues followed by methods and resources for quantifying transcriptomic data.

2 Materials

2.1 RNA Isolation

Ensure all solutions are nuclease free. Using contaminated reagents can yield degraded RNA.

1. Tri-reagent (Sigma-Aldrich 93,289) or TRIzol (Invitrogen 15596026).
2. Pestle motor and disposable pestle (Sigma-Aldrich Z359971 & Z359947).
3. Phase Lock Gel 1.7 ml tubes (optional, Quanta 2302830).
4. Chloroform (Sigma-Aldrich C2432).
5. Phenol–chloroform pH 6.0 (Sigma-Aldrich P2069).
6. Glycogen 20 mg/ml (Omega BioTek AC122).
7. Isopropanol.
8. 70% ethanol in nuclease-free water
9. Nuclease-free water.
10. DNase (optional, Ambion AM2222).

2.2 Quantification of Transcriptomic Data

Bioinformatic tools and publicly available datasets are ever evolving. Below are some useful resources that are current at the time of writing.

1. Alignment reference: As of writing there are several publicly available resources that can be used to quantify transcriptomic data (*see* Table 1). These include the *Nematostella vectensis* draft genome assembly, Nemve1, and its associated gene models [6]. The genome, and gene models can be downloaded from <https://mycocosm.jgi.doe.gov/Nemve1/Nemve1.home.html> under the Download tab (account required). Fredman and colleagues [18] used transcriptomic data in conjunction with the Nemve1 assembly to generate a second set of gene models identified with the prefix NVE. These datasets can be downloaded from https://figshare.com/articles/Nematostella_vectensis_transcriptome_and_gene_models_v2_0/807696.
2. Recently, Warner and colleagues generated a transcriptome assembly from a large dataset of RNAseq samples spanning embryogenesis and regeneration [17]. The browser for this resource is found at <http://nvertx.kahikai.org> and the transcriptome assembly can be downloaded at <http://ircan.unice.fr/static/downloads/NvERTx.4.fasta.gz> with the associated annotations found at <http://ircan.unice.fr/static/downloads/NvERTx.4.annotations.full.txt.gz>.

Table 1
Genomic and Transcriptomic resources for *Nematostella vectensis*

Moniker	Type	URL	Reference
Nemvel	Genome, gene models, annotations	https://mycocosm.jgi.doe.gov/Nemvel/Nemvel.home.html	[6]
NVE genes	Transcriptome, gene models, annotations	https://figshare.com/articles/Nematostella_vectensis_transcriptome_and_gene_models_v2_0/807696	[18]
NvERTx	Transcriptome, annotations	http://nvertx.kahikai.org	[17]

3. Sequence read quality assurance (optional): Sequence read quality can be assessed with fastQC [19]. If the reads need to be trimmed, both Trimmomatic [20] and Cutadapt [21] work well to trim common adapter sequences and low-quality bases.
4. Sequence read aligners: Choice of the alignment program depends on the type of reference used. For alignments to a genome, splice aware aligners will increase read mapping as they can accommodate introns. HISAT2 and STAR are two popular programs [22, 23]. For alignments to a transcriptome, end to end aligners perform well. Popular programs include Bowtie2 and BWA [24, 25].
5. Alignment counting: After alignment, reads need to be quantified. Featurecounts works well but requires a General Feature Format (GFF) file which is often unavailable for transcriptomes [26]. Other reliable counting programs include RSEM [27].
6. Differential expression testing: The two most popular programs for differential expression analysis are the R packages DEseq2 [28] and edgeR [29].

3 Methods

3.1 RNA Isolation and Quality Control

When isolating RNA, it is important to work quickly. For this reason, we do not advise extracting a large number of samples at one time, rather we recommend planning the experiments in advance so as to work in batches.

1. Lyse tissues: Collect 100–400 embryos or juveniles in a 1.7 ml tube. Gently spin sample down in a mini benchtop centrifuge. Add 300 μ l of Tri-reagent and homogenize with a pestle for 15 s. Rinse the tip of the pestle with 200 μ l of Tri-reagent. Immediately freeze sample in liquid nitrogen (*see* **Notes 1–3**).
2. Samples can be stored at -80°C or proceed with extraction.

