**Housekeeping Gene Expression Normalization in Transcriptomics Mitigates Data Leakage in Machine Learning Models**

Guilherme T. Ribas1,2, Cristian V. Riella3, Dieval Guizelini2, Maurício M. Rigo4, Leonardo V. Riella1\*†,Thiago J. Borges1\*†

1\*Center for Transplantation Science, Massachusetts General Hospital/Harvard Medical School, 149 13th St, Boston, 02129, MA, USA.

2Professional and Technological Education Sector, Federal University of Paraná, R. Dr. Alcides Vieira Arcoverde, 1225, Curitiba, 81520-260, PR, Brazil.

3Nephrology Division, Department of Medicine, Beth Israel Deaconess Medical Center, 330 Brookline Ave, Boston, 02215, MA, USA.

4 Center for Discovery and Innovation, Hackensack Meridian Health, Street, Nutley, XXXXXX, NJ, USA.

†These authors contributed equally to this work.

\*Corresponding author(s). E-mail(s): [tdejesusborges@mgh.harvard.edu;](mailto:gtabordaribas@mgh.harvard.edu) [lriella@mgh.harvard.edu;](mailto:lriella@mgh.harvard.edu)

Contributing authors: [gtabordaribas@mgh.harvard.edu;](mailto:gtabordaribas@mgh.harvard.edu;) [criella@bidmc.harvard.edu;](mailto:criella@bidmc.harvard.edu) [dievalg@gmail.com;](mailto:dievalg@gmail.com;) [mauriciomr1985@gmail.com;](mailto:mauriciomr1985@gmail.com)

**Abstract**

Inappropriate normalization can lead to data leakage and overfitting in machine learning models. Accurately identifying housekeeping genes (HKGs) can overcome this problem and is crucial for normalizing gene expression data, particularly in RNA-Seq experiments. Using cohorts of kidney transplant recipients, we demonstrate that the gene expression of commonly used HKGs experience significant alterations over time due to immunosuppressive treatments. By analyzing large public RNA-Seq and microarray datasets, we identified nine stable and better-suitable HKG candidates. Our results demonstrate that these HKGs improve the robustness and generalizability of machine learning models by minimizing data leakage, as evidenced by superior performance compared to benchmark methods like Median Ratio Normalization **(**MRN) and trimmed mean of M values (TMM). This work underscores the importance of niche-specific HKG in developing reliable biomarkers and diagnostic tools, providing a standardized Python package for broader applications.

**Keywords:** Data leakage, machine learning, normalization, housekeeping genes, kidney transplantation.

# Introduction

To calculate the differential expression of genes, discover biomarkers, and model machine learning predictors, it is fundamental to normalize the data expression using baseline levels of control genes, also named housekeeping genes (HKGs). The current benchmark methods to normalize transcript expression data in RNA-Seq experiments are MRN and TMM, included in packages DESeq2 [[1](#_bookmark1)] and EdgeR [[2](#_bookmark2)], respectively. Those methods are based on cross-sample normalizations [[3](#_bookmark3)][[4](#_bookmark4)][[5](#_bookmark5)], which can be a source of data leakage and following overfitted machine learning models [[6](#_bookmark6)][[7](#_bookmark7)], compromising the reproducibility and generalizations of the models on new datasets [[8](#_bookmark8)]. An alternative to cross-sample normalization is defining housekeeping genes for a specific niche, such as kidney transplanted recipients, and normalizing each sample by them independently of other samples.

Although the biological definition of housekeeping genes is still not fully determined, one can define them as genes with constant expression in all conditions, with an essential role in cellular maintenance and a conserved sequence in evolutionary history [[9](#_bookmark9)][[10](#_bookmark10)][[11](#_bookmark11)]. Nevertheless, since the expression levels of a gene can vary with drug treatment [[12](#_bookmark12)] and clinical conditions [[13](#_bookmark13)][[14](#_bookmark14)][[15](#_bookmark15)][[16](#_bookmark16)], different diseases and treatments can lead to a different set of HKGs. Thus, they should be defined for cohorts subject to specific treatments and conditions and be carefully analyzed before their application in experiments and diagnosis.

Transplantation is the treatment of choice for patients with end-stage kidney disease. In this context, it is essential 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 genes, including traditional housekeeping genes [[17](#_bookmark17)][[18](#_bookmark18)]. The commonly used housekeeping genes *GAPDH*, *ACTB,* and *B2M* have their expression dysregulated in the biopsy of kidney transplanted recipients [[19](#_bookmark19)], underlying the need for a better definition of housekeeping genes in this population in different tissues, including blood.

To address the lack of information about appropriate housekeeping genes to serve as controls in experiments using peripheral blood of kidney transplanted recipients, we initially develop a workflow based on large public datasets of NGS RNA-Seq and RNA microarrays with different time points and different clinical outcomes (Table [1](#_bookmark0)). We initially explore well-established methods to define non-differential expression genes. We then apply statistical concepts like coefficient of variance, Gini coefficient, and pairwise stability in an extensive dataset and use machine learning concepts to find the best housekeeping gene candidates. Moreover, we investigate their conservation across vertebrate species and explore their biological functions and processes. Importantly, all the findings in NGS RNA-Seq were validated in RNA microarrays, demonstrating our method's consistency and ability to generalize the findings across multiple kidney transplant datasets.

Finally, we demonstrate that the normalization by housekeeping genes found using our method minimizes data leakage during cross-validation in machine learning modeling due to its independence of cross-sample information. It can also improve the area under the ROC curve compared with cross-sample normalizations, MRN and TMM, even in independent sample normalization (TPM) comparison. The housekeeping gene mining process is organized in a Python package and can be easily used in other datasets.

# Results

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

We assess whether commonly used housekeeping genes [[19](#_bookmark19)] keep constant expression levels in peripheral blood from kidney transplant recipients at different time points (one week, three months, and six months after transplantation) using the public dataset GSE86884 [[20](#_bookmark20)]. Most genes are differentially expressed at least one time point in post-transplant when compared to a pre-transplant baseline (Fig 1A). The first week post-transplant shows a bigger absolute change for most genes. For example, *GAPDH* expression is downregulated by a log2FoldChange of -1.1 at one week post-transplant compared to pre-transplant levels. Six months post-transplant, it is downregulated by a log2FoldChange of -0.3, indicating differences in its expression dynamics through time (Fig 1A). We also analyzed the normalized expression profile at each time point and found that all analyzed genes varied expression levels at different time points (Fig 1B). Other traditional housekeeping genes, *TBP*, *PPIA*, *RPS13*, and *RPL13*,have a high expression value in pre-transplant. In contrast, other genes have high expression levels after transplantation, mainly in the first week and three months (Fig 1B).

We next analyze those genes in the context of non-rejection and different types of kidney

rejection events: T-cell-mediated rejection (TCMR) and antibody-mediated rejection (ABMR) from a different data set (GSE120649 [[21](#_bookmark21)]). None of those genes present a statistically significant difference. However, *B2M* presented a log2FoldChange of -0.7 with a p-adjusted = 0.08 in ABMR; a similar behavior is seen in *HPRT1* in ABMR and *TUBA1A* for TCMR in terms of log2FoldChange absolute values (FIG 1C-D).

Analogous to our results, all those commonly housekeeping genes presented variability through time and different conditions in kidney biopsy of transplanted recipients [[22](#_bookmark22)][[19](#_bookmark19)], breast cancer [[14](#_bookmark14)], brain stroke [[13](#_bookmark13)], and diabetes [[15](#_bookmark15)]. 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 RNA-Seq

To determine potential alternatives for the traditional housekeeping genes, we reprocess and analyze 731 samples of public NGS RNA-Seq from the peripheral blood of kidney transplant recipients (Table 1?). We set a pipeline to keep only genes with low variability in time and for different clinical outcomes: pre-transplant, one week, three or six months after transplantation; non-rejection, T-cell-mediated rejection (TCMR) and antibody-mediated rejection (ABMR) with or without donor-specific antibodies (DSA). First, we exclude differentially expressed genes at three different time points, considering pre-transplant as the baseline and controlling for gender, ethnicity, and age. We calculate the non-differential expression genes between pre-transplant samples versus one week (Fig 2A), three months (Fig 2B), and six months (Fig 2C) after transplantation in 96 non-rejection patients from study GSE86884 [[20](#_bookmark20)]. We then select the genes that exhibit a log2FoldChange between -0.5 and +0.5 with adjusted p-value *≤* 0.05 across all comparisons: pre-transplant vs. one week, three months, and six months post-transplant. In total, 3,563 genes are considered equivalent in all time points (Fig 2D).

