**Background**

**###CITE THIS PMC7778946**

To calculate the differential expression of genes, find biomarkers and even model machine learning predictors, it is fundamental to normalize the data expression using baseline levels of control genes, also named housekeeping genes. The current benchmark methods to normalize transcript expression data in RNA-Seq experiments are MRN and TMM, included in packages DESeq2 (PMID: 25516281) and EdgeR (PMC2796818), respectively. Those methods are based on cross-sample normalizations (PMC8220791 PMID: 30041017 PMC6209231), which can be a source of data leakage and following overfitted machine learning models (PMC10499856 PMC4634915), compromising the reproducibility and generalizations of the models on new datasets (PMC4495301). An alternative to cross-sample normalization is defining housekeeping genes for a specific niche, such as kidney transplanted patients, and normalizing each sample by them independently of other samples.

Although the biological definition of housekeeping genes is still not fully determined; those are usually defined as genes with constant expression in all conditions, with an essential role in cellular maintenance and their sequence is conserved in evolutionary history (PMC9312424, PMID: 11773595). Nevertheless, since the expression levels of a gene can vary with drug treatment (PMC10368838) and clinical conditions (PMC9296577, PMC4986254, PMC9831022, PMC1976390), different diseases and treatments can lead to a different set of housekeeping genes. Thus, the definitions of ubiquitously and constant expression across tissues are not enough to define universal housekeeping genes (PMID: 23810203) and they should be defined in specific cohorts of patients. Therefore, control genes must be carefully analyzed before their application in experiments and diagnosis. This will allow us to overcome the data leakage and overfitting problems in normalization methods.

Transplant is the treatment of choice for patients with end-stage kidney disease. In this context, it is important to define housekeeping genes to support biomarkers discovery and to build tools based on RNA-seq to indicate potential diagnosis of rejection. In general, kidney transplant recipients are treated with immunosuppressive medications to avoid rejection by the immune system. Some of them can affect some crucial cell maintenance processes, such as DNA repair, leading to alterations in cell physiology and the expression of some gene’s physiology (PMID: 22244961, PMID: 8588225). Those side effects can change the RNA profile and, perhaps, some traditional housekeeping genes can have their expression dysregulated in several conditions. For example, the commonly used housekeeping genes *GAPDH*, *ACTB* and *B2M* did not present a constant expression in the biopsy of kidney transplanted recipients (PMC6580566), underlying the need for a better definition of housekeeping genes in this population.

To address the lack of information about what appropriate housekeeping genes to serve as controls in experiments using peripheral blood of kidney transplanted patients, we initially develop a workflow-based on large public datasets of NGS RNA-seq and RNA Microarrays. We search for genes highly expressed with low coefficient of variance in different time points (pre-transplant, 1 week, 3 and 6 months after transplantation) and with different clinical outcomes. We mine housekeeping genes in high-throughput RNA-sequencing using different approaches. We then explore well-established methods to define non-differential expression genes, we apply basic statistical concepts like coefficient of variance, gini coefficient and pairwise stability in a big dataset and use machine learning concepts to find housekeeping gene candidates. Moreover, we investigate their conservation across species in vertebrates, and explore their biological functions. Importantly, all the finds in NGS RNA-seq were validated in RNA microarrays, which demonstrates the consistency of our method, and an ability to generalize the finds across multiple kidney transplant datasets.

Finally, we demonstrate that the normalization by housekeeping genes found using our method minimize data leakage during cross-validation in machine learning modeling due to its independence of cross-samples information. It can also improve the area under the ROC curve in comparison of cross-samples normalizations, MRN and TMM, even as in independent sample normalization (TPM) comparison.

