Evaluating differential gene expression using RNA-Sequencing data: a case study in host-pathogen interaction upon *Listeria monocytogenes* infection

Joana Rita Gonçalves da Cruz
joana.cruz@ist.utl.pt
Dep. of Bioengineering, Instituto Superior Técnico, Lisbon, Portugal

November 2013

Abstract

Motivation: Understanding the cell transcriptome is essential for interpret the functional elements of the genome. Recent developments of high-throughput DNA sequencing technologies have provided a new method to sequence RNA at unprecedented high resolutions. This method is termed RNA-Seq and has been emerging as the preferred technology for characterise the cell transcripts.

Results: We propose a pipeline to compare two RNA-Seq samples. This pipeline permits to obtain biological insight about the analysed samples by extracting the main cellular processes that are differentially active on both samples. Additionally we propose a novel methodology to inspect the activation of a given cellular pathway in a time-course RNA-Seq dataset. Finally, we hypothesise a novel approach to statistically reinforce the inference of differentially expressed genes among two RNA-Seq samples, using publicly available RNA-Seq datasets.

Conclusions: The evaluation of a *Listeria monocytogenes* RNA-Seq dataset with the developed tools testified its proper functioning. Employing the RNA-Seq analysis pipeline, it was concluded which cell processes were differentially active between a set of trancriptomes acquired from cells with different growth environments and along a time-course. In the same way, taking advantage of published relationships between genes upon infection with *Listeria monocytogenes*, it was confirmed the existence of these connections on the *Listeria monocytogenes* dataset. Finally, it was tested if public RNA-Seq data obtained in similar conditions as the control of the *Listeria monocytogenes* dataset could be used to improve the statistical confidence on the inference of differentially expressed genes. The results prove that this methodology is not reliable.

Keywords: Next generation sequencing, RNA-Sequencing, gene expression, gene networks, RNA-Seq analysis pipeline, RNA-Seq validation, *Listeria monocytogenes*.

1 Introduction

Since the chemistry Nobel prize was awarded to Fred Sanger and Walter Gilbert, in 1980, for their crucial contribution concerning the determination of base sequences in nucleic acids,[1,2] many were the developments of DNA sequencing technologies.[3] The improvement of these techniques brought a revolutionized approach to biological questions and, nowadays, the use of these technologies in order to access the cell transcriptome has become an integral part of biological research – RNA-Sequencing (RNA-Seq) .

Following up these improvements, in the bioinformatics field, a wide number of methods have been developed to cope with different aspects of RNA-Seq data analysis. However, combining them into a congruent analysis pipeline is not a trivial challenge, in the sense that their aggregation must address the needs and specificities of each problem. With this in mind, we propose a sequence of tools (pipeline) which is suitable to perform a comparative study between

two RNA-Seq samples (Section 2). As a case study, we used the implemented pipeline to process a RNA-Seq dataset acquired from human HeLa cells infected with *Listeria monocytogenes* (*L. monocytogenes*) (Section 2.2).

The available tools which infer differential signal among RNA-Seq samples use statistical tests to decide whether, for a given gene, an observed difference in read counts is significant. However, these methods are not able to aggregate the data along a time-course, only do individual analysis between two time-points. To address this limitation, we formulated a new methodology that uses known gene regulatory networks to investigate if a certain biological process is active on a RNA-Seq time-course dataset (Section 3). To test the proposed methodology, we studied how well a gene network describing the cell response upon *L. monocytogenes* infection is modelling the *L. monocytogenes* RNA-Seq dataset (Section 3.2).

Finally, we tested the hypothesis of using public RNA-Seq data, extracted from cell populations in similar conditions as the control of the *Listeria monocytogenes* dataset, to improve the statistical confidence on the inference of differentially expressed genes (Section 4).

2 RNA-Seq analysis pipeline

In this Section we describe a pipeline developed to analyse a time-course RNA-Seq dataset. Next, we use this pipeline to process a RNA-Seq dataset acquired from human HeLa cells infected with *L. monocytogenes*. Finally, we discuss the limitations of this pipeline.

2.1 Methods

Data analysis begins with the input of the raw read files and the reference files. Once this data is gathered, the reads are processed with FastQC.[4] This software ensures that the RNA-Seq data is qualitatively good and that there are no biases in it. Problems detected by a quality control analysis may derive from the sequencer or from the starting library material. Some abnormalities may be resolved by trimming base pairs from the raw read. The pipeline supports this, containing a script that is able to trim a given number of base pairs from the RNA-Seq data.

Afterwards, the clean reads will be aligned with a reference sequence using Bowtie 2, a well established mapper.[5] This is a fast and memory-efficient mapping tool that is particularly suitable for the alignment of small reads, as the ones from RNA-Seq technology, to long reference sequences, as the human genome. The output is a Sequence Alignment/Map (SAM) file which stores the informations about the read alignments against the reference sequence.

After the RNA-Seq reads are mapped to a reference, the number of reads that map a certain gene will be measured by a script integrated in the HTSeq Python package, named HTSeq-count.[6] Due to the nature of this study, where high-level pathway analysis was the goal, it was not considered relevant to distinguish multiple isoforms of the same gene and, thus, features in this analysis are equivalent to genes. Moreover, the pipeline has a conservative approach to reads that map to overlapping genes, discarding them.

