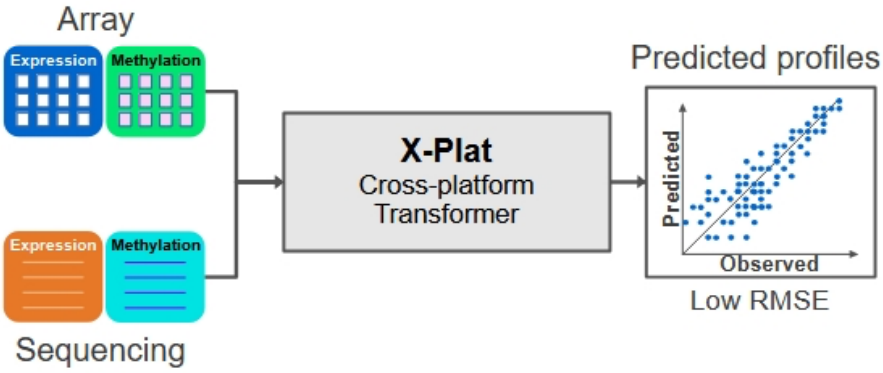
X-Plat: a polynomial regression-based tool for cross-platform transformation of expression and methylation data

Neeraja M. Krishnan[1,*], Sarah I Rahman[1], Lars Røn Olsen[2] and Binay Panda[1,*]

[1] School of Biotechnology, Jawaharlal Nehru University, New Mehrauli Road, New Delhi 110067, India
[2] Department of Immunology and Microbiology, University of Copenhagen, Blegdamsvej 3B, DK 2200 Copenhagen, Denmark

* To whom correspondence should be addressed. Email: neeraja@jnu.ac.in, binaypanda@jnu.ac.in

GRAPHICAL ABSTRACT



ABSTRACT

Many biological studies could benefit from combining data from legacy microarray and high-throughput sequencing platforms, especially in clinical domains where collecting additional samples is not possible. However, incompatibility between platforms makes legacy data difficult to integrate, owing to differences in platform design, target preparation, and dependence on prior annotations. Here, we describe X-Plat, a cross-platform data transformation tool for both expression and methylation assays that inter-converts data between microarray and sequencing platforms using per-gene second-degree polynomial regression. X-Plat learns cross-platform conversion rules from paired microarray–sequencing datasets spanning multiple conditions, sample sources, organisms, and platforms, and evaluates performance using cross-validated root mean square error (RMSE) per gene. In rat, *Arabidopsis*, and human datasets, X-Plat achieved lower cross-validated RMSE than HARMONY and TDM for the vast majority of genes (≥95% in all sequencing-to-array transformations and most array-to-sequencing transformations, with ~82% in the *Arabidopsis* array-to-sequencing setting), and these findings were confirmed using RMSE on held-out test samples from the first cross-validation fold. X-Plat also achieved low RMSE (≤0.2) for the majority of CpG regions in paired human array–sequencing methylation datasets. Using X-Plat, users can transform data between microarray and high-throughput sequencing platforms, enabling cross-platform comparison and reuse of legacy cohorts.

INTRODUCTION

DNA microarrays revolutionized high-throughput data generation and were the primary mode of measuring genome-wide gene expression, genetic polymorphisms, and promoter methylation for the better part of the last three decades. Over the last decade, next-generation DNA sequencing technology has increasingly replaced microarrays for genome-wide measurement of DNA, RNA, and epigenetic changes and is favored over array-based platforms because of its unbiased coverage and reduced dependence on prior sequencing and annotation information (1, 2). However, a full transition from array-based methods to sequencing-based ones, although imminent, faces hurdles due to higher cost, limited access to instruments and assay know-how, and the need for a skilled workforce trained to perform data analysis. Even today, many laboratories rely on microarrays to generate data while gradually increasing their use of sequencing-based assays (3, 4).

Legacy datasets generated using microarrays and deposited in public databases could be immensely useful for validating results obtained with sequencing-based assays, especially in clinical domains where sample number and accessibility are major constraints. For this reason, the ability to convert data from microarray-based formats to sequencing-based formats, and vice versa, is not just beneficial but increasingly necessary to continue using legacy data (5, 6).

Past efforts have produced cross-platform normalization and transformation tools to leverage legacy data, including methods for batch correction, probe-level alignment, transcriptomic distribution matching and methylation remapping (7–10). In one prior work, a two-stage Bayesian mixture modeling strategy was applied to assimilate and analyze four independent microarray studies to derive an inter-study validated "meta-signature" that unified disparate gene expression data on a common probability scale, enabling robust prognostic signatures (11). ComBat is a widely used batch correction tool for array data that integrates data from multiple experiments and platforms (7). ComBat can also be used for cross-platform integration by treating data from different platforms as different batches; however, it merges mean and variance across batches (inter-batch) while ignoring within-batch mean and variance, and an extension that accounts for single-batch mean and variance was subsequently proposed (12). Probe mapping across microarray platforms has been used as a crucial step for data integration (13). The methyLiftover method maps DNA methylation data from bisulfite sequencing approaches to CpG sites measured with Illumina methylation bead-array platforms, thereby enabling cross-platform DNA methylation data integration (9). A comparison of microarray platform integration approaches revealed that methods relying on actual probe-set signal intensities are superior to those considering only biological characteristics of probe sequences and that cross-platform integration improves correlation with RNA-seq data (14). Methods that transform RNA-seq data to make it compatible with microarray-based models have also been proposed (8, 15, 16), and bidirectional conversion is possible and has been demonstrated (5).

