Cell PRESS

Human housekeeping genes, revisited

Eli Eisenberg[1] and Erez Y. Levanon[2]

[1] Raymond and Beverly Sackler School of Physics and Astronomy, Tel-Aviv University, Tel Aviv 69978, Israel
[2] Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan 52900, Israel

Housekeeping genes are involved in basic cell maintenance and, therefore, are expected to maintain constant expression levels in all cells and conditions. Identification of these genes facilitates exposure of the underlying cellular infrastructure and increases understanding of various structural genomic features. In addition, housekeeping genes are instrumental for calibration in many biotechnological applications and genomic studies. Advances in our ability to measure RNA expression have resulted in a gradual increase in the number of identified housekeeping genes. Here, we describe housekeeping gene detection in the era of massive parallel sequencing and RNA-seq. We emphasize the importance of expression at a constant level and provide a list of 3804 human genes that are expressed uniformly across a panel of tissues. Several exceptionally uniform genes are singled out for future experimental use, such as RT-PCR control genes. Finally, we discuss both ways in which current technology can meet some of past obstacles encountered, and several as yet unmet challenges.

The concept of housekeeping genes

Housekeeping genes are genes that are required for the maintenance of basal cellular functions that are essential for the existence of a cell, regardless of its specific role in the tissue or organism. Thus, they are expected to be expressed in all cells of an organism under normal conditions, irrespective of tissue type, developmental stage, cell cycle state, or external signal. From a fundamental point of view, full characterization of the minimal set of genes required to sustain life is of special interest [1,2]. In addition, housekeeping genes are widely used as internal controls for experimental as well as computational studies [3–7]. Furthermore, many studies have highlighted unique genomic and evolutionary features of this special group of genes. For example, housekeeping genes were shown to have shorter introns and exons [8–11], a different repetitive sequence environment [enriched in short interspersed elements (SINEs) and depleted in long interspersed elements (LINEs)] [12,13], more simple sequence repeats in the 5′ untranslated region (UTR) [14], lower conservation of the promoter sequence [15], and lower potential for nucleosome formation in the 5′ region of these genes [16]. Protein products of housekeeping genes are enriched in some domain families [17]. These studies shed light on general aspects of gene structure and evolution.

Corresponding author: Eisenberg, E. (elieis@post.tau.ac.il).
Keywords: housekeeping genes; RNA-seq; gene expression patterns; internal control; next generation sequencing.

0168-9525/$ – see front matter
 http://dx.doi.org/10.1016/j.tig.2013.05.010

Early detection schemes for housekeeping genes

The notion of housekeeping genes has been in use in the literature for nearly 40 years. In particular, several mammalian genes have been used widely as internal controls in experimental expression studies, such as glyceraldehyde-3-phosphate dehydrogenase (GAPDH), tubulins, cyclophilin, albumin, actins, 18S rRNA or 28S rRNA. Yet, only at the turn of the 21st century, with the advancement of transcriptome profiling technology, did it become possible to identify, systematically, a set of housekeeping genes. These first attempts used large-scale expression data [18–20] or, more often, microarray profiling to look at the expression levels of many genes across a panel of tissue samples. Typically, they resulted in lists of hundreds to thousands of genes [8,19–25], many more than the dozen or so commonly used control genes.

Generally, the many lists produced show a considerable level of consistency. Typically, the intersection of any two of them yields approximately 50% coverage [8,24,26], suggesting that the sets are enriched in housekeeping genes but still lacking in specificity and selectivity. This could be partly attributed to the limited number of tissues examined in each separate analysis and the differences between the tissues across analyses. However, it is likely that technological limitations affecting the underlying data have contributed much to the quality and reproducibility of the results.

In particular, first-generation microarray technology is known to have had many problematic nonspecific probes [27]. Even the improved versions of microarrays are typically assumed to achieve only an approximately twofold accuracy in expression level measurement, and they are limited in their dynamical range. These inaccuracies could have large effects on deciding whether a gene is expressed (regardless of the rather arbitrary expression cutoff used to determine which probe set is 'expressed').

A second, more fundamental, issue relates to the very definition of housekeeping genes. Should one look for genes merely being expressed in all tissues, or should the gene also be expressed at a constant level across tissues? Early studies generally adopted the first definition and, in fact, GAPDH and other popular housekeeping genes for experimental controls have been found to vary considerably across tissues [3,28–30]. This choice was the pragmatic one to make, because it enabled the use of the binary present or absent calls of the microarray and rendered normalization issues unnecessary. However, this approach has two shortcomings. First, measurement errors and stochastic noise make it difficult to distinguish genes absent from the sample from those weakly expressed. Second, and more importantly, it was later appreciated

that a large part of the genome is expressed at a low basal level in all tissues [31]. Thus, most genes are expressed at some background level in all tissues. In light of this observation, and to make the concept of housekeeping genes more useful, one should either modify the definition of housekeeping genes to 'genes that are expressed above some cutoff level', which necessarily introduces an arbitrary parameter explicitly, or rather adopt the second option above and look for genes that are expressed at a constant level across all normal tissues.

