HALO - A Java framework for precise transcript half-life determination

Caroline C. Friedel\textsuperscript{1,1}, Stefanie Kaufmann\textsuperscript{1,1}, Lars Dölken\textsuperscript{2} and Ralf Zimmer\textsuperscript{1}

\textsuperscript{1}Institut für Informatik, Amalienstr. 17 and \textsuperscript{2}Max von Pettenkofer-Institut, Pettenkofer Str. 9a
Ludwig-Maximilians-Universität München, München, Germany

Associate Editor: Dr. Joaquin Dopazo

ABSTRACT

Summary: Recent improvements in experimental technologies now allow measurements of de novo transcription and/or RNA decay at whole transcriptome level and determination of precise transcript half-lives. Such transcript half-lives provide important insights into the regulation of biological processes and the relative contributions of RNA decay and de novo transcription to differential gene expression. In this article, we present HALO, the first software for the precise determination of transcript half-lives from measurements of de novo transcription or decay determined with microarrays or RNA-seq. In addition, methods for quality control, filtering and normalization are supplied. HALO provides a graphical user interface, command line tools and a well-documented Java API. Thus, it can be used both by biologists to determine transcript half-lives fast and reliably with the provided user interfaces as well as software developers integrating transcript half-life analysis into other gene expression profiling pipelines.

Availability: Source code, executables and documentation are available at http://www.bio.ifi.lmu.de/software/halo
Contact: Caroline.Friedel@bio.ifi.lmu.de

1 INTRODUCTION

Transcript levels in a cell are determined by a constant turnover driven by de novo synthesis by polymerases and degradation by nucleases. While constant transcript levels are due to an equilibrium between RNA synthesis and decay, changes in transcript levels reflect alterations in either of them (Ross, 1995). While standard gene expression profiling methods now allow measurements of total RNA levels on a genome-wide scale, they cannot distinguish whether changes in total RNA levels are due to changes in de novo transcription or RNA decay.

Previously, RNA decay rates have been determined by arresting transcription and subsequently monitoring ongoing RNA decay over time (Bernstein et al., 2002; Wang et al., 2002; Raghavan et al., 2002; Yang et al., 2003; Redon et al., 2005; Narzai et al., 2007). This is based on the assumption that RNA decay is not affected by transcription inhibition. However, it has been shown that a stress response, such as e.g. induced by transcriptional arrest, can lead to a substantial stabilization of individual transcripts (Gorospe et al., 1998; Blattner et al., 2000). Alternatively, nuclear run-on assays have been used to measure de novo transcription and the relative contributions of transcription rate and mRNA decay to steady-state mRNA levels (Fan et al., 2002; García-Martínez et al., 2004; Molina-Navarro et al., 2008). Recently, new methods have been introduced for the non-disruptive quantification of de novo transcription based on biosynthetic labeling (RNA tagging) of newly transcribed RNA with 4-thiouracil (4U) (Cleary et al., 2005) or 4-thiouridine (4sU) (Kenzelmann et al., 2007; Dölken et al., 2008). After labeling newly transcribed RNA, total RNA can be separated into the labeled newly transcribed RNA and the unlabeled pre-existing RNA and de novo transcription and decay can be quantified in a single experimental setting.

Using any of these methods, transcript half-lives, i.e. the speed of RNA turnover, can be determined for individual genes and the relative contributions of transcription rate and RNA decay to gene expression regulation can be evaluated. In this way, alterations of RNA decay have been confirmed as an important regulatory mechanism for many genes (Fan et al., 2002; Molina-Navarro et al., 2008). Moreover, as transcript half-lives are correlated to gene function and regulation (Yang et al., 2003; Redon et al., 2005; Narzai et al., 2007; Friedel et al., 2009), they can provide interesting insights into differential regulatory mechanisms for closely related genes (Friedel et al., 2009) and the regulation of protein complexes (Wang et al., 2002; Friedel et al., 2009).

In this article, we present HALO (Half-life Organizer), a software environment for the calculation of transcript half-lives from measurements of total RNA levels, de novo transcription or RNA decay obtained with any of the above described methods. As it is independent of the quantification method used in the process, it can be applied both to microarray and RNA-seq measurements. In addition, it provides methods for filtering and normalization of data and quality control of the microarray or RNA-seq experiments. Moreover, these algorithms are not restricted to RNA half-lives. Recent improvements in proteomics technologies, particularly SILAC (stable isotope labeling with amino acids in cell culture) (Mann, 2006; Ong et al., 2002) will soon be suitable to measure protein half-lives based on newly synthesized total protein ratios. To support these developments, HALO has been designed to be easily extendable to new methods. Thus, it will prove valuable for further studies on the role of half-lives in the regulation of many biological processes.

