HT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets

Bidossessi Wilfried Hounkpe[1], Francine Chenou[1], Franciele Lima[1], Erich Vinicius de Paula[1,2]

Affiliations: [1] School of Medical Sciences, University of Campinas, Campinas, SP, Brazil; [2] Hematology and Hemotherapy Center, University of Campinas, Campinas, SP, Brazil

Corresponding author:

Bidossessi Wilfried Hounkpe, MsC

School of Medical Sciences

University of Campinas

Rua Carlos Chagas, 480

Campinas, SP, Brazil

13083-878

Phone: +55 19 3521 8627

Email: bidossessi1@live.fr

Abstract

Housekeeping (HK) genes are constitutively expressed genes that are required for the maintenance of basic cellular functions. Despite their importance in the calibration of gene expression, as well as the understanding of many genomic and evolutionary features, important discrepancies have been observed in studies that previously identified these genes. Here, we present Housekeeping Transcript Atlas (HT Atlas v1.0, www.housekeeping.unicamp.br) a web-based database which addresses some of the previously observed limitations in the identification of these genes, and offers a more accurate database of human and mouse HK genes and transcripts. The database was generated by mining massive human and mouse RNA-seq data sets, including 12,482 and 507 high-quality RNA-seq samples from 82 human non-disease tissues/cells and 15 healthy tissues/cells of C57BL/6 wild type mouse, respectively. User can visualize the expression and download lists of 2,158 human HK transcripts from 2,176 HK genes and 3,024 mouse HK transcripts from 3,277 mouse HK genes. HT Atlas also offers the most stable and suitable tissue selective candidate reference transcripts for normalization of qPCR experiments. Specific primers and predicted modifiers of gene expression for some of these HK transcripts are also proposed. All of these resources can be accessed and downloaded from any computer or small device web browsers.

Introduction

Housekeeping (HK) genes have been classically defined as genes that are required for the maintenance of basic cellular functions, important for the existence of any cell type. Hence, they are expected to be constitutively expressed in all cell types of the organism in normal physiological condition regardless of specific cell function, cell cycle step or developmental stage (1, 2). Due to these characteristics, HK genes are useful as references of gene expression in molecular biology and computational experiments (3–9), as well as in our understanding of various structural and functional genomics and evolutionary features (10–13). In biomedical research, the importance of the precise identification of HK genes stems from the fact that these genes are used as internal controls for the calibration of quantitative PCR (qPCR), a workhorse technique in molecular biology and biotechnology laboratories used to quantitatively estimate the expression of any gene of interest under different experimental conditions (9, 14). However, the fact that HK genes are used as calibrators for these analyses implicates that the accuracy of qPCR results can be severely jeopardized if an inadequate HK gene is used. Accordingly, the selection of HK genes can well be considered one of the factors associated with the reproducibility crisis that affect biomedical science.

In the early days of qPCR, only a limited number of genes such as *GAPDH*, *B2M*, *HPRT*, actins, tubulins or rRNA genes were considered HK genes, based on their functions associated with cell maintenance, and on the observation that they tended to be constitutively expressed (15–17). Since then, these genes have been extensively used as reference genes for qPCR normalization (18–20). However, more recently, the expression of some of these genes have been shown to vary considerably across cell types and conditions, suggesting that their widespread use might not be the most accurate strategy to calibrate qPCR results (21–23). The advent of high-throughput transcriptomic technologies allowed the identification of larger lists, encompassing hundreds to thousands of putative HK genes (1, 15, 24–27). However, despite representing an improvement in the identification of HK genes, the concordance between these studies is still low (1, 15, 24–26), and false positives remain a problem (2).

Problems with these currently available HK gene lists (1, 15, 24–26) could be associated with limitations in studies that generated them, with the first being the very definition of what a HK gene is. HK genes are normally defined according to two definitions. The first one assumes that HK genes are constitutively expressed in every tissue, with a level above an arbitrary cutoff level, used to distinguish candidate HK genes from noise and/or from weakly expressed parts of the genome (2, 15, 25). The second definition emphasizes that a HK gene should present a

constant and stable expression across all tissue, instead of using a universal cutoff of expression level (1, 2). However, as genes can be expressed at different level in different tissues, the former definition excludes genes that are stably expressed at different levels in different tissues. Even though the second definition extends the first one, both failed to consider alternative splicing, a fundamental aspect of transcriptome complexity. As genes may have one or more isoforms variably expressed across tissues, instead of the "one gene, one polypeptide" concept, it is possible that one gene stably expresses one transcript in a set of tissues or cells, and another transcript in other sets of tissues (1, 28, 29). Therefore, we propose that a refined definition of HK genes should be formulated as: a single constitutive gene that expresses at least one of its protein-coding transcripts at a non-zero expression level, with low variability, which may either be constitutively expressed in all, or in a subset of tissues. In the case that any of these transcripts are constitutively expressed across all tissues, they can be referred to as a HK transcript. Following the standardization of qPCR terminology provided by the minimum information for publication of quantitative real-time PCR experiments (MIQE) guidelines (14), genes used for normalization should be referred to as reference genes, not as HK genes. Accordingly, we use the term "reference transcript" to define transcripts that are suitable for qPCR normalization (i.e., fulfill the criteria of HK transcript and are stably expressed in the specific experimental condition or tissue of interest).

