

**Do Context-specific Housekeeping Genes exist?**

by

HAN Ye Ri (UID: 3035603370)

Supervisor: Dr. David SHIH Jen Hao

A dissertation submitted in partial fulfilment of the requirements for the Degree of Bachelor of Biomedical Sciences at The University of Hong Kong

April 2023

**ACKNOWLEDGMENT**

I would like to express my deepest and sincere gratitude to Dr David Shih, who guided me through the project. He has been an invaluable mentor, teacher and sometimes even a map on lost days throughout the past year. I would like to take this opportunity to thank you and tell you that I mean everything I said. This work would not have been possible without your patience, support, encouragement, and guidance. As a beginner, I may probably not have been the most skilful student of yours, but I will never forget what you taught me. I am so grateful to have met a passionate teacher, who keeps thinking about and coming up with new ways to improve the teaching and learning experience. I learned the attitude of a good learner from him.

Thank you to all of the members of SHIH lab. Especially Noel Yue, I appreciate your kindness and offering of your hand to me every day. Thank you to Sum Yi Wong, my lab mate. I am happy that we met and walked this path together.

Thank you to Dr Jason Wong, my academic advisor, for your support and practical advice. Thank you for making me feel that I always have someone that welcomes me to find him and consult in HKU.

A special thanks to JM and my best friend Grace Yang for your love and support.

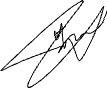
Dear my family, my beginning and end, thank you for your love and your presence.

Last but not least, I would like to thank all of the professors and faculty members that helped me complete this chapter of my undergraduate life.

**DECLARATION**

I confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis. The report has not been accepted for any degree and is not currently used for publication.

Signature:



Name: HAN Ye Ri

Date: 28th April 2023

Table of Contents

[ABSTRACT 4](#_Toc134034494)

[INTRODUCTION 4](#_Toc134034495)

[I.A NORMALIZING CONTROL 6](#_Toc134034496)

[(1) Purpose of measuring expression level 6](#_Toc134034497)

[(2) qPCR needs housekeeping genes as internal control 6](#_Toc134034498)

[(3) Western blot needs housekeeping genes as loading control 7](#_Toc134034499)

[II.HOUSEKEEPING GENES 7](#_Toc134034500)

[(1) Role of housekeeping gene 7](#_Toc134034501)

[(2) Common housekeeping genes 8](#_Toc134034502)

[(3) Housekeeping genes are widely used as internal control in gene expression analysis 8](#_Toc134034503)

[III.CONTEXT-SPECIFIC HOUSEKEEPING GENES 9](#_Toc134034504)

[(1) There is variation in the housekeeping gene expression 9](#_Toc134034505)

[(2) A need for identifying a context-specific housekeeping gene 9](#_Toc134034506)

[IV.SIGNIFICANCE OF THIS STUDY 10](#_Toc134034507)

[MATERIALS AND METHODS 11](#_Toc134034508)

[I.SELECTION OF DATASET 11](#_Toc134034509)

[II.Computational analysis 11](#_Toc134034510)

[(1) Pre-processing 11](#_Toc134034511)

[(2) Statistical analysis 13](#_Toc134034512)

[(3) Interpretation of the results to find the best candidate 14](#_Toc134034513)

[RESULTS 15](#_Toc134034514)

[I.Trend in between-group standard deviation and within-group standard deviation 15](#_Toc134034515)

[II.trend between within-group standard deviation and mean expression 18](#_Toc134034516)

[III.Selection of candidate context-specific housekeeping genes 19](#_Toc134034517)

[IV.Analysis of Housekeeping genes to find the best candidate 21](#_Toc134034518)

[(1) GAPDH 21](#_Toc134034519)

[(2) LDHA 26](#_Toc134034520)

[(3) PTBP1 27](#_Toc134034521)

[(4) HNRNPC 28](#_Toc134034522)

[(5) ACTB and UBC 29](#_Toc134034523)

[(6) g:Profiler results for candidate housekeeping genes 33](#_Toc134034524)

[DISCUSSION 34](#_Toc134034525)

[I.Interpretation of the results 34](#_Toc134034526)

[II.limitations and future directions 35](#_Toc134034527)

[CONCLUSION 36](#_Toc134034528)

[ABBREVIATIONS 37](#_Toc134034529)

[LIST of FIGURES 38](#_Toc134034530)

[LIST of TABLES 40](#_Toc134034531)

[REFERENCES 41](#_Toc134034532)

# ABSTRACT

Housekeeping genes refer to a group of genes in living organisms that are ubiquitously expressed at a certain level regardless of any conditions. Their expression levels are maintained as they perform the very basic and vital functions of the cell, including cell structure integrity maintenance, movement, energy metabolism and apoptosis. Their consistent expression at relatively high levels across diverse tissue, regardless of the specialization of the cell, makes them suitable references against which other gene expression measurements can be normalized in qPCR, Western blotting, RNA-seq and more molecular biology experiments. However, traditionally-used reference housekeeping gene expressions are, in fact, found to vary across tissue types, diseases, development stages and experiment conditions in recent studies. Context-specific housekeeping genes are the housekeeping genes that are expressed at consistent levels in a particular cell type and stage of development of the cell and invariant across the tissues and disease status. 10982 samples from the TCGA PanCan dataset are divided into groups of different cancer and tissue types and the expressions of 20386 genes are statistically analysed to find a particular tissue type the gene is oddly upregulated or downregulated from the mean. The standard deviations within and between the groups are calculated. The genes with low within-group and between-group standard deviation and high mean expressions are filtered. The genes whose expression is consistent within the group while showing deviated expression from the mean in some of the phenotypic groups were sorted out. After eliminating these groups, the recalculated between-group standard deviation dropped. 123 new candidates of housekeeping genes including *PTBP1* and *HRNNPC* are found and they all showed a pattern of improved consistency in expression level by becoming context-specific. Identifying and characterizing more context-specific housekeeping genes will help build a more valid design of quantification experiments in the biomedical field.

**Keywords:** housekeeping genes, context-specific, within-group standard deviation, between-group standard deviation

# INTRODUCTION

Human DNA contains genetic information and biological instructions for transcription and protein synthesis. DNA content in each cell is the same but the cell can make use of epigenetic mechanisms to develop all sorts of shapes and functions. The gene expression can be flexible depending on the time of the measurement and the type of cell or tissue. Variation in gene expression at the cellular differentiation level commits to deciding the type of the cell and the function of the tissue. The activity or expression of the genes is switched on or off or regulated by complicated interactions and the balance of different transcription factors. Successful regulation of the genes ensures normal cell cycle and functioning and survival of the cell.

On the other hand, in the case of genes essential for the survival of the cell, they are expressed consistently in any tissue or cell. Such genes are called housekeeping genes. They are involved in cell basal metabolic activities including expression, DNA replication, cell division and energy metabolism(1). In other words, the housekeeping gene is a gene that is crucial for maintaining cell life activities and functions regardless of certain conditions, and has very little change in expression, even in various conditions and developmental stages.

Measuring and comparing gene expression levels to an appropriate control is an important approach to finding the disease pathology and studying potential treatments. Problems in the expression of housekeeping genes destine cells to death. Their stable, always-everywhere property makes them useful internal controls to ensure accurate gene expression measurements and comparison of other genes of interest.