Secondly, since some public studies cannot provide complete patient metadata information, we cluster all samples using a Louvain unsupervised clustering to create homogeneous subgroups (Fig S1). For all 3,563 genes, we calculate the coefficient of variation (CV) with equation (1), coefficient of variation of pairwise stability (cvSTB) (equation 2), and Gini coefficient (equation 3). The ideal housekeeping genes must have a low coefficient of variation through all samples, showing that the expression levels are at the same level independently of the patient’s condition. The same is true for pairwise stability; the log ratio of all paired gene combinations is calculated to assess how much the expression of one gene varies with another in different conditions [[23](#_bookmark23)][[24](#_bookmark24)]. Following the same idea, a lower Gini coefficient indicates more stability of the gene expression in other conditions [[25](#_bookmark25)][[26](#_bookmark26)][[27](#_bookmark27)].

To select the genes in an unbiased way, we perform another unsupervised cluster on the CV, cvSTB, and Gini metrics per gene (Fig 2E, G, I) and choose the group of genes with the lowest values of those metrics in 635 samples with rejection and non-rejection samples. Group 14 has the best candidates for housekeeping genes since this group has values for the 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 2F, H, J). This strategy solves the known problem of defining arbitrary cutoff levels of housekeeping gene expressions [[11](#_bookmark11)]. A total of 85 candidates were selected for the following filtering step.

After selecting the genes consistently expressed in all 635 samples, we ask whether some of those 85 targets are differentiated expressed by comparing rejection and non-rejection states in 384 posttransplant samples from the GSE175718 [[28](#_bookmark28)] study. We performed the Kruskal-Wallis H-test to identify genes with significant differences in expression across non-rejection, ABMR, and TCMR (Table S1). The genes *AKT2, ANKRD11, BTG1, CYLD, EWSR1, FUS,* and *PRRC2C* have p-values *≤* 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-Sided-Test (TOST) with the non-parametric Brunner-Munzel test to verify whether the 78 genes have Cohen’s d effect size less than 0.3 when compare non-rejection vs ABMR and non-rejection vs TCMR (Table S2). Only the genes *AP2B1, CCNI, FBXO7, GUK1, UBB, UBXN6, VPS28, YBX3,* and *MT-CO1* have a significant adjusted p-value *≤* 0.05 for this equivalence test (Fig 2K).

In sum, a total of 9 genes (*AP2B1, CCNI, FBXO7, GUK1, UBB, UBXN6, VPS28, YBX3* and *MT-CO1*) are stable in different conditions and different time-pionts, presenting low CV, low cvSTB, low Gini coefficient end are statistically equivalent. *CCNI, GUK1, UBB, UBXN6,* and *VPS28* were also defined as housekeeping genes in another cross-non-disease tissue study [[29](#_bookmark29)]. Those selected genes are considered for validation in the Microarray platform.

## Validation of housekeeping genes in microarray

To validate the selected housekeeping genes in a microarray platform, we calculate the coefficient of variation, Gini coefficient, and coefficient of variation of stability for all genes in four RNA microarray datasets (Table 1). As observed in NGS RNA-Seq, the genes *AP2B1, CCNI, FBXO7, GUK1, UBB, UBXN6, VPS28, YBX3,* and *MT-CO1* have low CV (Fig 3A-D), low Gini coefficient (Fig 3E-H), as well as low variability in stability metric (Fig 3I-L). 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 to evaluate how informative they are to distinguish rejection and non-rejection classes in four microarray studies. Genes with high stability between classes are not able to distinguish between these two conditions. We perform dimensionality reduction and semi-supervised Kmeans clustering (n clusters = 2) using the nine genes and calculate each cluster's entropy based on each sample's class. High entropy means high heterogeneity within clusters. Fig 3M-P shows for all microarray datasets that the HKGs information does not reduce the entropy of predicted clusters, which is expected of non-informative features [[30](#_bookmark30)]. Thus, our results consistently demonstrate that the proposed housekeeping genes have low variability in different clinical outcomes.

## Candidate housekeeping genes participate in critical pathways and have highly conserved sites in 630 species of vertebrates

We perform Gene Set Enrichment Analysis [[31](#_bookmark31)] for the nine potential housekeeping genes to verify their molecular function and involvement in biological process. The selected genes are involved in critical molecular functions and biological processes 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 (Fig 4A).

An important characteristic of housekeeping genes is their conservation across species to verify their importance in evolutionary history [[9](#_bookmark9)][[10](#_bookmark10)]. To evaluate the conservation sites in vertebrates’ homologs, we retrieve the orthologue genes of each proposed housekeeping gene from HomoloGene Database [[32](#_bookmark32)]. In the upset plot in Fig. 4B, it is possible to verify the total of species that share orthologues. The gene *FBXO7* is present in 607 species of 630 analyzed, while *YBX3* is present only in 360 species. 536 species share at least seven of the nine housekeeping gene candidates. The upset plots with all combinations for one Phylum, eight Classes, 111 Orders, 288 Families, 500 Genera, and 630 Species are available in Fig S2.

Further, to quantify the conservation of the amino acids given the *Homo sapiens* sequence as a reference, we perform multiple alignments and calculate the normalized Shannon entropy for each group of homologs. Low normalized entropy values are highly concentrated in essential motifs of each sequence; those regions are highlighted in blue in Fig 4C. For example, in *MT-CO1* between regions 32-69, there is an important Calcium binding region ([33] cd00054), while in UBB region 1-228, there is a ubiquitin-like region ([[33](#_bookmark33)] cd01803). All regions and entropy calculations are described in supplemental (Table S3).

These data 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-sample 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 [[1](#_bookmark1)] and EdgeR [[2](#_bookmark2)] packages, respectively. These methods are the most common and recommended NGS RNA-Seq normalizations in bioinformatics [[4](#_bookmark4)][[34](#_bookmark34)][[3](#_bookmark3)][[5](#_bookmark5)]. We also compare HKG normalization with TPM and the inverse hyperbolic sine of TPM. Unlike MRN and TMM, TPM normalizes independently of other samples [[3](#_bookmark3)].

After excluding outlier samples flagged by the HDBSCAN algorithm, we define a standard Random Forest model pipeline to cross-validate 25 times in training and test datasets under the same conditions for each normalization. We use the differential expressed genes findings from the work of Loon V. et al. [28] 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 with imblearn package [[35](#_bookmark35)]. All steps are applied through a scikit-learning [[36](#_bookmark36)] pipeline to compliance the best practices in machine learning [[37](#_bookmark37)]. We perform the analysis on 378 samples of GSE175718 samples [[28](#_bookmark28)].

To normalize the expression data based on our findings, we calculate the pairwise correlation distance [38] between the nine proposed HKGs and cluster them to find groups that could be used as normalization factors (Fig 5A and Fig S2). 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 arithmetic mean (equation 4), geometric mean (equation 5), harmonic mean (equation 6), and the root means square (equation 7) of expression of all HKG together for each gene clusters and for pairs within each cluster (Fig 5A). *UBB* and *AP2B1* are used individually as normalization factors since they are not clustered closely to any other gene. 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 genes *VPS28* and *MT-CO1* presents the higher AUC metric distribution in the test dataset for the Random Forest model, followed by the harmonic mean of genes *FBXO7* and *CCNI*. The harmonic mean of all proposed housekeeping genes also presents a higher AUC than MRN and TMM.

Since MRN and TMM are cross-sample methods, we test two approaches to observe the impact of data leakage in cross-sample 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 adapted MRN and TMM to be compatible with the scikit-learning pipeline and minimize data leakage. This way, the training dataset is normalized separately from the test dataset. While these methods calculate scaling factors for each dataset sample, using the scaling factor from training to testing is not possible since they may have different samples. Therefore, we mitigate the data leakage for MRN and TMM but do not eliminate it. We name these normalizations as TMM\_adapted and MRN\_adapted.

It is possible to infer that MRN and TMM can deliver an overfitted model because of data leakage, based on Fig 5A. When we mitigate the data leakage using MRN\_adapted and TMM\_adapted, we observe the distribution of the area under the ROC curve (AUC) reduce significantly in both the train and test datasets (p-adjusted *≤* 0.05 in Fig 5B). Therefore, they should not be used for modeling predictors but only in investigating differential expressions, in which data leakage is not a problem. Even our versions of MRN\_adapted and TMM\_adapted should be avoided. As observed in Fig 5C-D, the differences between AUC distributions from the train to test data set have higher values, showing a reduction of around 0.4 in the AUC median between train and test for the adapted version of TMM and MRN. Conversely, the normalization based on the arithmetic mean of *VPS28* and *MT-CO1* presented a reduction of around 0.2 (Fig 5C-D). 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-B) and a low AUC difference between train and test results (Fig 5C-D). When we compare the AUC metrics with MRN and TMM, the normalizations based on pairs of genes *VPS28* with *MT-CO1* and *FBXO7* with *CCNI* deliver a better test AUC distribution (FIG 5A-B) and the lowest difference between train and test (Fig 5C-D). Since housekeeping gene normalization is a non-cross-sample method, there is no risk of data leakage related to cross-sample normalization. Our method performs better in the test dataset even when we compare housekeeping genes normalization with TPM, which is also a non-cross-sample normalization. This may be related to the technical effect that can be mitigated when housekeeping genes are used in normalization. Since the technical effects affect the entire sample, using the HKGs as normalization factors can reduce the technical differences between samples. We propose that this method can be a simpler solution for batch effect correction for machine learning applications, leading to better predictors.