A second limitation of current HK gene lists refers to the fact that they were identified in studies involving a relatively low diversity of tissues and cell types, as well as low sampling, which can introduce false positives (1, 15, 24–26). And finally, technical biases from first generation microarray and Expressed Sequence Tag (EST) sequencing data could have also affected the accuracy of HK gene identification in previous studies. Briefly, besides the problem of hybridization and cross-hybridization artifacts, microarrays are also affected by dye-based detection issues and the capability to detect low abundant genes that can be physiologically relevant. Furthermore, the coverage of all possible genes by first-generation platforms is also very limited (15, 27). EST sequencing data are mainly affected by biases of cDNA libraries cloning processes which determine what transcript sequences are represented (30). In contrast to these technologies, RNA-seq technology overcomes microarray and EST limitation and provides a more comprehensive way to measure transcriptome by ultra-high-throughput sequencing (31). Together these limitations may explain the large discrepancy of currently available HK gene lists, and justify efforts to refine their definition and identification.

To fill this gap we present HT Atlas v1.0, a web-based tool that provides access to a reliable database of human and mouse HK genes and transcripts. By combining high-quality gene

expression data generated with RNA-seq from a large diversity of tissues deposited at two large public databases (GTEx and ARCHS4) (32), with a stringent detection strategy based on a refined and strict definition of what a HK gene is, HT Atlas v1.0 offers to the research community a valuable and accurate tool to the identification of these genes and transcripts for use as calibrators of qPCR experiments, as well as for other research questions.

Data collection

The workflow for the generation of our database (Figure 1) was based on data generated in the GTEx project (version 7) and ARCHS4 (32). These projects provide expression data from RNA-seq in a useful processed format that enables the reuse of their datasets. For the identification of human HK genes, non-normalized transcript level read counts and meta-data from 11,281 samples, including 52 non-disease sites tissues were downloaded from GTEx portal. These data were also used to identify the most reliable candidate reference genes for each cell and tissue. In addition, other 1,201 samples from 30 human tissues and cell types were obtained from ARCHS4 to increase the candidate reference genes database following a rigorous filtering strategy: (i) sequencing depth equal or higher than 20,000,000 reads; (ii) alignment rate provided by ARCHS4 higher than 70%; (iii) library generated from mRNA enrichment protocol; (iv) and construction of paired end read library. These strategies were designed to enable an accurate detection of transcript isoforms. For the identification of mouse HK genes, 507 high-quality RNA-seq data sets from 15 tissues and cells types from wild type healthy C57BL/6 control mice were manually curated from ARCHS4 following the same criteria described above.

Data processing

Identification of HK genes

After downloading of the datasets, read counts were library-size normalized for each cell and tissue types, and RPKM (reads per kilobase million) were calculated using edgeR package in R environment (33, 34). The identification of HK genes was based on the previously mentioned definition of a HK gene: specifically, to be considered a HK gene, a single gene must have fulfilled the following criteria: (i) the gene must be expressed at non-zero level in all tissue and cell types included in the analysis (that is, at least one of its protein-coding transcripts must have an expression level higher than 1 RPKM); (ii) the variability of transcript expression should be low within all tissues and cell types, as evidence by a standard deviation of the log2 RPKM <

1; (iii) the maximum fold change (MFC), represented by the ratio between maximum and minimum log2 RPKM of the transcript ((maximum log2 RPKM)/(minimum log2 RPKM)), must be lower than 2. Only transcripts with well-supported transcript models were included in the database (transcripts with Refseq and/or the Consensus Coding Sequence (CCDS) project annotation). Finally, transcripts belonging to these HK genes were considered HK transcripts if they were, in addition, constitutively expressed in all of the analyzed tissues and cells types that were analyzed.

Identification of candidate reference transcripts

The conventional strategy of qPCR normalization used the expression of reference genes to calibrate the expression levels of genes of interest. Here we observed that only a few transcripts per HK gene fulfilled our criteria and were defined as HK transcripts. Thus, a more accurate strategy would be the normalization of qPCR experiments using reference transcripts instead of the currently used "gene model". In order to test this hypothesis, we summarized the transcript RPKM into gene model as previously described (35). Then, genes RPKM were further submitted to our HK genes identification criteria. Overall, 1,003 genes passed our criteria, of which only 823 were common with HT Atlas v1.0 HK genes list (Table S1). This reinforces the adequacy of our suggested normalization strategy. Further qPCR experiments are being performed to demonstrate the reliability of the transcript-based method for determining candidate reference genes.

So, to refine the list of transcripts that can be further validated as reference transcripts in qPCR experiments, reliable transcripts were selected for each tissue. Briefly, the list of identified HK transcripts was used to provide suitable reference transcripts for selected cells and tissues, for qPCR experiment normalization. Candidate reference transcripts had presented a standard deviation of the $\log_2$ RPKM < 0.5 and mean of RPKM greater or equal to 30. Finally, to ensure reliability of qPCR experiments, isoforms of genes with known pseudogenes were also excluded from the reference transcripts lists (36, 37).