Once gene expression has been quantified, a differential expression (DE) test is performed between the RNA-Seq samples. This test has as main goal the detection of differentially expressed genes among the conditions in study. The inference of differential signal in such data is done using DESeq,[7] an R/Bioconductor package. This software uses a negative binomial distribution to model the gene expression distribution. To select the most significant entries in this analysis, the pipeline is pre-set to trim genes with *p-value* higher than 0.1.

Lastly, the differentially expressed genes are associated with the GO terms using a Bioconductor package called GOStats.[8] This software uses a standard hypergeometric test in order to relate a given gene list with the controlled vocabulary in the GO database. The GO project provides a controlled vocabulary of terms in an effort for consistent gene product descriptions in different databases. For the entries generated through the hypergeometric test, it is defined a *p-value* cut-off of 0.1. Thus, in this final step, it is possible to have a biological insight about the samples being compared. Specifically, from the obtained table one can conclude about the most differentially active processes when the cell is subjected to distinct growth circumstances.

2.2 *Listeria monocytogenes* case study

The pipeline described above was used to process a RNA-Seq dataset composed by three populations: *Control*, HeLa cells not infected growing in

a healthy medium; *LM1*, HeLa cells infected with wild-type *L. monocytogenes* strain EGDe; *LM2*, HeLa cells infected with mutant *L. monocytogenes* strain EGDe, to which was removed *hly* gene that encodes for listeriolysin O (LLO), a *L. monocytogenes* virulence factor.[9] Total RNA was extracted from the cells of each population at four time-points (20, 60, 120 and 240 minutes) with the purpose of having represented specific stages in the bacterium lifecycle. Extracted RNA was sequenced using Illumina platform. From this procedure resulted a paired-ended dataset in which each DNA fragment is constituted by 90 base pairs.

A brief summary of the statistically significant biological processes GO terms that differ between non-infected and *L. monocytogenes* infected cells is represented in table 1. Analysis of these terms shows that at an early stage (time-point 20) the cell is reacting to the binding of an extracellular ligand to a receptor on its surface. This evidences the binding of the *L. monocytogenes* receptors to the host cell surface proteins. The main differences between control *versus* LM1 and control *versus* LM2 analysis arise in time-points 60 and 120, with opposite active processes. Particularly, for the first comparison, the terms evidence that the cell is already responding to the bacterial invasion by reducing the frequency of its cellular and biological processes and promoting apoptosis. The process of cell suicide works as a natural mechanism to prevent the dissemination of infection to the healthy neighbour cells, functioning as a host defense against the pathogen invasion.[10] Contrary, for the second analysis, the terms evidence the cell proliferation. Finally, after 240 minutes, the most noteworthy terms for the analysis control *versus* LM1 is related with cell communication, which could be explained by the bacterium invasion of neighbouring cells. For control *versus* LM2 the population of cells continues its proliferation.

The results evidence that for the first case, the cells population responds to the *L. monocytogenes* infection. Contrasty, in the second case the cells did not respond to the presence of the pathogen and continue their proliferation processes along the four acquisition points. This may be explained by the hypothesis that for the last case (LM2 samples) the bacteria could not be free in the host cytosol. Due to the non-existence of *hly* gene in the mutant bacteria, which synthesizes an important toxin for the bacteria to escape its internalization vacuole, *L. monocytogenes* is incapable of disrupting its phagosome and, therefore, to use the host cell machinery. Bearing in mind these results, we formulated the following hypothesis: when LLO, a protein codified by *hly* gene, is not produced *L. monocytogenes* loses its virulence. Particularly, without LLO the bacteria is not able to disrupt the internalization vacuole and use the cell machinery to replicate. These conclusions are congruent with what Portnoy *et al.*, 2002,[11] describe. Moreover, due to the fact that the internalization process is performed in a membrane-bound phagosome by inducing local cytoskeletal rearrangements in the host cell,[9] with no disruption of this vacuole the cell cannot detect any foreign body.

2.3 Discussion

The tools integrated in the developed pipeline are just one alternative to perform a given task (table 2). Concerning the sequencing quality assessment, the methodology used for alternative applications is very similar to the one used by FastQC. However, FastQC output has significant added value for its clearness and simplicity combined with the display of all important information. This allows the user to easily conclude about the data sequencing quality.

Table 1: Main GO terms for the Biological processes ontology along the four acquisition time-points for control *versus* LM1 and Control *versus* LM2 analysis.

Time-point	**Control *vs.* LM 1**		**Control *vs.* LM2**	
	Definition	*P-value*	Definition	*P-value*
20	Enzyme linked receptor protein signalling pathway	2.06e-07	Enzyme linked receptor protein signaling pathway	0.000112
60	Programmed cell death	1,05e-11	Regulation of endothelial cell proliferation	1.96e-06
120	Negative regulation of cellular process	2.11e-20	Positive regulation of cell proliferation	4.62e-06
240	Regulation of cell communication	5.05e-11	Regulation of cell proliferation	8,31e-06

With respect to the alignment tool, Bowtie 2 is a state-of-the-art mapper that is specialized at aligning short reads (from 50-100bp) with long reference genomes.[5] Relatively to Bowtie 2 alignment quality, as stated in Friedel *et al.*, 2012,[12] "Bowtie 2 alignment algorithm shows a remarkable tolerance to sequencing errors (...) making hash-based aligners obsolete" (like MAQ or SOAP2). When comparing Bowtie 2 algorithm with BWA or the old Bowtie, results confirm that Bowtie 2 has a higher ratio between the number of correct alignments and incorrect ones.[5]

Regarding gene expression profiling, HTSeq-count is a simple script that takes advantage of the tools that the python package HTSeq contains. The output of this tool is ideal for the next pipeline step.