*To whom correspondence should be addressed
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors
Fig. 1. Outline of the different steps involved in transcript half-life calculation and the methods implemented in HALO.

2 METHODS

HALO takes as input whole transcriptome data of total, newly transcribed and/or pre-existing RNA either from microarray experiments or RNA-Seq which have been pre-processed using standard methods. For microarray data, pre-processing includes background correction, a first normalization and probe-level summarization, e.g. with GC-RMA (Wu and Irizarry, 2004). For RNA-seq data, this requires read mapping and calculation of RNA levels from read counts. Since the input to HALO are the final quantified RNA levels for either genes, individual exons or probe sets, it is independent of the experimental method used to obtain these levels. As transcript half-lives can be calculated from either newly transcribed/total RNA ratios, newly transcribed/pre-existing RNA ratios or pre-existing/total RNA ratios, measurements of any two of the three RNA fractions are sufficient. However, to fully exploit the quality control and normalization methods of HALO, all three RNA fractions have to be measured. See Figure 1 for an outline of the individual steps involved in half-life calculation.

In the first step, a quality control can be applied to check for an experimental bias in measuring newly transcribed or pre-existing RNA (Quality control (1)). For RNA tagging, for instance, capture of newly transcribed RNA depends on sufficient incorporation of 4U or 4S. Thus, insufficient labeling of short transcripts with low uracil content may lead to reduced capture rates for these transcripts and a bias in calculating transcript half-lives. To check for this kind of bias, the number of uracils for each transcript calculated from transcript sequences can be compared against newly transcribed/total RNA ratios. A potential bias can then be corrected with HALO using the regression method proposed by Miller et al. (2009). Another quality control can be applied based on the relationship between newly transcribed, pre-existing and total RNA if all three RNA fractions have been measured. Since newly transcribed and pre-existing RNA should sum up to total RNA, a negative linear correlation should be observed between newly transcribed/total RNA and pre-existing/total RNA ratios.

Using this negative correlation, ratios of newly transcribed/total RNA and pre-existing/total RNA can be easily normalized using linear regression analysis (Dölken et al., 2008) (Normalization (2)). This second normalization step is necessary as amounts of template RNA differ between newly transcribed RNA, pre-existing RNA and total RNA samples. If only two of the RNA fractions have been measured (e.g. newly transcribed and total RNA), ratios can be normalized based on median transcript half-life.

Before normalization, unreliable measurements for individual genes, exons or probe sets can be filtered based on absent calls and/or low expression values (Filtering (3)). In addition, filtering can be performed after normalization using a quality score calculated for each gene, exon or probe set based on the distance of the corresponding ratios from the linear regression line. The smaller the distance, the higher the quality score. Again, this requires measurements for all three RNA fractions. Both filtering and normalization steps can be easily extended by new methods.

In the final step, transcript half-lives are calculated from the normalized ratios using an exponential decay model (Half-life calculation (4)). Half-lives can be calculated either from pre-existing/total RNA, newly transcribed/total RNA or newly transcribed/pre-existing RNA ratios. The latter method combines the overall high precision of the newly transcribed/total RNA ratios with the higher precision for short transcript half-lives of pre-existing/total RNA ratios which otherwise are unreliable for medium-to-long-lived transcripts (Friedel and Dölken, 2009). Normalization and transcript half-life calculation can be performed individually for each replicate or for average measurements.

To analyze the results, plots can be created for quality control, the normalization with linear regression analysis and the distribution of quality scores and final transcript half-lives (Visualization (5)).

HALO is implemented in Java and provides several user interfaces to perform all of these steps in a straightforward way. The graphical user interface (GUI) is also available as a Java webstart version which can be started directly from the webbrowser without installing the package. Furthermore, a set of command line tools are included for script and batch usage and all methods of HALO are accessible via a well-documented Java API.

3 CONCLUSION

HALO is an extensible Java framework providing state-of-the-art methods for transcript half-life calculation as well as quality control, filtering and normalization. It can be used as a standalone tool or can be integrated into other gene expression profiling frameworks. The intuitive graphical user interface makes it accessible for users without programming skills aiming to calculate transcript half-lives in a fast and straightforward way from new measurements of de novo synthesis and/or decay both for microarrays or RNA-Seq. In addition, researchers developing new software programs and methods can easily incorporate and extend the available methods of HALO in their own programs. Thus, HALO will be a valuable tool for the analysis of transcript half-lives and, thus, an important prerequisite for the analysis of the regulation of biological systems.

Requirements

HALO requires the Java 6.0 (or higher) Runtime Environment (freely available at http://www.java.com). Visualization (optional) requires the JFreeChart library (available under the GNU Lesser General Public License at http://www.jfree.org/jfreechart/).

REFERENCES

Bernstein, J. A. et al. (2002). Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA
microarrays. Proc Natl Acad Sci USA, 99(15), 9697–9702.