Candidate reference transcript ranking system: Score Product

Following candidate reference transcript identification, an algorithm based on the determination of a "Score product" was developed to rank these transcripts in each tissue or cell type. The algorithm was built to prioritize the most reliable transcripts for qPCR normalization. Candidate reference transcript stability was ranked based on mean RPKM,

standard deviation of $\log_2$ RPKM and MFC value. Giving *n* transcripts, three intermediate scores (*k=3*) with equal weight were generated (score-1, score-2, and score-3 for RKPM, standard deviation and MFC, respectively for each transcript). Score-1 assigns *n* to the highest expressed transcript, estimated by mean RPKM, and 1 to the lowest expressed. Similarly, for score-2 and score-3 the highest scoring transcripts are the ones that have the lowest standard deviation or the lowest MFC. Given $sc_{t,i}$ the score of the transcript t for the *i*-th rank, we express the Score Product (SP) via the geometric mean:

$$\text{SP (t)} = \left(\prod_{i=1}^{k} sc_{t,i}\right)^{\frac{1}{k}}$$

The highest scoring transcript is one that has the highest SP and is assigned to the first position (rank = 1). That way, the *n*-th transcript is the one that has the lowest SP. The Score Product is proportional to the variability.

Primer design

In addition to the identification and ranking of candidate reference transcripts, we also designed specific primers that can be used for qPCR normalization. These candidate reference transcript-specific primers were designed to facilitate the reference transcript-based normalization strategy that we are suggesting. Primer-pairs were designed to target amplicons with length spanning between 75 and 200bp, and templates with long repeats of single bases (>4) were avoided. Each primer of a primer-pair was quality checked with the following criteria: (i) GC content must be between 50 and 60%; (ii) the melting temperature (Tm) must be between 50°C and 65°C; and (iii) secondary structure and primer-dimer formation were avoided. A least one of the primer-pairs was designed to span two exon junctions, and their transcript level specificity was checked using In-Silico PCR (38).

Integration of HT Atlas with predicted gene expression modifiers

In order to further guide the choice of suitable candidate reference genes for qPCR normalization, our list of human HK genes was integrated with the following resources extracted from the Harmonizome platform (39), which were downloaded and manually curated: *GEO Signatures of Differentially Expressed Genes for Diseases, GEO Signatures of Differentially Expressed Genes for Small Molecules and Connectivity Map Signatures of*

Differentially Expressed Genes for Small Molecules. These datasets were used to provide users with a list of small molecules and diseases which have been previously shown to modify the expression of each HK gene, when available.

Implementation and data access

All of the generated data were loaded into SQLite database. HT Atlas v1.0 web tool was developed using Shiny package (40). The application front-end was implemented using HTML, CSS and Bootstrap 4. All of the components and a Shiny Server were encapsulated in a deployable Docker container running on a Unix-based operating system with Apache HTTP server. User instructions communicated to the server were processed by R and the requests are sent to the SQLite database mounted as the container volume. Figure 1 schematically represents the general workflow of the resource (Figure 1A). HT Atlas v1.0 offers a user friendly and reactive interface designed to provide a nice using experience. The web tool was tested in Firefox, Chrome, Explorer, Safari as well as android and iOS small devices.

Database contents and features

HT Atlas v1.0 statistics

HT Atlas analyzed 12,482 and 507 human and mouse samples respectively, involving 82 human cell and tissue types and 14 mouse cell and tissue types. After processing, 2,176 unique genes fulfilled our criteria and were defined as HK genes, of which 2,158 HK transcripts were constitutively expressed. In the mouse database, 3,277 HK transcripts and 3,024 HK genes fulfilled our criteria. At least two HK transcript-specific primers were designed per tissue. Overall, 44 and 23 primer-pairs were designed for human and mouse candidate reference genes, respectively.

Furthermore, 52% of the human HK genes present their orthologous in the mouse HK list (the full lists can be downloaded from Download page). We compared our list of HK genes with previously published lists which were also based on RNA-seq data sets (1, 25, 26). As shown in figure 2A, only 91 human HK genes were exclusively identified in the present study (Fig 2A). Similarly, a high overlap was also observed between previous mouse HK genes lists (41, 42) and our HT Atlas database (Fig 2B). Importantly, thousands of genes identified in these previous studies didn't fulfill our stringent criteria. These results support our hypothesis that

these previous lists of HK genes, generated with low sample size and heterogeneous definitions of HK genes, might suffer from low specificity.

Web interface

From the home interface users can access human and mouse candidate reference genes databases by clicking on the human or mouse icons. The search interface allows searching for reliable candidate reference transcripts for qPCR normalization by selecting one of the 82 tissues or cell types included in our database. By clicking on the search box, the full list of cells and tissues is automatically displayed on the screen and user can select one option to interrogate the SQLite database. From the same interface users can also access link to visualize expression of HK transcripts across cells/tissues, or to download the complete list. A similar interface is provided for mouse database.

In each of these interfaces, after searching for a specific cell or tissue type, the results page shows the list of candidate reference transcripts. Each transcript is identified by its Ensembl ID, followed by the gene symbol. By default, results are ranked by the SP criteria described above. In addition, a color code is used to categorize transcripts according to expression levels (highly and moderately expressed in green and turquoise respectively). Alternatively, transcripts can be ranked at the convenience of the user by clicking in the arrow next to any of these parameters (Figure 3A). Users can select the number of candidate reference transcripts to be shown per page, and then download the full list displayed on the screen in their preferred format (csv, Excel or pdf). Specific primer lists can also be downloaded.

