Chapter 14

Preparing cDNA Libraries from Lytic Phage-Infected Cells for Whole Transcriptome Analysis by RNA-Seq

Bob Blasdel, Pieter-Jan Ceyssens, and Rob Lavigne

Abstract

Whole genome wide analysis of transcription using RNA-Seq methods is a powerful way to elucidate differential expression of gene features in bacteria across different conditions as well as for discovering previously exotic RNA species. Indeed, RNA sequencing has revolutionized the study of bacterial transcription with the diversity and quantity of small noncoding RNA elements that have been found and its ability to clearly define operons, promoters, and terminators. We discuss our experience with applying RNA sequencing technology to analyzing the lytic cycle, including extraction, processing, and a guide to the customized statistical analysis necessary for analyzing differential host and phage transcription.

Key words Bacteriophage, RNA-Seq, Library preparation, Transcriptome, RNA, Gene expression

1 Introduction

RNA sequencing (RNA-Seq), also known as Whole Transcriptome Shotgun Sequencing, is the use of second-generation platforms to sequence cDNA libraries that have been reverse transcribed from RNA populations present in target cells. When applied to phage-infected cells it allows for the identification of both phage and host encoded mRNAs, tRNAs, and sRNAs while quantifying them in relation to each other in a single experiment. RNA-Seq presents a number of significant advantages over microarray-based techniques, as it is not biased by hybridization efficiencies between oligonucleotides and allows the precise definition of RNA species to the single nucleotide level for both host and phage. It is also able to capture a faithful sample of the target population of RNAs across a much wider range of expression levels as it does not rely on direct detection methods like radioactivity or light, which can become oversaturated when enough material is used to detect low abundance transcripts (1).

As the number of both phage and bacterial genomes published in public databases continues to increase exponentially, our ability

Martha R.J. Clokie et al. (eds.), *Bacteriophages: Methods and Protocols, Volume 3*, Methods in Molecular Biology, vol. 1681, https://doi.org/10.1007/978-1-4939-7343-9_14,

to understand and annotate the gene features that give those genomes useful function has not kept pace. Published gene features have almost entirely been predicted in silico, based on the presence of open reading frames and often distant orthology to other often hypothetical features. By experimentally defining the shape and location of transcripts in both phage and host, directional RNA-Seq has the ability to discover novel coding sequences, particularly for small phage peptides falling below gene prediction thresholds (2), and refine annotations of existing coding sequences. Additionally, directional RNA-Seq allows detection of a plethora of noncoding RNA species. For example, it can define *cis* antisense encoded RNA, which has been described in N4-like phage (3), that are not possible to predict in silico and exist on conditionally bidirectionally transcribed regions and block translation or other functions of sense transcripts (4).

It is important to consider that, excluding the smallest types, phage typically progressively express multiple transcriptional schemes—changing expression over time to fit the temporally distinct needs of the phage. Where, classically according to the T4 model, phage will first transcribe genes involved in shutting down the host's self-defense capability while converting its metabolism toward viral production in an "early phase" of expression. Next, genes involved in genome replication and the production of structural proteins are transcribed in a "middle phase" before genes related to assembly, packaging, and lysis in a "late phase." When RNA-Seq is performed on a synchronously infected population of cells, each phase can be captured individually in separate samples and compared quantitatively.

With the biological replicates necessary to demonstrate statistical significance, RNA-Seq can also qualitatively evaluate differences in gene expression imposed on the host relative to phage-negative controls. Even as phage transcripts rapidly replace host RNA species, RNA-Seq will detect both host operons specifically targeted by the phage for modulation as well as the host mediated response to phage infection. Whether differential expression is mediated by the host or phage can be distinguished by performing RNA-Seq on multiple phage infecting the same host.

When assessing whether RNA-Seq of the phage lytic cycle is adaptable to a given phage-host model system, it is important to consider that it requires accurately sequenced genomes for both phage and host to align RNA-Seq reads to. Additionally, producing synchronously infected cultures requires the ability to generate high titers of phage that adsorb quickly relative to the timespan of infection. Moreover, developing an educated guess for when to take samples requires controlled infection parameters such as when the latent phase ends and when lysis occurs in the system being sampled from.

