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

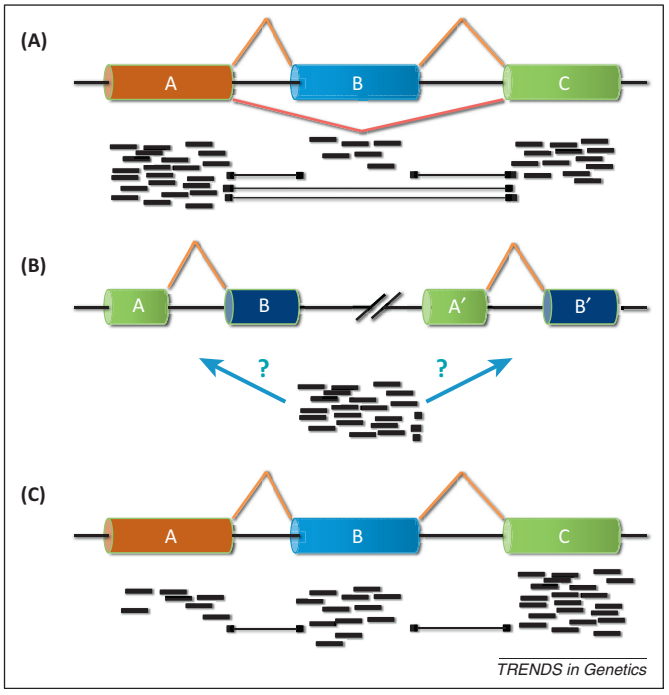

Figure 1. Examples of challenges in housekeeping gene detection. **(A)** Genes having several splice variants could have different expression levels [indicated by the number of reads (black bars)] for different parts of the gene. **(B)** Duplicative regions, due to pseudogenes and other duplications, 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].

There is still no consensus as to the best way to account for all of these in a standard and consistent way.

In terms of housekeeping gene identification, RNA-seq data indeed show explicitly that basal (leaky) low expression levels can be found throughout the genome. Therefore, any definition of housekeeping genes should refer to the quantitative expression level. This can be done using a cutoff, or by adding the requirement of low variability in expression across tissues. Here, we promote the latter course of action. Setting a cutoff value as the main criteria for defining the housekeeping genes is undesirable for three reasons. First, there seems to be no natural cutoff value, thus forcing one to make an arbitrary choice. Second, due to the lack of a proper intergene normalization scheme, the same RPKM values for different genes could indicate different expression levels [4,46]. Third, using the expression level as a measure of importance for cell function is also questionable: cells are likely to require different gene products at different concentrations. There is no good reason to exclude genes that are constantly expressed at a mid rather than a high level. Thus, we feel that low variability should be used as the main criteria for selecting housekeeping genes.

Another advantage of RNA-seq data is that they measure the expression along the gene (similar to the older exon arrays) and can thereby provide expression at the exon level. Some software tools try to extract transcript expression levels from RNA-seq data (e.g., [47]). However, there is still much to be desired in terms of reliability within the limits of current technology [43]. This is expected to improve significantly, as read length increases. Note that recent findings [48] show significant variability in exon boundaries, making even the comparison of exon expression imperfect. An interim partial solution, which we adopt below, is to measure expression at the more basic exon level and aim to define a set of housekeeping exons.

Extracting a set of housekeeping genes from Human BodyMap data

Here, we demonstrate the power of the new technology for identifying housekeeping genes by analyzing expression data from the Human BodyMap (HBM) 2.0 Project. This includes publicly available RNA-Seq data (GEO accession number GSE30611, HBM), generated on HiSeq 2000 instruments, providing expression profiling in 16 normal human tissue types: adrenal, adipose, brain, breast, colon, heart, kidney, liver, lung, lymph, ovary, prostate, skeletal muscle, testes, thyroid, and white blood cells. Two different read lengths were used for each tissue (2×50-bp paired-end and 1×75-bp single-read data), each of which was sequenced in a separate HiSeq 2000 lane.

We aligned the reads to the genome using the Bowtie2 aligner [49] and measured the read coverage of each of the coding exons of the (uniquely aligned) RefSeq sequences [50], in normalized RPKM units. For exons that were partly coding, only the coding part was considered. Short exons (<50 bp) are prone to alignment problems and were discarded. We compared the RPKM values obtained from the paired-end data and the single-read data to assess the technical reproducibility of the RPKM measure, and found that the typical fold-ratio between the two was 1.5 (Figure 2A). We observed a bias against the upstream exons of transcripts, which tended to have a lower expression levels. This effect might result from imperfect reverse transcription resulting in cDNA missing the upstream part of the transcript (Figure 1C).