As diseases and small molecules are known to affect the stability of housekeeping genes, HT Atlas is integrated with three manually curated functional association data sets from Harmonizome database that predict expression stability modifiers based on differential expression, and are intended to guide users in the choice of the best candidate reference genes for qPCR normalization. A gene must be avoided for normalization of qPCR if the disease or the molecule of interest appears in its modifiers list (Figure 3B). These datasets can be accessed from the results page. Alternatively, when clicking on “Expression Modifiers” in the navigation bar, users can access the same information for any single gene.

Users can access the “Gene expression” menu from the navigation bar to visualize gene expression across tissue and cell types. This interface also provides the description of a selected HK gene, its official name and its synonyms retrieved from the NCBI's portal and

based on RefSeq project. When a gene was identified as a reliable candidate reference gene for a subset of tissues and cells, a list of that tissues and cell types is also provided (Fig 4).

Finally, by clicking on the "Download" menu from the navigation bar, users can download the complete list of human and mouse HK genes and HK transcripts. From the same page users can download a list of the most stable genes across all non-disease analyzed tissues. These genes can be validated in cell types not included in our database for further normalization purposes.

In conclusion, HT Atlas v1.0 represents a valuable tool for researchers from a wide range of fields in biomedical research, due to its capability to refine the identification of a critical parameter (i.e. the gene used to calibrate expression level reads) in one of the most commonly used techniques of molecular biology studies. The database can also be used to assist in research questions about structural and functional genomics that require a more precise identification of human and mouse HK genes and transcripts. Our strategy for the future will focus in including more cells and tissues into the database. We are also planning to analyze, using the same workflow, samples from other model organisms. Finally, the prediction of gene modifiers will be improved in future versions. Because there is no way to know whether gene expression will change under all experimental and disease conditions, we recommend that every research group validate empirically the proposed candidate reference transcripts before using in their own qPCR experiments.

Funding

This study was financially supported by the Sao Paulo Research Foundation. grants # 2016/14172-6, 2014/0984-3 and 2015/24666-3; CNPq Brazil. grant # 309317/2016.

Acknowledgments

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Part of the data used for the analyses described in this manuscript was obtained from GTEx v7.

Data availability

All data are available from HT Atlas (http://www.housekeeping.unicamp.br). Processing codes and source codes are available from github (https://github.com/Bidossessih/housekeepingAtlas).

References

1. Eisenberg,E. and Levanon,E.Y. (2013) Human housekeeping genes, revisited. *Trends Genet.*, 10.1016/j.tig.2013.05.010.

2. Zhang,Y., Li,D. and Sun,B. (2015) Do housekeeping genes exist? *PLoS One*, 10.1371/journal.pone.0123691.

3. Teng,M., Love,M.I., Davis,C.A., Djebali,S., Dobin,A., Graveley,B.R., Li,S., Mason,C.E., Olson,S., Pervouchine,D., *et al.* (2016) A benchmark for RNA-seq quantification pipelines. *Genome Biol.*, 10.1186/s13059-016-0940-1.

4. Zyprych-Walczak,J., Szabelska,A., Handschuh,L., Górczak,K., Klamecka,K., Figlerowicz,M. and Siatkowski,I. (2015) The Impact of Normalization Methods on RNA-Seq Data Analysis. *Biomed Res. Int.*, 10.1155/2015/621690.

5. Ou,J., Liu,H., Yu,J., Kelliher,M.A., Castilla,L.H., Lawson,N.D. and Zhu,L.J. (2018) ATACseqQC: A Bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC Genomics*, 10.1186/s12864-018-4559-3.

6. Crow,M., Lim,N., Ballouz,S., Pavlidis,P. and Gillis,J. (2019) Predictability of human differential gene expression. *Proc. Natl. Acad. Sci.*, 10.1073/pnas.1802973116.

7. Monaco,G., Lee,B., Xu,W., Mustafah,S., Hwang,Y.Y., Carré,C., Burdin,N., Visan,L., Ceccarelli,M., Poidinger,M., *et al.* (2019) RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. *Cell Rep.*, 10.1016/j.celrep.2019.01.041.

8. Ratnapriya,R., Sosina,O.A., Starostik,M.R., Kwicklis,M., Kapphahn,R.J., Fritsche,L.G., Walton,A., Arvanitis,M., Gieser,L., Pietraszkiewicz,A., *et al.* (2019) Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. *Nat. Genet.*, 10.1038/s41588-019-0351-9.

9. Pfaffl,M.W. (2001) A new mathematical model for relative quantification in real-time RT-

PCR. *Nucleic Acids Res.*

10. Nie,H., Crooijmans,R.P.M.A., Lammers,A., van Schothorst,E.M., Keijer,J., Neerincx,P.B.T., Leunissen,J.A.M., Megens,H.J. and Groenen,M.A.M. (2010) Gene expression in chicken reveals correlation with structural genomic features and conserved patterns of transcription in the terrestrial vertebrates. *PLoS One*, 10.1371/journal.pone.0011990.

11. Kouadjo,K.E., Nishida,Y., Cadrin-Girard,J.F., Yoshioka,M. and St-Amand,J. (2007) Housekeeping and tissue-specific genes in mouse tissues. *BMC Genomics*, 10.1186/1471-2164-8-127.