Other cross-platform normalization tools, such as MatchMixeR, model platform differences using gene-specific linear mixed-effects regression estimated from paired profiles and can effectively reduce cross-platform discrepancies for expression data (17). In contrast, our goal here is to evaluate a simple per-gene polynomial mapping that supports explicit bidirectional conversion between platforms and is readily extensible to non-expression modalities such as DNA methylation. However, neither explicit, reusable gene-specific conversion rules nor a universal conversion rule between platforms has been developed for routine forward and reverse transformation of both expression and methylation profiles. Solving this key problem and providing a platform to integrate legacy data with modern high-throughput data would greatly benefit the community.

Here, we describe a cross-platform data transformation tool, X-Plat, for bidirectional conversion of microarray and sequencing data for both expression and methylation assays. X-Plat builds per-gene predictive models by fitting second-degree polynomial regressions to paired array–sequencing values and evaluates these models using 10-fold cross-validated RMSE. It derives cross-platform conversion rules from paired mixed-source microarray–sequencing datasets spanning multiple conditions, sample sources, and organisms, and multiple array and sequencing platforms.

MATERIAL AND METHODS

Collecting data

We selected datasets from NCBI Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo) for rat, *Arabidopsis*, and human expression studies and for human methylation studies. Datasets with at least 50 paired samples, where the same sample was assayed using both the same microarray platform and the same high-throughput sequencing platform, were chosen to build the model. The sample and gene numbers, including platform details are summarized in Table 1, and sample-level details are provided in Supplementary Table S1.

Table 1: Description of paired microarray–sequencing data used to build and test X-Plat.

Assay	Organism	Array platform	Seq platform	No. of paired samples	No. of probes/regions	No. of genes
Expression	Rat (*Rattus norvegicus*)	Affymetrix Rat Genome 230 2.0 Array	Illumina HiScanSQ	148	21 887	11 012
			Illumina HiSeq 2000			
			Illumina HiSeq 500			
	Arabidopsis (*Arabidopsis thaliana*)	Affymetrix Arabidopsis ATH1 Genome Array	Illumina NextSeq 500	111	22 812	8 785
	Human (*Homo sapiens*)	Affymetrix HG-U133 Plus 2.0 Array	Illumina HiSeq 2000	62	54 675	14 323
Methylation	Human (*Homo sapiens*)	Illumina HumanMethylation450 BeadChip Array	Illumina HiSeq 2500	99	119 535	19 607

Data pre-processing

The pipelines used for pre-processing the expression and methylation data, including quality assessment using the Galaxy framework (stable release v22.05), are illustrated in Figure 1 (18).

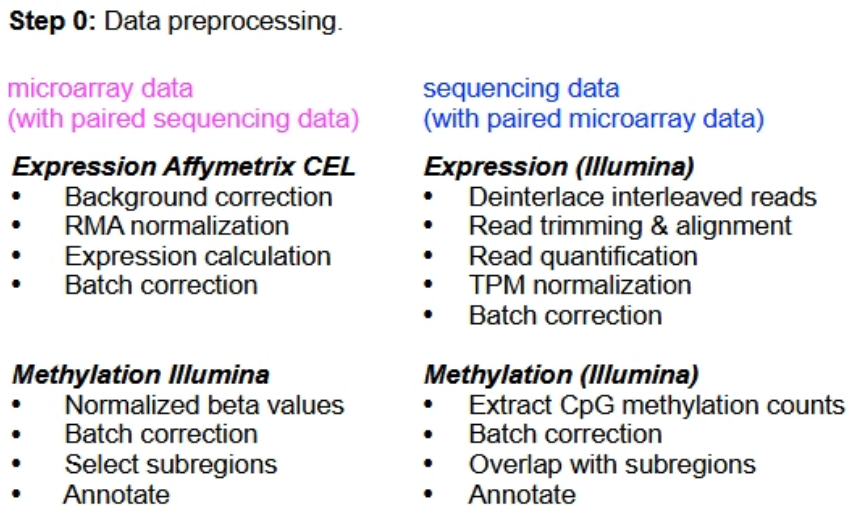


Figure 1: Pre-processing of expression and methylation data. The schematic summarizes platform-specific preprocessing steps for paired microarray and sequencing assays, including normalization, batch correction, and region-level annotation for both expression and methylation data.

Expression data. We analyzed the Affymetrix microarray expression data (.CEL) using the R/Bioconductor package affy (19). We first normalized the data using the robust multi-array average (RMA) procedure to correct for systematic, non-biological variation in three steps: (a) background correction, (b) adjustment for hybridization effects unrelated to the probe–target interactions, and (c) normalization. This was followed by summarization, in which multiple probe intensities within a probe set were combined to yield a single value per gene representing the RNA transcript expression level. At this stage, we corrected for batch effects using ComBat (7) and then performed probe-to-gene annotation.

We implemented the pre-processing pipeline for RNA-seq data on the Galaxy server (18). First, we extracted reads using the NCBI SRA Toolkit fastq-dump utility (v2.11.0) as single interleaved FASTQ files. We then used the Galaxy "FASTQ de-interlacer" tool (v1.1.5) to split interleaved reads into two FASTQ files for the paired reads. Quality control was performed using FastQC (v0.73) (20), followed by adaptor and Illumina-specific sequence trimming using Trimmomatic v0.38 (21) using the ILLUMINACLIP and SLIDINGWINDOW parameters. Next, we aligned reads to the reference genomes (mRatBN7.2 for rat, TAIR10 for *Arabidopsis*, and T2T-CHM13v2.0 for human) using HISAT2 (v2.2.1) (22). To quantify the number of reads mapped to each gene, we used featureCounts (v2.0.1) on the aligned BAM files together with the corresponding gene annotation for each genome. We then adjusted the counts for batch effects using the R package Combat-Seq (23) to eliminate batch-related technical variation, before normalizing to transcripts per kilobase million (TPM) using the R package DGEobj.utils (v1.0.6) (24).