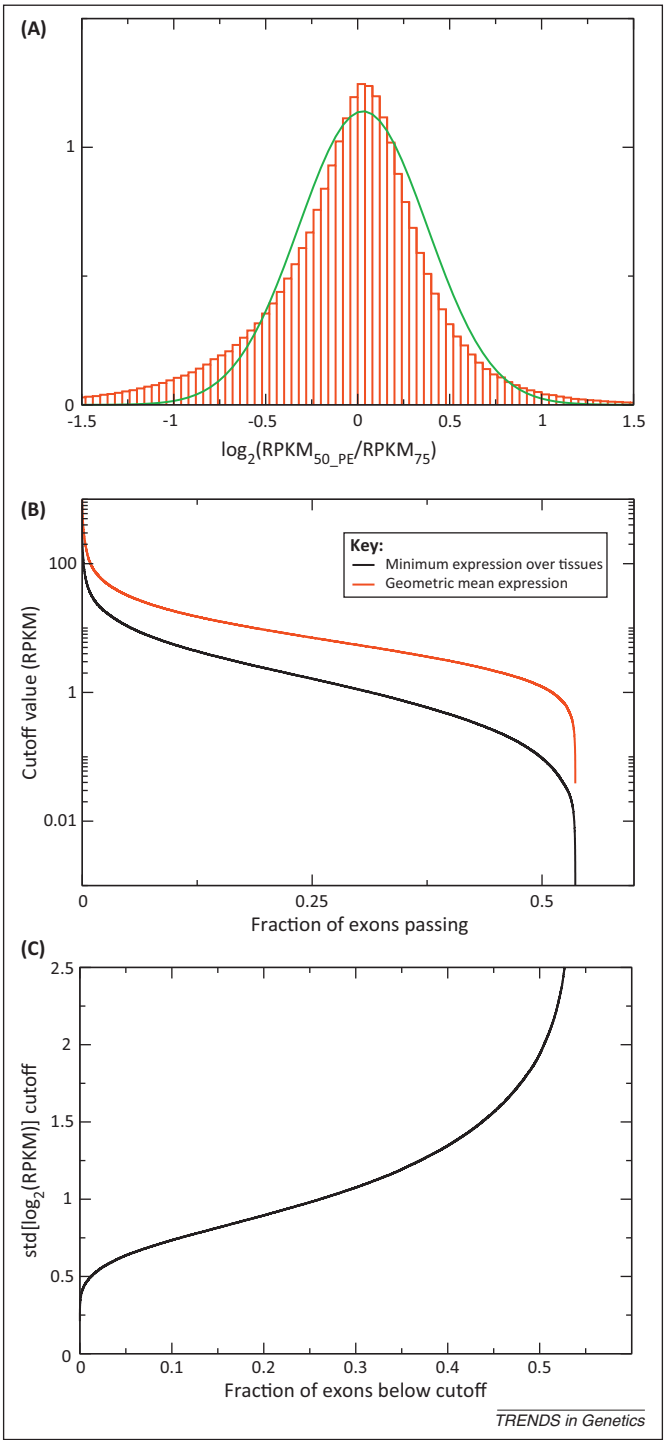

Figure 2. Characterization of the expression profile in Human BodyMap (HBM) data. **(A)** Reproducibility of the measured reads per kilobase per million mapped reads (RPKM) levels per exon, as assessed by comparing the 50-bp paired-end and the 75-bp single-read data. The continuous line is the best fit for a Gaussian distribution, added to accentuate the fat tails of the actual distribution. The width of the distribution is approximately 0.55 ($\log_2$ units), leading to a typical variability of 1.5-fold. **(B)** Fraction of exons expressed above a cutoff value in all 16 tissues, for different cutoff values. In total, 55% of all exons are expressed to a detectable level in the HBM data set. **(C)** Cumulative distribution of the exon expression variance. Most of the exons being expressed in all tissues have standard-deviation [$\log_2$(RPKM)] values between 0.7 and 1.5.