12. Zhang,L. and Li,W.H. (2004) Mammalian Housekeeping Genes Evolve More Slowly than Tissue-Specific Genes. *Mol. Biol. Evol.*, 10.1093/molbev/msh010.

13. Zhu,J., He,F., Hu,S. and Yu,J. (2008) On the nature of human housekeeping genes. *Trends Genet.*, 10.1016/j.tig.2008.08.004.

14. Bustin,S.A., Benes,V., Garson,J.A., Hellemans,J., Huggett,J., Kubista,M., Mueller,R., Nolan,T., Pfaffl,M.W., Shipley,G.L., *et al.* (2009) The MIQE guidelines: Minimum information for publication of quantitative real-time PCR experiments. *Clin. Chem.*, 10.1373/clinchem.2008.112797.

15. Warrington,J.A., Nair,A., Mahadevappa,M. and Tsyganskaya,M. (2000) Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol. Genomics*, 10.1152/physiolgenomics.2000.2.3.143.

16. Rifkind,R.A., Marks,P.A., Bank,A., Terada,M., Maniatis,G.M., Reuben,R. and Fibach,E. Erythroid differentiation and the cell cycle: some implications from murine foetal and erythroleukemic cells. *Ann. Immunol. (Paris).*, **127**, 887–93.

17. Perfetti,V., Manenti,G. and Dragani,T.A. (1991) Expression of housekeeping genes in Hodgkin's disease lymph nodes. *Leukemia*, **5**, 1110–2.

18. Pallisgaard,N., Clausen,N., Schrøder,H. and Hokland,P. (1999) Rapid and sensitive minimal residual disease detection in acute leukemia by quantitative real-time RT-PCR exemplified by t(12;21) TEL-AMLI fusion transcript. *Genes Chromosom. Cancer*, 10.1002/(SICI)1098-2264(199912)26:4<355::AID-GCC10>3.0.CO;2-3.

19. Cance`f,W.G., Craven`$,R.J. and Liuft,E.T. (1992) Expression polymerase chain reaction : a sensitive method for analysis of gene expression in human tumours.

20. Laurendeau,I., Bahuau,M., Vodovar,N., Larramendy,C., Olivi,M., Bieche,I., Vidaud,M. and Vidaud,D. (1999) TaqMan PCR-based gene dosage assay for predictive testing in individuals from a cancer family with INK4 locus haploinsufficiency. *Clin. Chem.*

21. Kosinová,L., Cahová,M., Fábryová,E., Týcová,I., Koblas,T., Leontovyè,I., Saudek,F. and Køíž,J. (2016) Unstable expression of commonly used reference genes in rat pancreatic islets early after isolation affects results of gene expression studies. *PLoS One*, 10.1371/journal.pone.0152664.

22. de Jonge,H.J.M., Fehrmann,R.S.N., de Bont,E.S.J.M., Hofstra,R.M.W., Gerbens,F., Kamps,W.A., de Vries,E.G.E., van der Zee,A.G.J., te Meerman,G.J. and ter Elst,A. (2007) Evidence based selection of housekeeping genes. *PLoS One*, 10.1371/journal.pone.0000898.

23. Dheda,K., Huggett,J.F., Bustin,S.A., Johnson,M.A., Rook,G. and Zumla,A. (2004) Validation of housekeeping genes for normalizing RNA expression in real-time PCR. *Biotechniques*.

24. Fagerberg,L., Hallström,B.M., Oksvold,P., Kampf,C., Djureinovic,D., Odeberg,J., Habuka,M., Tahmasebpoor,S., Danielsson,A., Edlund,K., *et al.* (2014) Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics*, 10.1074/mcp.M113.035600.

25. Ramsköld,D., Wang,E.T., Burge,C.B. and Sandberg,R. (2009) An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.*, 10.1371/journal.pcbi.1000598.

26. Uhlén,M., Fagerberg,L., Hallström,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,Å., Kampf,C., Sjöstedt,E., Asplund,A., *et al.* (2015) Tissue-based map of the human proteome. *Science (80-.).*, 10.1126/science.1260419.

27. HSIAO,L.-L., DANGOND,F., YOSHIDA,T., HONG,R., JENSEN,R. V., MISRA,J., DILLON,W., LEE,K.F., CLARK,K.E., HAVERTY,P., *et al.* (2015) A compendium of gene expression in normal human tissues. *Physiol. Genomics*, 10.1152/physiolgenomics.00040.2001.

28. Gingeras,T.R. (2007) Origin of phenotypes: Genes and transcripts. *Genome Res.*, 10.1101/gr.6525007.

29. Gerstein,M.B., Bruce,C., Rozowsky,J.S., Zheng,D., Du,J., Korbel,J.O., Emanuelsson,O., Zhang,Z.D., Weissman,S. and Snyder,M. (2007) What is a gene, post-ENCODE? History

and updated definition. *Genome Res.*, 10.1101/gr.6339607.

30. Sultan,M., Schulz,M.H., Richard,H., Magen,A., Klingenhoff,A., Scherf,M., Seifert,M., Borodina,T., Soldatov,A., Parkhomchuk,D., *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science (80-.).*, 10.1126/science.1160342.

31. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, 10.1038/nmeth.1226.