1.1 Design

Performing RNA Seq can be divided into three distinct parts, collecting nucleic acid samples from various phases of a synchronous infection (**Part A**, *see* Fig. 1), processing those samples into a collection of sequencing reads that are representative of the RNA population in the infected culture (**Part B**, *see* Fig. 1), and aligning those reads to both the host and phage genomes (**Part C**, *see* Fig. 1).

A. To collect data that is specific to the various phases of phage transcription, a synchronous infection must first be prepared. To do this, a culture of ~1 × 10^8 cells growing in the early exponential phase is infected at a high MOI under conditions that allow fewer than 5% of bacterial survivors to be remaining within 5 min (*see* **Note 1**). Then, at time points selected to represent early, middle, and late transcription, one third of the infected culture is removed and halted by rapid cooling in diluted phenol, which also temporarily stabilizes the RNA population. Generating statistically significant differential expression data requires that this be performed in triplicate to create biological replicates.

B. To process collected samples into cDNA libraries for sequencing, cells are first lysed and RNases present in the cells and media are inactivated to produce a stable suspension of nucleic acids (**step 1**). Then all genomic DNA from both the phage and host must be enzymatically removed from the suspension (**step 2**). Optionally, rRNA may then be removed from the sample with commercially available kits to better economize on available sequencing depth (**step 3**). The RNA population is then reverse transcribed using commercially available kits into a cDNA library that can be shotgun sequenced (**step 4**).

C. The obtained sequencing reads for each sample must then be processed to remove adaptors and low quality reads before they can be aligned to the genomes of both phage and host using either open source programs or commercially available pipeline software. Once aligned to phage genomes, the distribution of reads can be used to correct gene annotations, define operons and upstream untranslated regions, as well as discover new gene features such as sRNA and small peptides that fall below ordinary gene prediction thresholds in size. Additionally, with replicates, the read counts that align to annotated gene features can be compared between samples to statistically test for differential expression.

This chapter is primarily focused on **Part B**, but discusses aspects of **Part C**. The methods required for **Part A** are described in detail by Kropinski in Chapter 2.