Figure 2B presents the fraction of exons being expressed above a certain cutoff RPKM value in all tissues. Note that approximately 55% of all exons are expressed at a detectable level in all HBM tissues, demonstrating why the old definition of housekeeping genes is not useful. In addition, it is hard to detect a natural expression cutoff value. The variation in expression level is estimated by the standard deviation of $\log_2$(RPKM) over samples. Figure 2C shows

Table 1. Genes proposed for calibration[a]

Gene symbol	RefSeq accession number	Gene name	Genomic coordinates (hg19) of exons passing the filters		
C1orf43	NM_015449	Chromosome 1 open reading frame 43	chr1	154192817	154192883
			chr1	154186932	154187050
			chr1	154186368	154186422
			chr1	154184933	154185100
			chr1	154184795	154184854
CHMP2A	NM_014453	Charged multivesicular body protein 2A	chr19	59065411	59065579
			chr19	59063625	59063805
			chr19	59063421	59063552
EMC7	NM_020154	ER membrane protein complex subunit 7	chr15	34382517	34382656
			chr15	34380253	34380334
			chr15	34376537	34376687
GPI	NM_000175	Glucose-6-phosphate isomerase	chr19	34857687	34857756
			chr19	34859487	34859607
			chr19	34868639	34868786
			chr19	34869838	34869910
			chr19	34872370	34872424
			chr19	34884152	34884213
			chr19	34884818	34884971
			chr19	34887205	34887335
			chr19	34887485	34887562
			chr19	34890111	34890240
			chr19	34890460	34890536
			chr19	34890623	34890690
PSMB2	NM_002794	Proteasome subunit, beta type, 2	chr1	36101910	36102033
			chr1	36096874	36096945
			chr1	36070833	36070883
PSMB4	NM_002796	Proteasome subunit, beta type, 4	chr1	151372456	151372663
			chr1	151372917	151373064
			chr1	151373239	151373321
			chr1	151373714	151373831
RAB7A	NM_004637	Member RAS oncogene family	chr3	128525214	128525433
			chr3	128526385	128526514
			chr3	128532169	128532262
REEP5	NM_005669	Receptor accessory protein 5	chr5	112256859	112256953
			chr5	112238076	112238215
			chr5	112222711	112222880
SNRPD3	NM_004175	Small nuclear ribonucleoprotein D3	chr22	24953642	24953768
			chr22	24963951	24964144
VCP	NM_007126	Valosin containing protein	chr9	35067887	35068060
			chr9	35066671	35066814
			chr9	35064150	35064282
			chr9	35062213	35062347
			chr9	35061999	35062135
			chr9	35061573	35061686
			chr9	35061011	35061176
			chr9	35060797	35060920
			chr9	35060309	35060522
			chr9	35059489	35059798
			chr9	35059060	35059216
			chr9	35057372	35057527
			chr9	35057116	35057219
			chr12	110930800	110931036
VPS29	NM_016226	Vacuolar protein sorting 29 homolog	chr12	110929812	110929927
			chr12	110929812	110929927

[a]Genes chosen have most of their exons showing geometrical mean expression exceeding RPKM = 50, standard deviation of log2(RPKM) <0.5, and no single tissue showing an expression level different from the geometrical mean by twofold or more. Genes with pseudogenes were excluded.

the cumulative distribution of these standard deviation values for the different exons. To define housekeeping exons, the exon must be expressed in all tissues at any nonzero level, and must exhibit a uniform expression level across tissues. Thus, we adopted the following criteria: (i) expression observed in all tissues; (ii) low variance over tissues: standard-deviation [$\log_2$(RPKM)]<1; and (iii) no exceptional expression in any single tissue; that is, no log-expression value differed from the averaged $\log_2$(RPKM) by two (fourfold) or more. These criteria resulted in a list of 37 363 unique exons (20% of studied exons), belonging to 11 648 RefSeq transcripts and 6289 genes. These included most of the stable housekeeping genes reported based on microarray data [30].

We define a housekeeping gene as a gene for which at least one RefSeq transcript has more than half of its exons meeting the previous criteria (thus being housekeeping exons). Altogether, we found 3804 such human housekeeping genes. The lists of housekeeping exons and housekeeping genes are available at http://www.tau.ac.il/~elieis/HKG/. In addition, we propose a short list of highly uniform and strongly expressed genes that may be used for calibration in future experimental settings (Table 1).