32. Lachmann,A., Torre,D., Keenan,A.B., Jagodnik,K.M., Lee,H.J., Wang,L., Silverstein,M.C. and Ma'ayan,A. (2018) Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.*, **9**, 1366.

33. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2009) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 10.1093/bioinformatics/btp616.

34. R Core Team (2018) R: A language and environment for statistical computing. R Foundation for Statistical Computing. *Vienna, Austria*, 10.1108/eb003648.

35. Zhao,S., Xi,L. and Zhang,B. (2015) Union exon based approach for RNA-seq gene quantification: To be or not to be? *PLoS One*, 10.1371/journal.pone.0141910.

36. Williams,T.K., Yeo,C.J. and Brody,J. (2008) Does this band make sense? Limits to expression based cancer studies. *Cancer Lett.*, 10.1016/j.canlet.2008.05.033.

37. Sun,Y., Li,Y., Luo,D. and Liao,D.J. (2012) Pseudogenes as weaknesses of ACTB (Actb) and GAPDH (Gapdh) used as reference genes in reverse transcription and polymerase chain reactions. *PLoS One*, 10.1371/journal.pone.0041659.

38. Rhead,B., Karolchik,D., Kuhn,R.M., Hinrichs,A.S., Zweig,A.S., Fujita,P.A., Diekhans,M., Smith,K.E., Rosenbloom,K.R., Raney,B.J., *et al.* (2009) The UCSC genome browser database: Update 2010. *Nucleic Acids Res.*, 10.1093/nar/gkp939.

39. Rouillard,A.D., Gundersen,G.W., Fernandez,N.F., Wang,Z., Monteiro,C.D., McDermott,M.G. and Ma'ayan,A. (2016) The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database (Oxford).*, 10.1093/database/baw100.

40. Ishimaru,S., Weppner,J., Poxrucker,A., Kunze,K., Lukowicz,P. and Kise,K. (2014) Shiny. In.

41. Li,B., Qing,T., Zhu,J., Wen,Z., Yu,Y., Fukumura,R., Zheng,Y., Gondo,Y. and Shi,L. (2017) A Comprehensive Mouse Transcriptomic BodyMap across 17 Tissues by RNA-seq. *Sci. Rep.*, 10.1038/s41598-017-04520-z.

42. Zeng,J., Liu,S., Zhao,Y., Tan,X., Aljohi,H.A., Liu,W. and Hu,S. (2016) Identification and analysis of house-keeping and tissue-specific genes based on RNA-seq data sets across 15 mouse tissues. *Gene*, 10.1016/j.gene.2015.11.003.

Figure legends

Figure 1: Housekeeping Transcript Atlas v1.0 workflow

The workflow of database generation and web-tool implementation are shown. After downloading and identification of HK genes using a specific algorithm, transcripts of genes with unknown pseudogenes were used to select suitable candidate reference transcripts. Some candidate reference transcript-specific primers were designed. The web-based tool was developed using Shiny package and encapsulated into docker image for deployment.

Figure 2: Overlap of HT Atlas with other HK genes lists. (A) The diagram demonstrates that only 91 human genes were exclusively detected in the present study. In contrast, thousands of putative HK genes that have been previously listed by other authors (listed in the diagrams) were not identified by our criteria and are not listed in our HT Atlas. (B) Similarly, when comparing HT Atlas with previous mouse HK genes studies, 141 genes were exclusively detected in the present study, while several others which were listed in previous studies did not reach the criteria established in HT Atlas.

Figure 3: Screenshot of the candidate reference transcript search result page. (A) The result page presents a tab with three clickable menus (Reference Transcripts, Expression Modifiers, and Specific Primers). When clicking into this menu, the specified information is displayed. Panel A displays a list of 10 HK transcripts, ranked according to HT Atlas algorithm. HK transcripts are identified by Ensembl ID and gene symbol. Quantitative parameters of gene expression are listed. (B) In the lower panel, a list of five potential modifiers (small molecules) of the expression of a specific HK gene is shown.

Figure 4: Screenshot of the expression visualization page. For each selected HK gene, information such as gene nomenclature, synonyms and a list of tissues with stable expression are listed. Box plots depicting transcript expression in different cells and tissues using data from our database are also shown. Boxplots were generated using plotly package (Carson Sievert (2018) plotly for R. https://plotly-r.com).