By quantifying mRNA and protein levels and normalizing the result to the expression of housekeeping genes, a better understanding of how genes are expressed and their role in the particular development stage of abnormal conditions can be achieved. The most commonly used reference housekeeping genes include *GAPDH, ACTB, LDHA* and *UBC.* This molecular-level study makes it possible to deduce different target genes or proteins for diseases and disorders based on the function and the interconnected regulation system of the genes.

However, numerous empirical evidence points have reported that there is significant variation in housekeeping gene expression across tissue types. During the analysis of the expression of a gene, the expression of the target gene in microarray and RNA-seq results has to be normalized through dividing them by the expression of the housekeeping gene in each condition and compared directly, but the relative difference may vary depending on the selection of the housekeeping gene. It necessitates the identification of universal housekeeping genes, of which expression level is inflexible across diverse tissues, and the characterization of context-specific housekeeping genes.

Here, the article will highlight the usefulness of housekeeping genes as an internal control in mRNA and protein expression studies and the impact of expression fluctuation based on various contexts.

## A NORMALIZING CONTROL

### **Purpose of measuring the expression level**

Understanding the central dogma and discovering the polymerase chain reaction (PCR) technique and Taq polymerase led to a drastic development in genetic studies (2, 3). Its usefulness and application range massively from the identification of alleles or genotyping to quantifying mRNA transcription by Real-Time PCR or qPCR. Quantifying the amount of mRNA transcripts and protein product of a gene respectively in qPCR and Western Blot assumes that the results represent their expression level.

Selecting reference genes for normalization simply from the most frequently used housekeeping gene without considering the characteristics of the tissues or cells under study can cause errors in the research results. It may lead to misinterpretation of the physiological mechanism of the research interest.

### qPCR needs housekeeping genes as an internal control

qPCR, an experimental method for gene quantification, can measure gene expression by monitoring changes in gene expression in real-time. SYBR Green is the most common fluorescent dye characterized by binding to the minor groove of dsDNA. It means the amount of DNA is proportional to the intensity of the fluorescent signal. Therefore, by measuring the increase of absorbance of the fluorescence, the amount of DNA is quantified in real-time.

The sensor detects the doubling intensity of fluorescence and draws in the form of an exponential function. The greater the amount of initial DNA, the fewer PCR cycles it takes to reach a certain fluorescence intensity. A baseline fluorescence intensity higher than the background should reflect a statistically significant amount of DNA. This intensity is called a threshold. The number of PCR cycles required to reach the threshold is called a threshold cycle (Ct), which is used to compare the initial amount of DNA quantitatively.

Based on the principle of relative quantitation, the delta-delta Ct (ΔΔCt) method is used to calculate the relative difference in gene expression between two samples. It uses a housekeeping gene as an internal control to normalize the gene expression data. The ΔΔCt method calculates the relative difference in gene expression by subtracting the Ct value of the housekeeping gene from the Ct value of the gene of interest (4, 5).

The types of housekeeping genes commonly used in qPCR methods include *ACTB*, *GAPDH*, *HPRT1* and *B2M*.

### Western blot needs housekeeping genes as a loading control

The Western blotting technique is used to detect and analyse proteins in a sample of tissue or cell lysate. The technique is based on immunodetection, which uses antibodies to detect specific proteins in the sample. Antibodies specifically bind to a particular antigen. The western blotting assay separates proteins by electrophoresis. The separated proteins are then transferred to a membrane followed by successive primary and secondary antibody incubation.

A housekeeping protein such as b-actin is commonly used in western blotting as a loading control. It is assumed the expression of b-actin is always constant regardless of intervention, and the amount of lysate loaded is considered appropriate when compared same with that of b-actin.

However, recent studies point out that significant variations in housekeeping gene expression were found across different tissues or in tissues at pathological events and a more reliable alternative reference control has to be examined and specified (6).

## HOUSEKEEPING GENES

### Role of housekeeping genes

Housekeeping genes refer to a group of genes that are expressed and maintained at a high level under any conditions for the maintenance of cell structure and integrity, movement and survival. The common examples of housekeeping genes include beta-actin which maintains the structure of cells, ribosomal proteins for producing intracellular proteins, glyceraldehyde 3-phosphate dehydrogenase involved in metabolic reactions to generate ATP in cells, and heat-shock proteins to protect the cell from apoptosis(1). The significance of housekeeping genes in the cell’s survival could not be more emphasized as it was confirmed that bacterial activity was still maintained in *Mycoplasma mycoides* of which genome was artificially synthesized by adding only 437 housekeeping genes out of a total of 917 genes in 2016(7).

|  |  |  |
| --- | --- | --- |
| Name | Gene | Function |
| 18S ribosomal RNA | *RRN18S* | Protein synthesis |
| Beta Actin | *ACTB* | Codes for beta-actin |
| Glyceraldehyde-3-phosphate dehydrogenase | *GAPDH* | Glycolysis |
| Beta-2-microglobulin | *B2M* | Compose MHC class I molecule |
| Ribosomal protein L13a | *RPL13A* | Codes for ribosomal protein |
| Ribosomal protein, large, P0 | *RPLP0* | Codes for ribosomal protein |
| Hypoxanthine phosphoribosyl-transferase 1 | *HPRT1* | Recycle purines |
| TATA box binding protein | *TBP* | Codes for TATA box binding protein |
| lactate dehydrogenase A | *LDHA* | Cellular respiration |
| Ubiquitin C | *UBC* | Synthesis and destruction of proteins |

### Common housekeeping genes

**Table 1.** List of important housekeeping genes in humans. The full name, the gene nomenclature and the function of each gene are described.

There are a number of different housekeeping genes, each with its own specific function. Glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) is an enzyme that catalyses glycolysis by converting glyceraldehyde 3-phosphate to D-glycerate 1,3-bisphosphate. TATA box binding protein (*TBP*) binds to the TATA box sequence in DNA. The TATA box is a regulatory element found in the promoter region of many genes, and *TBP* is essential for the transcription of these genes. *RRN18S* codes for 18S ribosomal RNA which is a component of the small ribosomal subunit. *ACTB* codes for β-actin which forms the cytoskeleton of the cell that supports cell integrity as well as cell motility.

### Housekeeping genes are widely used as an internal control in gene expression analysis

Due to stable expression at high level of housekeeping genes regardless of tissue and cell types, housekeeping genes have been used as an internal standard or control of gene expression studies.

## CONTEXT-SPECIFIC HOUSEKEEPING GENES

### There is variation in the housekeeping gene expression

|  |  |  |
| --- | --- | --- |
| Traditional standard(s) | Internal standard(s) | Tissue/cell line |
| *GAPDH, ACTB* | *RPLP0, EEF1A1 (8)* | Synovial fibroblast |
| *GAPDH, ACTB* | *RPL0 (9)* | Synovial fibroblast |
| *GAPDH, ACTB, PUM1* | *CCSER2, SYMPK, ANKRD17 and PUM1(10)* | Epithelial cell of breast (cancer) |
| *GAPDH, ACTB* | *RPL30, RPL31 (11)* | Endothelial colony-forming cells |
| *GAPDH, ACTB* | *Atp5f1, Pgk1, GAPDH (12)* | Induced pluripotent stem cell |

In fact, recent studies have found that the expression levels of commonly-used housekeeping genes are not perfectly consistent universally. There are several factors that are accused to cause variation in their expression level: gene, developmental stage, cell type, experimental condition, disease and infection (13).