In order to perform the differential expression test, it was intended a method that, firstly, had a conservative approach with samples that contain a high number of outliers and, secondly, that was capable of performing the comparative analysis without replicated reads. Bearing in mind the considerations stated in Soneson *et al.*, 2013,[13] DESeq, edgeR and NBPSeq are the best methods to process datasets with low number of replicate samples, as the *L. monocytogenes* case study. Among these methods, Soneson *et al.* concluded that they have similar results, with DESeq having the most conservative behaviour.

At last, to determine which pathways are overrepresented in the set of differentially expressed genes it is pretended to perform a over-representation analysis, following the classification stated in Khatri *et al.*, 2012.[14] As referred in Emmert-Streib *et al.*, 2011,[15] GOStats is the principal method to perform this sort of analysis and, hence, it was chosen to integrate the developed pipeline.

Table 2: Alternative tools for each pipeline step.

Used tool	Alternative tools
FastQC	RSeqQC, htSeqTools HTQC, QC-Chain
Bowtie2	BWA, SOAP2, MAQ, Bowtie
HTSeq-count	bedtools multicov
DESeq	edgeR, TSPM, baySeq, NOISeq
GOStats	Onto-Express, GenMAPP, GoMiner, GO:TermFinder

Moreover, there are also available pre-build pipelines which take the raw RNA-Seq reads and process it to extract biological conclusions. Processing RNA-Seq data using already implemented pipelines can be advantageous to, for instance, biologists with poor computational knowledge. However, to users with computational knowledge, analysing the data with such tools restricts the way in which the processing can be performed and, moreover, is frequently harder and less effective. Pre-build pipelines may not contain the tool which best suits the data in analysis and are not so versatile. Building a pipeline from scratch means that it is designed for the dataset in analysis and, furthermore, even if the dataset changes it easy to adapt the same pipeline to perform its analysis.

Nevertheless, the developed tool has also some limitations: 1) Bowtie 2 does not perform spliced alignment. Nevertheless, permits gapped alignments; 2) if a read maps to multiple places on the genome with equal score, Bowtie 2 chooses randomly the genome portion where that read is going to be mapped; 3) Bowtie 2 mapping process does not assures that the alignment reported is the best possible in terms of alignment score; 4) the pipeline has a conservative approach to reads which only one mate of the paired-ended read was aligned to the reference, discarding those alignments; 5) reads that map to overlapping genes are also discarded; 6) the developed pipeline is only suitable to process data from Illumina's sequencing protocol.

In addition, the experimental data can also decrease the confidence of the results. Particularly for the *L. monocytogenes* case study, the data is unreplicated. As Fisher[16] noted, without an estimate of variability there is no basis for inference (between treatment groups). Although we can test for differential expression between treatment groups from unreplicated data, the results of the analysis only apply to the specific subjects included in the study (i.e., the results cannot be generalized). Furthermore, the RNA-Seq data is extracted from a pool of infected and non-infected cells. The *L. monocytogenes* infection was done on a population of cells. Some cells get infected and some do not. The transcriptional responses between these two types of cells may vary considerably. The result is that, at the level of the population (which is what RNA-Seq translates), the gene may not translate the response of infected cells. This phenomena is called bystander effect.[17]

3 Gene networks to prove the existence of a given biological response in RNA-Sequencing data

Understanding the complex genes interactions implicated in the cell response to an external stimulus is extremely important to contextualize and validate the results from the pipeline referred in Section 2.

Statistical methods such as the ones provided by DESeq and GOStats packages are employed in the developed pipeline to extract from the RNA-Seq samples the cell differential active processes. However, these methods are not able to aggregate the data along a time-course, only do individual analysis for each acquired time-point. In other words, they are not capable to integrate the data in a systems biology view. Systems biology does not investigate individual genes and rather focus on the behaviour and relationships of all the elements in a particular biological system while is functioning, along a time course. To address this constraint, we developed an algorithm that is capable of inspecting the activation of a given biologic pathway on a RNA-Seq dataset.

3.1 Methods

This methodology is based on previously known gene relationships on a given cellular pathway. This information is usually concatenated in gene regulatory networks. Gene regulatory networks are graphic diagrams that are used to visualize the regulatory relationships between genes upon a certain stimulus or just along the cell life-cycle. These networks are comprised of *nodes*, the genes and their regulators, joined together by *edges*, which represent physical and/or regulatory interactions.

To inspect if a given gene network is modelling well a RNA-Seq dataset in analysis, the developed methodology assumes that this gene network is translating the relations that best modulate the genes interactions of that dataset. Therefore, the number of genes, as well as, the interactions between the genes are considered as described in the literature and, thus, are known *a priori*.