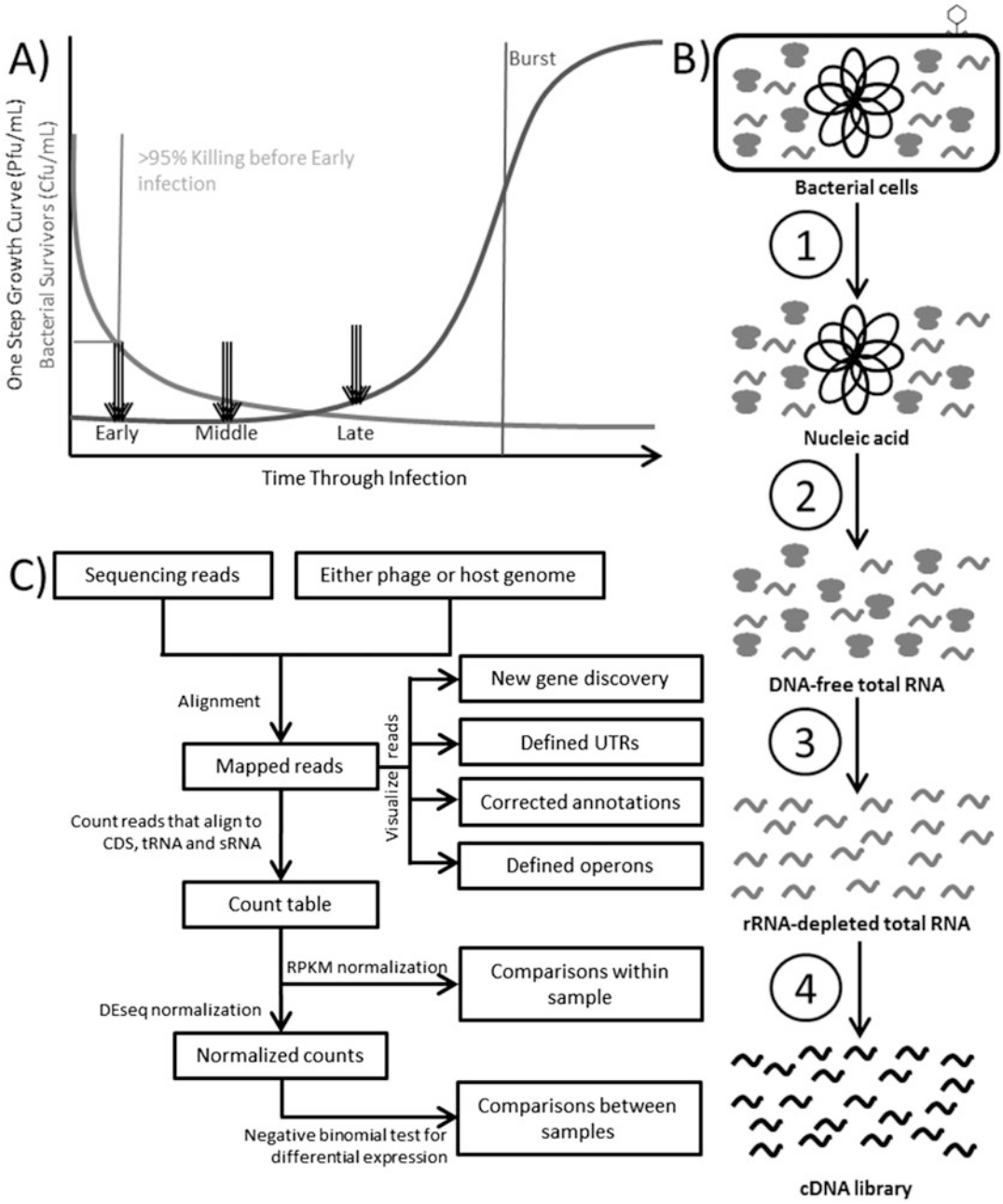

Fig. 1 Workflow for RNA-Seq analysis of cells infected by lytic phage. Biologically relevant samples are first collected in triplicate from a phage-negative control as well as time points in a synchronous infection (**Part A**). The samples must then be processed independently to liberate nucleic acids, remove both phage and host genomic DNA, deplete rRNA, and convert the remaining RNA into cDNA libraries for sequencing (**Part B**). Once sequencing is complete, the resulting reads must be aligned to their relevant reference genomes where they can be visualized to show the shape of the transcriptome. Additionally, they can be counted to make statistical comparisons between the abundance of reads that align to different gene features within a sample or between samples (**Part C**)

2 Materials

TRIzol® (Life Technologies)

Chloroform.RNase-free ethanol.RNase-free water.RNase-free 3 M NaOAc pH 5.2.RNase-free DNase.RNase-free disposables such as pipette tips and microcentrifuge tubes.Titers of phage in excess of 1×10^{11}/ml.Stop Solution: One part RNA buffered phenol to nine parts absolute ethanol by volume, kept ice cold. Lysis Buffer: Solution of lysozyme prepared according to manufacturer's instructions to 4 mg/ml.

3 Methods (Part B)

3.1 Organic Extraction of RNA (Step 1)

1. Before infection, prepare one centrifuge tube large enough to contain 1/3 of the infection per time point to be taken, each with one part Stop Solution for every nine parts of cell suspension that RNA is being extracted from and place on ice.
2. Over the course of infection, remove samples of ~2.5×10^7 cells at the desired time points and pipette them into one tenth volume of prepared Stop Solution before immediately shaking vigorously and placing back on ice.
3. After infection, centrifuge at 5000 × *g* for 15 min, to securely pellet the stopped cells and remove the supernatant.
4. Resuspend pellet in 400 μl of Lysis Buffer before transferring to a 1.5 ml microcentrifuge tube.
5. Incubate for 10 min, but not longer, at room temperature before freezing with liquid nitrogen and thawing in a water bath at 45 °C. Repeat the freeze–thaw cycle three times and look under microscope to confirm cell lysis (*see* **Note 2**).
6. Add 500 μl TRIzol® to pellet, thoroughly pipette mix, and incubate for 10 min at room temperature (*see* **Note 3**).
7. Add 200 μl chloroform and mix before incubating for 10 min at room temperature.
8. Centrifuge ~16,000 × *g* for 15 min 4 °C.
9. Carefully transfer aqueous upper phase into new tube without disturbing the organic phase or the protein layer found at the interphase.
10. Add 1/10 volume RNase-free 3 M NaOAc and 2 volumes 96% EtOH, splitting the sample into multiple tubes.
11. Store at −20 °C overnight to ensure precipitation of small RNA species and centrifuge ~16,000 × *g* for 1 h at 4 °C.
12. Remove supernatant and wash pellet with ice cold 70% EtOH before centrifuging again ~16,000 × *g* for 15 min at 4 °C.