**Table 2**. Choice of reference housekeeping genes in various gene expression studies. The traditional standard refers to the canonical and previously popular choice of a reference housekeeping gene (14). The internal standard refers to a novel choice of reference housekeeping gene the authors re-evaluated and implemented in their studies. Tissues and cell lines that are experimented in the particular study are listed.

### A need for identifying a context-specific housekeeping gene

As shown in Table 2, gene expression patterns can be variable both within the tissue and between different tissues(15). Studies in Table 2 re-evaluated housekeeping genes specific for their own experiment tissue. *GAPDH* and *ACTB* were often used as an internal control in traditional studies in the same field. They were considered good choices because it was easy to measure experimentally due to their large gene expression values. It was then found out that these housekeeping genes had a wider differential expression distribution between tissues due to the same reason (i.e. high expression level). Therefore, for gene expression research, the stability of the housekeeping gene should be evaluated under various conditions to select an appropriate reference gene, or context-specific housekeeping gene (CSHKG), for each condition.

## SIGNIFICANCE OF THIS STUDY

Housekeeping genes are used as a reference gene or a control group for gene expression research because they are considered to be expressed in all types of cells and tissues at constant level regardless of tissue type, development state, stage in the cell cycle, or external environmental factor. Over-expression or inhibition of the housekeeping genes is also related to various diseases and cancers as they are involved in vital cellular reactions. Due to mentioned reasons, housekeeping gene has always been used to support the foundation of genetic research and the validity of the results. However, increasing numbers of studies are reporting variations in the classic housekeeping gene expressions across different tissues and contexts have the potential to discredit their experiment results. Evaluation of a housekeeping gene for being the reference gene in gene expression analysis has now become a requirement, not a choice. In the current era, various evaluation approaches including direct experimental verification and software analysis are encouraged to find out the suitable housekeeping gene of the most stable expression. The advancement in sequencing technology allowed the detection of new housekeeping genes of which expression is stable at a relatively middle level, while repeats in highly duplicative genes like housekeeping genes may result in bias caused by alignment problems(1). What is clear is that a novel profiling and more comprehensive characterization of CSHKGs will be needed to catch up with improvements in gene expression studies. This study aims to find new candidates of housekeeping genes that can be used across broad contexts and reveal the patterns of the differential expressions in them.

# MATERIALS AND METHODS

## SELECTION OF DATASET

Gene expression data was obtained from the Pan-Cancer Atlas (PanCan). As a part of the Cancer Genome Atlas (TCGA), PanCan is run and provided by the National Cancer Institute (NCI) and the National Human Genome Research Institute to set up an integrated accumulation of genetic variation data on cancer and analyze biological information. PanCan profiles 33 different tumour types at DNA, RNA, protein, and epigenetic levels. The dataset used in the study contains RNA-seq data of 10982 samples and 20386 different genes.

## Computational analysis

The pre-processing of the data, statistical analysis, and construction of all the plots that are incorporated in the paper were done using RStudio on Linux. R packages used in the project include BioConductor, BiocManager, annotate, tidyverse and io. The documentation for each package can be found online at <https://www.rdocumentation.org>. GitHub was used as a form of recording and tracking the progress of the computational analysis.

### Pre-processing

The raw expression data file was pre-processed by several steps to get ready for further analysis. Each column of the raw file contains project information, participant number, sample type code, cancer type, and tissue source site as a single-code sample ID. All samples were divided into phenotypic groups based on their cancer type and sample type. For instance, the sample ID TCGA-OR-A5J1-01A-11R-A295-07 is an adrenocortical carcinoma and a primary solid tumour. TCGA-OR-A5J1-01A-11R-A295-07 is then assigned to the phenotypic group ACC-TP based on NCI sample type codes and TCGA study abbreviation shown in Table 3(16). Initially, 84 groups were created. Among 84 groups, the groups that have less than 10 samples or group frequency of less than 10 were appointed as rare groups and removed before calculation since it is difficult to say the mean is correct with only few samples. After filtering out the rare groups, 52 groups proceeded to further analysis (Figure 1). 21 classic housekeeping genes were selected and additionally labelled as common housekeeping genes (Table 4). They are: *ACTB, UBC, GAPDH, TBP, RPS18, G6PD, HPRT1, LDHA, RPLP1, RPL19, RPL18, RPL11, RPL32, PGK1, PPIA, SDHA, ASNS, ATP284, RPL13A, PEX19* and *RXRA.* Labelling previously listed common housekeeping genes will be useful in the later steps where the selection of candidate CSHKGs is carried out.

|  |  |
| --- | --- |
| Study Abbreviation | Study Name |
| LAML | Acute Myeloid Leukemia |
| ACC | Adrenocortical carcinoma |
| BLCA | Bladder Urothelial Carcinoma |
| LGG | Brain Lower Grade Glioma |
| BRCA | Breast invasive carcinoma |
| CESC | Cervical squamous cell carcinoma and endocervical adenocarcinoma |
| CHOL | Cholangiocarcinoma |
| LCML | Chronic Myelogenous Leukemia |
| COAD | Colon adenocarcinoma |
| CNTL | Controls |
| ESCA | Esophageal carcinoma |
| FPPP | FFPE Pilot Phase II |
| GBM | Glioblastoma multiforme |
| HNSC | Head and Neck squamous cell carcinoma |
| KICH | Kidney Chromophobe |
| KIRC | Kidney renal clear cell carcinoma |
| KIRP | Kidney renal papillary cell carcinoma |
| LIHC | Liver hepatocellular carcinoma |
| LUAD | Lung adenocarcinoma |
| LUSC | Lung squamous cell carcinoma |
| DLBC | Lymphoid Neoplasm Diffuse Large B-cell Lymphoma |
| MESO | Mesothelioma |
| MISC | Miscellaneous |
| OV | Ovarian serous cystadenocarcinoma |
| PAAD | Pancreatic adenocarcinoma |
| PCPG | Pheochromocytoma and Paraganglioma |
| PRAD | Prostate adenocarcinoma |
| READ | Rectum adenocarcinoma |
| SARC | Sarcoma |
| SKCM | Skin Cutaneous Melanoma |
| STAD | Stomach adenocarcinoma |
| TGCT | Testicular Germ Cell Tumors |
| THYM | Thymoma |
| THCA | Thyroid carcinoma |
| UCS | Uterine Carcinosarcoma |
| UCEC | Uterine Corpus Endometrial Carcinoma |
| UVM | Uveal Melanoma |

|  |  |  |
| --- | --- | --- |
| Code | Definition | Short Letter Code |
| 1 | Primary Solid Tumor | TP |
| 2 | Recurrent Solid Tumor | TR |
| 3 | Primary Blood Derived Cancer - Peripheral Blood | TB |
| 4 | Recurrent Blood Derived Cancer - Bone Marrow | TRBM |
| 5 | Additional - New Primary | TAP |
| 6 | Metastatic | TM |
| 7 | Additional Metastatic | TAM |
| 8 | Human Tumor Original Cells | THOC |
| 9 | Primary Blood Derived Cancer - Bone Marrow | TBM |
| 10 | Blood Derived Normal | NB |
| 11 | Solid Tissue Normal | NT |
| 12 | Buccal Cell Normal | NBC |
| 13 | EBV Immortalized Normal | NEBV |
| 14 | Bone Marrow Normal | NBM |
| 15 | sample type 15 | 15SH |
| 16 | sample type 16 | 16SH |
| 20 | Control Analyte | CELLC |
| 40 | Recurrent Blood Derived Cancer - Peripheral Blood | TRB |
| 50 | Cell Lines | CELL |
| 60 | Primary Xenograft Tissue | XP |
| 61 | Cell Line Derived Xenograft Tissue | XCL |
| 99 | sample type 99 | 99SH |