The formulated methodology needs to know the following data *a priori*: the count table containing the genes expression level (output of HTSeq-count tool referred in Section 2.1), the experimental metadata (such as the library type and the time-points where the reads were acquired) and, finally, the network of interest which describes how a certain set of genes are related in the biological process in study. This set of genes corresponds to the genes in the network nodes and will be referred in this document as node genes (NG).

The developed algorithm begins by extracting the NG expression level information from the count table. This data is, then, normalized in two steps: the first normalization is performed using DESeq's approach, which is considered a robust method.[18] Then, this gene expression level is subtracted by the normalized genes expression of the control sample and transformed to a logarithmic scale. After this pre-processing, the kinetics of these genes is determined taking into account the expression level information. In order to do so, and following the model proposed in several surveys, the gene regulatory network is modelled by a system of linear differential equations, which general mathematical form is described by equation 1.[19–21]

$$\frac{dx_i(t)}{dt} = \sum_{j=1}^{C} w_{i,j} \, .x_j(t) + b_i.u(t). \qquad (1)$$

where

- $x_i(t)$: Log-ratio of the gene i expression level at time t;
- $w_{i,j}$: Entry of the gene-gene interaction matrix in column i and row j;
- b_i : External stimulus response vector for gene i;
- $u(t)$: Heaviside step function: $u(t < 0) = 0$ and $u(t \geq 0) = 1$.

The column vector b_i numerically translates the influence that an external stimulus, such as infection, has in the gene *i*. The Heaviside step function represents the influence of the external stimulus as constant over time. Each entry of the gene-gene interaction matrix W describes the influence that gene i has on gene *j*. Knowing the network *a priori*, it is possible to determine the number of entries that are not null as well as its sign in the interaction matrix W. Therefore, the linear differential equations can be solved in order to $w_{i,j}$ by the least square method. For that, an overdetermined system of equations is constructed.

This system has into account the normalized expression values and the time derivatives for the node gene i, both along the several acquisition time-points. The time derivatives values are estimated from the experimental data points by linear interpolation as it is done in Guthke *et. al*, 2005.[19] This system is assumed to be at equilibrium prior to stimulation, i.e. $dx_i(t < 0)/dt = x_i(t < 0) = 0$. Moreover, in order to solve that system, the expression level of a certain gene i needs to be not null. Otherwise the differential equation will be impossible to solve. The approach chosen to cope with this situation was to trim the non-expressed genes from the analysis. After solving this system, the linear differential equations that modulate the variation of the NG along the time are completely defined. By numerical integrating those differential equations the algorithm finds the kinetics associated with the NG along the acquisition time-points.

At this point, aiming to test the statistical significance of the network of interest in the data in analysis, the algorithm performs a permutation test. In order to do so, the algorithm chooses, randomly, from a trimmed gene universe a set of n genes, where n is the number of genes in the nodes of the known network. It is important to refer that this trimmed gene universe corresponds to the set of genes on the raw count table that do not have a null expression level for any of the acquisition time-points. Afterwards, the ratio between the mean square error calculated from the simulated and the experimental expression values and the variance of the experimental values along the time (MSE/Var) is measured (equation 2). This procedure is repeated t times (with t greater than 1000) and the statistical values measured compared with the same value for the genes in the network nodes (NG). The *p-value* associated with the existence of the network's gene interactions in the RNA-Seq dataset is calculated by measuring the number of random sets of genes that have lower MSE/Var than the NG.

$$\frac{MSE}{Var}(gene_i) = \frac{\sum_{i=1}^{n_{tp}}(\hat{x_i} - x_i)^2}{\sum_{i=1}^{n_{tp}}(x_i - \mu)^2}. \qquad (2)$$

where

n_{tp} : Number of acquisition time-points;

x_i : Vector with experimental values of gene i expression level in the acquisition points;

$\hat{x_i}$: Vector with predictions of gene i expression level in the acquisition time-points;

μ : Mean value of the x_i vector.

Statistical value MSE/Var measures how far the calculated kinetics is from the experimental values. For clusters of genes with low variance, the measurement of the absolute MSE will be low even if they are being described by a bad model. Therefore, the MSE value is normalized by the variance to get a fraction of explained variance. In this way, we are able to measure properly how well a set of genes are being modelled by the gene network of interest, independently if that set of genes has a high or a low variance associated.

In addition, we designed a HTML interface where it is easy for users with no computational know-how to gather the necessary data and proceed analysis, without the need of using the command line.

3.2 *Listeria monocytogenes* case study

Considering that regulatory networks controlling gene expression function as decision-making circuits within the cell, it is intended with the subsequent analysis to understand which pathways of the cell circuits are activated when the cell is infected by *L. monocytogenes*. Given the characteristics of the case study in analysis, the gene regulatory network extracted from published data describes the cell immunological response during *L. monocytogenes* infection (figure 1).