Methylation data. We downloaded Illumina HumanMethylation450 (450K) microarray methylation data from NCBI GEO as normalized β-values and batch-corrected these using

ComBat (7). We downloaded all bisulfite-seq methylation data as CpG-level methylation counts from NCBI GEO and batch-corrected these using ComBat as well. The raw files (.IDAT for microarrays and .fastq for sequencing) were not available for all methylation studies; therefore, full preprocessing from raw intensities or reads was not possible.

X-Plat algorithm workflow

The workflow of the X-Plat algorithm is illustrated in Figure 2. We prepared data for each platform as matrices, where the rows represented samples and columns represented genes. We ensured a one-to-one correspondence between the microarray (array) and sequencing (seq) matrices in terms of both the order and exact identity of samples and genes before proceeding with the analysis. For brevity, we refer to these as "array" and "seq" and denote the two prediction directions as array→seq and seq→array. For each gene, we considered all available paired array–seq samples and worked on their log-transformed values. The array and sequencing data were transformed using Python's NumPy function log1p (natural logarithm of one plus the input value) element-wise to handle small values (25).

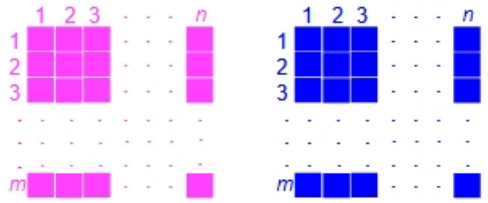
For array-to-sequencing (array → seq) conversion, a second-degree polynomial regression model is fitted per gene g to the paired log-transformed values ($array_{ig}$, seq_{ig}), and the polynomial coefficients are estimated for each gene g. The fitted model is then used to predict log-transformed sequencing values seq_{jg} for held-out samples j from their corresponding log-transformed array values, $array_{jg}$. The prediction errors e_{jg} between predicted and observed sequencing values are computed for each held-out sample, and a root mean square error (RMSE) is calculated over the held-out set for that gene as

$$RMSE_g = \sqrt{\sum_{j=1}^{n}\left(e_j^2\right)/n} \qquad (1)$$

where n is the number of held-out samples for gene g.

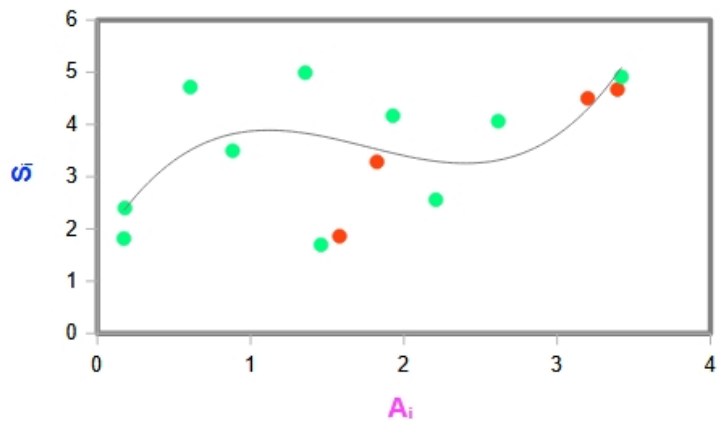
For sequencing-to-array (seq → array) conversion, the roles of the array and seq axes are reversed in the above procedure, and a second-degree polynomial is fitted to predict log(1 + array) from log(1 + seq).

Step 1: Construct Array (A) and sequencing (S) data matrices with *n* genes and *m* samples, in the same order and identity, and transform all expression values using log1p



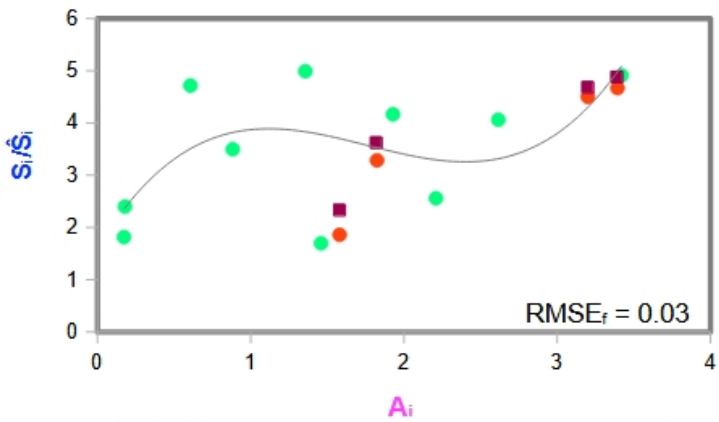
Step 2: For each gene *i* and prediction direction, fit a second-degree polynomial regression on the training folds.

A_i	S_i	Set
0.17	1.82	Tr
0.18	2.40	Tr
0.61	4.72	Tr
0.88	3.50	Tr
1.36	4.99	Tr
1.46	1.70	Tr
1.93	4.17	Tr
1.58	1.86	Te
1.82	3.28	Te
3.20	4.50	Te
3.39	4.67	Te

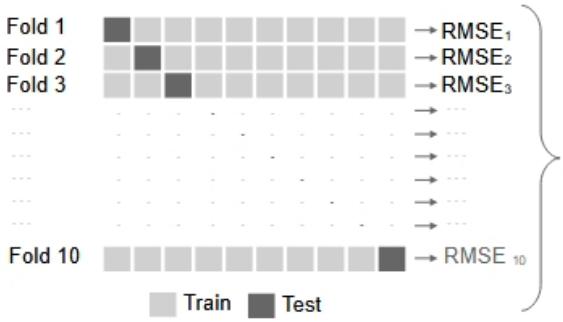


Step 3: Use the fitted polynomial model to predict $\hat{S}_i$ from A_i or $\hat{A}_i$ from S_i in the held-out test samples, and compute $RMSE_f$ between predicted and observed values for the held-out fold *f*.