# Discussion

This study presents new normalization factors based on identifying more robust housekeeping genes in transcriptomics datasets to avoid data leakage and mitigate technical differences between samples. Our method outperforms traditional cross-sample normalization methods such as MRN and TMM, which can deliver over-optimistic results [[6](#_bookmark6)]. As well as the TPM, an independent sample method that cannot correct for technical artifacts [3]. To construct these normalization factors, we mine through thousands of genes, choosing the highly expressed ones with low coefficients of variation, high stability, and low Gini coefficient across various conditions and treatments in high-throughput RNA sequencing. Besides, those genes participate in crucial cellular maintenance processes and are conserved in vertebrate species, following standard definitions of housekeeping genes [[9](#_bookmark9)].

Previous studies based on the transcriptional profile of tissue biopsies also showed that extensively used housekeeping genes such as *GAPDH, B2M, RER1, RPL13, TUBA1A,* and *ACTG1* are not good references for kidney transplantation patients [[19](#_bookmark19)] as well as in other kidney conditions [[22](#_bookmark22)]. 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 at different time points or in distinct clinical outcomes (e.g., rejection or non-rejection). Furthermore, differently from other studies, we investigate the biggest dataset of kidney transplanted recipients; we insert the Gini coefficient along with other metrics to measure stability; we use unsupervised method to define the stale genes; we perform the equivalence test (TOST) to make sure that the difference is within a specified interval, while one study considered stable genes the genes that were not differentially expressed, which a mistake, since the absence of differential expression evidence, is not evidence of non-differentially.

A refined normalization method based on consistently stable genes is essential to create machine learning models to predict rejection responses from peripheral blood transcriptomics with high specificity and sensitivity. We demonstrate that predictors modeled on data normalized by our factors can deliver high Area Under the Curve ROC values in the test dataset avoiding data leakage. Mitigating this source of overfitting is crucial to lead to better models which can contribute to decision makers in health. It can meliorate predictors that can lead to minimal invasive diagnostics in liquid biopsy, such as peripheral blood. This approach has the potential to enhance precision and personalized medicine significantly, allowing for earlier diagnosis and continuous monitoring at a low cost to the patients [[39](#_bookmark39)][[40](#_bookmark40)][[41](#_bookmark41)][[42](#_bookmark42)].

While our findings are promising, they are currently limited to peripheral blood samples from kidney transplant recipients. Although our methodology can be easily applied to any cohort through the Python package developed in this work, further research is necessary to validate these HKGs across other tissues and in different transplant contexts. Additionally, the lack of demographic data in public RNA-Seq datasets poses a challenge in generalizing these results to broader populations.

Future studies should focus on validating these HKGs in larger, more diverse cohorts, including different demographic groups and conditions. Additionally, exploring the application of these HKGs or the methodology in other organ transplants could provide insights into their broader utility. Investigating the normalization based on these genes in different machine-learning models would also be beneficial, potentially leading to the development of robust, clinically applicable algorithms for early rejection detection.

Developing more reliable normalization methods using niche-specific HKGs marks a significant step forward in integrating machine learning with clinical genomics. By minimizing data leakage and enhancing model generalizability, our approach could pave the way for more accurate and reproducible biomarker discovery, ultimately contributing to better patient outcomes in transplantation medicine.

We derive a method to normalize transcriptomic data based on identifying more robust housekeeping genes. Using this method, we overcome the data leakage pitfall in cross-sample normalizations like MRN and TMM, which can give an over-optimistic result [[6](#_bookmark6)]. Additionally, we mitigate the technical differences between samples, which are not considered when TPM normalization is applied [[3](#_bookmark3)]. Our methodology mines stably expressed genes across various conditions and treatments in high-throughput RNA sequencing. We use well-established methods to define non-differential expression genes, coefficient of variance, variance of pairwise stability, and gini coefficient in large datasets in addition to machine learning approaches such as unsupervised clustering. We identify nine housekeeping genes from peripheral blood samples of kidney transplant recipients that follow standard definitions of this category of genes [[9](#_bookmark9)]. For different time points and clinical conditions, they present a low coefficient of variance, low variation in stability, and low Gini coefficient. They are highly expressed, participate in crucial cellular maintenance processes, and are conserved in vertebrate species. Therefore, these house keeping genes are ideal…

Previous studies based on the transcriptional profile of tissue biopsies also showed that extensively used housekeeping genes such as *GAPDH, B2M, RER1, RPL13, TUBA1A,* and *ACTG1* are not good references for kidney transplantation patients [[19](#_bookmark19)] as well as in other kidney conditions [[22](#_bookmark22)]. 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 at different time points or in distinct clinical outcomes (e.g., rejection or non-rejection). Utilizing non-stable genes as housekeeping genes may lead to erroneous normalization and affect a study’s results and conclusions

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, retesting the variance and stability of all the proposed housekeeping genes before using them is important since there is a lack of demographic and other information on RNA-Seq public data.

# Methods

## Public NGS RNA-Seq processing

We reprocess 731 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 and controlling for the preprocessing confounders. The GEO ID and summary of the samples are in Table 1.

**Table 1** Public peripheral blood transcriptomics data deposited in NCBI.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| GEO id | Pos-tx | Pre-tx | Technology | Rejection | Non-rejection |
| GSE175718 [[28](#_bookmark28)] | 384 | 0 | RNA-Seq | 136 | 248 |
| GSE120649 [[21](#_bookmark21)] | 16 | 0 | RNA-Seq | 10 | 6 |
| GSE86884 [[20](#_bookmark20)] | 32 | 64 | RNA-Seq | 0 | 96 |
| GSE112927 [[43](#_bookmark43)] | 0 | 235 | RNA-Seq | 83 | 152 |
| GSE14346 [[44](#_bookmark44)] | 59 | 0 | RNA microarray | 31 | 28 |
| GSE15296 [[45](#_bookmark45)] | 75 | 0 | RNA microarray | 51 | 24 |
| GSE46474 [[46](#_bookmark46)] | 40 | 0 | RNA microarray | 20 | 20 |
| GSE129166 [[47](#_bookmark47)] | 117 | 0 | RNA microarray | 47 | 70 |
| TOTAL | 723 | 299 | - | 374 | 648 |

All the raw fastq files are downloaded using sra-toolkit 3.0.0 [[48](#_bookmark48)]. We use Fastqc

0.11.9 with standard parameters [[49](#_bookmark49)] to evaluate the quality of reads. We trimmed the adapters with Fastp 0.22.0 [[50](#_bookmark50)] and quantified the abundance of transcripts with Salmon 1.9.0 [[51](#_bookmark51)]. We used the human transcriptome and the annotation named GRCh38.p14 v.44:2023-03-01 from Gencode as transcriptome reference [[52](#_bookmark52)].

We exclude pseudogenes from this analysis to keep only transcripts that could be found in microarray. To guarantee the experiment's reproducibility, we developed a Snakemake [[53](#_bookmark53)] workflow to run this entire preprocessing step. The pipeline is on GitHub.

The dataset GSE86884 is a longitudinal study with 96 samples and four time points 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 six 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 present rejection, 86 presented ABMR, 68 presented TCMR, and 18 had concomitant ABMR and TCMR. The dataset GSE112927 contains 235 pretransplant samples with 152 non-rejection and 83 patients with rejection.

## Public microarray processing

We download the CEL files of the following microarray studies based on the peripheral blood of transplanted recipients GSE14346, GSE15296, GSE46474, and GSE129166 (Table 1). The study GSE14346 contains samples from peripheral blood leukocytes, which were excluded, and samples from peripheral blood, where 28 didn’t reject 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 70 non-rejection samples and 47 acute rejections, where 21 were diagnosed with ABMR, 17 with TCMR, and 9 with both types of rejection. Only the last study provides ABMR and TCMR classification.

All studies were run on Affymetrix Human Genome U133 Plus 2.0 Array platform. We extract the expression levels with the R 4.4.0 package affy 1.82.0 [[54](#_bookmark54)].