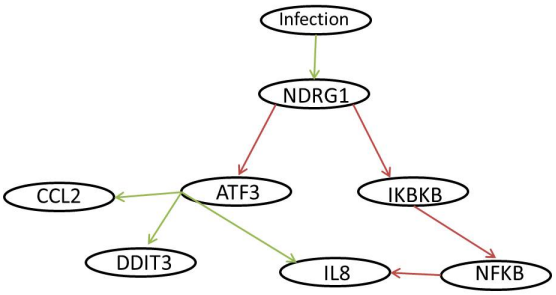

Figure 1: Network constructed from literature review on *L. monocytogenes* infection. The arrows in green represent stimuli and the arrows in red represent inhibitions. This network was kindly provided by Loretta Magagula from Mhlanga laboratory at CSIR, in South Africa.

Firstly, the developed algorithm determined how these NG kinetics varies in the *L. monocytogenes* dataset. This simulation is represented, for both LM1 and LM2 samples, in figures 2(a) and 2(b) according to equations 3 and 4, respectively. Next, the permutation test was performed, with $t = 5000$. For

this specific case, the result evidences that the NG conjunction is well suited for the *L. monocytogenes* dataset with a considerable low *p-value*. Particularly, for LM1 sample only three set of randomly chosen genes had lower ratio between the mean square error and the variance (MSE/Var) and, hence, the *p-value* is 0.059%. Concerning LM2 sample, the existence of the network of interest in the data has associated a *p-value* of 10.137%. The differences between the two *p-values* might arise from the fact that in LM1 sample the bacteria is able to disrupt its internalization vacuole and use the cell machinery to replicate. Contrasting, in LM2 sample, lacking the gene for LLO, the bacteria is not capable to invade the cell citosol. Therefore, in the first sample it is expected a stronger immunological cell response. This is translated by the *p-values* differences, which statistically describes how well the known network is fitting the dataset.

Therefore, considering the previously referred results, it is possible to conclude that the data being analysed is consistent with the published network. The *p-value* associated with the statistical measured value (MSE/Var) is quite low for the NG, particularly for LM1 sample. Hence, the described conjunction of data is not random and it is present in the analysed dataset in the same way that is described in the analysed genes network. Furthermore, from this analysis we are also able to support the idea that in sample LM1 a stronger cellular immunological response occurred than in LM2 samples. These results were evidenced in the analysis of the *L. monocytogenes* dataset by the developed pipeline and are congruent with what we were expecting.

3.3 Discussion

The analysis of RNA-Seq data in the systems biology context gives insight about the active cell pathways upon a certain stimulus and along a time-course. The methodology proposed here takes advantage of this information to validate the RNA-Seq data. The validation performed by this method aims to prove that the genes in the RNA-Seq dataset are well modelled by a published gene networks. The network used can describe a pathway that is expected to be active (by knowing which stimulus was applied in the cell upon transcriptome sequencing). This is done when it is intended to investigate if the data describes well a biological process in study. And can also be applied when a pathway is thought not to be occurring. The developed methodology allows to investigate if an adjacent pathway is activated when the cell is submitted to different conditions.

Moreover, this methodology may prove particularly useful for RNA-Seq datasets that do not have replicates. Without replicates it is hard to understand if the RNA-Seq reads are translating well the cell transcriptome or are just a consequence of sampling noise. This method provides a way to test that a certain model of a biological pathway is, in fact, a good model for a RNA-Seq dataset in study. Ultimately, the gene kinetics information can evidence that a certain cellular pathway is described in a next generation sequencing data, supporting a comparative analysis such as the one performed in Section 2.2.

This methodology can also be useful when a RNA-Seq dataset contains replicate reads. The available tools that statistically compare RNA-Seq reads are able to investigate the congruency between single samples but not the biological processes that they are describing, in a systems biology way. The use of this novel methodology not only allows to investigate if a process is described along the RNA-Seq data but also, by investigating how well a certain network is modelling a dataset, clarifies about the similarly between the NG variation among the samples. This new methodology intends to complements the available tools that statistically compare RNA-Seq reads by consolidating its conclusions.

Nonetheless, this methodology is based on several assumptions that may limit the obtained results. For instance, if a gene that is described in the network in analysis is not expressed, the methodology trims it from the analysis. In addition, the time derivatives dx_i/dt need also to be determined by linear interpolation, and, therefore, this value corresponds only to an estimate. However, this methodology was adopted for several researchers to perform the modulation of their data in the gene networks context and proved to be an effective approach.[19,21] Finally, the ratio between the mean square error and the variance (MSE/Var) might not be accurate in cases where the variance associated with the NG is very low. This occurs because in the permutation test will appear set of random genes with high variances. This high values imply that the measured statistical value will be low, even if the MSE is high.

$$\begin{cases} \frac{dx_1}{dt} = -6.9x_1 + 10.3.u(t) \\ \frac{dx_2}{dt} = -29.9x_1 - 20.4x_2 + 11.02.u(t) \\ \frac{dx_3}{dt} = 44.1x_2 - 49.7x_3 + 24.9x_7 - 19.7.u(t) \\ \frac{dx_4}{dt} = -3.8x_2 - 8.5x_4 + 4.1.u(t) \\ \frac{dx_5}{dt} = -3.6x_2 - 21.7x_5 - 15.8.u(t) \\ \frac{dx_6}{dt} = 2.2x_1 - 39.1x_6 + 5.2.u(t) \\ \frac{dx_7}{dt} = 88.7x_6 - 11.7x_7 + 3.4.u(t) \end{cases} \quad (3)$$