A_i	S_i	$\hat{S}_i$	Set
0.17	1.82		Tr
0.18	2.40		Tr
0.61	4.72		Tr
0.88	3.50		Tr
1.36	4.99		Tr
1.46	1.70		Tr
1.93	4.17		Tr
1.58	1.86	2.33	Te
1.82	3.28	3.62	Te
3.20	4.50	4.68	Te
3.39	4.67	4.88	Te



Step 4:Repeat Steps 2–3 across all 10 folds so that each sample serves once as test, and average the fold-specific $RMSE_f$ values to obtain the cross-validated $RMSE_{cv}$ for each gene *i* and prediction direction. .



$$RMSE_{cv} = \sum_{f=1}^{10} (RMSE_f)/10$$

Figure 2: X-Plat algorithm workflow. The schematic illustrates the construction of paired array (A) and sequencing (S) matrices with matched genes and samples, log1p transformation, fitting per-gene second-degree polynomial regression models on training samples for each prediction direction, using the fitted models to predict expression values in the held out test samples for each fold, and computation of cross-validated RMSE across folds for both array→seq and seq→array prediction directions.

We evaluated prediction performance using 10-fold cross-validation, performed independently for each prediction direction across genes. All paired samples were randomly partitioned into 10 approximately equal folds (with shuffling and a fixed random seed). In each iteration, a polynomial regression model was trained on 9 folds and evaluated on the remaining fold, and we computed the RMSE between predicted and observed log-transformed values on the held-out fold. This process was repeated so that each fold served once as the test set. For each gene and prediction direction, we thus obtained 10 RMSE values $\{RMSE_1,\ldots,RMSE_{10}\}$ and computed their mean as

$$RMSE_{cv} = \sum_{f=1}^{10} RMSE_f / 10 \qquad (2)$$

where $RMSE_{cv}$ denotes the cross-validated RMSE and f indexes the fold.

These mean 10-fold RMSE values were used to summarize gene-wise performance distributions and for all figures that report X-Plat RMSE alone. For direct comparisons between X-Plat and other cross-platform normalization tools, we used the RMSE from the first cross-validation fold (i.e., the RMSE on the test set of that fold) so that all methods were evaluated on the same held-out samples.

To assess robustness to training sample size, we performed down-sampling analyses on the rat expression dataset, randomly reducing the original training set to 90%, 70%, 50%, or 30% of its size while keeping the same held-out test set and computing cross-validated RMSE as described above.

Comparison with other tools

We compared the performance of X-Plat to two other tools, Training Distribution Matching (TDM; v0.9.9; 8) and Shambhala (26), using the held-out test set samples from the first cross-validation fold so that all methods were evaluated on the same samples. As in X-Plat, we applied the log1p transform to both microarray and sequencing data for the test samples using Python's NumPy implementation prior to running the tools.

TDM is an R package that adjusts the distribution of RNA-seq data to make it more compatible with models trained on microarray data while preserving between-sample relationships, thereby improving recognition of features learned from microarrays. It uses the tdm_transform

function on the matrix of RNA-seq values (set as the target) and the microarray expression values (set as the reference) to produce transformed RNA-seq values, which we used for RMSE estimation.

Shambhala is a cross-platform data harmonization tool that compares multiple gene expression datasets by transforming them into a common, platform-independent representation. It harmonizes a target dataset to a reference definitive dataset *Q* using an auxiliary calibration dataset and makes the target more comparable to the *Q* platform (26). Shambhala is implemented in the R package, HARMONY, which is available at https://github.com/oncobox-admin/harmony. We used the harmony function with paired microarray and sequencing data from the test set and extracted harmonized sequencing values for RMSE estimation.

In addition to RMSE, we compared sample-wise prediction errors across X-Plat, TDM, and HARMONY. For all three organisms (rat, *Arabidopsis* and human), and for each gene, we computed the difference (Δ) in prediction errors (*E*) between tools for each test sample, separately for array (a) and sequencing (s) expression values. Specifically, for each test sample we calculated E_a(X-Plat) - E_a(TDM), E_a(X-Plat) - E_a(HARMONY), E_s(X-Plat) - E_s(TDM), and E_s(X-Plat) - E_s(HARMONY) which we refer to as ΔE_a(XT), ΔE_a(XH), ΔE_s(XT) and ΔE_s(XH), respectively. A negative ΔE_a(XT) for a given sample–gene combination indicates that TDM has a smaller prediction error than X-Plat, whereas a positive value indicates the reverse.

For each gene, we recorded the maximum negative and maximum positive Δ*E* values across samples (max Δ) for each of the ΔE_a(XT), ΔE_a(XH), ΔE_s(XT) and ΔE_s(XH) categories. In each category, we then selected the five genes with the largest positive and the five with the largest negative max Δ values and examined their expression profiles in detail.

RESULTS

Performance trends of X-Plat across platforms, assays, and organisms

The performance of X-Plat in predicting sequencing values from microarray values and vice versa was primarily measured using cross-validated RMSE (see Materials and Methods). For ease of discussion, we refer to the RMSE for predicted microarray values as array RMSE, and the RMSE for predicted sequencing values as seq RMSE.