13. Remove supernatant, centrifuge again for 1 min to, remove remaining supernatant and air-dry pellet for 5 min.
14. Resuspend the pellets for each sample in 200 μl total RNase-free water and combine into a single tube.
15. Analyze sample on NanoDrop spectrophotometer (Thermo Scientific, Wilmington, DE) to ensure adequate concentration and purity: $OD_{260/280\ \&\ 260/230} > 1.8$ (*see* **Note 4**).
16. Store at −20 °C.

3.2 Removal of Genomic DNA (Step 2)

Complete removal of both phage and host gDNA is essential to obtain accurate sequence information, as cDNA and gDNA reads will be indistinguishable. Eliminating genomic DNA contamination can be challenging, as the high concentrations of RNA present will act as a competitive inhibitor to commercially available DNase enzymes, impeding their function. It is also important to consider that the DNase used may be sensitive to noncanonical nucleotides commonly present in phage DNA. Using standard DNase according to manufacturers' instructions may work, though below is an expanded protocol that optimizes enzyme function, and even this may need to be repeated several times.

1. Add RNase-free DNase buffer to 1× concentration. Incubate for 5 min at 65 °C to ensure remnant DNA is fully in solution (*see* **Note 5**).
2. Return to room temperature before adding the recommended amount of RNase-free DNase and incubating for 1 h at 37 °C.
3. Add the same amount of DNase a second time and incubate for another 1 h at 37 °C.
4. Analyze each sample for residual phage and host DNA by performing PCR using primers that amplify small products and have been verified to be sensitive to low concentrations.
5. Store at −20 °C.

3.3 rRNA Depletion (Step 3—Optional)

Depending on the sequencing resources available, it may be desirable to use commercially available rRNA depletion kits for bacterial total RNA to increase the depth of coverage for desired RNA species (Table 1). We have had variable success using the

Ribo-Zero kit available from Illumina (San Diego, California), which captures rRNA using oligo hybridization to beads that are then removed with a strong magnet. Although commercially available kits are typically regarded as less reliable than advertised, even a reduction of the rRNA fraction from ~95% to ~50% of the sample can result in enrichment of the output for other RNA species by an order of magnitude.

Table 1
Commercially available rRNA removal methods appropriate for bacteria

Name	Supplier	Catalog Number
Ribo-Zero™ rRNA Removal Kit (Bacteria)	Illumina	#MRZMB126
MICROBExpress™ Bacterial mRNA Enrichment Kit	Life Technologies (Thermo Fisher Scientific)	#AM1905
Terminator™ Exonuclease (*see* **Note 6**)	Epicenter Biotechnologies	#TER51020

3.4 cDNA Library Preparation and Sequencing (Step 4)

DNA and rRNA depleted RNA is typically transformed into double stranded cDNA libraries through a process that uses random hexamer primed reverse transcription, followed by synthesis of a second strand. Through this process, both strands of cDNA are then sequenced identically in a way that scrambles the natural strand specificity inherent to transcription. However, particularly with the extraordinary coding density of phage genomes the various strand specific methods that have been devised for cDNA library preparation have special value for understanding the transcriptomes of phage (5). Indeed, there have been significant amounts of antisense RNA that have been characterized (3) that would be impossible to distinguish with un-stranded RNA-Seq, and transcript features often overlap in ways that strand specificity can aid in defining appropriately.