Equation 3: Differential equations describing genes kinetics for sample infected with wild-type *L. monocytogenes* relatively to network from figure 1. Each entry is multiplied by a factor of 1000.

$$\begin{cases} \frac{dx_1}{dt} = -8.0x_1 + 0.8.u(t) \\ \frac{dx_2}{dt} = -14.8x_1 - 50.3x_2 + 121.9.u(t) \\ \frac{dx_3}{dt} = 32.2x_2 - 28.3x_3 + 22.8x_7 - 47.8.u(t) \\ \frac{dx_4}{dt} = -2.6x_2 - 7.1x_4 + 21.9.u(t) \\ \frac{dx_5}{dt} = -13.9x_2 - 33.5x_5 - 28.2.u(t) \\ \frac{dx_6}{dt} = -7.7x_1 - 37.2x_6 - 3.2.u(t) \\ \frac{dx_7}{dt} = 6.6x_6 - 4.8x_7 + 2.0.u(t) \end{cases} \quad (4)$$

Equation 4: Differential equations describing genes kinetics for sample infected with mutant *L. monocytogenes* relatively to network from figure 1. Each entry is multiplied by a factor of 1000.

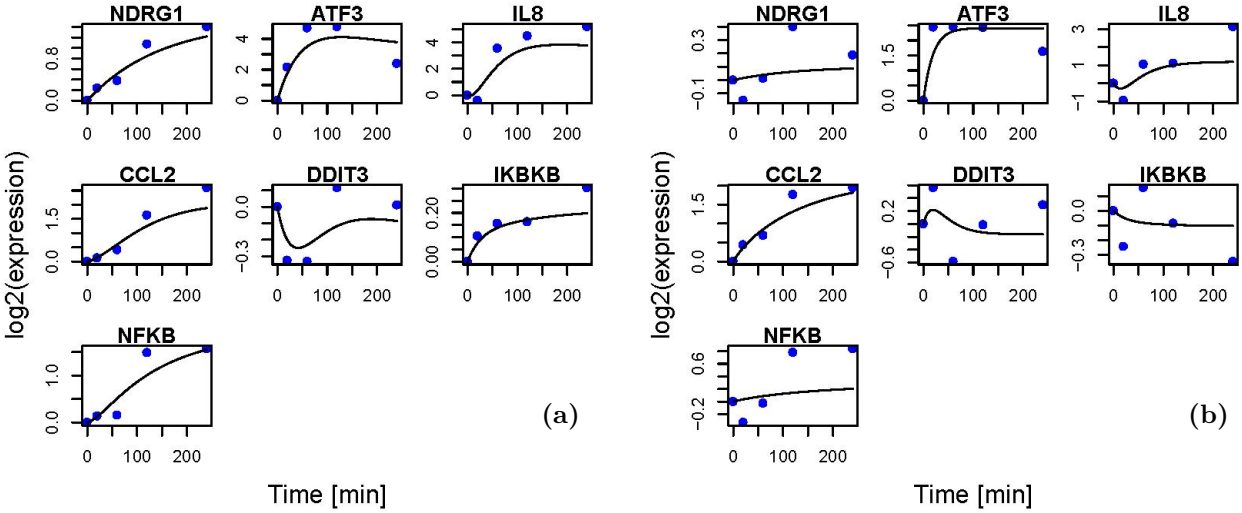

Figure 2: Measured and simulated expression kinetics in log-ratios of the genes in the nodes of network 2, for both **(a)** wild-type and **(b)** mutant infected datasets.

4 Using publicly available data as RNA-seq replicates

When applying statistical methods, such as the ones provided by DESeq package, it is extremely important to have proper replicates in order to validate the obtained results. In fact, without replicates, it is not possible to assert if differences between conditions are just a consequence of experimental and biological noise.

4.1 Methods

The methodology here proposed intends to support the DE analysis, by using publicly available data and bioinformatics tools. This idea was a consequence of the challenge that it is validate biological conclusions from RNA-Seq datasets without replicates. Nevertheless, it is important to keep in mind that, ultimately, this approach was proven not to be feasible. However, this analysis is not worthless. In fact, allows to conclude that the approach defined in this section is not a reliable way to reinforce the DE inference among RNA-Seq samples.

This methodology starts by extracting RNA-Seq samples similar to the ones in analysis from public databases, such as Gene Expression Omnibus or ArrayExpress. Then, to access the similarly between the samples, it is used the pipeline described in Section 2 until the DESeq step. At this point, it is possible to understand if the public sample is suitable to be used as replicate of the RNA-Seq sample in analysis, by both analysing the output table of differentially expressed genes and the MA-plots illustrating the distribution of the genes log fold change along the several expression levels. If the sample downloaded from a public database is similar to the RNA-Seq that is intended to support, the MA-plot will have the shape of a narrow funnel.