As expected, the housekeeping genes are enriched in gene ontology (GO) categories associated with basic cellular activity, such as gene expression and biogenesis of nucleotides and amino acids, catabolic processes, protein localization, and so on [51]. The overlap with previous lists is partial, due to the different definition of housekeeping genes used. In particular, GAPDH and actin beta (ACTB) do not appear in our new list, because these genes vary across tissues [3,28–30]. Nevertheless, some of the most pronounced features previously reported for housekeeping genes, such as the much shorter introns [8–11] and more duplications [52], also characterize the new set.

Concluding remarks

Current technology enables global measurement of expression levels with unprecedented accuracy. This advancement has revealed that large parts of the genome are normally expressed at a low level. Accordingly, we found that most human exons are expressed at some level in all the human tissues studied. This new technological era calls the community to reevaluate the concept of a housekeeping gene. Here, we have presented our own perspective, suggesting the use of low expression variation as the main criteria for defining housekeeping genes. We also provide sets of exons and genes that are ubiquitously and uniformly expressed, as well as a short list of genes suitable for experimental calibration.

More high-quality deep-sequencing transcriptome profiling data are expected to emerge in the near future, enabling improvements of the analysis described here using better statistics for the tissues studied and adding more tissue types. Furthermore, including extreme pathological conditions relevant for various tissues could further purify the housekeeping genes list [53]. A significant advance should come from new experiments currently being done on single-cell transcriptome profiling [54]. This could improve the specificity in detecting housekeeping genes, narrowing the list to genes that are expressed in each and every single cell. In addition, accumulation of tissue-specific epigenetic data, such as histone marks and nucleotide methylations, could be used in the future to better distinguish regulated expression from low-level noise.

As discussed above, normalization (within a sample and across samples) is still an unresolved issue. Advancement in this direction could greatly improve housekeeping gene detection. In addition, usage of longer reads is expected to decrease alignment errors and reduce bias. Longer reads (and improved analysis tools) are expected to raise considerably the sensitivity of expression level measurement at the transcript level, enabling direct evaluation of the housekeeping splice-variants list.

In conclusion, the dramatic advancement of sequencing technologies calls for a reassessment of the notion of housekeeping genes, and allows for improving quantitatively and qualitatively the resolution. We thus provide updated lists of housekeeping exons and genes for public use, available at http://www.tau.ac.il/~elieis/HKG/. It is expected that emerging technologies could very soon facilitate meeting the yet open challenges, allowing for better and more accurate housekeeping gene profiling.

Acknowledgments

We thank Ami Haviv and Gilad Finkelstein for help with reads' alignments, and Lily Bazak for help in gene lengths' analysis. This work was supported by Israel Science Foundation 379/12 (EE), by the I-CORE Program of the Planning and Budgeting Committee and the Israel Science Foundation (grant No 41/11) and by the Marie Curie Integration Grant 256593(EYL).

References

1 Fraser, C.M. *et al.* (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* 270, 397–403
2 Koonin, E.V. (2000) How many genes can make a cell: the minimal-gene-set concept. *Annu. Rev. Genomics Hum. Genet.* 1, 99–116
3 Thellin, O. *et al.* (1999) Housekeeping genes as internal standards: use and limits. *J. Biotechnol.* 75, 291–295
4 Robinson, M.D. and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11, R25
5 Dheda, K. *et al.* (2004) Validation of housekeeping genes for normalizing RNA expression in real-time PCR. *Biotechniques* 37, 112–114, 116, 118–119
6 Rubie, C. *et al.* (2005) Housekeeping gene variability in normal and cancerous colorectal, pancreatic, esophageal, gastric and hepatic tissues. *Mol. Cell. Probes* 19, 101–109
7 Vandesompele, J. *et al.* (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* 3, RESEARCH0034
8 Eisenberg, E. and Levanon, E.Y. (2003) Human housekeeping genes are compact. *Trends Genet.* 19, 362–365
9 Vinogradov, A.E. (2004) Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet.* 20, 248–253
10 Carmel, L. and Koonin, E.V. (2009) A universal nonmonotonic relationship between gene compactness and expression levels in multicellular eukaryotes. *Genome Biol. Evol.* 1, 382–390
11 Castillo-Davis, C.I. *et al.* (2002) Selection for short introns in highly expressed genes. *Nat. Genet.* 31, 415–418
12 Eller, C.D. *et al.* (2007) Repetitive sequence environment distinguishes housekeeping genes. *Gene* 390, 153–165
13 Versteeg, R. *et al.* (2003) The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.* 13, 1998–2004
14 Farré, D. *et al.* (2007) Housekeeping genes tend to show reduced upstream sequence conservation. *Genome Biol.* 8, R140
15 Lawson, M.J. and Zhang, L. (2008) Housekeeping and tissue-specific genes differ in simple sequence repeats in the 5′-UTR region. *Gene* 407, 54–62