While there are many established techniques for accomplishing strand specificity in RNA-Seq a comprehensive comparative analysis of strand-specific RNA sequencing methods has convincingly argued that the Illumina RNA ligation methods (6) and the dUTP second strand marking methods (7) provide better results for the effort expended (8), *see* (5) for additional discussion. We have had success with Illumina's TruSeq® Stranded Total RNA Sample Prep Kit, which uses a method similar to the dUTP second strand marking method (Catalog #: RS-122-2201). However, once a cDNA library is generated it can be sequenced using standard high throughput platforms to generate the list of millions of short reads that will be used in the next section.

4 Experimental Analysis (Part C)

4.1 Mapping Reads

The first step in making sense of the millions of short reads generated by RNA-Seq is to turn those reads into a quantification of localized transcript abundance by aligning them to either the phage or host genomes. This involves attempting to match each read to a corresponding sequence in each potential reference genome, a process that is complicated by short reads aligning to multiple

locations, RNA-Seq sequencing errors, reference genome sequencing mistakes, and RNA editing events. Current protocols use either the Burrows Wheeler transform or hash table based methods to assemble a list of candidate matches available in the reference for each read and then pick between them. Alignment can be performed using various free and open-source software packages such as the Burrows Wheeler Aligner (9) or TopHat (10).

4.2 Generating Transcription Maps

Once aligned to both phage and host genomes, the reads form a map revealing the abundance of RNA transcribed from any given locus in the infected cell accurate to the single nucleotide level. These maps can be used to precisely determine transcription start and end sites allowing promoters and terminators to be predicted and their operons to be characterized. With defined operons, 5′ upstream untranslated regions can be annotated and their effects on translation hypothesized. Additionally, unannotated yet transcribed regions can be scrutinized for peptides that are too small to definitively predict from sequence alone, indeed using RNA-Seq Ceyssens et al. (2) updated the ΦKZ genome with 63 (20.5%) additional coding sequences. Noncoding sRNAs will also be highlighted as transcribed features without plausible open reading frames while both *cis* and *trans*-encoded antisense RNAs will map to the antisense strand of coding sequences. These maps can also point out faulty annotations, when previously defined open reading frames lack sense transcripts or the start of transcription indicates a different start codon.

4.3 Differential Expression Analysis

Differential analysis of the number of sequencing reads that align to specifically annotated host gene features between an uninfected control, sampled immediately before infection, and various time points after infection also provides a valuable window into how phage infection affects host transcript abundance. This is accomplished by summarizing expression data into a table of the number of reads that map to each host CDS, ncRNA, and tRNA with three biological replicates before infection and comparing them statistically to three biological replicates after infection. To perform this differential analysis we recommend using DESeq as a R/Bioconductor package to normalize read counts between samples and then to test for differential expression and thus infer signal within the noise inherent to RNA-Seq. DESeq uses a method based on the negative binomial distribution to model the differences that would be expected due to natural variation and thus determine if an observed difference in read counts is statistically significant. This is more appropriate than other methods based on the Poisson distribution for modeling the variance inherent to phage infection (11).

While the alignment of sequencing reads to either host or phage has remained clear and distinct in our experience,

determining whether the host or the phage is causing observed changes in the host transcript abundance during phage infection can become muddied. Differential analysis highlights changes in the abundance of specific transcripts imposed on the host by the phage such as the promotion and repression of particular transcripts or targeted degradation. However, depending the presence or success of phage mechanisms for shutting down host systems, it will also highlight host responses to phage infection for defense or as a reaction to various stresses that are inherent to phage infection. The difference between the two can be potentially distinguished by context and other sources of data, but can also be highlighted by performing RNA-Seq on infections by several diverse phages in the same host. As taxonomically divergent phages are unlikely to affect even a common host in the same way, a similar transcriptomic response to many phages will indicate that it is performed as a host response.