The metadata was extracted from an *in-home* Python 3.12.1 algorithm. We transformed the names of the probes 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 3.2.2 [[55](#_bookmark55)] and grouped the duplicated gene symbols by the median of each sample. MyGene package didn’t translate the probe ID to genes MT-CO1 and UBNX6; we manually searched for their probe codes on the GEO platform database (h[ttps://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL570).](http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL570))

## Statistical Analysis

*Discovering:* We use Tximport 1.32.0 [[56](#_bookmark56)] to import the abundance files and analyze them with DESeq2 1.44.0 [[1](#_bookmark1)]. We test the alternative hypothesis as “greater than” and log2FoldChange equal 0 for FIG 1. In FIG 2A-C, we use the alternative hypothesis “less than” and log2FoldChange equals 0.5. The genes of interest have differential expression between pre-transplant vs one week, three months, and six months in maximum +0.5 and minimum -0.5 of log2FoldChange. We consider an adjusted p-value *≤* 0.05 as a significant level. For each study, we use the normalized expression counts from DESeq2 and calculate the coefficient of variation (CV) (equation 1), coefficient of variation of pairwise stability (cvSTB) (equation 2), and Gini coefficient (equation 3) for all *n* genes of *m* samples with *in-home* Python 3.12.1 script.

A math equations on a white background

Description automatically generated

Where e is the gene expression value, µ is the average gene expression, and *σ* is the standard deviation.

Since the majority of the public dataset does not inform batch sample process, demography, and health conditions of the patients, we cluster the samples of each study with a Louvain algorithm from scikit-network 0.31.0 python package [[57](#_bookmark57)], we calculate CV and cvSTB for each group and pool the values giving the same weight for each cluster.

After calculating the metrics for all genes, we clustered the genes to group those with similar patterns using the Louvain algorithm. We tested if the best candidates have different distributions with the Kruskal-Wallis H-test and performed an equivalent Brunner-Munzel One-Two-Sided-Test (TOST) with a Cohen’s d equal 0.3. Kruskal- Wallis H-test is calculated with SciPy 1.11.4 [58], while TOST is calculated with an *in-home* algorithm.

*Microarray validation*: We calculate the coefficient of variation, coefficient of variance of stability, and Gini coefficient of all genes in each microarray study. We used UMAP from umap-learn 0.5.5 python package [[59](#_bookmark59)] to visualize the samples of the nine housekeeping genes in two dimensions. We calculate the K-means clustering for k=2 with scikit-learning 1.4.0. We calculate the Shannon entropy for each cluster with SciPy 1.11.4[[58](#_bookmark58)].

Pathway Analysis: We used GSEApy 1.1.1 [[60](#_bookmark60)] to filter pathways related to housekeeping genes. We performed the enrichment analysis against GO Molecular Function and GO Biological Process 2023.

*Orthology Analysis*: We search for orthologues in the HomoloGene database from NCBI. We identify unique Phylum, Classe, Order, Family, Genus, and Species with an *in-home* script. We download all the vertebrate amino acid sequences with NCBI Datasets tools 16.22.1 [[61](#_bookmark61)] and align them using the Clustal Omega 1.2.4 algorithm [[62](#_bookmark62)]. For each site of each aligned gene, we calculate the normalized Shannon entropy with SciPy 1.11.4 [[58](#_bookmark58)] to quantify the conservation of the amino acids given the homo sapiens sequence as a reference.

*Machine learning to evaluate normalizations*: We analyze samples of GSE175718 [[28](#_bookmark28)]. We detect outliers with HDBSCAN 0.8.36 python package [[63](#_bookmark63)] on non-cross-sample normalization TPM. We calculate the pairwise correlation distance [[38](#_bookmark38)] between genes and keep only one gene from groups of highly correlated genes. We calculate the pairwise R Pearson correlation between the nine housekeeping genes and cluster them based on distance correlation and the average hierarchical linkage method. We chose a cophenetic distance for 4 clusters based on Silhouette, Calinski Harabasz, and Davies Bouldin scores (FIG S3). We used SciPy 1.11.4 and scikit-learning 1.4.0 [[36](#_bookmark36)] to perform all the previous analyses.

We calculate the original MRN and TMM normalizations based on the package conorm (https://pypi.org/project/conorm/), as well as create two Python classes to integrate an adapted version of MRN and TMM transformations to a scikit-learning pipeline [[36](#_bookmark36)]. We used the following equations for each sample to calculate the normalization factors based on different means, where e is the expression value of *n* genes used as housekeeping genes.

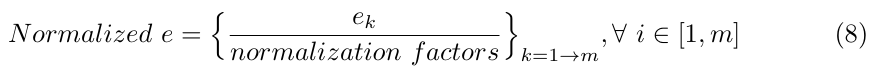
A math equations and formulas

Description automatically generated with medium confidence

A math equations with numbers

Description automatically generated with medium confidence

To normalize the expression dataset, we divided each gene expression value by the normalization factors calculated by the sample (equation 8), where *ek* is a gene expression value in each *m* sample.



To avoid imbalanced classes, we used the method RandomUnderSampler from the imbalanced-learn 0.12.0 Python package [[35](#_bookmark35)]. We construct a pipeline in scikit-learning 1.4.0 [[36](#_bookmark36)] to undersampling, select features, and train/test a Random Forest for different normalization methods for the same subsamples. We apply cross-validation and calculate the AUC for each set of training and tests.

*Data manipulation and plotting*: We use Pandas [[64](#_bookmark64)], anndata [[65](#_bookmark65)], and Numpy

1.26.4 [[66](#_bookmark66)] to manipulate the data and Matplotlib 3.8.0 [[67](#_bookmark67)], Seaborn 0.11.2 [[68](#_bookmark68)] and upsetplot [[69](#_bookmark69)] to plot them.

The entire coding process is available online, including Snakemake workflows, bash scripts, notebooks, and a new Python package we developed during this investigation.

## Data availability

**Supplementary information.** All codes and results of *in silico* analysis are deposited in Harvard Dataverse under the

# Declarations

The authors declare no competing interests.