**Table 3.** NCI TCGA study abbreviation table showing cancer type code (on the left hand side) and sample type code (on the right hand side). Abbreviations and their full names and definitions are listed.

A screenshot of a computer

Description automatically generated with low confidence

**Figure 1.** Number of samples (at the bottom) in each phenotypic group (on top) after filtering out the rare groups. The name of the group is a combination of study abbreviation in TCGA study abbreviation and NCI sample type code.

|  |
| --- |
| **21 Common housekeeping genes** |
| *ACTB, UBC, GAPDH, TBP, RPS18, G6PD, HPRT1, LDHA, RPLP1, RPL19, RPL18, RPL11, RPL32, PGK1, PPIA, SDHA, ASNS, ATP284, PEX19, RPL13A, RXRA* |

**Table 4.** 21 common housekeeping genes are chosen and labelled.

### Statistical analysis

Within-group standard deviation (SD) and between-group standard deviation are calculated as well as the mean of the groups. Within-group SD is a measure of variance among the samples within a group and between-group SD measures how the same gene expression varies between different groups.

1. Within-group standard deviation

Within-group SD

expression of sample *i* in group *g*

mean expression in all the samples within group *g*

number of samples in the group *g*

number of groups

1. Between-group standard deviation

= overall mean of samples

total number of groups

mean expression of group *k*

number of genes

1. Mean expression of the group

For gene A, the mean of the expression of all of the samples is calculated by groups. Expressions of gene A are combined and divided by the number of samples in the group.

### Interpretation of the results to find the best candidate

Candidate CSHKGs were selected based on the mean expression, within-group SD and between-group SD of 21 common housekeeping genes. Candidates should have sufficiently high mean expression which is at least as same as the lowest mean expression of 21 housekeeping genes. Candidates should also have lower within-group SD than that of 21 housekeeping genes. Between-group SDs of candidates have to be larger than their within-group SDs. The expression of the gene of interest in each sample was plotted for each group. A blue horizontal line was added to the plot to show the overall mean expression in all of the samples. The groups that show particularly upregulated or downregulated expression of the gene were manually observed and sorted out. The groups that were manually categorized to have dysregulated gene expressions were excluded. An additional plot was then constructed using the new set of groups and between-group SD was recalculated and compared to the original. Among the candidates, the CSHKGs of which expression do not vary in different tissue types could be deduced. The Gene Ontology (GO) enrichment analysis was carried out on web g:Profiler to confirm that the functions of the candidate genes were fundamental to the maintenance and survival of the cell(17).

# RESULTS

## Trend in between-group standard deviation and within-group standard deviation

Chart, scatter chart

Description automatically generated

**Figure 2.** Between-group standard deviation is plotted against within-group standard deviation for all of 20386 genes expressed in samples in the non-rare groups. The straight line that crosses the overall plot depicts line of slope = 1.0. 21 common housekeeping genes are highlighted and labelled in red. The opacity of the plot was set to 10% to visualise where the plots are concentrated clearly.

Although it ranges, all highlighted housekeeping genes tend to cluster on the bottom left corner of the graph. *TBP* has the lowest within-group SD and *ATP2B4* has the highest within-group and between-group SD. This is expected as housekeeping genes have very stable expression in different types of tissues and samples. However, it is also interesting to notice that other genes also seem they tend to follow the straight line of slope = 1.0. The following mathematical steps may explain the rationale behind this observation.

,

The common housekeeping genes of which between-group SD is greater than within-group SD are *GAPDH, LDHA, RPL13A, RPS18* and *SDHA* (Table 5). The genes that are plotted above the identity line, indicating chance of differential expression depending on the type of the tissue.

|  |  |  |  |
| --- | --- | --- | --- |
| **Gene** | **Within-group SD** | **Between-group SD** | **Mean** |
| *ACTB* | 0.5181342 | 0.4293657 | 16.531571 |
| *ASNS* | 0.9691193 | 0.9257248 | 9.227693 |
| *ATP2B4* | 0.8497256 | 0.7945115 | 11.522266 |
| *G6PD* | 0.8546989 | 0.7176015 | 10.372851 |
| *GAPDH* | 0.6988123 | 0.7234513 | 15.927636 |
| *HPRT1* | 0.6565984 | 0.598318 | 9.441051 |
| *LDHA* | 0.7370976 | 0.7867797 | 13.599979 |
| *PEX19* | 0.4666227 | 0.4292843 | 10.807102 |
| *PGK1* | 0.6645354 | 0.5245667 | 13.199975 |
| *PPIA* | 0.5089595 | 0.4356253 | 12.288541 |
| *RPL11* | 0.6120978 | 0.5229753 | 13.822263 |
| *RPL13A* | 0.6711144 | 0.7115543 | 13.992058 |
| *RPL18* | 0.710385 | 0.6012657 | 13.813475 |
| *RPL19* | 0.6360119 | 0.5158671 | 14.351875 |
| *RPL32* | 0.6789193 | 0.5466778 | 13.866778 |
| *RPLP1* | 0.7000465 | 0.6222899 | 14.555316 |
| *RPS18* | 0.7316804 | 0.7979357 | 14.542232 |
| *RXRA* | 0.6706583 | 0.6094941 | 10.903959 |
| *SDHA* | 0.5607891 | 0.6620696 | 11.445043 |
| *TBP* | 0.3962313 | 0.2434846 | 8.071081 |
| *UBC* | 0.4619325 | 0.3767266 | 14.99872 |

**Table 5.** Within-group and between-group standard deviation, and mean expression of the highlighted 21 common housekeeping genes

## trend between within-group standard deviation and mean expression

Figure 3 shows the mean expression of each of the 20386 genes plotted against within-group SD. A short, slightly fanning-out dark line is formed at the bottom left part of the graph. This tells us that some of the genes are barely expressed and as a result of this, their within-group SD was calculated to be very low as well. A housekeeping gene is defined as a gene with sufficiently high expression, and therefore, the genes that are positioned here are not housekeeping genes. Such misleading data has to be managed before further analysis as we are only looking for housekeeping genes. The 21 common housekeeping genes are expressed sufficiently high, while their mean expressions could still vary by multiple folds as the expression is in the log scale. For instance, the expression of *ACTB* and *TBP* are both high but the mean expression of *ACTB* is much greater than that of *TBP*.