**Results**

**Gene expression of commonly used housekeeping genes varies in transplanted patients over time.**

We assess whether commonly used housekeeping genes keep constant expression levels in peripheral blood from kidney transplant recipients at different time points (1 week, 3 months and 6 months after transplantation) using the public dataset GSE86884 (PMC4422721). The genes *GAPDH*, *B2M*, *ACTB*, *YWHAZ,* and others, have different log2FoldChange in different post-transplant time points when compared to a pre-transplant baseline (Fig 1A), demonstrating that these are not ideal candidates to be used in normalization methods. The first week post-transplant shows the highest absolute change in most genes. For example, GAPDH expression is downregulated by a log2FoldChange of -1.1 at one-week post-transplant compared to pre-transplant levels. At six months post-transplant, it is downregulated by a log2FoldChange of -0.3, indicating a reduction in downregulation (Fig 1A). We also analyzed the normalized expression profile in each time point and found that all analyzed genes vary their expression levels in different time points (**Fig 1B**). Other traditional housekeeping genes *TBP*, *PPIA*, *RPS13,* and *RPL13* have a high expression value in pre-transplant, while other genes have high expression levels after transplantation, mostly in the first week and three months (**Fig 1B**). All those genes presented variability trough time and conditions in kidney biopsy of transplanted recipients (PMC8659319, PMC6580566), breast cancer (PMC4986254), brain stroke (PMC9296577) and diabetes (PMC9831022). Our data demonstrate that commonly used housekeeping genes should be validated before being used as reference genes in transcriptional studies involving peripheral blood of kidney transplant recipients, as their expression can be influenced by immunosuppression or surgical stress.

**Identification of housekeeping genes in NGS RNAseq**

To find refined alternatives for the housekeeping genes, we reprocess and analyze 635 samples of public NGS RNAseq from the peripheral blood of kidney transplant recipients. We set some parameters to keep only genes with low variability in time and for different clinical outcomes: pre-transplant, 1 week, 3 or 6 months after transplantation; non-rejection, T-cell-mediated rejection (TCMR) and antibody-mediated rejection (ABMR) with or without donor-specific antibodies (DSA). The first rule excludes differentially expressed genes in three different time points considering pretransplant as the baseline, controlling for gender, ethnicity and age. We calculated the non-differential expression genes between pretransplant samples versus one week (**Fig 1C**), three months (**Fig 1D**), or six months (**Fig 1E**) after transplantation in non-rejection patients. We then select the genes that exhibit a log2FoldChange between +0.5 and -0.5 with a significant adjusted p-value across all comparisons testing for equivalent expression: pre-transplant vs. 1 week, 3 months, and 6 months post-transplant. In total, 3563 genes follow these rules in all time points (**Fig 1F**).

Secondly, since some public studies don’t provide fully metadata information about patients, we cluster all samples using a Louvain unsupervised clustering (supp fig X) creating homogeny subgroups. For all 3563 genes we calculate the coefficient of variation (Eq. 1), coefficient of variation of pairwise stability (Eq. 2) and gini coefficient (Eq. 3) for each Louvain group and pooled the results giving the same weight for each cluster.

The ideal housekeeping genes must have a low coefficient of variation through all samples, showing that the expression levels are in the same levels independently of the patient’s condition. As well as the pairwise stability, where the log ratio of all paired gene combinations is calculated to assess how much the expression of one gene varies in respect to another in different conditions (PMC126239 PMC5728501). Following the same idea, lower gini coefficient indicates more stability of the gene expression in different conditions (PMC5840522 PMC6884504 PMC11139953).

To select the genes in an unbiased way, we perform another unsupervised cluster on the pooled metrics per gene (Fig 2ACE) and select the group of genes with lowest values of those metrics. The group 14 has the best candidates for housekeeping genes, since this group has values for coefficient of variation, gini coefficient and coefficient of variance of stability lower than the medians of each of these metrics in all 635 samples (**Fig 2BDF**). With this strategy, we solve the known problem of defining arbitrary cutoff levels of housekeeping genes expression (PMID: 23810203). A total of 85 candidates were selected for the following filtering step.

After selecting the genes that were consistently expressed in all samples 635 samples, we ask whether some of those 85 targets are differentiated expressed rejection and non-rejection states in 384 posttransplant samples from GSE175718 study. We performed Kruskal-Wallis H-test to helps us identify genes with significant differences in expression across non-rejection, ABMR and TCMR (**Sup table 1**) . The genes *AKT2*, *ANKRD11*, *BTG1*, *CYLD*, *EWSR1*, *FUS* and *PRRC2C* have p-values less than 0.05, indicating their expression levels were significantly different among the outcomes. We exclude them from the set of housekeeping genes candidates.