3. Thaw samples on ice for 10 min. If using Phase Lock tubes, spin them for 60s at >12,000 rcf (*see Note 4*).
4. Add 500 μ l of Tri-reagent and mix by pipetting up and down. The sample should be thawed and no longer translucent.
5. Add 200 μ l of chloroform to the sample. If using Phase Lock tubes (optional), transfer the sample to a prespun Phase Lock tube. Shake the sample for 15–20 s. Do not vortex, especially if using Phase Lock tubes!
6. Incubate the samples for 3 min on ice.
7. Spin the samples at >12,000 rcf for 15 min at 4 °C.
8. Carefully pipet the aqueous (upper) phase into a new 1.7 ml tube or prespun Phase Lock tube.
9. Perform a second extraction adding 500 μ l of phenol–chloroform (pH 6.0) and repeating **steps 6–8** (*see Notes 5 and 6*).
10. Precipitate the RNA by adding 2 μ l of glycogen (20 mg/ml), followed by 500 μ l of 100% isopropanol. Shake tube for 10 s and incubate at room temperature for 2 min.
11. Spin samples at >12,000 rcf for 15 min at 4 °C.
12. Remove the liquid from the tube by carefully pipetting. Avoid disturbing the translucent whitish-tan pellet (*see Notes 7 and 8*). Once the liquid is removed, spin briefly at >12,000 rcf for 10 s and remove remaining liquid with a p10 pipette.
13. Wash pellet by adding 1 ml of 70% ethanol. Mix by inverting the tube 3 times.
14. Spin samples at >12,000 rcf for 15 min at 4 °C.
15. Repeat **steps 12–14**.
16. Remove the remaining liquid as described in **step 12**. Invert tube on clean surface or delicate task wipe and allow to dry for approximately 5 min (*see Note 9*).
17. Resuspend in 15–30 μ l of nuclease free water (*see Note 10*).
18. Optional: treat RNA with DNase (*see Notes 11 and 12*).
19. Assess RNA integrity by running 1 μ l on an RNA gel or using a microfluidic electrophoresis machine (e.g., Agilent Bioanalyzer, Fig. 1) and quantify using a spectrophotometer (e.g., NanoDrop) (*see Note 13*).
20. Proceed with Next Generation Sequencing library preparation protocol that is appropriate for the downstream application (*see Note 14*).

3.2 Quantification of Transcriptomic Data

The following describes a standard workflow for quantifying Next Generation Sequencing (NGS) reads from a transcriptomic analysis. As programs, tools, and file-formats are ever changing we include here a general workflow describing tools that are current

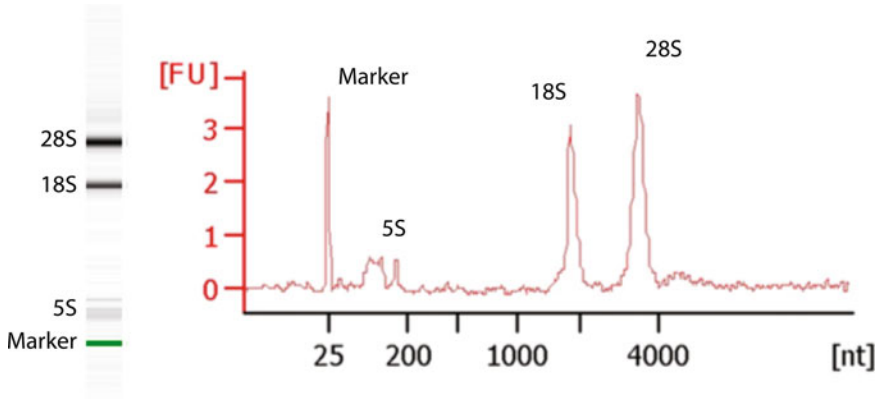


Fig. 1 Bioanalyzer results from total RNA extraction of *Nematostella vectensis*. The capillary electrophoresis gel is pictured on the left with the fluorescence intensity quantification on the right. The 5S, 18S, and 28S ribosomal subunits show up as clear, well separated bands indicating a high-quality extraction

at the time of writing. The programs below are intended to be run on a Linux server or Linux/Unix based PC.