Introducing an expression cutoff requires a quantitative comparison of expression levels of different genes in the same sample. This is known to be a complex problem, due to questions of bias in PCR amplification, different probe affinities, and so on. Furthermore, normalizing the values obtained from different experiments is also a nontrivial challenge. Early microarrays studies generally used linear normalization, setting the mean expression level, or the trimmed mean, constant. Later, the more sophisticated quantile normalization was introduced [32]. These and other normalization procedures generally assume similar expression-value distributions for all samples studied. This could be justified for samples coming from identical or highly similar biological conditions, perhaps even for healthy and diseases samples of the same tissue. However, it is not yet clear how accurate this assumption is for cross-tissue comparisons, and how much it skews the results [33].

A third issue that was not fully addressed in previous studies of housekeeping genes is alternative splicing. It has been appreciated for more than a decade that most human genes have more than one isoform [34,35]. Thus, one could envision a situation in which one splice variant is constitutively expressed, making it a housekeeping transcript, whereas another transcript from the same gene exhibits a more complex expression profile (Figure 1A). Moreover, it is possible that a single gene expresses one transcript in one set of tissues and another transcript in other tissues, such that the gene, as such, is always expressed, but each transcript is specific to a subset of tissues. In principle, then, one would like to define the set of housekeeping transcripts. Early microarray technology did rather poorly in distinguishing between transcripts and, thus, some studies deliberately 'zoomed out' to the gene level.

Housekeeping genes in the deep-sequencing era

New horizons are opening as deep-sequencing technology takes over microarrays as the method of choice for transcriptome profiling [36]. RNA-seq was found to be preferable to microarrays as a tool for expression measurement. Unlike microarrays, RNA-seq does not require pre-knowledge of the genomic sequence (although it is helpful for analysis), and requires smaller amounts of RNA. It provides information at the single-base level, enabling better assessment of alternative splicing and even allelic variation. Background levels in RNA-seq are lower, due to the better specificity and improved control of *in silico* sequence alignment compared with probe hybridization. Consequently, a wider dynamic range is accessible. Importantly, RNA-Seq is also more accurate in quantifying spike-in RNA controls of known concentration, and produces

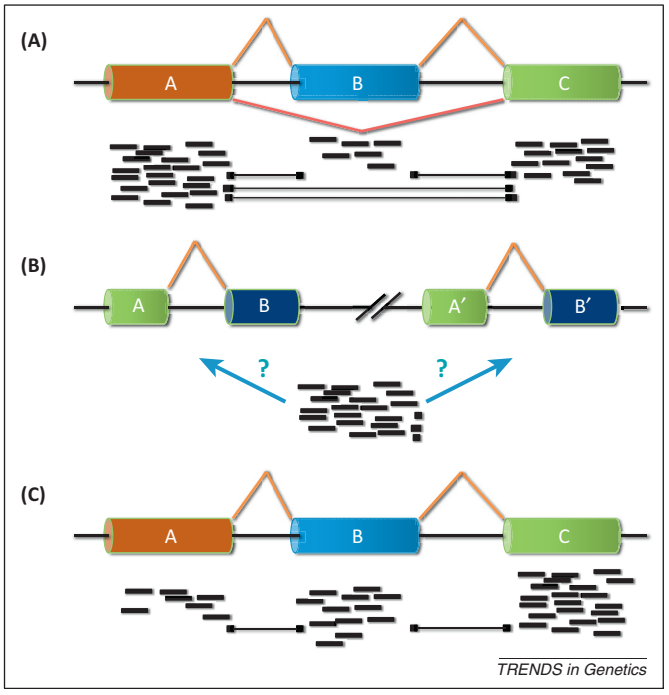

Figure 1. Examples of challenges in housekeeping gene detection. **(A)** Genes having several splice variants could have different expression levels [indicated by the number of reads (black bars)] for different parts of the gene. **(B)** Duplicative regions, due to pseudogenes and other duplications, complicate unique read alignments, thus biasing expression-level measurement. **(C)** Expression measurement has several biases, including the lower expression (on average) of the upstream exons due to imperfect reverse transcription resulting in partial cDNA molecules.

expression values that correlate better with quantitative PCR (qPCR) results [36] and protein levels [37]. This new and improved platform enables some of the challenges to be met that have been standing for many years, but it also opens up new questions.

In terms of normalization, read coverage generally provides a rather robust measure for comparing different genomic regions within the same sample. Exceptions to this are generally a result of alignment problems in repetitive or duplicative regions (Figure 1B). For the task of housekeeping gene identification, these can be partly avoided by limiting analysis to the nonrepetitive coding regions of the exons [33] and using long reads. Note, however, that highly expressed coding exons (e.g., GAPDH) are prone to having more duplications [38], resulting in alignment problems. Small-scale PCR biases are expected to be washed out when looking at the averaged expression level over whole exons. By contrast, the issue of cross-tissue normalization is still open. The popular reads per kilobase per million mapped reads (RPKM) measure takes care of normalizing for the two most obvious factors affecting the raw number of reads per gene, transcript, or exon: the total number of reads produced and their length [39]. The RPKM measure is simple and straightforward, but does not fully solve the between-sample normalization issue. More subtle biases, resulting from variations in transcript length distribution in the sample, coverage dependence on local sequence due to GC content, priming and other biases, and variability in mappability of different regions were detected [40–45].