Chart, scatter chart

Description automatically generated

**Figure 3.** Mean expression of all of 20386 genes plotted against within-group standard deviation. Each dot represents one gene. 21 common housekeeping genes from Table 4 is highlighted and labelled red. They are concentrated at the top left part of the overall shape of the plots. Opacity of the plot is set to 10% and clusters can be identified more clearly (as concentrated, darker shade regions).

## Selection of candidate context-specific housekeeping genes

In order to select the candidate CSHKGs, as previously mentioned in the results section II, primarily only genes with sufficiently high expression have to be filtered. Among 20386 genes, the lowest mean expression of 21 housekeeping genes was *TBP*, approximately at 8.07. Therefore, the threshold expression applied for the candidate housekeeping genes selection was 8.00, slightly adjusted from that of *TBP*. All the other genes whose expression is less than 8 were excluded as they were disqualified for being housekeeping genes. In addition, the genes with within-group SD greater than 1 are also excluded from consideration. These two primary conditions for searching for candidate CSHKGs were both contingent on the 21 common housekeeping genes and their figures. Among the genes that satisfy the conditions, 11 genes that are highly expressed and have lower within-group SD than that of 21 common housekeeping genes are highlighted and labelled in light blue colour (Figure 4). They are *HNRNPK, HNRNPU, PCBP1, HNRNPC, PTBP1, SF3B2, HNRNPL, SF1, SRSF1, EWSR1,* and *UBE2D3*. After applying the conditions to the dataset, the graph of between-group SD against within-group SD of 7785 genes was newly constructed (Figure 5).

**Chart, scatter chart

Description automatically generated**

**Figure 4.** Mean expression of genes whose within-group SD is less than 1.00 are plotted against within-group standard deviation as black dots of opacity 10%. Each dot represents one gene. Mean expression of 21 common housekeeping genes (red dot) and 11 candidate context-specific housekeeping genes (blue dot) are highlighted in different colours.

Among the 11 candidate genes, 2 genes, *PTBP1* and *HNRNPC*  were chosen as candidate CSHKGs as their variances between groups are greater than the variance within the group (Figure 5). The other two candidate CSHKGs are *HNRNPL* and *EWSR1*. *LDHA* and *GAPDH* also have between-group SD greater than within-group SD, therefore, they were also analysed.

**Chart, scatter chart

Description automatically generated**

**Figure 5.** Between-group standard deviation of highly expressed genes (expression > 8.0) are plotted against within-group standard deviation (n = 7785 genes). Each dot represents one gene. 21 classic housekeeping genes are highlighted and labelled in red. 11 candidate context-specific housekeeping genes are highlighted and labelled in blue. The highlighted genes that are above the identity line are PTBP1, HNRNPC, HNRNPL, EWSR1, SDHA, RPL13A, GAPDH, LDHA, and RPS18.

## Analysis of Housekeeping genes to find the best candidate

### *GAPDH*

Each dot in Figure 6 represents expression of *GAPDH* of one sample and one graph for each of 52 groups is constructed. The blue horizontal line shows the mean expression of *GAPDH* in all of the samples. The mean expression of *GAPDH* is approximately 15.9276 and between-group SD is 0.723. Most of the groups had their plots lined with the overall mean expression, however, it is important to note that the expressions in solid normal tissues (NT) are lower than other groups and *GAPDH* is more highly expressed in tumour tissues.

**Timeline

Description automatically generated with medium confidence**

**Figure 6.** The expressions of GAPDH in the samples in the groups are plotted. One black dot represents one sample. Number of groups is 52. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of the graph.

NT groups are then removed and the expressions of the rest of the groups are plotted (Figure 7). 36 groups were plotted and most of them were primary tumour tissues. The between-group SD was recalculated and it appeared to decrease greatly from 0.723 to 0.514 upon exclusion. Among the tumour groups, adrenocortical carcinoma (ACC-TP), testicular germ cell tumours (TGCT-TP) and lung squamous cell carcinoma (LUSC-TP) had higher gene expression than the mean while other tissues remained around the mean.

**A picture containing text

Description automatically generated**

**Figure 7.** The expressions of GAPDH in the groups excluding normal solid tissue groups. One black dot represents one sample. 16 NT groups were excluded and the total number of groups are 36. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of the graph. After exclusion, between-group standard deviation of GAPDH has decreased from 0.723 to 0.514.

**Graphical user interface, application

Description automatically generated**

**Figure 8.** Expression of GAPDH in different solid normal tissue groups (NT). One black dot represents one sample. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of the graph. The mean of the 16 NT groups was 14.979.After separating the NT groups from the rest of the groups, the SD between NT groups was calculated to be 0.619.

The expression of *GAPDH* in 16 NT groups were plotted independently from the rest of the groups (Figure 8). The between-group SD of 16 NT groups was lowered from 0.723 to 0.619. The expression level in the tumour groups seems to be more uniform than that in normal tissues. Meanwhile, low sample numbers in NT groups has to be put into consideration before concluding expression levels in the normal tissues are much more unstable. Two lung tissue groups, lung adenocarcinoma (LUAD-NT) and lung squamous cell carcinoma (LUSC-NT), had most of their samples plotted below the blue line, indicating downregulated expression compared to the rest of the NT groups. As continuation of the current idea, these two groups were excluded and the between-group SD is calculated again (Figure 9). NT groups without previously mentioned two lung tissue groups showed much uniform expression. The between-group SD of 14 NT groups was pulled further down from 0.619 to 0.529. The mean expression is 15.136.

**Graphical user interface, application

Description automatically generated**

**Figure 9.** Expression of GAPDH in non-lung solid normal tissue groups. One black dot represents one sample. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of the graph. The mean of the groups is 15.136. The between-group standard deviation in the 14 NT groups is 0.529.

### *LDHA*

**A picture containing timeline

Description automatically generated**

**Figure 10.** Expression of LDHA genes in each of the 52 groups. One black dot represents one sample. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of the graph.

**A picture containing graphical user interface

Description automatically generated**

**Figure 11.** Expression of LDHA in 48 groups after excluding following 4 upregulated or downregulated groups: KIRC-TP, HNSC-TP, LGG-TP, THCA-TP. One black dot represents one sample. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of the graph.

The expressions of *LDHA* in different groups are plotted. The mean of the entire sample is shown as a blue horizontal line on each graph (Figure 10). In Figure 10, *LDHA* gene expressions in KIRC-TP and HNSC-TP groups are upregulated as most of the plots lie above the blue line while that of LGG-TP and THCA-TP are downregulated in the same context. Using the same logic in the results section IV.I., these 4 groups are excluded to calculate a new between-group SD. By removing 4 groups from the set, the between-group SD decreased from 0.787 to 0.681 (Figure 11).

### *PTBP1*

Table

Description automatically generated

**Figure 12.** Expression of PTBP1 in all of the samples were plotted by groups. One black dot represents one sample. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of the graph. The mean of the expressions is 12.428.