1. Obtain reference for read alignment: Download a *Nematostella vectensis* genome or transcriptome assembly from one of the sources in Table 1.
2. Trim reads (optional): While read trimming is not required, it can increase the read mapping rate of lesser quality sequence data. Both Trimmomatic and Cutadapt work very well for trimming NGS sequence reads. Ensure that the appropriate sequencing adapters are used to trim (*see Note 15*).
3. Align reads to the reference: Read alignment can be performed with any number of popular programs including Bowtie2 for end-to-end alignment (e.g., for alignment to a transcriptome) or HISAT2 for splice aware alignments (e.g., for alignment to a genome). If the alignment rate is less than 50%, assess read quality with fastQC and consider trimming low quality reads.
4. Count successfully mapped reads: Read quantification can be performed with a variety of tools including RSEM [27], or Featurecounts [26]. The former works well with alignments to complex transcriptomes while the later works very well with genomic alignments.
5. Gene level quantification: When aligning to a transcriptome in which multiple transcripts per gene are present, gene level quantification can be estimated by summing the mapped reads of all transcripts per gene. The R functions `aggregate()`, and `summarize()` are very useful for this. NvERTx, includes Nemve1 annotations that can be used for gene level quantification (*see Note 16*). De novo transcriptome assemblies can be annotated using a draft genome or closely related species to identify putative genes.

6. Remove lowly expressed genomic features (optional but recommended): Before normalizing, transcripts or genes with very low counts should be removed as these features, depending on the number of replicates, will often have very high variance and affect dispersion estimates. A detection limit can be calculated to determine the cutoff threshold. As a rule of thumb, when analyzing a result done in triplicate, genomic features with 5 counts or fewer in more than 25% of the samples will be too low to accurately quantify.
7. Normalization and differential testing: The Bioconductor R packages both DEseq2 and edgeR perform very well with transcriptomic data and can accommodate a variety of experimental designs. For both of these, there are extensive tutorials, documentation, and knowledgebase (*see* **Note 17**). We refer the reader to these resources to develop a workflow appropriate to their experimental design.
8. Quality assurance: While the DEseq2 and edgeR workflows produce differential expression analyses with corrected p -values, it is still recommended to ensure the global structure of your results are reasonable. Two methods for doing this are to use principal component analysis (PCA) or hierarchical clustering to visualize sample clustering. The R functions `prcomp()` and `hclust()` work well for these types of analyses (*see* **Note 18**). Ideally the results would show tight clustering of the replicates, with separation of samples by other biological conditions as is reasonable, for example treatment and control samples grouping together (*see* Fig. 2).

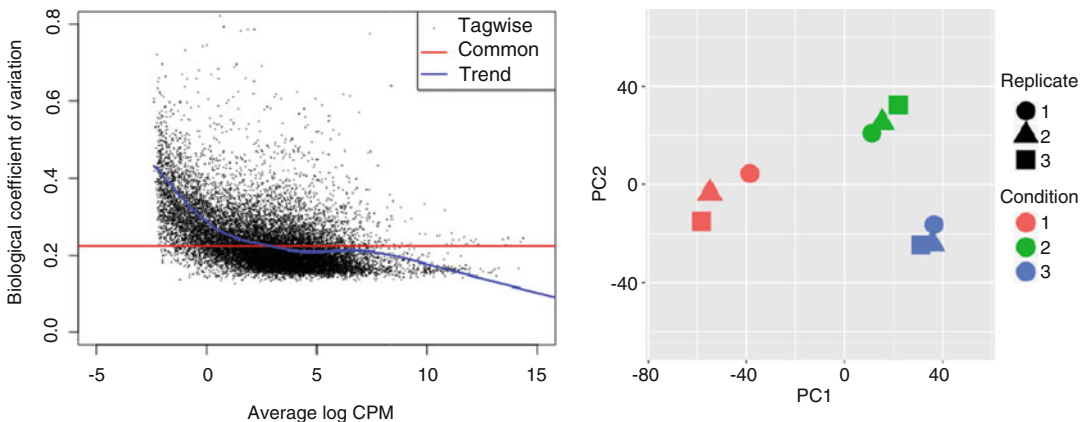


Fig. 2 (a) Mean–variance plot of reanalyzed RNAseq data from Warner et al. 2018. Y-axis is the Biological coefficient of variation (BCV), x-axis shows the average log counts per million of each feature CPM across all samples. The common BCV is ~ 0.21 . Note the higher variance of lowly expressed features (average log CPM < 0). (b) A principal component analysis (PCA) of the same samples with PC1 shown on the x-axis and PC2 on the Y-axis. Note the grouping of the three replicates and clear separation of the samples by condition