Since the previous test doesn’t guarantee equivalence of the 78 remaining candidates, we perform the Two-One-Side-Test (TOST) with the non-parametric Brunner-Munzel test to verify if the 78 genes have Cohen’s d effect size less than 0.3 (**Sup table 2**). Only the genes *AP2B1*, *CCNI*, *FBXO7*, *GUK1*, *UBB*, *UBXN6*, *VPS28*, *YBX3*, *MT-CO1* have significant adjusted p-value less than 0.05 for this equivalence test (**Fig 2G**). In sum, a total of 9 genes are considered for validation in the Microarray platform (*AP2B1, CCNI, FBXO7, GUK1, UBB, UBXN6, VPS28, YBX3* and *MT-CO1*). Genes *CCNI*, *GUK1*, *UBB, UBXN6* and *VPS28* were also defined as housekeeping genes in another cross non-disease tissue study (PMC7778946).

**Validation of the set of genes in Microarray**

Using different datasets from different technical platforms is crucial to develop robust biomarkers discovery and to build better tools for the diagnosis of rejection. To validate the selected housekeeping genes in a microarray platform, we calculate the Coefficient of variation , gini coefficient and coefficient of variation of stability of the 9 candidates housekeeping genes in four RNA microarray datasets. Like observed for NGS RNAseq, we observe that the genes have low CV (**Fig 3A-D**), low gini coefficients (**FIG 3F-I**), as well low variability in stability metric (**FIG 3J-M**). The gene UBB has the lowest value for all metrics in all studies, followed by the gene MT-CO1. The gene AP2B1 has the highest metric values in the microarray studies.

Another way to validate these genes as housekeeping genes is evaluating how informative they are to distinguish rejection and non-rejection classes in four Microarray studies. To do that, we perform dimensionality reduction and semi-supervised Kmeans clustering (n clusters = 2) using the 9 genes and calculate the entropy of each cluster based on the classes of the samples. High entropy means high heterogeneity within clusters. **Fig 3N-Q** shows for all Microarray datasets that the HKG information does not contribute to reducing the entropy of predicted clusters, which is expected of non-informative features (PMC7943624). Thus, our results consistently demonstrate that the proposed housekeeping genes have low variability in different conditions.

**Candidate housekeeping genes participate in important pathways and have highly conserved sites in 630 species of vertebrates**

We perform Gene Set Enrichment Analysis (PMC3106198) for the 9 potential housekeeping genes to verify their molecular function and what biological process they are involved in. Fig 4A shows the statistically significant process and functions that these genes participate in. Those genes are involved in important molecular functions and biological processes that are crucial to cell living, like maintenance of chromatin structure, DNA repair, stress response, mortality, protein digestion, motility, proliferation, expression and splicing regulation and cell energy.

One important characteristic of housekeeping genes is their conservation across species to verify their importance in evolutionary history (PMC9312424, PMID: 11773595). To evaluate the conservation sites in vertebrates’ homologues we retrieve the orthologue genes of each proposed housekeeping gene from HomoloGene Database (PMID: 34850941). In the upset plot in **Fig 4B**, is possible to verify the total of species that share orthologues. The gene FBXO7 is present in 607 species of 630 species analyzed, while YBX3 is present only in 360 species. In total, 536 species share at least 7 of the 9 housekeeping genes candidates.

For each group of homologues, we perform multiple alignment and calculate the normalized Shannon entropy to quantify the conservation of the amino acids given the *Homo sapiens* sequence as reference. Low values of normalized entropies are highly concentrated in important motifs of each sequence, those regions are highlighted in blue in **Fig 4C**. For example, in MT-CO1 between region 32-69 there is an important Calcium binding region (PMC7378889 cd00054), while in UBB region 1-228 there is a ubiquitin-like region (PMC7378889 cd01803). All regions are described in supplemental (supp XX).

These underline their role as housekeeping genes being highly conserved and playing important molecular functions and biological processes, confirming the ability of our method to identify new candidates to be used in non-cross-samples normalization methods.

**Normalization by housekeeping genes mitigates data leakage and improves rejection predictor**

Finally, to evaluate the impact of the normalization based on housekeeping genes in machine learning modeling, we compare our method with MRN and TMM from the DESeq2 (PMID: 25516281) and EdgeR (PMC2796818) packages, respectively. These methods are the most common and recommended NGS RNAseq normalizations in bioinformatics (PMID: 30041017 PMC10776424 PMC8220791 PMC6209231). We also compare HKG normalization with TPM and the inverse hyperbolic sine of TPM, normalization calculated by the Salmon, the software used in this work to estimate the transcription expression abundance. Differently of MRN and TMM, TPM normalizes independently of other samples (PMC8220791).