# References

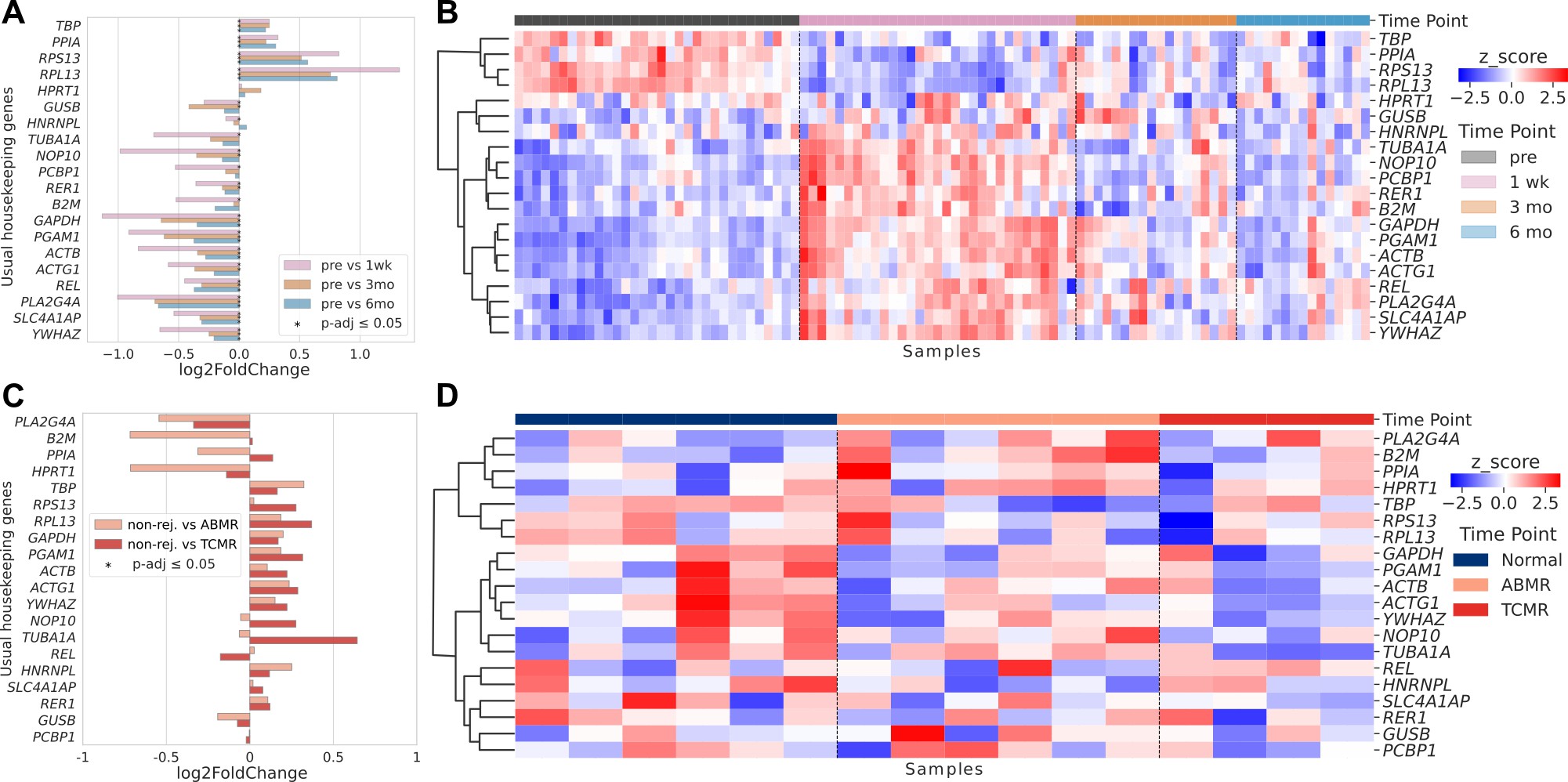
1. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome biology **15**, 1–21 (2014)
2. Robinson, M.D., McCarthy, D.J., Smyth, G.K.: edger: a bioconductor package for differential expression analysis of digital gene expression data. bioinformatics **26**(1), 139–140 (2010)
3. Zhao, Y., Li, M.-C., Konat´e, M.M., Chen, L., Das, B., Karlovich, C., Williams, P.M., Evrard, Y.A., Doroshow, J.H., McShane, L.M.: Tpm, fpkm, or normalized counts? a comparative study of quantification measures for the analysis of rna- seq data from the nci patient-derived models repository. Journal of translational medicine **19**(1), 269 (2021)
4. Han, H., Men, K.: How does normalization impact rna-seq disease diagnosis? Journal of Biomedical Informatics **85**, 80–92 (2018) [https://doi.org/10.1016/j.](https://doi.org/10.1016/j.jbi.2018.07.016) [jbi.2018.07.016](https://doi.org/10.1016/j.jbi.2018.07.016)
5. Abbas-Aghababazadeh, F., Li, Q., Fridley, B.L.: Comparison of normaliza- tion approaches for gene expression studies completed with high-throughput sequencing. PloS one **13**(10), 0206312 (2018)
6. Kapoor, S., Narayanan, A.: Leakage and the reproducibility crisis in machine- learning-based science. Patterns **4**(9) (2023)
7. Blagus, R., Lusa, L.: Joint use of over-and under-sampling techniques and cross-validation for the development and assessment of prediction models. BMC bioinformatics **16**(1), 1–10 (2015)
8. Patil, P., Bachant-Winner, P.-O., Haibe-Kains, B., Leek, J.T.: Test set bias affects reproducibility of gene signatures. Bioinformatics **31**(14), 2318–2323 (2015)
9. Joshi, C.J., Ke, W., Drangowska-Way, A., O’Rourke, E.J., Lewis, N.E.: What are housekeeping genes? PLoS computational biology **18**(7), 1010295 (2022)
10. Butte, A.J., Dzau, V.J., Glueck, S.B.: Further defining housekeeping, or “main- tenance,” genes focus on “a compendium of gene expression in normal human tissues”. Physiological genomics **7**(2), 95–96 (2001)
11. Eisenberg, E., Levanon, E.Y.: Human housekeeping genes, revisited. TRENDS in Genetics **29**(10), 569–574 (2013)
12. Irie, N., Warita, K., Tashiro, J., Zhou, Y., Ishikawa, T., Oltvai, Z.N., Warita, T.: Expression of housekeeping genes varies depending on mevalonate pathway inhibition in cancer cells. Heliyon **9**(7) (2023)
13. Casas, A.I., Hassan, A.A., Manz, Q., Wiwie, C., Kleikers, P., Egea, J., L´opez, M.G., List, M., Baumbach, J., Schmidt, H.H.: Un-biased housekeeping gene panel selection for high-validity gene expression analysis. Scientific Reports **12**(1), 12324 (2022)
14. Tilli, T.M., Castro, C.d.S., Tuszynski, J.A., Carels, N.: A strategy to identify housekeeping genes suitable for analysis in breast cancer diseases. BMC genomics **17**, 1–11 (2016)
15. Hazarika, A., Nongkhlaw, B., Mukhopadhyay, A.: Identification of stable reference genes in peripheral blood mononuclear cells from type 2 diabetes mellitus patients. Scientific Reports **13**(1), 486 (2023)
16. Jonge, H.J., Fehrmann, R.S., Bont, E.S., Hofstra, R.M., Gerbens, F., Kamps, W.A., Vries, E.G., Zee, A.G., Meerman, G.J., Elst, A.: Evidence based selection of housekeeping genes. PloS one **2**(9), 898 (2007)
17. Herman-Edelstein, M., Rozen-Zvi, B., Zingerman, B., Lichtenberg, S., Malachi, T., Gafter, U., Ori, Y.: Effect of immunosuppressive drugs on dna repair in human peripheral blood mononuclear cells. Biomedicine & Pharmacotherapy **66**(2), 111– 115 (2012)
18. Thomson, A., Bonham, C., Zeevi, A.: Mode of action of tacrolimus (fk506): molecular and cellular mechanisms. Therapeutic drug monitoring **17**(6), 584–591 (1995)
19. Wang, Z., Lyu, Z., Pan, L., Zeng, G., Randhawa, P.: Defining housekeeping genes suitable for rna-seq analysis of the human allograft kidney biopsy tissue. BMC medical genomics **12**, 1–11 (2019)
20. Dorr, C., Wu, B., Guan, W., Muthusamy, A., Sanghavi, K., Schladt, D.P., Maltz- man, J.S., Scherer, S.E., Brott, M.J., Matas, A.J., *et al.*: Differentially expressed gene transcripts using rna sequencing from the blood of immunosuppressed kidney allograft recipients. PloS one **10**(5), 0125045 (2015)
21. Matz, M., Heinrich, F., Zhang, Q., Lorkowski, C., Seelow, E., Wu, K., Lachmann, N., Addo, R.K., Durek, P., Mashreghi, M.-F., *et al.*: The regulation of interferon type i pathway-related genes rsad2 and etv7 specifically indicates antibody- mediated rejection after kidney transplantation. Clinical transplantation **32**(12), 13429 (2018)
22. Strauss, P., Mikkelsen, H., Furriol, J.: Variable expression of eighteen common housekeeping genes in human non-cancerous kidney biopsies. Plos one **16**(12), 0259373 (2021)
23. Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., Speleman, F.: Accurate normalization of real-time quantitative rt-pcr data by geometric averaging of multiple internal control genes. Genome biology **3**, 1–12 (2002)
24. Wan, Q., Chen, S., Shan, Z., Yang, Z., Chen, L., Zhang, C., Yuan, S., Hao, Q., Zhang, X., Qiu, D., *et al.*: Stability evaluation of reference genes for gene expression analysis by rt-qpcr in soybean under different conditions. PloS one **12**(12), 0189405 (2017)
25. O’Hagan, S., Muelas, M.W., Day, P.J., Lundberg, E., Kell, D.B.: Genegini: Assess- ment via the gini coefficient of reference “housekeeping” genes and diverse human transporter expression profiles. Cell systems **6**(2), 230–244 (2018)
26. Wright Muelas, M., Mughal, F., O’Hagan, S., Day, P.J., Kell, D.B.: The role and robustness of the gini coefficient as an unbiased tool for the selection of gini genes for normalising expression profiling data. Scientific reports **9**(1), 17960 (2019)
27. Tung, K.-F., Pan, C.-Y., Lin, W.-c.: Housekeeping protein-coding genes inter- rogated with tissue and individual variations. Scientific Reports **14**(1), 12454 (2024)
28. Van Loon, E., Lamarth´ee, B., Loor, H., Van Craenenbroeck, A.H., Brouard, S., Danger, R., Giral, M., Callemeyn, J., Tinel, C., Calabuig, A´.C., *et al.*: Biological pathways and comparison with biopsy signals and cellular origin of peripheral blood transcriptomic profiles during kidney allograft pathology. Kidney International **102**(1), 183–195 (2022)
29. Hounkpe, B.W., Chenou, F., Lima, F., De Paula, E.V.: Hrt atlas v1. 0 database: redefining human and mouse housekeeping genes and candidate reference tran- scripts by mining massive rna-seq datasets. Nucleic acids research **49**(D1), 947–955 (2021)
30. K¨allberg, D., Vidman, L., Ryd´en, P.: Comparison of methods for feature selection in clustering of high-dimensional rna-sequencing data to identify cancer subtypes. Frontiers in Genetics **12**, 632620 (2021)
31. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsd´ottir, H., Tamayo, P., Mesirov, J.P.: Molecular signatures database (msigdb) 3.0. Bioinformatics **27**(12), 1739–1740 (2011)
32. Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., *et al.*: Database resources of the national center for biotechnology information. Nucleic acids research **36**(suppl 1), 13–21 (2007)
33. Yang, M., Derbyshire, M.K., Yamashita, R.A., Marchler-Bauer, A.: Ncbi’s con- served domain database and tools for protein domain analysis. Current protocols in bioinformatics **69**(1), 90 (2020)
34. Elahimanesh, M., Najafi, M.: Differentially expressed genes of rna-seq data are suggested on the intersections of normalization techniques. Biochemistry and Biophysics Reports **37**, 101618 (2024)
35. LemaA˜Zˇtre, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of machine learning research **18**(17), 1–5 (2017)
36. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al.*: Scikit-learn: Machine learning in python. the Journal of machine Learning research **12**, 2825–2830 (2011)
37. Cawley, G.C., Talbot, N.L.: On over-fitting in model selection and subsequent selection bias in performance evaluation. The Journal of Machine Learning Research **11**(1), 2079–2107 (2010)
38. Sz´ekely, G.J., Rizzo, M.L., Bakirov, N.K.: Measuring and testing dependence by correlation of distances. The Annals of Statistics **35**(6), 2769–2794 (2007) <https://doi.org/10.1214/009053607000000505>
39. Agborbesong, E., Bissler, J., Li, X.: Liquid biopsy at the frontier of kidney dis- eases: application of exosomes in diagnostics and therapeutics. Genes **14**(7), 1367 (2023)
40. Connal, S., Cameron, J.M., Sala, A., Brennan, P.M., Palmer, D.S., Palmer, J.D., Perlow, H., Baker, M.J.: Liquid biopsies: the future of cancer early detection. Journal of translational medicine **21**(1), 118 (2023)
41. Noor, J., Chaudhry, A., Noor, R., Batool, S.: Advancements and applications of liquid biopsies in oncology: a narrative review. Cureus **15**(7) (2023)
42. Lone, S.N., Nisar, S., Masoodi, T., Singh, M., Rizwan, A., Hashem, S., El-Rifai, W., Bedognetti, D., Batra, S.K., Haris, M., *et al.*: Liquid biopsy: a step closer to transform diagnosis, prognosis and future of cancer treatments. Molecular cancer **21**(1), 79 (2022)
43. Zhang, W., Yi, Z., Wei, C., Keung, K.L., Sun, Z., Xi, C., Woytovich, C., Farouk, S., Gallon, L., Menon, M.C., et al.: Pretransplant transcriptomic signature in peripheral blood predicts early acute rejection. JCI insight **4**(11) (2019)
44. Li, L., Khatri, P., Sigdel, T.K., Tran, T., Ying, L., Vitalone, M., Chen, A., Hsieh, S.-c., Dai, H., Zhang, M., *et al.*: A peripheral blood diagnostic test for acute rejection in renal transplantation. American Journal of Transplantation **12**(10), 2710–2718 (2012)
45. Kurian, S., Williams, A., Gelbart, T., Campbell, D., Mondala, T., Head, S., Horvath, S., Gaber, L., Thompson, R., Whisenant, T., *et al.*: Molecular classi- fiers for acute kidney transplant rejection in peripheral blood by whole genome gene expression profiling. American Journal of Transplantation **14**(5), 1164–1172 (2014)
46. Gu¨nther, O.P., Shin, H., Ng, R.T., McMaster, W.R., McManus, B.M., Keown, P.A., Tebbutt, S.J., Lˆe Cao, K.-A.: Novel multivariate methods for integration of genomics and proteomics data: applications in a kidney transplant rejection