As shown in Figure 12, the expression of *PTBP1* in TGCT-TP, COAD-TP, HNSC-TP, and UCEC-TP was upregulated. The expression in LGG-TP, LGG-TR, BRCA-NT, and GBM-TP was downregulated. The mean of the samples is 12.428. After removing these 8 groups from the set, a new graph was plotted (Figure 13). The between-group SD of the gene *PTBP1* decreased from 0.410 to 0.317.

A picture containing table

Description automatically generated

**Figure 13.** Expression of PTBP1 in all of the samples plotted by groups excluding the following 8: TGCT-TP, COAD-TP, HNSC-TP, UCEC-TP, LGG-TP, LGG-TR, BRCA-NT, and GBM-TP. One black dot represents one sample. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of the graph.

### *HNRNPC*

**A picture containing graphical user interface

Description automatically generated**

**Figure 14.** The expressions of HNRNPC in all the samples are plotted by 52 phenotypic groups. One black dot represents one sample. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of the graph.

Figure 14 shows the expressions of the *HNRNPC* gene in all of the samples by phenotypic groups. Like the rest of the candidate genes, in most of the groups, the expression of *HNRNPC* is close to average. However, it was upregulated in THYM-TP, OV-TP and TGCT-TP and downregulated in ESCA-NT and SARC-TP. By removing the 5 groups and reconstructing the expression plots, the between-group SD of *HNRNPC* was decreased from 0.342 to 0.289 (Figure 15).

**A picture containing text

Description automatically generated**

**Figure 15.** The expression of all of the samples and their between-group standard deviation of HNRNPC gene after excluding 5 groups whose overall expressions are highly deviated from the overall mean: THYM-TP, OV-TP, TGCT-TP, ESCA-NT and SARC-TP. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of graph. The between-group standard deviation decreased from 0.342 to 0.289.

### *ACTB* and *UBC*

Classic housekeeping genes with low between-group SDs were also analysed to find the difference in the change of between-group SD after excluding groups to make the gene more context-specific. *ACTB* and *UBC* have within-group SD of 0.518 and 0.462, and between-group SD of 0.429 and 0.377, respectively. Their within-group SDs are greater than their between-group SDs. The expressions of *ACTB* in every group are plotted (Figure 16) and another graph was plotted with the 7 groups where the gene was generally upregulated or downregulated excluded (Figure 17). The upregulated groups were: DLBC-TP, LUAD-NT, and LUSC-NT. The downregulated groups were: PCPG-TP, PRAD-TP, KIRC-NT and LIHC-NT. The same approach was taken at *UBC.* HNSC-TP and KIRC-TP groups showed upregulated expression and UVM-TP, LAML-TB, and READ-TP groups showed downregulated expression of *UBC* (Figure 19). These groups were excluded and a new between-group SD was calculated (Figure 20). The between-group SD decreased from 0.377 to 0.317 upon the exclusion of previously mentioned groups.

A picture containing timeline

Description automatically generated

**Figure 16.** The expressions of ACTB in all the samples are plotted by 52 phenotypic groups. One black dot represents one sample. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of the graph. The mean of the expression of ACTB is 16.532.

A picture containing text

Description automatically generated

**Figure 17.** The expression of all of the samples and their between-group standard deviation of ACTB gene after excluding 6 groups whose overall expressions are highly deviated from the overall mean: DLBC-TP, LUSC-NT, PRAD-TP, PCPG-TP, KIRC-NT, and LIHC-NT. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of graph. The between-group standard deviation of ACTB was pulled down to 0.352 from 0.429.

A picture containing graphical user interface

Description automatically generated

**Figure 18.** The expressions of UBC in all the samples are plotted by 52 phenotypic groups. One black dot represents one sample. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of the graph. The mean of the expression of UBC is 14.999.

A picture containing text

Description automatically generated

**Figure 19.** The expression of all of the samples and their between-group standard deviation of UBC gene after excluding 5 groups whose overall expressions are highly deviated from the overall mean: HNSC-TP, KIRC-TP, UVM-TP, LAML-TB, and READ-TP. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of graph. The between-group standard deviation of UBC was decreased from 0.377 to 0.317.

### g:Profiler results for candidate housekeeping genes

123 candidate genes whose within-group and between-group SD are less than 0.5 and mean expression is greater than 12 were run on g:Profiler to confirm their functions are essential housekeeping functions of the cell (Table 6).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Candidate context-specific housekeeping genes | | | | | | | | | |
| ACTR2 | CALM3 | DDB1 | FAM120A | HNRNPA3 | KPNB1 | PCBP2 | SEC61A1 | SNRNP200 | XRCC5 |
| ADAR | CAP1 | DDOST | FUS | HNRNPC | LAPTM4A | PPP2R1A | SEPTIN2 | SPTAN1 | XRCC6 |
| ADD1 | CAPRIN1 | DDX17 | GANAB | HNRNPF | LARP1 | PRKCSH | SERBP1 | SRP14 | YWHAB |
| AP2M1 | CAPZB | DDX5 | GDI2 | HNRNPH1 | MAP4 | PRPF8 | SET | ST13 | YWHAE |
| APH1A | CCT7 | DYNC1H1 | GLG1 | HNRNPK | MATR3 | PSMD2 | SF1 | SURF4 | YWHAG |
| ARF1 | CDC42 | EIF1 | GNB1 | HNRNPL | MLF2 | PTBP1 | SF3B1 | TMED10 | YWHAQ |
| ARF3 | CHD4 | EIF2AK1 | GTF2IP1 | HNRNPM | MORF4L2 | PTGES3 | SF3B2 | TMED2 |  |
| DDX39B | CNBP | EIF2S3 | H3-3B | HNRNPU | MRFAP1 | RAB1B | SFPQ | TUFM |  |
| PRRC2A | COPA | EIF3D | HADHA | HNRNPUL1 | MTCH1 | RAB5B | SRSF1 | UBA1 |  |
| BAG6 | COPG1 | EIF4G1 | HDLBP | HSPA9 | NACA | RAB7A | SRSF3 | UBC |  |
| BRD2 | CSDE1 | EIF4G2 | HMGB1 | ILF3 | NCL | RAC1 | SKP1 | UBE2D3 |  |
| C1orf43 | CSNK1A1 | EIF4H | HNRNPA1 | IRF2BP2 | NONO | RHOA | SLC25A3 | VCP |  |
| CALM2 | DAZAP2 | EWSR1 | HNRNPA2B1 | KDELR1 | PCBP1 | RPN1 | SND1 | WDR1 |  |

**Table 6.** List of 123 candidate housekeeping genes of within-group standard deviation < 0.5, between-group standard deviation < 0.5 and mean expression > 12.

The g:Profiler results (Figure 20) found many enriched GO terms related to essential cellular maintenance functions. The enriched molecular function terms include RNA binding, helicase activity and protein complex binding. The enriched biological processes were regulation of nitrogen compound metabolic process, intracellular transport and more. Cellular component terms enriched involved extracellular exosomes and spliceosomal complexes.