Before starting training models, we exclude outliers samples flagged by HDBSCAN algorithm. We exclude genes with a quantity of zero expressions higher than half of total number of samples for each condition. We calculate the pairwise distance correlation to exclude high correlated genes expressions to reduce dimensionality (Measuring and testing dependence by correlation of distances). After that, the remaining number of samples are 378 and 161 genes.

To evaluate only the differences between normalization methods, we define a standard Random Forest model pipeline to be trained and tested in the same condition for each normalization. We perform 25 grouped cross validation for training and testing the models. We use the differential expressed genes findings from PMC9231008 to select features for the modeling *(HOMER3, CD14, IFI27, ZEB2, IL18R1, DAAM2, GBP5, NKG7, PATL2, DGKH* and *SLAMF7*). All classes in train and test datasets are balanced under sampling method with imblearn package (http://jmlr.org/papers/v18/16-365.html). All steps are applied through a scikit-learning (doi/10.5555/1953048.2078195) pipeline to compliance the best practices in machine learning (doi: 10.5555/1756006.1859921). We perform the analysis on 378 samples of GSE175718 samples (PMC9231008).

To calculate normalization factors based on 9 housekeeping genes, we calculate the pairwise correlation distance and cluster them to find groups that could be used as normalization factors **Fig 5A**. We use the inverse hyperbolic sine of TPM to reduce the differences in scales between genes and to transform their distribution close to the gaussian distribution. We calculate the average expression for each pair of clustered genes: *VPS28* with *MT-CO1*, *UBXN6* with *GUK1*, *FBXO7* with *CCNI*. The genes *YBX3*, *UBB* and *AP2B1* were used individually as normalization factors, since they were not clustered closely to any other gene. We also calculate another three normalization factors based on arithmetic mean, quadratic mean and harmonic mean of 9 housekeeping genes combined. After that we create a new dataset for each normalization factor subtracting it from all expression values. The normalization calculated based on the arithmetic mean of gene *VPS28* with *MT-CO1* presents the higher AUC metrics for the Random Forest model. We used this normalization to compare MRN and TMM.

Since MRN and TMM are cross-samples methods, we test two approaches to observe the data leakage impact in cross-samples normalizations. First, we apply the normalization for the entire dataset before splitting it into train and test. We name these approaches MRN and TMM. Secondly, we adapt MRN and TMM to be compatible with scikit-learning pipeline and minimize the data leakage. In this way, the train dataset is normalized separately from the test dataset. Since these methods calculate scaling factors for each sample of the dataset, it is not possible to use the scaling factor from training to testing, since they may have different number of samples. Therefore, we mitigate the data leakage for this method, but do not eliminate it. We name these normalizations as TMM\_adapted and MRN\_adapted.

In **Fig 5B** is possible to verify that when we mitigate the data leakage using MRN\_adapted and TMM\_adapted to apply the MRN and TMM normalizations in train and test datasets separately, we observe the metric area under the ROC curve [AUC] reduce significantly (FDR <= 0.05) in both, train and test dataset. It is possible to infer that MRN and TMM can deliver an overfitted model because of data leakage. Therefore, they should not be used for modeling predictors, but only in investigation of differential expressions, which data leakage is not a problem. Even our versions of MRN\_adapted and TMM\_adapted should be avoided. As observed in **Fig 5B**, the differences between AUC distributions from train to test data set have medians close to 0.4 of median, showing a reduction of 0.4 in AUC between train and test, while VPS28\_MT-CO1 presents a difference median equal 0.25. Besides that, the cross-validation for these adaptations does not normalize the test dataset based on training fitting, but in a new calculation of normalization factors in testing.

Importantly, our proposed housekeeping genes normalization performed better than both MRN\_adapted and TMM\_adapted, delivering a statistically higher AUC in training and testing datasets (**Fig 5A**), as well as a low AUC difference between train and test results. Since housekeeping genes normalization is a non-cross-sample method, there is no risk of data leakage related to cross-sample normalization. When we compare housekeeping genes normalization with TPM, also a non-cross-sample normalization, our method performs better in test dataset. It may be related to the technical effect that can be mitigate when housekeeping genes are used in normalization. Since the technical effects affect the entire sample, using the arithmetic mean can reduce the technical differences between samples. What can be a simple solution for batch effect correction for machine learning applications, leading to better predictors.