study. Omics: a journal of integrative biology **18**(11), 682–695 (2014)

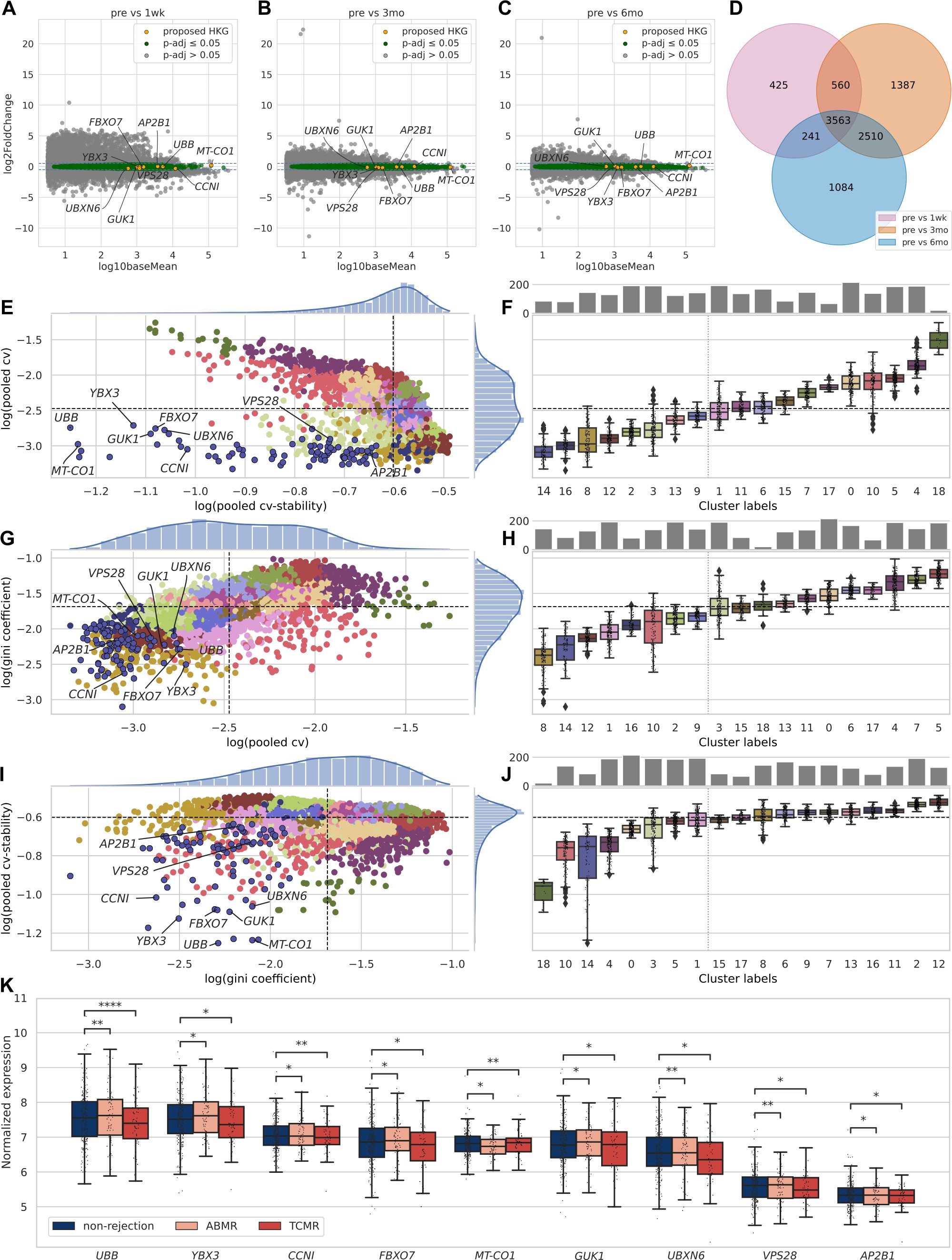
1. Van Loon, E., Gazut, S., Yazdani, S., Lerut, E., Loor, H., Coemans, M., No¨el, L.- H., Thorrez, L., Van Lommel, L., Schuit, F., *et al.*: Development and validation of a peripheral blood mrna assay for the assessment of antibody-mediated kidney allograft rejection: a multicentre, prospective study. EBioMedicine **46**, 463–472 (2019)
2. Leinonen, R., Sugawara, H., Shumway, M., Collaboration, I.N.S.D.: The sequence read archive. Nucleic acids research **39**(suppl 1), 19–21 (2010)
3. Andrews, S., et al.: FastQC: a quality control tool for high throughput sequence data. 2010 (2017)
4. Chen, S., Zhou, Y., Chen, Y., Gu, J.: fastp: an ultra-fast all-in-one fastq preprocessor. Bioinformatics **34**(17), 884–890 (2018)
5. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., Kingsford, C.: Salmon provides fast and bias-aware quantification of transcript expression. Nature methods **14**(4), 417–419 (2017)
6. Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J.E., Mudge, J.M., Sisu, C., Wright, J.C., Armstrong, J., Barnes, I., *et al.*: Gencode 2021. Nucleic acids research **49**(D1), 916–923 (2021)
7. M¨older, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., et al.: Sustainable data analysis with snakemake. F1000Research **10** (2021)
8. Gautier, L., Cope, L., Bolstad, B.M., Irizarry, R.A.: affy—analysis of affymetrix genechip data at the probe level. Bioinformatics **20**(3), 307–315 (2004)
9. Wu, C., MacLeod, I., Su, A.I.: Biogps and mygene. info: organizing online, gene- centric information. Nucleic acids research **41**(D1), 561–565 (2013)
10. Soneson, C., Love, M.I., Robinson, M.D.: Differential analyses for rna-seq: transcript-level estimates improve gene-level inferences. F1000Research **4** (2015)
11. Bonald, T., De Lara, N., Lutz, Q., Charpentier, B.: Scikit-network: Graph analysis in python. Journal of Machine Learning Research **21**(185), 1–6 (2020)
12. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cour- napeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., *et al.*: Scipy 1.0: fundamental algorithms for scientific computing in python. Nature methods **17**(3), 261–272 (2020)
13. McInnes, L., Healy, J., Saul, N., Großberger, L.: Umap: Uniform manifold approx- imation and projection. Journal of Open Source Software **3**(29), 861 (2018)