16 Ganapathi, M. *et al.* (2005) Comparative analysis of chromatin landscape in regulatory regions of human housekeeping and tissue specific genes. *BMC Bioinformatics* 6, 126
17 Lehner, B. and Fraser, A.G. (2004) Protein domains enriched in mammalian tissue-specific or widely expressed genes. *Trends Genet.* 20, 468–472
18 Velculescu, V.E. *et al.* (1999) Analysis of human transcriptomes. *Nat. Genet.* 23, 387–388
19 Zhu, J. *et al.* (2008) How many human genes can be defined as housekeeping with current expression data? *BMC Genomics* 9, 172
20 Zhu, J. *et al.* (2008) On the nature of human housekeeping genes. *Trends Genet.* 24, 481–484
21 Chang, C-W. *et al.* (2011) Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. *PLoS ONE* 6, e22859
22 Hsiao, L.L. *et al.* (2001) A compendium of gene expression in normal human tissues. *Physiol. Genomics* 7, 97–104
23 Lee, S. *et al.* (2007) Identification of novel universal housekeeping genes by statistical analysis of microarray data. *J. Biochem. Mol. Biol.* 40, 226–231
24 She, X. *et al.* (2009) Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. *BMC Genomics* 10, 269
25 Warrington, J.A. *et al.* (2000) Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol. Genomics* 2, 143–147
26 Butte, A.J. *et al.* (2001) Further defining housekeeping, or 'maintenance', genes Focus on 'A compendium of gene expression in normal human tissues'. *Physiol. Genomics* 7, 95–96
27 Irizarry, R.A. *et al.* (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* 31, e15
28 Barber, R.D. *et al.* (2005) GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. *Physiol. Genomics* 21, 389–395
29 Lee, P.D. *et al.* (2002) Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. *Genome Res.* 12, 292–297
30 De Jonge, H.J.M. *et al.* (2007) Evidence based selection of housekeeping genes. *PLoS ONE* 2, e898
31 Kapranov, P. *et al.* (2007) Genome-wide transcription and the implications for genomic organization. *Nat. Rev. Genet.* 8, 413–423
32 Bolstad, B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 185–193
33 Ramsköld, D. *et al.* (2009) An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.* 5, e1000598
34 Modrek, B. and Lee, C. (2002) A genomic view of alternative splicing. *Nat. Genet.* 30, 13–19
35 Johnson, J.M. *et al.* (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* 302, 2141–2144
36 Wang, Z. *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63
37 Fu, X. *et al.* (2009) Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* 10, 161
38 Zhang, Z. *et al.* (2003) Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res.* 13, 2541–2558
39 Mortazavi, A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628
40 Wagner, G.P. *et al.* (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* 131, 281–285
41 Dillies, M-A. *et al.* (2012) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* http://dx.doi.org/10.1093/bib/bbs046
42 Dohm, J.C. *et al.* (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* 36, e105
43 Schwartz, S. *et al.* (2011) Detection and removal of biases in the analysis of next-generation sequencing reads. *PLoS ONE* 6, e16685
44 Li, J. *et al.* (2010) Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol.* 11, R50
45 Jones, D.C. *et al.* (2012) Compression of next-generation sequencing reads aided by highly efficient de novo assembly. *Nucleic Acids Res.* 40, e171
46 Roberts, A. *et al.* (2011) Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* 12, R22
47 Trapnell, C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515
48 Pelechano, V. *et al.* (2013) Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* 497, 127–131
49 Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359
50 Pruitt, K.D. *et al.* (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* 40, D130–D135
51 Huang, D.W. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57
52 Zhang, Z. *et al.* (2004) Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet.* 20, 62–67
53 Chen, M. *et al.* (2013) Identification of human HK genes and gene expression regulation study in cancer from transcriptomics data analysis. *PLoS ONE* 8, e54082
54 Tang, F. *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6, 377–382