4.2 *Listeria monocytogenes* case study

To understand if the methodology here presented is valid, the control sample from the

L. monocytogenes dataset was used to evaluate it. The control sample was generated by deep sequencing all poly-(A) tailed mRNAs of a HeLa cell population growing on a plate. Therefore, the first step was to obtain from the available databases a RNA-Seq dataset that was obtained from a population of HeLa cells growing in a healthy medium.[22,23] The public dataset that is was used in this analysis contains two biological replicates and one technical replicate for each biological replicate. Aiming to understand how these samples are correlated with our control data, reads were processed as explained in the previous subsection. The genes' dispersion in figure 3(a) is very high, with a portion of the genes having a fold change of approximately -10, which corresponds to a difference of 1024 in the fold change between the two conditions being analysed. Oppositely, the genes' fold change in the comparison between the two replicates (figure 3(b)) is concentrated near to 0, which is what we were expecting, and have associated a very low dispersion.

4.3 Discussion

Even though this methodology could be an improvement on the RNA-Seq DE inference in datasets without replicates, the result from the analysis between the our control sample and the public transcriptome data was not good. A high batch effect was found among these two datasets. In fact, there is a high limitation associated with this methodology: is highly depended of the data acquisition protocol. Even when the cell is simply growing on a plate, the transcriptome can be deeply influenced by the external environment or the chosen growing medium. Moreover, variability can be also associated with the library construction from the mature cell population or even the sequencing process. All these parameters can be prominent in the fallibility of this methodology.

5 Conclusion and Future work

Recent technological advances in genomics and proteomics are generating data at unprecedented high resolution. One example is high-throughput sequencing of RNA (RNA-Seq) which allows the simultaneous measurement of RNAs sequence and expression at whole cell level. With the introduction of these novel technologies, new bioinformatic approaches are required. This article describes two useful tools to analyse RNA-Seq data. And, moreover, we also tested a new approach to reinforce the genes DE inference. First, we implemented a pipeline to analyse RNA-Seq data. This tool is able to find,from the raw RNA-Seq samples, which are the differential active cellular process between them. This enables the user to conclude about the influence that a certain cell environment or stimulus has into the cell active processes. Subsequently, we developed an innovative tool that is able to investigate if a certain biological phenomena is well described by the dataset in analysis. In order to perform that analysis, the developed methodology takes advantage of previous knowledge and confirms if the gene interactions in a published gene network are modelling well the RNA-Seq dataset in analysis. Finally a novel methodology was proposed to overcome the limitations of RNA-Seq datasets with no replicated samples. This methodology is based on the use of already published RNA-Seq samples as replicates of that poor dataset.

In order to evaluate the developed tools, it was studied a *L. monocytogenes* RNA-Seq dataset. Regarding the analysis of this dataset with the developed pipeline, it was possible to extract biological meaningful considerations. Namely, what were the cell active processes when submitted to different environments. Similarly, it was confirmed that this dataset is well fitted by a gene regulatory network describing the cell immunological response upon *L. monocytogenes* infection. Given that the *L. monocytogenes* RNA-Seq dataset does not contains any replicates, we attempted to improve its statistical confidence by using public RNA-Seq data from HeLa cells as a replicate of the control sample. However, this methodology was proven not to be reliable.

In order to strengthen the reliability of the new tools described above, a new dataset, with more robust information, for instance, should be processed resorting to the methodologies here proposed. It would be important to perform the analysis of a RNA-Seq dataset which has replicate reads for each collected condition and, moreover, this new dataset should study a cell process that it is not related with the immunological response. To understand the RNA-Seq analysis pipeline reliability, the analysis of this new dataset should be performed in two ways: one including all the replicates reads and another considering only one read from each condition. From this analysis, we will be able to compare the results with and with-

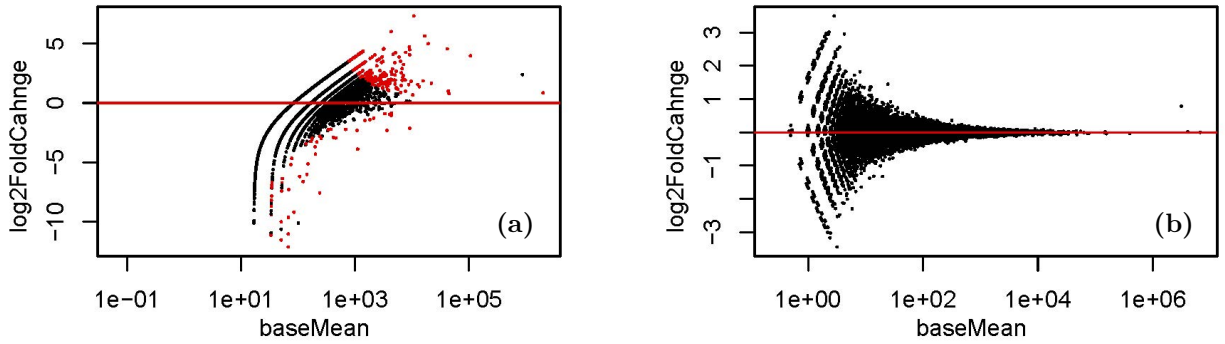

Figure 3: Plot of normalized counts mean *versus* log_2 fold change for the contrast between **(a)** *L. monocytogenes* dataset control and samples from dataset 3; and **(b)** first and second replicate from dataset 3. Outliers genes represented in red.

out replicates and, subsequently, confirm or refute the similarity between the cellular processes found to be differentially active among both analysis. Additionally, in order to investigate the acuity of the RNA-Seq validation algorithm described in Section 3, it should be analysed a dataset which already was tested by other tools and from which are known the active cell pathways. Having this information, the inspection of the *p-value* value gives insight about the algorithm reliability. Moreover, for the RNA-Seq analysis pipeline, the obtained results could be supported by confirming the expression variation of the genes which are thought to have a greater influence in the cell immunological response to the *L. monocytogenes* invasion, using, for instance, qRT-PCR technique. Concerning the methodology described in Section 4, in the future, it may be worthwhile to revisit this approach when the technology has improved.