Graphical user interface, application

Description automatically generated

**Figure 20.** g:Profiler results for 123 candidate housekeeping genes. The enriched Gene Ontology terms are shown.

# DISCUSSION

## Interpretation of the results

After removing NT groups from the expression plots of *GAPDH,* the between-group SD decreased greatly from 0.723 to 0.514. This indicates that *GAPDH* may not be the best choice as a reference housekeeping gene in the study that compares a particular gene expression difference between the tumour tissues and the normal tissues. The decrease in between-group SD of NT groups after getting excluded further supports the idea of differential expression of *GAPDH* in different types of tissues. Moreover, a special expression pattern (i.e. downregulation) was observed in the lung solid normal tissue groups. A huge variation in some of the groups, for example, in oesophageal carcinoma (ESCA-NT), was observed but this observation was not treated significantly as the number of samples was too small. *GAPDH*, being one of the most commonly used reference housekeeping genes in the vast ranges of biological studies,is a critical gene for energy metabolism thus, it is highly expressed in tumour tissues (18). Meanwhile, even among the same type of tissues - primary tumour tissue, some of them can also have expressions near the normal level. Therefore, it is very important to find the housekeeping gene that is stable and specific to the tissues of study interest.

Another classic housekeeping gene, *LDHA*, also showed improvement in the between-group SD after removing some of the groups with upregulation and downregulation. Two newly identified non-canonical housekeeping genes, *PTBP1* and *HNRNPC* as well were proven to give inconsistent expression in some cancer and tissue types. *PTBP1* codes for polypyrimidine tract-binding protein 1 and *HNRNPC* codes for heterogeneous nuclear ribonucleoprotein C. Both of the proteins are RNA-binding proteins that are involved in splicing of pre-mRNA.

As the decline magnitude in the between-group SD of ACTB and UBC was less than or about the same as the lowest of that of GAPDH, LDHA, PTBP1, and HNRNPC, the change in the between-group SD of the genes in the process of making the gene more context-specific by eliminating a highly dysregulated group of tissues from calculation seemed to be more obvious in the gene whose original between-group SD was greater than its within-group SD.

The GO term enrichment analysis showed that the functions of the candidate genes were essential for the maintenance and survival of the cell, confirming their identity as new housekeeping genes.

## limitations and future directions

Besides, there are some limitations and places for improvement. First of all, the entire study can be repeated using different, bigger datasets such as the Genotype-Tissue Expression (GTEx) dataset(19) to reinforce the reliability of the work. Secondly, PanCan’s main focus is on the biology of cancer(20). Therefore, it will also be very interesting to look at different diseased states and developmental stages like inflammation and stem cell and embryonic development.

One source of error to be taken into account is that the upregulation and downregulation of the gene in the expression plots by phenotypic groups were manually spotted and judged by eye.

One major limitation of this study was that it was difficult to calculate the statistical significance. This is because most of the data plots in the first graph where between-group SD was plotted against within-group SD were below the identity line. This means the assumption of the analysis of variance (ANOVA) was violated. In order for the primary assumption of homogeneity of variance in ANOVA to be correct, most of the plots should have been on the line of identity.

Non-parametric methods like the Kruskal-Wallis test that does not require equal variances can be used, but rank-based results are not preferred in terms of the purpose of the study. Instead, a permutation study can be attempted to generate the null distribution and compared it with observations. The focus of this research is on revealing the patterns in regard to variation in gene expressions.

# CONCLUSION

This study has found a list of novel candidates of housekeeping genes that can be used as a new internal control in scientific research and pointed out the patterns of non-constitutive expressions numerous different housekeeping genes show depending on the context of the tissue. Through expanding the current research, at the end, it is anticipated that a more accurate and advanced quantification paradigm in the biomedical and bioinformatic field will be developed.

# ABBREVIATIONS

|  |  |
| --- | --- |
| **Abbreviation** | **Meaning** |
|  |  |
| **ANOVA** | Analysis of variance |
| **CSHKG** | Context-specific housekeeping gene |
| ***GAPDH*** | Glyceraldehyde-3-phosphate dehydrogenase |
| **GO** | Gene Ontology |
| **NCI** | National Cancer Institute |
| **NT** | Solid normal tissue |
| **MHC** | Major Histocompatibility Complex, |
| **PanCan** | Pan-Cancer Atlas |
| **SD** | Standard deviation |
| ***TBP*** | TATA box binding protein |
| **TCGA** | The Cancer Genome Atlas |

# LIST OF FIGURES

[**Figure 1.** Number of samples (at the bottom) in each phenotypic group (on top) after filtering out the rare groups. The name of the group is a combination of study abbreviation in TCGA study abbreviation and NCI sample type code. 13](#_Toc134036489)

[**Figure 2.** Between-group standard deviation is plotted against within-group standard deviation for all of 20386 genes expressed in samples in the non-rare groups. The straight line that crosses the overall plot depicts line of slope = 1.0. 21 common housekeeping genes are highlighted and labelled in red. The opacity of the plot was set to 10% to visualise where the plots are concentrated clearly. 15](#_Toc134036490)

[**Figure 3.** Mean expression of all of 20386 genes plotted against within-group standard deviation. Each dot represents one gene. 21 common housekeeping genes from Table 4 is highlighted and labelled red. They are concentrated at the top left part of the overall shape of the plots. Opacity of the plot is set to 10% and clusters can be identified more clearly (as concentrated, darker shade regions). 19](#_Toc134036491)

[**Figure 4.** Mean expression of genes whose within-group SD is less than 1.00 are plotted against within-group standard deviation as black dots of opacity 10%. Each dot represents one gene. Mean expression of 21 common housekeeping genes (red dot) and 11 candidate context-specific housekeeping genes (blue dot) are highlighted in different colours. 20](#_Toc134036492)

[**Figure 5.** Between-group standard deviation of highly expressed genes (expression > 8.0) are plotted against within-group standard deviation (n = 7785 genes). Each dot represents one gene. 21 classic housekeeping genes are highlighted and labelled in red. 11 candidate context-specific housekeeping genes are highlighted and labelled in blue. The highlighted genes that are above the identity line are PTBP1, HNRNPC, HNRNPL, EWSR1, SDHA, RPL13A, GAPDH, LDHA, and RPS18. 21](#_Toc134036493)

[**Figure 6.** The expressions of GAPDH in the samples in the groups are plotted. One black dot represents one sample. Number of groups is 52. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of the graph. 22](#_Toc134036494)

[**Figure 7.** The expressions of GAPDH in the groups excluding normal solid tissue groups. One black dot represents one sample. 16 NT groups were excluded and the total number of groups are 36. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of the graph. After exclusion, between-group standard deviation of GAPDH has decreased from 0.723 to 0.514. 23](#_Toc134036495)

[**Figure 8.** Expression of GAPDH in different solid normal tissue groups (NT). One black dot represents one sample. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of the graph. The mean of the 16 NT groups was 14.979.After separating the NT groups from the rest of the groups, the SD between NT groups was calculated to be 0.619. 24](#_Toc134036496)

[**Figure 9.** Expression of GAPDH in non-lung solid normal tissue groups. One black dot represents one sample. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of the graph. The mean of the groups is 15.136. The between-group standard deviation in the 14 NT groups is 0.529. 25](#_Toc134036497)