Expression data. X-Plat achieved array RMSE ranges of 0–0.25, 0.01–0.18, and 0.01–0.35, and seq RMSE ranges of 0–1.21, 0.04–3.51, and 0.04–1.87 for expression data across 11 012, 8 785, and 14 323 genes from rat, *Arabidopsis* and human, respectively (Figure 3). The corresponding

median array RMSE values were 0.03, 0.06, and 0.04, and the median seq RMSE values were 0.17, 1.25, and 0.34, respectively. These ranges correspond to relatively small errors on the log scale compared with the overall expression dynamic range, indicating that most genes are predicted with modest absolute deviation. Overall, the RMSE range was consistently smaller for the microarray platform than for the sequencing platform.

X-Plat's prediction trends varied across organisms. For rat (Figure 3A–B), array RMSE was ≤0.05 for most genes (10 576/11 012) and 0.05–0.1 for the remainder. Among these 10 576 genes, seq RMSE was ≤0.2 for 7 354 genes. Only one gene had a high array RMSE of 0.25, and 18 genes had relatively high seq RMSE values ≥0.8.

For *Arabidopsis* (Figure 3C–D), array RMSE was in an intermediate-low range of 0.03–0.09 for most genes (7 645/8 785). Within this subset, seq RMSE was 0.6–1.2 for 3 235 genes and 1.2–1.8 for 3 256 genes; 18 genes had higher array RMSE (≥0.15) and 16 genes had higher seq RMSE (≥0.24).

For human (Figure 3E–F), array RMSE was ≤0.05 and 0.05–0.10 for most genes (9 100 and 5 223 of 14 323 genes, respectively). Among these 9 100 and 5 223 genes, seq RMSE was 0.27–0.54 for 5 557 and 3 628 genes, respectively, and ≤0.27 for 2 963 and 487 genes, respectively. Only one gene, *SCGB2A2*, had a higher array RMSE (0.35), and 6 genes had higher seq RMSE (>1.62).

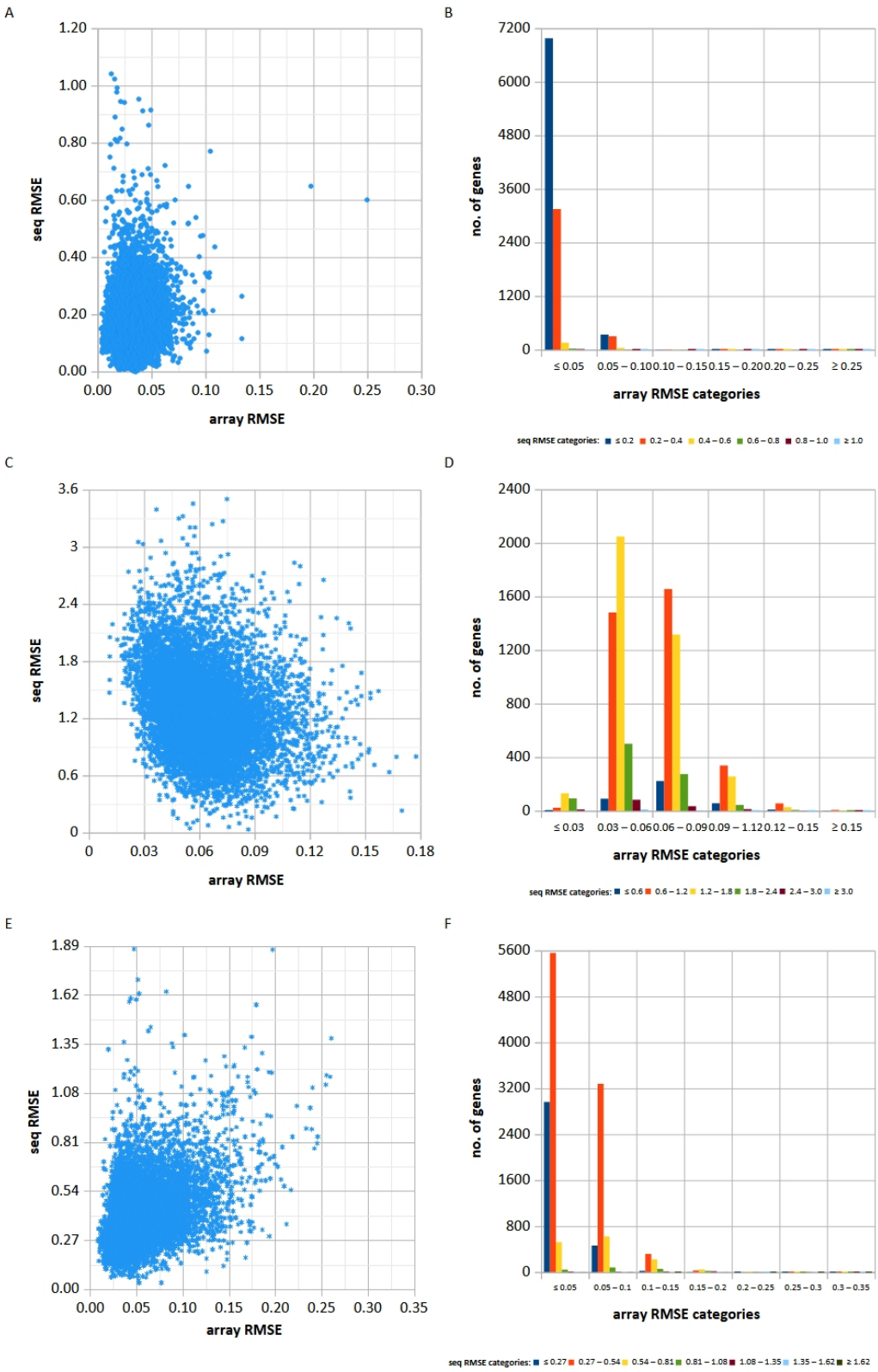


Figure 3: Root mean square errors (RMSE) for X-Plat-predicted microarray and RNA-seq expression values (array RMSE and seq RMSE) in rat (A–B), *Arabidopsis* (C–D), and human (E–F). Scatterplots of array RMSE vs seq RMSE (A, C, E) and histograms of gene counts within predefined array RMSE and seq RMSE categories (B, D, F) are shown.