<https://doi.org/10.21105/joss.00861>

1. Fang, Z., Liu, X., Peltz, G.: Gseapy: a comprehensive package for performing gene set enrichment analysis in python. Bioinformatics **39**(1), 757 (2023)
2. O’Leary, N.A., Cox, E., Holmes, J.B., Anderson, W.R., Falk, R., Hem, V., Tsuchiya, M.T., Schuler, G.D., Zhang, X., Torcivia, J., *et al.*: Exploring and retrieving sequence and metadata for species across the tree of life with ncbi datasets. Scientific Data **11**(1), 732 (2024)
3. Madeira, F., Pearce, M., Tivey, A.R., Basutkar, P., Lee, J., Edbali, O., Mad- husoodanan, N., Kolesnikov, A., Lopez, R.: Search and sequence analysis tools services from embl-ebi in 2022. Nucleic acids research **50**(W1), 276–279 (2022)
4. McInnes, L., Healy, J., Astels, S.: hdbscan: Hierarchical density based clustering. The Journal of Open Source Software **2**(11), 205 (2017)
5. McKinney, W., *et al.*: Data structures for statistical computing in Python. In: SciPy, vol. 445, pp. 51–56 (2010)
6. Virshup, I., Rybakov, S., Theis, F.J., Angerer, P., Wolf, F.A.: anndata: Annotated data. BioRxiv, 2021–12 (2021)
7. Harris, C.R., Millman, K.J., Van Der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., *et al.*: Array programming with numpy. Nature **585**(7825), 357–362 (2020)
8. Hunter, J.D.: Matplotlib: A 2d graphics environment. Computing in science & engineering **9**(03), 90–95 (2007)
9. Waskom, M.L.: Seaborn: statistical data visualization. Journal of Open Source Software **6**(60), 3021 (2021)
10. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., Pfister, H.: Upset: visu- alization of intersecting sets. IEEE transactions on visualization and computer graphics **20**(12), 1983–1992 (2014)

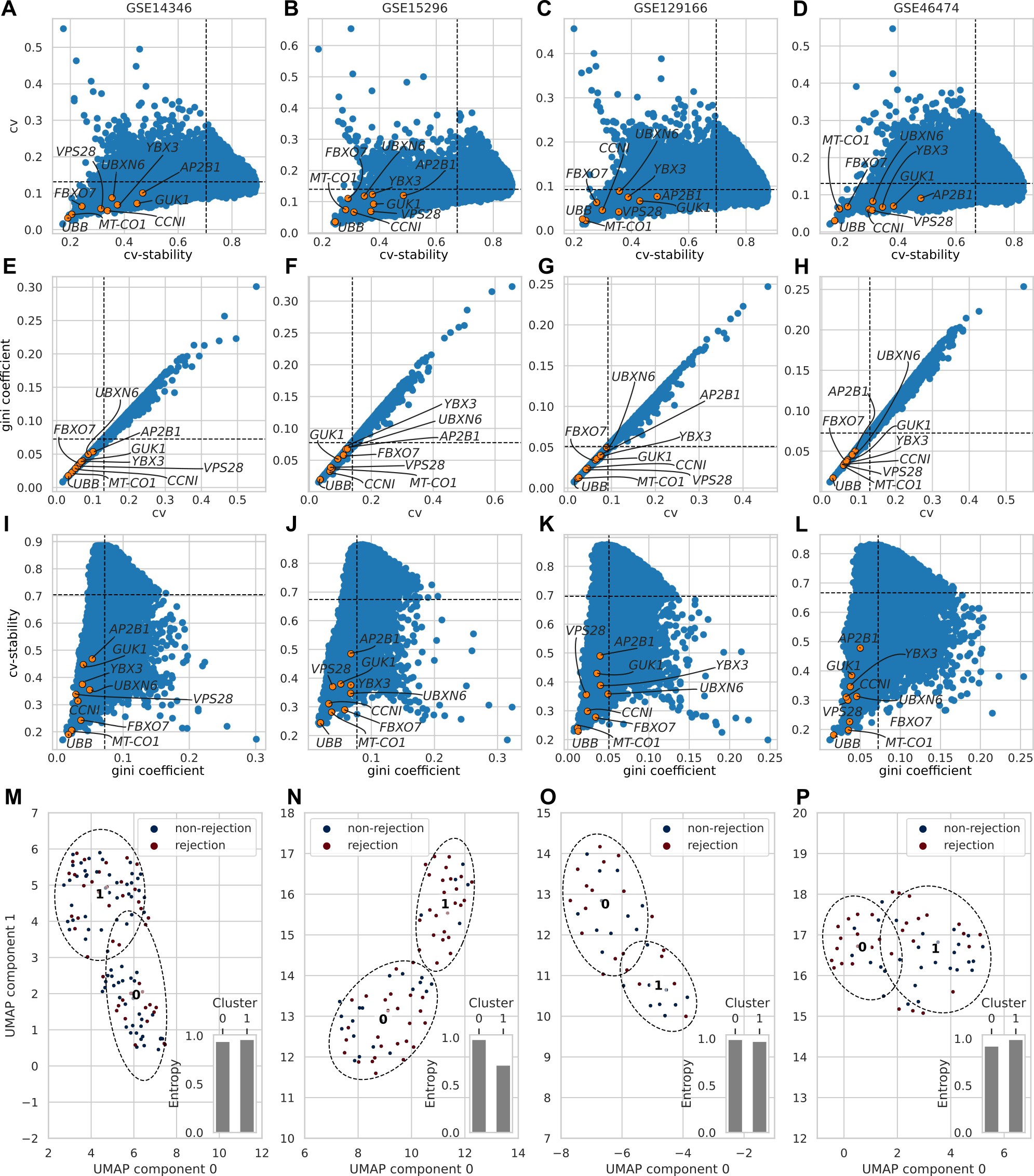


**Figure 1. Commonly used housekeeping genes vary in expression according to the patient's clinical outcome and time-point in transplantation** (**A)** Differential expression log2FoldChange and (**B)** Heatmap of normalized expression of HKGs in non-rejection transplanted recipients in different time points. (**C)** Differential expression log2FoldChange and (**D)** Heatmap of normalized expression of HKG in non-rejection, ABMR, and TCMR of transplanted patients.