References

[1] W Gilbert et al. **"The nucleotide sequence of the lac operator."** In: *Proceedings of the National Academy of Sciences of the United States of America* 70.12 (Dec. 1973), pp. 3581–4. ISSN: 0027-8424.

[2] F. Sanger et al. **"Nucleotide sequence of bacteriophage ϕX174 DNA"**. In: *Nature* 265.5596 (Feb. 1977), pp. 687–695. ISSN: 0028-0836.

[3] Elaine R Mardis. **"A decade's perspective on DNA sequencing technology."** In: *Nature* 470.7333 (Feb. 2011), pp. 198–203. ISSN: 1476-4687.

[4] Simon Andrews. ***Babraham Bioinformatics - FastQC***. 2010.

[5] Ben Langmead et al. **"Fast gapped-read alignment with Bowtie 2."** In: *Nature methods* 9.4 (Apr. 2012), pp. 357–9. ISSN: 1548-7105.

[6] S. Anders. ***HTSeq: Analysing high-throughput sequencing data with Python***.

[7] Simon Anders et al. **"Differential expression analysis for sequence count data."** In: *Genome biology* 11.10 (Jan. 2010), R106. ISSN: 1465-6914.

[8] S Falcon et al. **"Using GOstats to test gene lists for GO term association"**. In: *Bioinformatics* 23.2 (Jan. 2007), pp. 257–258. ISSN: 1367-4811.

[9] Mélanie Hamon et al. **"Listeria monocytogenes: a multifaceted model."** In: *Nature reviews. Microbiology* 4.6 (June 2006), pp. 423–34. ISSN: 1740-1526.

[10] Erich Gulbins et al. **"Pathogens, Host-Cell Invasion and Disease"**. In: *American scientist* 89.5 (2001), p. 406.

[11] Daniel a Portnoy et al. **"The cell biology of Listeria monocytogenes infection: the intersection of bacterial pathogenesis and cell-mediated immunity."** In: *The Journal of cell biology* 158.3 (Aug. 2002), pp. 409–14. ISSN: 0021-9525.

[12] Robert Lindner et al. **"A comprehensive evaluation of alignment algorithms in the context of RNA-seq."** In: *PloS one* 7.12 (Jan. 2012). Ed. by Steven L. Salzberg, e52403. ISSN: 1932-6203.

[13] Charlotte Soneson et al. **"A comparison of methods for differential expression analysis of RNA-seq data."** In: *BMC bioinformatics* 14.1 (Jan. 2013), p. 91. ISSN: 1471-2105.

[14] Purvesh Khatri et al. **"Ten years of pathway analysis: current approaches and outstanding challenges."** In: *PLoS computational biology* 8.2 (Jan. 2012), e1002375. ISSN: 1553-7358.

[15] Frank Emmert-Streib et al. **"Pathway analysis of expression data: deciphering functional building blocks of complex diseases."** In: *PLoS computational biology* 7.5 (May 2011), e1002053. ISSN: 1553-7358.

[16] R. A. Fisher. ***The design of experiments.*** Oxford, England: Oliver & Boyd, 1935, p. 251.

[17] Christoph Alexander Kasper et al. **"Cell-cell propagation of NF-κB transcription factor and MAP kinase activation amplifies innate immunity against bacterial infection."** In: *Immunity* 33.5 (Nov. 2010), pp. 804–16. ISSN: 1097-4180.

[18] Marie-Agnès Dillies et al. **"A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis."** In: *Briefings in bioinformatics* (Sept. 2012). ISSN: 1477-4054.

[19] Reinhard Guthke et al. **"Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection."** In: *Bioinformatics* 21.8 (Apr. 2005), pp. 1626–34. ISSN: 1367-4803.

[20] Hailong Zhu et al. **"Reconstructing dynamic gene regulatory networks from sample-based transcriptional data."** In: *Nucleic acids research* 40.21 (Nov. 2012), pp. 10657–67. ISSN: 1362-4962.

[21] M K Stephen Yeung et al. **"Reverse engineering gene networks using singular value decomposition and robust regression."** In: *Proceedings of the National Academy of Sciences of the United States of America* 99.9 (Apr. 2002), pp. 6163–8. ISSN: 0027-8424.

[22] Kathi Zarnack et al. **"Direct Competition between hnRNP C and U2AF65 Protects the Transcriptome from the Exonization of Alu Elements"**. In: *Cell* 152.3 (2013), pp. 453–466.

[23] ***ArrayExpress: E-MTAB-1147***. http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1147/. 2013.