**Discussion**

We identify 11 housekeeping genes from peripheral blood samples of kidney transplant recipients that follow standard definitions of this category of genes (PMC9312424). For different time points and clinical conditions, they present low coefficient of variance, high stability, are highly expressed, participate in important cellular maintenance processes and are conservate in vertebrate’s species. More importantly, by identifying these genes with our method, we overcame the data leakage pitfall present in cross-samples normalizations like MRN and TMM, what can give an overoptimistic result (PMC10499856). Additionally, we mitigate the technical differences between samples which are not considered when TPM normalization is applied (PMC8220791).

Previous studies based on the transcriptional profile of tissue biopsies also showed that extensively used housekeeping genes such as *GAPDH*, *B2M*, *RER1*, *RPL13*, *TUBA1A*, *TBP* and *ACTG1* are not good references for kidney transplantation patients (PMC6580566) as well as in other kidney conditions (PMC8659319). These genes commonly used as housekeeping genes have high coefficient of variation and low stability in different conditions. However, there has been no study until now defining housekeeping genes in peripheral blood for this population or confirming that usual housekeeping genes have a high expression variation in different time points or in distinct clinical outcomes (e.g. rejection or non-rejection). Utilizing housekeeping genes with a high CV may lead to erroneous normalization and consequently would affect a study's results and conclusions. A refined normalization method is essential to create machine learning models to predict rejection response from peripheral blood transcriptomics with high specificity and sensitivity. These predictors can lead to minimal evasive diagnostics in liquid biopsy, such as peripheral blood, improving the precision and personalized medicine providing earlier diagnosis and continuous monitoring with a low cost to the patients (PMC10379367 PMC9922467 PMC10466971 https://doi.org/10.1186/s12943-022-01543-7).

Our methodology mine reference genes in high-throughput RNA-sequencing using well established methods to define non-differential expression genes in addition to machine learning approaches as Birch unsupervised clustering. We applied basic statistical concepts like coefficient of variance and pairwise stability in large datasets and used machine learning to find housekeeping genes stably expressed across various conditions and treatments.

We propose that the use of these housekeeping genes must be refined for different tissues and clinical conditions. Therefore, our findings are limited to studies involving peripheral blood from kidney transplanted recipients. Also, it is important to retest the variance and stability of all the proposed housekeeping genes before using them, since there is a lack of information of demographic and other information of RNA-seq public data.

**Conclusions**

In summary, our data mining workflow delivered 11 housekeeping genes based on peripheral blood from kidney transplanted patients. Those genes are highly expressed in all analyzed samples with a low variance and high stability trough different time point, outcomes and different RNA-seq technologies. They play important roles in cell maintenance and are well conserved in vertebrates. We showed that the traditional housekeeping genes are dysregulated in this population, sometimes varying in time or in patient outcome. We also demonstrate that the use of the housekeeping genes to normalize the transcriptome expression data can reduce the overfitting and increase the AUC of machine learning model. Future studies could apply better feature selection, parametrization, and other machine learning algorithms to improve the metrics of the predictors. The results are limited to the peripheral blood of kidney recipients, but our method can be easily applied to different clinical cohorts and the entire workflow is available at [www.github.com](http://www.github.com), including the raw data preprocessing.

**Methods**

**Public NGS RNA-seq processing**

We reprocess 496 samples from three NGS RNA-seq bulk studies from the peripheral blood of kidney transplanted patients. We guarantee that all samples were submitted to the same pipeline and the same reference genome, preserving the reproducibility of the *in silico* experiment, as well as controlling for the preprocessing confounders.

All the raw fastq files are downloaded using sra-toolkit (PMC3013647). We use Fastqc with standard parameters (https://www.bibsonomy.org/bibtex/f230a919c34360709aa298734d63dca3) to evaluate the quality of reads. We trimmed the adapters with Fastp (PMC6129281) and quantified the abundance of transcripts with Salmon (PMC5600148). We used the human transcriptome and the annotation named GRCh38.p14 v.44:2023-03-01 from Gencode as transcriptome reference (PMC7778937).

We excluded pseudogenes from this analysis to keep only transcripts that could be found in microarray. To guarantee the reproducibility, we developed a Snakemake (PMC8114187) workflow to run this entire preprocessing step. The pipeline is in GitHub [LINK].

The dataset GSE86884 is a longitudinal study with 4 timepoints of collected data: pretransplant, one week, 3 and 6 months after transplantation. None of the patients rejected the transplant at the time of collection. The dataset GSE120649 contains 16 samples where 6 patients had stable graft function (non-rejection), 6 with antibody-mediated rejection (ABMR), and 4 with T-cell mediated rejection (TCMR). The dataset GSE175718 contains 384 samples where 248 patients didn’t reject, 86 presented ABMR, 68 presented TCMR, and 18 had concomitant ABMR and TCMR.