When interpreting your results it is important to consider that, aside from dramatic examples such as those produced by prophages in the host sensing infection and attempting to escape (2), most host transcripts will downshift in abundance relative to the total RNA in the cell during a successful infection due to the rapid synthesis of phage transcripts. Specific modulation of host transcripts needs to be tested for independently of this global depletion, which is done when normalizing only host read counts in one sample to host read counts in another sample while excluding phage reads. This will faithfully highlight how the distribution of reads transcribed from the host genome changes, but will not on its own show changes in abundance relative to the total transcript population in the cell as it hides the natural relative decrease in host reads.

5 Notes

1. If phage binding efficiency proves inadequate, the addition of 1–20 mM $CaCl_2$ and/or $MgCl_2$ to the medium may be needed to assist the phage (12).
2. If this proves inadequate to lyse the host, additional preferred methods can supplement or replace incubation with lysozyme such as beating with microbeads. It is important to ensure that the time that the sample spends at room temperature before **step 6** is kept to an absolute minimum.
3. The RNA samples suspended in TRIzol® at **step 6** can be safely frozen at −20 °C for up to 3 months. It is important to note that this stage is intended to inactivate the RNases made by the host and present in the media. From this point on, all materials that interact with the sample after this point must be RNase-

free, gloves must be worn, and the samples should be kept on ice when worked with on the bench.

4. A poor $OD_{260/280}$ ratio indicates an unacceptable level of either phenol contamination, from the phenol contained in the TRIzol®, or protein contamination, from the white interphase layer in **step 9**, relative to the RNA concentration. It can be addressed by starting again from **step 6**. A poor 260/230 concentration indicates an unacceptable level of salt contamination, from the NaOAc used in **step 10** as well as the media, relative to the RNA concentration. It can be addressed by starting again from **step 10**.
5. The 2′-OH group of RNA is capable of catalyzing autocleavage of RNA strands at high temperatures and high pH.
6. Epicenter Bioscience's Terminator™ 5′-Phosphate-Dependent Exonuclease is an inexpensive and especially effective way to deplete rRNA, which are posttranscriptionally modified with a 5′-monophosphate, but will also remove any other RNA species that could have been similarly modified.

References

1. Oshlack A, Robinson MD, Young MD (2010) From RNA-Seq reads to differential expression results. Genome Biol 11:220. doi:10.1186/gb-2010-11-12-220
2. Ceyssens P, Minakhin L, Van den Bossche A, Yakunina M, Klimuk E, Blasdel B, De Smet J, Noben J, Bläsi U, Severinov K, Lavigne R (2014) Development of giant bacteriophage ΦKZ is independent of the host transcription apparatus. J Virol 88(18):10501–10510
3. Wagemans J, Blasdel B, Van den Bossche A, Uytterhoeven B, De Smet J, Paeshuyse J, Cenens W, Aertsen A, Uetz P, Delattre A, Ceyssens P, Lavigne R (2014) Functional elucidation of antibacterial phage ORFans targeting Pseudomonas aeruginosa. Cell Microbiol 16(12):1822–1835
4. Georg J, Hess WR (2011) cis-antisense RNA, another level of gene regulation in bacteria. Microbiol Mol Biol Rev 75(2):286–300
5. Mills JD, Kawahara Y, Janitz M (2013) Strand-specific RNA-Seq provides greater resolution of transcriptome profiling. Curr Genomics 14 (3):173–181
6. Croucher NJ, Fookes MC, Perkins TT, Turner DJ, Marguerat SB, Keane T, Quail MA, He M, Assefa S, Bähler J, Kingsley RA, Parkhill J, Bentley SD, Dougan G, Thomson NR (2009) A simple method for directional transcriptome sequencing using Illumina technology. Nucleic Acids Res 37(22):e148
7. Zhang Z, Theurkauf WE, Weng Z, Zamore PD (2012) Strand-specific libraries for high throughput RNA sequencing (RNA-Seq) prepared without poly(A) selection. Silence 3:9. doi:10.1186/1758-907X-3-9
8. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods 7(9):709–715
9. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. Bioinformatics 1(5):589–595
10. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-Seq experiments with TopHat and Cufflinks. Nat Protoc 1(3):562–578
11. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11:10. doi:10.1186/gb-2010-11-10-r106
12. Rountree PM (1951) The role of certain electrolytes in the adsorption of staphylococcal bacteriophages. Microbiology 5(4):673–680