9. Batch correction (optional): If the samples being analyzed were sequenced on different platforms, or if they were produced from different research groups, the results can be strongly skewed by batch effects. These effects will be evident in the quality assurance step above. If this is the case, batch effect correction can be employed. The R package sva [30] includes tools for batch correction including the function `ComBat()` (see **Notes 19** and **20**).
10. GO term enrichment: Differential testing will yield a list of genes enriched in a sample of interest (e.g., timepoint or condition). Browsing this list can help develop hypotheses but often these lists are large and difficult to interpret. These gene lists can be further mined by testing for significant enrichment of gene ontology (GO) terms. To do this, the gene IDs or transcript IDs must be mapped to GO terms. Nemvel includes associated GO terms on the download tab of the web browser (<https://mycocosm.jgi.doe.gov/Nemvel/Nemvel.home.html>). An updated GO terms list for Nemvel gene IDs, based on best blast hit to the Uniprot databases can be found at (http://ircan.unice.fr/ER/ER_plotter/about). These GO terms can be used as a background set for enrichment testing where the differentially expressed gene list is used as a test set. The R package topGO can be used to calculate GO term enrichment (see **Note 21**).

4 Notes

1. Tri reagent, phenol, and chloroform should be used under an appropriate fume extraction hood.
2. If harvesting RNA from multiple samples, work each sample one at a time, not in parallel. This helps minimize the time between Tri-reagent addition, lysis, and freezing helping to prevent RNA degradation.
3. If harvesting RNA from adult tissues, use a larger amount of Tri-reagent. The sample should not exceed 10% of the Tri-reagent volume.
4. Phase Lock tubes are optional but can increase yield and reduce contamination.
5. This second extraction increases RNA purity.
6. For certain commercial phenol–chloroform preparations, *not* adding the supplied buffer results in the required pH of 6.0.
7. An RNA pellet will appear as a whitish-tan opaque speck on the bottom of the tube. Holding the tube up to a window while removing the liquid can help avoid aspirating the pellet.

8. Excessive white precipitate can indicate contamination by salts or other debris during the extraction. These samples may be rescued by resuspending in nuclease free water and performing an additional phenol–chloroform extraction followed by precipitation.
9. When dry, the pellet will appear glass-like and lose its color. Hold tube up to a window or light to help visualize the pellet.
10. Pellets can occasionally be difficult to resuspend if dried for longer than necessary. Be patient and pipet until small particles are no longer observed when pipetting up and down.
11. For DNase treatment we strongly recommend using a DNase kit with bead-based inactivation as heat inactivation can result in RNA degradation.
12. If using a DNase treatment set aside 1 μ l of untreated RNA for quality assurance. This hold back sample can be analyzed during trouble shooting in the event the final RNA is degraded and rule out the DNase treatment as a source of the degradation.
13. Typically formaldehyde-based gels are used to assess RNA quality. An appropriate substitute, however, is to use a “Bleach gel” as described in [31].
14. There are many Next Generation Sequencing library preparation kits commercially available. The major considerations in selecting a kit include whether or not the kit produces “stranded” libraries, whether or not the kit selects for mRNA or other types of RNA, the size of the fragments produced, and the compatibility with sequencing technologies.
15. If the transcriptomic data is to be used for a de novo transcriptome assembly, read trimming is essential.
16. Transcript clustering algorithms that aim to reduce redundancy such as Cd-hit can be used to approximate “gene level” quantification [32]. It should be noted, however, that these programs rarely reduce transcriptome redundancy to the number of “true genes,” often resulting in several contigs per gene. This can be problematic for gene level quantification as it inflates gene copy number and will skew downstream results.
17. As of writing, there are two excellent tutorials for DESeq2 and edgeR which include detailed explanations and code examples: (<http://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html>; <https://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf>).
18. Log₂ normalization of the data can help reduce the distances/heights in PCA and hierarchical clustering analyses.

19. Any batch correction should be performed as conservatively as possible to avoid introducing inappropriate bias to the data. For example, batches should be assigned as categorical covariates and other phenotypic information including timepoint, treatment or condition should not be included.
20. The `ComBat()` function does not perform well with excessive missing or 0 values. These should be removed prior to batch correction.
21. topGO includes related GO terms for testing. As such, terms may be identified as significant that are not included in the GO mappings list. The “depth” of these related terms can be adjusted in the analysis. For more information please see: <http://geneontology.org/docs/ontology-relations/>.