**Public Microarray RNA-seq processing**

We download the series matrix of the following microarray studies based on peripheral blood of transplanted recipients GSE14346, GSE15296, GSE46474 and GSE129166. The study GSE14346 contains samples from peripheral blood leukocyte, which were excluded, and samples from peripheral blood where 28 didn’t rejected and 31 presented acute rejection. The study GSE15296 contains 51 patients with rejection and 24 non-rejections. The study GSE46474 classifies patients in acute rejection (n = 20) and non-rejection (n = 20). The study GSE129166 contains 88 non-rejection samples and 29 acute rejections, where 10 were diagnosed with ABMR, 17 with TCMR and two with both types of rejection. Only the last study classifies patients in ABMR and TCMR.

All studies were run on Affymetrix Human Genome U133 Plus 2.0 Array platform. And the expression levels and metadata were extracted from an *in-home* python algorithm available in supplementary material. We transformed the probes names to symbol genes based on the same reference genome used in NGS RNA-seq processing. To update the probes to that genome, we used MyGene.py (PMC3531157) and grouped the duplicated gene symbols by the median of each sample. We only log-transformed the expressions values of GSE14346 and GSE15296 studies because they were stored non log transformed in NCBI.

**Statistical Analysis**

Discovering:We use Tximport (PMC4712774) to import the abundance files to analyze it with DESeq2 (PMC4302049). We perform the test of the alternative hypothesis as “less than” and log2Fold-Change equal 0.5. The genes of interest have differential expression between pretransplant vs one week, 3 months and 6 months in maximum +0.5 and minimum -0.5 of log2Fold-Change. We consider an adjusted p-value less or equal to 0.05 as a significant level.

For each study we use the normalized expression counts from DESeq2 and calculate the coefficient of variation (CV) (Eq-1) and the mean of expression counts (µ) (Eq-2) for all genes. We rescale the µ and the CV with standard scaler and clustered the genes by CV and µ with Birch algorithm (BIRCH: An efficient data clustering method for large databases.) from scikit-learning (Scikit-learn: Machine Learning in {P}ython).

We calculate pairwise log ratio of genes expression per sample from the first five clusters with lowest CV median to filter genes by stability. We calculate the coefficient of variation of stability (CVS) (PMC126239 PMC5728501) (CVS) (Eq-3) keeping the CVS lower than the first quartile of all CVS combined pairs. Then, we count the number of pairs of each gene with the remaining stability pairs. A lower head-to-head coefficient of variation of stability (CVS) to multiple genes means the better housekeeping gene candidate (PMC126239 PMC5728501).

We chose the genes with the lowest CVS for highest number of pairs to calculate Pearson correlation and test the hypotheses with Kruskal-Wallis H-test to verify whether the genes expression in different conditions have same distribution. Both Pearson correlation and Kruskal-Wallis are calculated with SciPy ({SciPy} 1.0: Fundamental Algorithms for Scientific Computing in Python).

Microarray validation: We use the Eq-1 to calculate the CV of all genes in each microarray study, and the Eq-2 to calculate the mean of expression. We used UMAP (UMAP: Uniform Manifold Approximation and Projection) to reduce the dimensionality of the housekeeping genes expression matrix to visualize the samples in two dimensions. We calculate the k-means for 2 clusters with scikit-learning. For each cluster we calculate the Shannon entropy with SciPy.

Pathway Analysis: We used GSEApy (GSEApy: a comprehensive package for performing gene set enrichment analysis in Python) to filter pathways related to the housekeeping genes. We performed the enrichment analysis against GO Molecular Function and GO Biological Process

Expression Atlas DB: We use the Expression Atlas to retrieve expression values of the housekeeping genes in GTEx studies. We searched for those genes and downloaded the FPKM matrix. We used seaborn to plot the heatmap of **Fig 3**.

Orthology Analisys: We search for orthologues in HomoloGene database from NCBI. We download all the vertebrate codon sequences and align them with Clustal Omega algorithm (PMC9252731). For each site of each aligned gene, we calculate the normalized Shannon entropy to quantify the conservation of the amino acids given the *homo sapiens* sequence as reference.