Figures

Figure 1: Housekeeping Transcript Atlas workflow

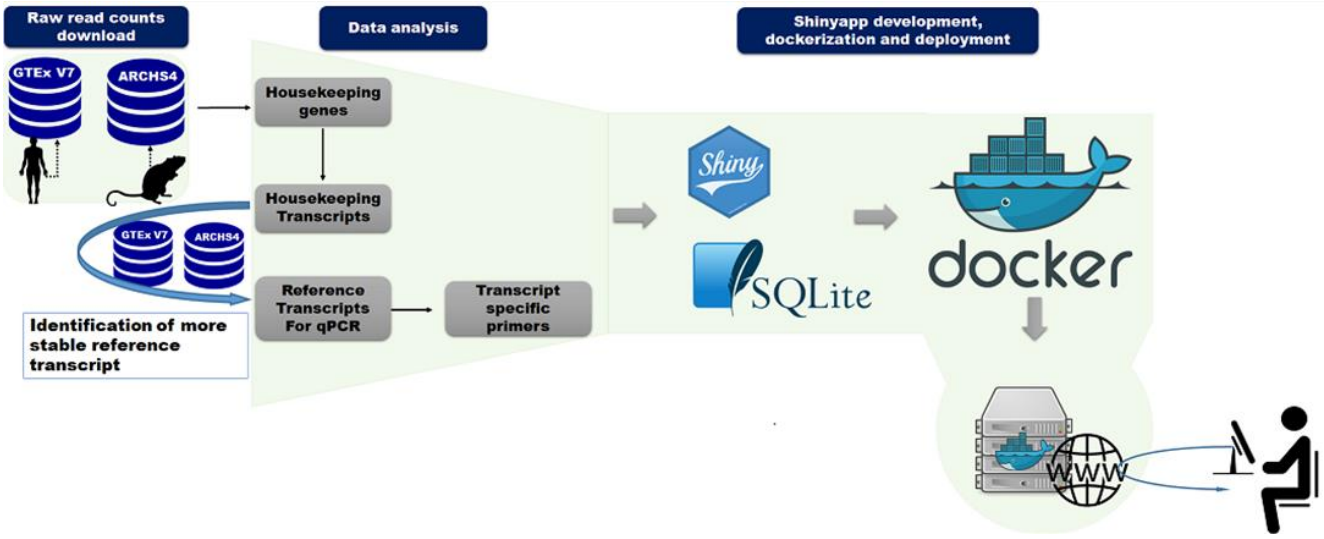


Figure 2: Overlap of HT genes between HT Atlas v 1.0 and other HK genes lists

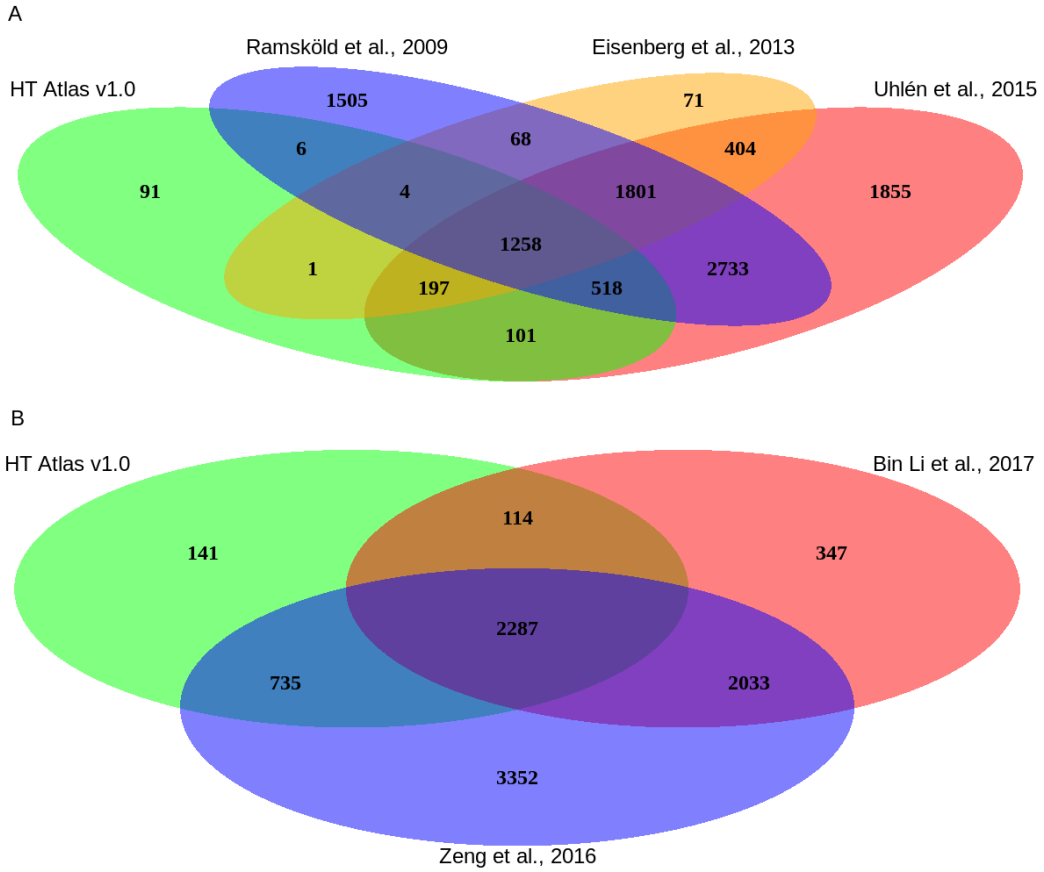


Figure 3: Screenshot of the candidate reference transcript search result page.

A

Reference Transcripts | Expression Modifiers | Specific Primers

Copy | Print | Download | Show 10 entries | Filter:

12 Transcripts found from 407 high quality Whole Blood samples. Highly expressed transcripts based on RPKM are highlighted in green and moderatly expressed transcripts in blue.