References

1. Hand C, Uhlinger KR (1992) The culture, sexual and asexual reproduction, and growth of the sea anemone *Nematostella vectensis*. *Biol Bull* 182(2):169–176
2. Reitzel A, Burton P, Krone C, Finnerty J (2007) Comparison of developmental trajectories in the starlet sea anemone *Nematostella vectensis*: embryogenesis, regeneration, and two forms of asexual fission. *Invertebr Biol* 126(2):99–112
3. Passamanek YJ, Martindale MQ (2012) Cell proliferation is necessary for the regeneration of oral structures in the anthozoan cnidarian *Nematostella vectensis*. *BMC Dev Biol* 12(1):1–1
4. Bossert PE, Dunn MP, Thomsen GH (2013) A staging system for the regeneration of a polyp from the aboral physa of the anthozoan cnidarian *Nematostella vectensis*. *Dev Dyn* 242:1320–1331
5. Amiel AR, Johnston HT, Nedoncelle K, Warner JF, Ferreira S, Röttinger E (2015) Characterization of morphological and cellular events underlying oral regeneration in the sea anemone, *Nematostella vectensis*. *Int J Mol Sci* 16(12):28449–28471
6. Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A et al (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* 317(5834):86–94
7. Schwaiger M, Schonauer A, Rendeiro AF, Pribitzer C, Schauer A, Gilles AF et al (2014) Evolutionary conservation of the eumetazoan gene regulatory landscape. *Genome Res* 24(4):639–650
8. Rentzsch F, Fritzenwanker JH, Scholz CB, Technau U (2008) FGF signalling controls formation of the apical sensory organ in the cnidarian *Nematostella vectensis*. *Development* 135(10):1761–1769
9. Röttinger E, Dahlin P, Martindale MQ (2012) A framework for the establishment of a Cnidarian Gene Regulatory Network for “Endomesoderm” specification: the inputs of β -catenin/TCF signaling. *PLoS Genet* 8(12):e1003164
10. Layden MJ, Röttinger E, Wolenski FS, Gilmore TD, Martindale MQ (2013) Microinjection of mRNA or morpholinos for reverse genetic analysis in the starlet sea anemone, *Nematostella vectensis*. *Nat Protoc* 8(5):924–934
11. Servetnick MD, Steinworth B, Babonis LS, Simmons D, Salinas-Saavedra M, Martindale MQ (2017) Cas9-mediated excision of *Nematostella brachyury* disrupts endoderm development, pharynx formation and oral-aboral patterning. *Development* 144(16):2951–2960
12. Kraus Y, Aman A, Technau U, Genikhovich G (2016) Pre-bilaterian origin of the blastoporal axial organizer. *Nat Commun* 7:11694
13. Ikmi A, McKinney SA, Delventhal KM, Gibson MC (2014) TALEN and CRISPR/Cas9-mediated genome editing in the early-branching metazoan *Nematostella vectensis*. *Nat Commun* 5:5486
14. Fischer A, Smith J. *Nematostella* high-density RNAseq time-course. 2013
15. Tulin S, Aguiar D, Istrail S, Smith J (2013) A quantitative reference transcriptome for *Nematostella vectensis* early embryonic development: a pipeline for de novo assembly in emerging model systems. *EvoDevo* 4(1):16

16. Helm RR, Siebert S, Tulin S, Smith J, Dunn CW (2013) Characterization of differential transcript abundance through time during *Nematostella vectensis* development. *BMC Genomics* 14(1):266
17. Warner JF, Guerlais V, Amiel AR, Johnston H, Nedoncelle K, Röttinger E (2018) NvERTx: a gene expression database to compare embryogenesis and regeneration in the sea anemone *Nematostella vectensis*. *Development* 145(10):dev162867
18. Fredman D, Schwaiger M, Rentzsch F, Rentzsch F, Technau U (2013) *Nematostella vectensis* transcriptome and gene models v2.0. 1–1
19. Wingett SW, Andrews S (2018) FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res* 7:1338
20. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120
21. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBNetjournal* 17:10–12
22. Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 12(4):357–360
23. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S et al (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1):15–21
24. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357–359
25. Li H, Durbin R (2009) Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 25(14):1754–1760
26. Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30(7):923–930
27. Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12(1):323
28. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol BioMed Central* 15(12):550
29. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1):139–140
30. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28(6):882–883
31. Aranda PS, LaJoie DM, Jorcyk CL (2012) Bleach gel: a simple agarose gel for analyzing RNA quality. *Electrophoresis* 33(2):366–369
32. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22(13):1658–1659