Machine learning to evaluate normalizations: We perform the analysis on samples of GSE175718 (PMC9231008). This public dataset doesn’t have batch information, so we calculated the K-neighbors (doi/10.5555/1953048.2078195) and clustered the samples by Louvain algorithm from scikit-network package (http://jmlr.org/papers/v21/20-412.html) to detect similar clusters and avoid that samples from the same group could be used in train and test for a cross-validation (supp FIG-X). We define 52 combinations to cross-validate between different groups. We adapted the MRN and TMM algorithm to be used in a scikit-learning pipeline. To avoid imbalanced classes, we used the method RandomUnderSampler from imblearn package (<http://jmlr.org/papers/v18/16-365.html>). We construct a pipeline in scikit-learning (doi/10.5555/1953048.2078195) to undersamplig, normalize, select features, and train/test a Random Forest for different normalization methods for the same set of subsamples. We apply cross-validation and calculate the AUC for each set of training and test. All the algorithms and pipelines are in supp and git hub.Data manipulation and plotting: We use Pandas (10.25080/Majora-92bf1922-00a) and Numpy (https://doi.org/10.1038/s41586-020-2649-2) to manipulate the data and Matplotlib (10.1109/MCSE.2007.55) and Seaborn (https://doi.org/10.21105/joss.03021) to plot them.

The third rule for filtering housekeeping genes is pairwise stability. The pairwise stability, defined in Eq. 3, is the log ratio of all paired gene combinations to assess how much the expression of one gene varies in respect to another in different conditions. A lower head-to-head coefficient of variation of stability (CVS) to multiple genes means the better housekeeping gene candidate (PMC126239 PMC5728501). We keep the CVS values lower than the first quartile and select the genes that have a low variance of stability between higher quantity of pairs. We propose 11 housekeeping genes (plus *UBB*, *MT-MDL4,* and *BTG1*) highlighted in yellow with low CVS (less than 0.3) between at least 1000 other genes in a pool of 1186 evaluated genes (**Fig 1I**).

We next confirm their stability using pairwise Pearson correlation between our proposed 14 genes. We set a cutoff of R Pearson correlation of ≥+0.5 and p-value ≤ 0.05. Only the gene *UBB* did not correlate well with any other stable genes since its R statistics value varies between -0.08 to +0.32 (**Fig 1J**). All other genes correlate with at least four different genes. Genes *PABPC1 and BTG1* are correlated with all other genes but *UBB*. Genes *TMSB10*, *COX4I1,* and *SRRM2* are correlated with most stable genes, except with UBB and *PCBP2*. It is important to note that the order of investigating first stability and after correlation were strategically decided based on the computational source consumption. Calculating the pairwise stability is much faster than the correlation of 1186 genes. This can be useful for other studies that need to mine a greater pool of genes.

**Expression of candidate housekeeping genes vary in different tissues**

To better characterize the 9 candidate genes selected by our method, we investigate their expression in other tissues using the Expression Atlas database (PMC4702781) of GTEx studies (PMID: 32913098). As we can see in **Fig 3 A**, the gene *TMSB10* has high expression in most different types of tissues, demonstrating its high ubiquitousness. Other genes are also present in several types of tissues with high expression. But the genes *RBM39*, *PCBP2*, and *COX4I1* are grouped in the cluster with lower expression levels throughout tissues. Besides those genes are expressed in different tissues showing their importance in different regions, their expressions levels vary in each tissue. This behavior reasserts the importance of defining housekeeping genes for different tissues and conditions and underlies the utility of our method exemplified here by peripheral blood from kidney transplant recipients.

The normalization of housekeeping genes is defined by dividing all genes in each sample by the median of housekeeping genes of the respective sample. TPM normalization is transformed in log10 to reduce the differences in scales between genes and to transform their distribution close to the gaussian distribution. Since MRN and TMM are cross-samples methods, we test two approaches to observe the data leakage impact in cross-samples normalizations. First, we apply the normalization for the entire dataset before splitting it into train and test. We name these approaches as MRN and TMM. Secondly, we adapt MRN and TMM to be compatible with scikit-learning pipeline and minimize the data leakage. In this way, the train dataset is normalized separately from the test dataset. since these methods calculate scaling factors for each sample of the dataset, it is not possible to use the scaling factor from training to testing, since they may have different number of samples. Therefore, we mitigate the data leakage for this method, but do not eliminate it. We name these normalizations as TMM\_adapted and MRN\_adapted.