Rank	Ensembl ID	Gene	Normalized Expression	Std Deviation	MFC
1	ENST00000309311	EEF2	302	0.5	1.4
2	ENST00000418115	RHOA	264	0.49	1.39
3	ENST00000303577	PCBP1	217	0.43	1.46
4	ENST00000237654	CCNI	99	0.46	1.6
5	ENST00000378609	GNB1	72	0.38	1.48
6	ENST00000295702	SSR2	108	0.48	1.65
7	ENST00000356674	HNRNPA2B1	87	0.47	1.62
8	ENST00000222305	USF2	51	0.44	1.59
9	ENST00000375882	CSNK2B	70	0.49	1.66
10	ENST00000298289	RPL36AL	79	0.5	1.93

Showing 1 to 10 of 12 entries

Previous 1 2 Next

B

Small molecule perturbations changing expression of EEF2 gene from the GEO Signatures of Differentially Expressed Genes for Small Molecules dataset.

GEO Signatures of Differentially Expressed Genes for small Molecules
amoxicillin_rattus norvegicus_gpl341_proximal small intestine_gds1273
arsenic trioxide_homo sapiens_gpl201_gse48441
candesartan_rattus norvegicus_gpl85_gds2070
candesartan_rattus norvegicus_gpl85_gse2739
captopril_mus musculus_gpl1261_gds3683

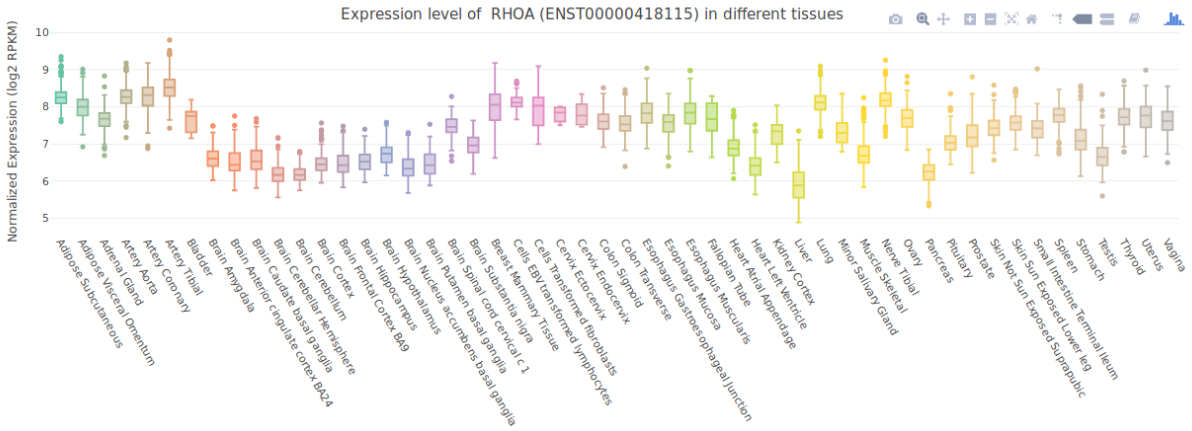
Showing 6 to 10 of 102 entries

Previous 1 2 3 4 5 ... 21 Next

Figure 4: Screenshot of the expression visualization page.

Select a human housekeeping gene to show its transcript expression in different tissues/cells

RHOA (ENST00000418115)

RHOA

Gene name	Ras homolog family member A provided by HGNC
Other names	ARHA; ARH12; RHO12; RHOH12
Description	This gene encodes a member of the Rho family of small GTPases, which cycle between inactive GDP-bound and active GTP-bound states and function as molecular switches in signal transduction cascades. Rho proteins promote reorganization of the actin cytoskeleton and regulate cell shape, attachment, and motility. Overexpression of this gene is associated with tumor cell proliferation and metastasis. Multiple alternatively spliced variants have been identified. [provided by RefSeq, Sep 2015]
ENST00000418115 Reference transcript Proposed for normalization of:	Adipose Subcutaneous; Adipose Visceral Omentum; Adrenal Gland; Artery Aorta; Artery Coronary; Artery Tibial; Bladder; Brain Amygdala; Brain Anterior cingulate cortex BA24; Brain Caudate basal ganglia; Brain Cerebellar Hemisphere; Brain Cerebellum; Brain Cortex; Brain Frontal Cortex BA9; Brain Hippocampus; Brain Hypothalamus; Brain Nucleus accumbens basal ganglia; Brain Putamen basal ganglia; Brain Spinal cord cervical c 1; Brain Substantia nigra; Breast Mammary Tissue; Cells EBV transformed lymphocytes; Cells Transformed fibroblasts; Cervix Ectocervix; Cervix Endocervix; Colon Sigmoid; Colon Transverse; Esophagus Gastroesophageal Junction; Esophagus Mucosa; Esophagus Muscularis; Heart Atrial Appendage; Heart Left Ventricle; Kidney Cortex; Liver; Lung; Minor Salivary Gland; Muscle Skeletal; Nerve Tibial; Ovary; Pancreas; Pituitary; Prostate; Skin Not Sun Exposed Suprapubic; Skin Sun Exposed Lower leg; Small Intestine Terminal Ileum; Spleen; Stomach; Testis; Thyroid; Uterus; Vagina; Whole Blood; Bladder cancer cell line; Caco2; Human Tracheobronchial Epithelial Cells; Normal primary human bronchial epithelial NHBE; Pulmonary microvascular endothelial cells; small airway epithelial cells; Dendritic