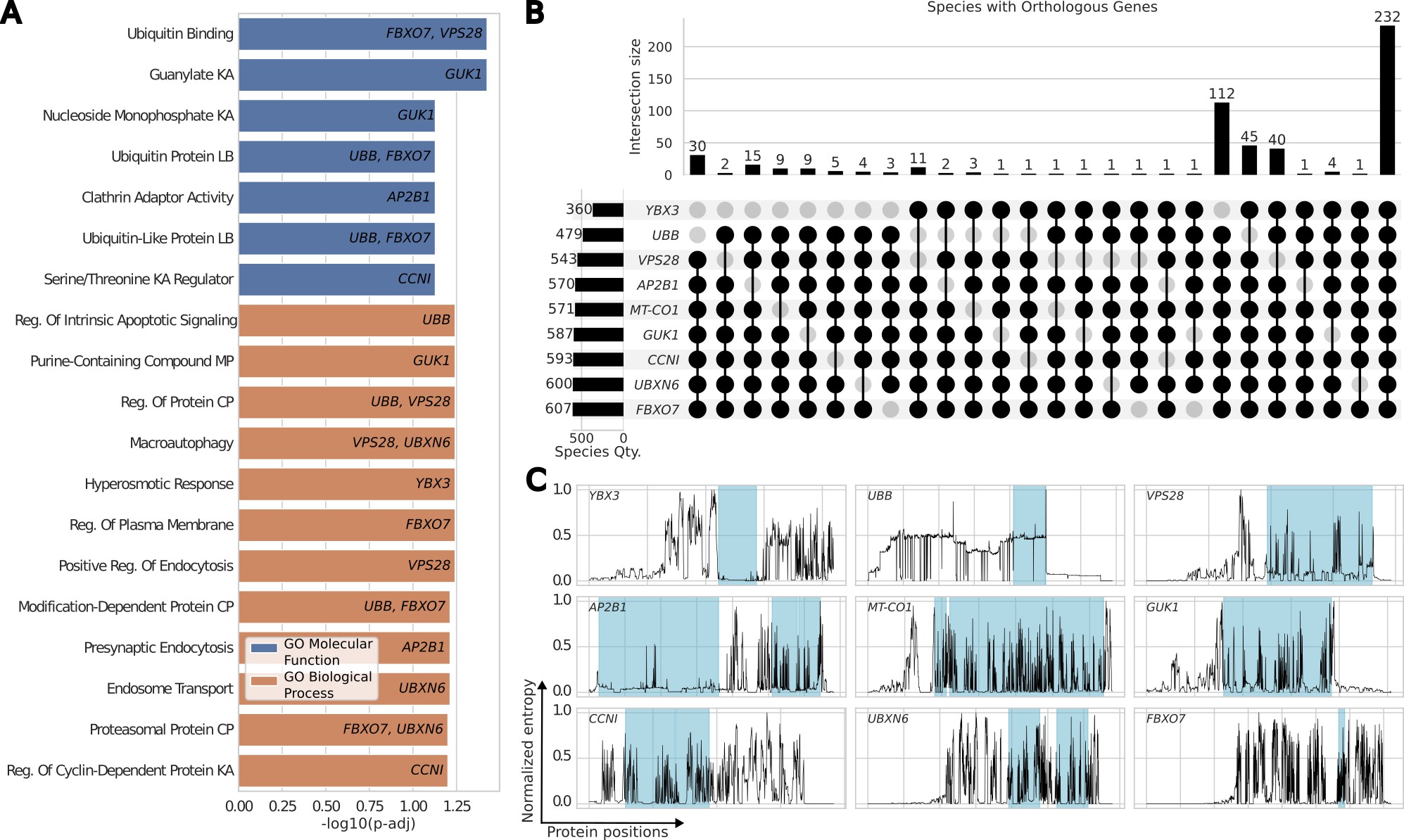


**Figure 2. Proposed housekeeping genes present stable expression over all transplanted patients.** Gene expressions are in between -0.5 and +0.5 of log2FoldChange in transplanted patients when compared to pre-transplant with different time points, including (**A)** one week, (**B)** 3 months, and (**C)** 6 months after transplantation). (**D**)Venn diagram showing the intersection of 3,563 genes with log2FoldChange between -0.5 and +0.5 in all time points. For the 3,563 intersection genes, we display: (**E)** Scatter plot of coefficient of variation and cv-stability; (**F)** Boxplot of coefficient of variation of genes clusters;(**G)** Scatter plot of gini coefficient and coefficient of variation; (**H)** Boxplot of cv-stability of genes clusters;. (**I)** Scatter plot of cv-stability and Gini coefficient; (**J)** Boxplot of cv-stability of genes clusters; (**K)** Two-One-Sided-Test comparing equivalence of distributions between non-rejection, ABMR, and TCMR.

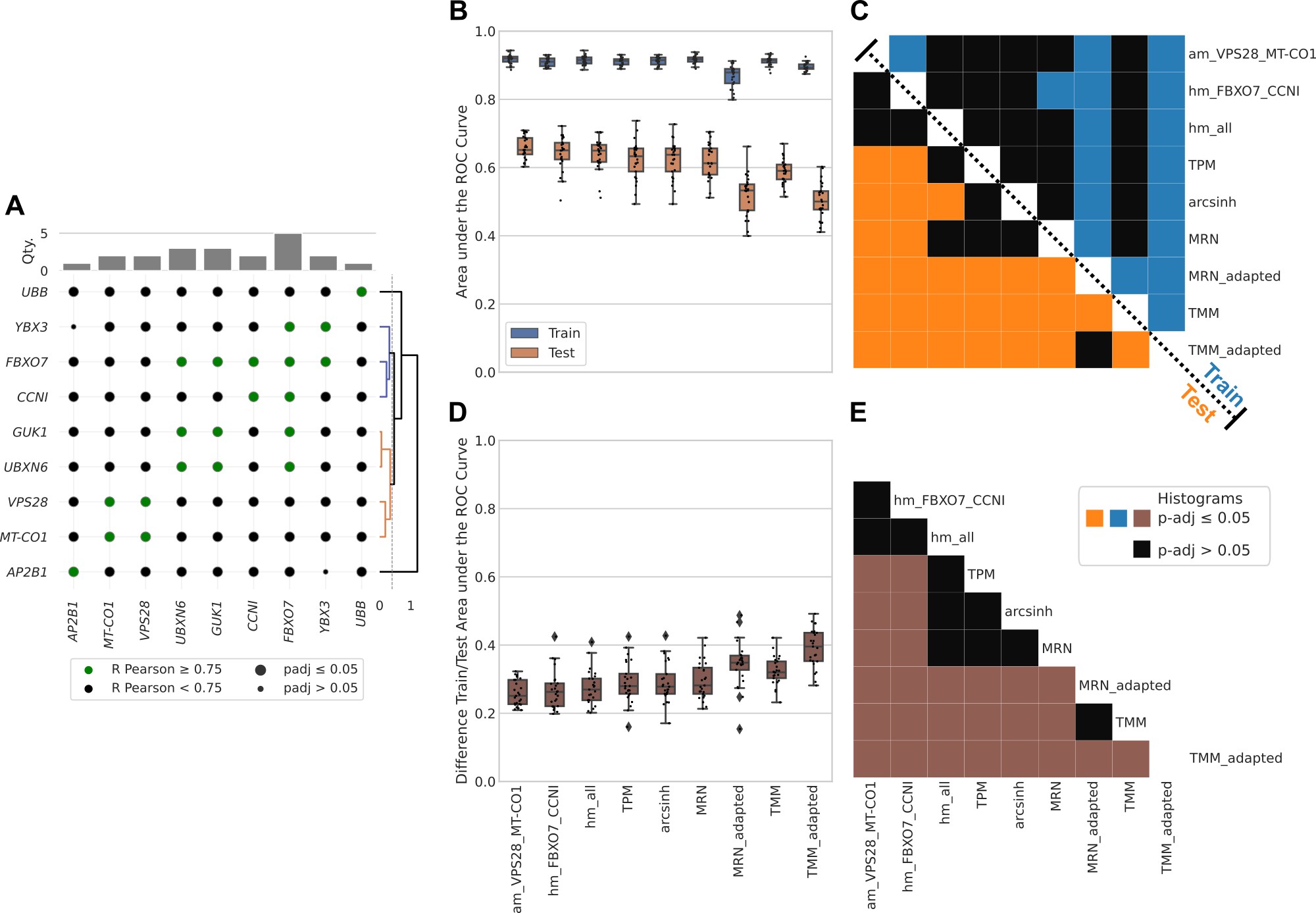
ABMR: antibody-mediated rejection; TCMR : T-cell-mediated rejection.



**Figure 3. Validation of proposed housekeeping genes in four microarray datasets.** (**A-D)** Coefficient of variation and cv-stability of……. (**E-H)** Gini coefficient and coefficient of variation of…….. (**I-L)** CV-stability and Gini coefficient. (**M-P)** Scatter plot of two UMAP dimensions of proposed housekeeping genes expression clustered in two Kmeans groups. At the right bottom of each scatter is a bar representing the entropy of each Kmeans cluster.



**Figure 4. Biological importance and conservation of proposed housekeeping genes. (A)** GSEA enrichment analysis for molecular function and biological process of the XXX genes from XXXX.. (**B)** Upset plot of species that share homologous genes. (**C)** Relative entropy of aligned homologous genes positions and significant motifs in blue.



**Figure 5. Analysis of data leakage in normalization based on proposed housekeeping genes and benchmark methods.** (**A)** Pairwise Pearson correlation of housekeeping gene expressions and hierarchical clustering to form groups of normalization factors. **B** Train and test values of AUC metric distributions for different normalization methods for 25 cross-validations. **C** Adjusted p-values of pairwise comparisons of AUC distributions between different normalization methods. **D** Distributions of differences between Train and Test AUC values for each normalization method. **E** Adjusted p-values of pairwise comparisons between differences in AUC distributions for different normalization methods.