Methylation data. For the human methylation data (Figure 4A), there were 127 031 CpG regions after annotation, pairing between microarray and sequencing data, filtering out regions for which more than 3% of samples had zero values, and restricting to chosen sub-regions for which X-Plat was applied. Of these, 15 895 regions were within TSS200 of genes, 27 401 within TSS1500, 71 837 within the 5′ UTR, 21 932 within the 3′ UTR, and 43 064 within the first exon. Overall, array RMSE values were in the low range (≤0.1) for most (120 707 out of 127 031) filtered CpG regions. Among these, 120 681 regions had low seq RMSE (≤0.2), 23 regions had seq RMSE between 0.2 and 0.4, and only two regions had seq RMSE ≥0.4: cg09872233 in the first exon of *ALOX15* (array RMSE 0.015, seq RMSE 0.61) and cg24043887 in the TSS200 region of *TMEM80* (array RMSE 0.014, seq RMSE 1.12) (Figure 4B). The highest array RMSE (0.49) was observed for cg03884082 in the 5′ UTR of *NBL1* (Figure 4B), followed by cg13449778 (array RMSE 0.48) in the first exon of *FAM163A* (Figure 4B).

Considering individual sub-regions, the dominant pattern of most CpGs falling within the combined ≤0.2 array RMSE and ≤0.2 seq RMSE range was preserved across TSS200, TSS1500, 5′ UTR, 3′ UTR, and first exon (Figure 4B). The 3′ UTR was a slight exception, with a modest decrease (to 92.71%) in the proportion of CpGs within the 0–0.1 array RMSE and 0–0.2 seq RMSE combined range, accompanied by an increase (to 7.26%) in the 0.1–0.2 array RMSE and 0–0.2 seq RMSE combined range (Figure 4B).

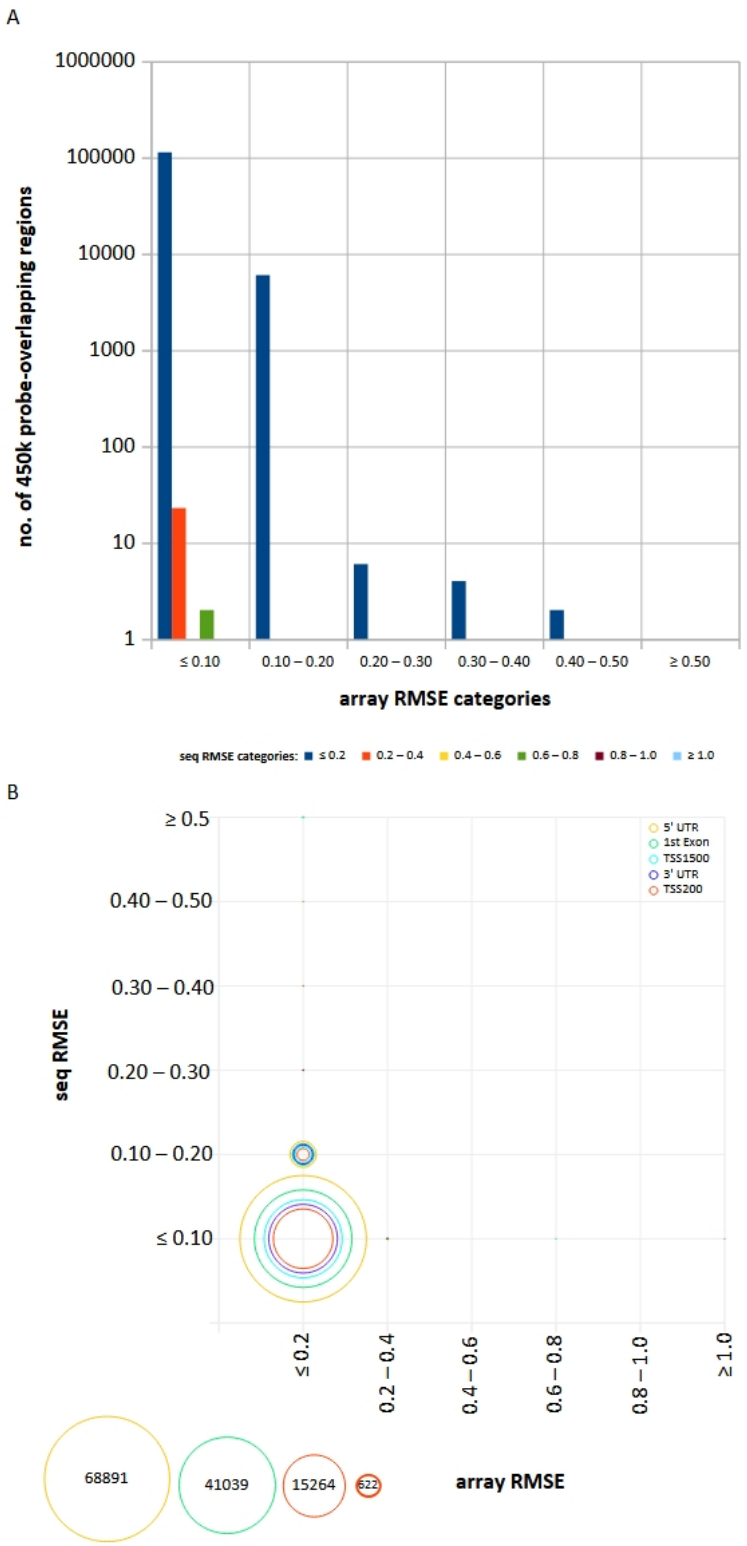


Figure 4: Root mean square error (RMSE) values for X-Plat-predicted methylation microarray and sequencing values (array RMSE and seq RMSE) in human breast cancer samples for probes overlapping various sub-regions (TSS200, TSS1500, 5′ UTR, 3′ UTR, and first exon) (A), and for each sub-region individually (B). Histograms show the number of overlapping regions within predefined array RMSE and seq RMSE categories; bubble sizes represent individual sub-regions within each region category.