[**Figure 10.** Expression of LDHA genes in each of the 52 groups. One black dot represents one sample. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of the graph. 26](#_Toc134036498)

[**Figure 11.** Expression of LDHA in 48 groups after excluding following 4 upregulated or downregulated groups: KIRC-TP, HNSC-TP, LGG-TP, THCA-TP. One black dot represents one sample. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of the graph. 26](#_Toc134036499)

[**Figure 12.** Expression of PTBP1 in all of the samples were plotted by groups. One black dot represents one sample. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of the graph. The mean of the expressions is 12.428. 27](#_Toc134036500)

[**Figure 13.** Expression of PTBP1 in all of the samples plotted by groups excluding the following 8: TGCT-TP, COAD-TP, HNSC-TP, UCEC-TP, LGG-TP, LGG-TR, BRCA-NT, and GBM-TP. One black dot represents one sample. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of the graph. 28](#_Toc134036501)

[**Figure 14.** The expressions of HNRNPC in all the samples are plotted by 52 phenotypic groups. One black dot represents one sample. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of the graph. 28](#_Toc134036502)

[**Figure 15.** The expression of all of the samples and their between-group standard deviation of HNRNPC gene after excluding 5 groups whose overall expressions are highly deviated from the overall mean: THYM-TP, OV-TP, TGCT-TP, ESCA-NT and SARC-TP. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of graph. The between-group standard deviation decreased from 0.342 to 0.289. 29](#_Toc134036503)

[**Figure 16.** The expressions of ACTB in all the samples are plotted by 52 phenotypic groups. One black dot represents one sample. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of the graph. The mean of the expression of ACTB is 16.532. 30](#_Toc134036504)

[**Figure 17.** The expression of all of the samples and their between-group standard deviation of ACTB gene after excluding 6 groups whose overall expressions are highly deviated from the overall mean: DLBC-TP, LUSC-NT, PRAD-TP, PCPG-TP, KIRC-NT, and LIHC-NT. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of graph. The between-group standard deviation of ACTB was pulled down to 0.352 from 0.429. 31](#_Toc134036505)

[**Figure 18.** The expressions of UBC in all the samples are plotted by 52 phenotypic groups. One black dot represents one sample. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of the graph. The mean of the expression of UBC is 14.999. 31](#_Toc134036506)

[**Figure 19.** The expression of all of the samples and their between-group standard deviation of UBC gene after excluding 5 groups whose overall expressions are highly deviated from the overall mean: HNSC-TP, KIRC-TP, UVM-TP, LAML-TB, and READ-TP. The horizontal axis of each plot has no value. The mean of the expressions is shown as an uniform horizontal blue line on each of graph. The between-group standard deviation of UBC was decreased from 0.377 to 0.317. 32](#_Toc134036507)

[**Figure 20.** g:Profiler results for 123 candidate housekeeping genes. The enriched Gene Ontology terms are shown. 34](#_Toc134036508)

# LIST OF TABLES

[**Table 1.** List of important housekeeping genes in human. The full name, the gene nomenclature and the function of each gene are described. 8](#_Toc134036509)

[**Table 2**. Choice of reference housekeeping genes in various gene expression studies. Traditional standard refers to canonical and previously popular choice of reference housekeeping gene (14). Internal standard refers to novel choice of reference housekeeping gene the authors re-evaluated and implemented in their studies. Tissues and cell lines that are experimented in the particular study are listed. 9](#_Toc134036510)

[**Table 3.** NCI TCGA study abbreviation table showing cancer type code (on the left hand side) and sample type code (on the right hand side). Abbreviations and their full names and definitions are listed. 12](#_Toc134036511)

[**Table 4.** 21 common housekeeping genes are chosen and labelled. 13](#_Toc134036512)

[**Table 5.** Within-group and between-group standard deviation, and mean expression of the highlighted 21 common housekeeping genes 17](#_Toc134036513)

[**Table 6.** List of 123 candidate housekeeping genes of within-group standard deviation < 0.5, between-group standard deviation < 0.5 and mean expression > 12. 33](#_Toc134036514)

# REFERENCES

1. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. TRENDS in Genetics. 2013;29(10):569-74.

2. Crick F. Central dogma of molecular biology. Nature. 1970;227(5258):561-3.

3. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science. 1988;239(4839):487-91.

4. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2− ΔΔCT method. methods. 2001;25(4):402-8.

5. Schmittgen TD, Livak KJ. Analyzing real-time PCR data by the comparative CT method. Nature protocols. 2008;3(6):1101-8.

6. Li R, Shen Y. An old method facing a new challenge: re-visiting housekeeping proteins as internal reference control for neuroscience research. Life sciences. 2013;92(13):747-51.

7. Hutchison III CA, Chuang R-Y, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, et al. Design and synthesis of a minimal bacterial genome. Science. 2016;351(6280):aad6253.

8. Nazet U, Schröder A, Grässel S, Muschter D, Proff P, Kirschneck C. Housekeeping gene validation for RT-qPCR studies on synovial fibroblasts derived from healthy and osteoarthritic patients with focus on mechanical loading. PloS one. 2019;14(12):e0225790.

9. Karouzakis E, Raza K, Kolling C, Buckley CD, Gay S, Filer A, et al. Analysis of early changes in DNA methylation in synovial fibroblasts of RA patients before diagnosis. Scientific Reports. 2018;8(1):1-6.

10. Tilli TM, Castro CdS, Tuszynski JA, Carels N. A strategy to identify housekeeping genes suitable for analysis in breast cancer diseases. BMC genomics. 2016;17(1):1-11.

11. McLoughlin KJ, Pedrini E, MacMahon M, Guduric-Fuchs J, Medina RJ. Selection of a real-time PCR housekeeping gene panel in human endothelial colony forming cells for cellular senescence studies. Frontiers in medicine. 2019;6:33.

12. Panina Y, Germond A, Masui S, Watanabe TM. Validation of common housekeeping genes as reference for qPCR gene expression analysis during iPS reprogramming process. Scientific reports. 2018;8(1):1-8.

13. Casas AI, Hassan AA, Manz Q, Wiwie C, Kleikers P, Egea J, et al. Un-biased housekeeping gene panel selection for high-validity gene expression analysis. Scientific Reports. 2022;12(1):1-12.

14. Thellin O, Zorzi W, Lakaye B, De Borman B, Coumans B, Hennen G, et al. Housekeeping genes as internal standards: use and limits. Journal of biotechnology. 1999;75(2-3):291-5.

15. Kozera B, Rapacz M. Reference genes in real-time PCR. Journal of applied genetics. 2013;54(4):391-406.

16. NIC. TCGA Study Abbreviations. National Cancer Institute. 2023.

17. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic acids research. 2019;47(W1):W191-W8.

18. Zhang J-Y, Zhang F, Hong C-Q, Giuliano AE, Cui X-J, Zhou G-J, et al. Critical protein GAPDH and its regulatory mechanisms in cancer cells. Cancer biology & medicine. 2015;12(1):10.

19. Consortium G. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369(6509):1318-30.

20. Hutter C, Zenklusen JC. The cancer genome atlas: creating lasting value beyond its data. Cell. 2018;173(2):283-5.

END OF THE DISSERTATION