Factors affecting X-Plat RMSE for expression data

We explored whether factors such as expression value range and transcript length could explain variability in X-Plat RMSE across genes, and observed that the expression data range correlated directly with RMSE for both array and sequencing transformations.

Data Range. We observed a strong positive correlation between array RMSE and the range of array expression values, and between seq RMSE and the range of sequencing expression values (Figure 5). The Pearson correlation coefficients (r) were 0.51 (array) and 0.50 (seq) for rat, 0.91 (array) and 0.91 (seq) for *Arabidopsis*, and 0.92 (array) and 0.85 (seq) for human, indicating a stronger trend in *Arabidopsis* and human than in rat.

Relationship with transcript length. The RMSE values bear no direct correlation with transcript length for either array or sequencing data in any of the organisms studied (Supplementary Figure S1).

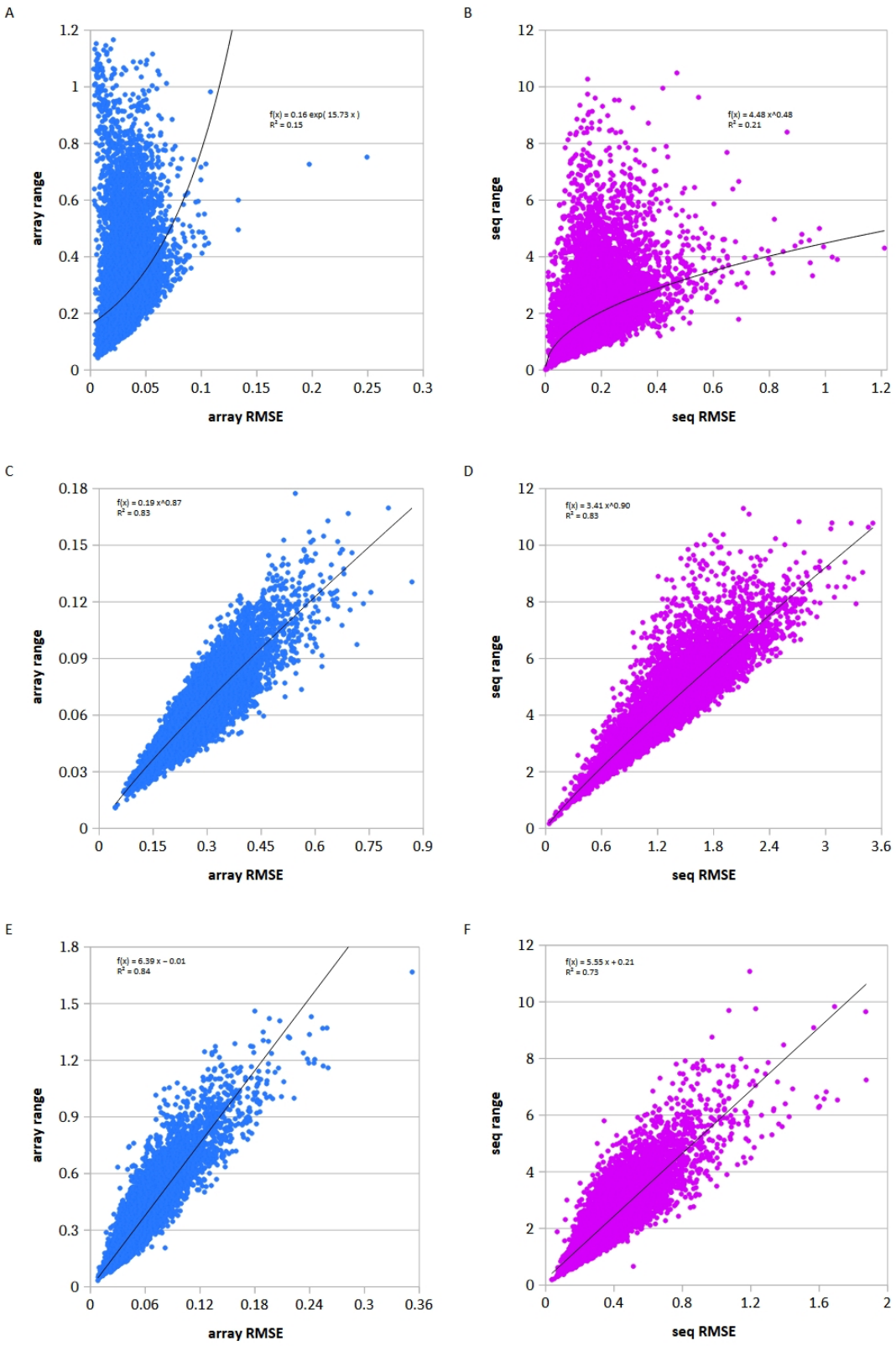


Figure 5: Root mean square errors (RMSE) for X-Plat-predicted microarray and RNA-seq expression values (array RMSE and seq RMSE) in rat (A–B), *Arabidopsis* (C–D), and human (E–F). Scatterplots of array RMSE vs array range (A, C, E) and seq RMSE vs seq range (B, D, F), where "range" denotes the difference between maximum and minimum expression values, are shown.

Comparing X-Plat with other transformation tools

We applied TDM and HARMONY to the expression datasets from all three organisms and compared the resulting RMSE values on the held-out test samples from the first cross-validation fold with those obtained using X-Plat (Figure 6).

Comparison using RMSE. In terms of average RMSE per gene, X-Plat outperformed HARMONY and TDM for the vast majority of genes across organisms and directions. Specifically, for the sequencing-to-array transformation, X-Plat achieved lower average RMSE than both tools for 100%, 100%, and 99.99% of genes in rat, *Arabidopsis*, and human, respectively, and for the array-to-sequencing transformation, for 99.65%, 82.53%, and 95.80% of genes in these organisms. The array RMSE spread for X-Plat was ≤0.5 across all three organisms. TDM showed a broader array RMSE spread (≤1.5), and HARMONY had the largest spread, with array RMSE extending beyond 4 in all three organisms (Figure 6A, C, E). In terms of seq RMSE, X-Plat values were ≤2 in rat and human and substantially lower than those of TDM and HARMONY, whereas in *Arabidopsis* they were generally ≥3; although X-Plat still outperformed TDM and HARMONY, the margin of improvement in RMSE was smaller than in rat and human (Figure 6B, D, F), with X-Plat performing best overall, followed by HARMONY and then TDM.

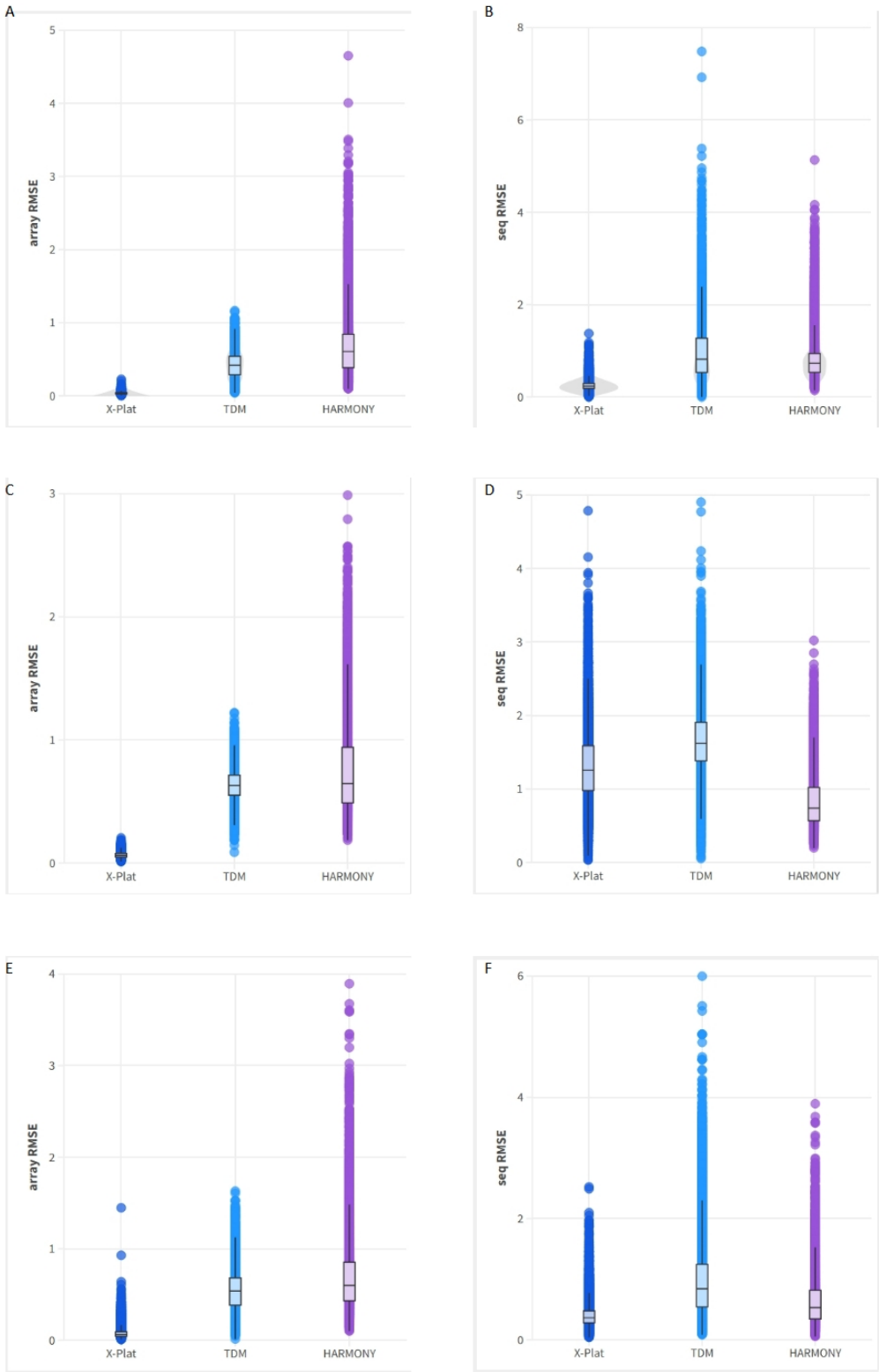


Figure 6: Root mean square error (RMSE) values for X-Plat, TDM, and HARMONY for array RMSE (A, C, E) and seq RMSE (B, D, F) across rat, *Arabidopsis*, and human.

Comparison using per-sample prediction error. For rat, X-Plat had the lowest prediction error for the seq → array transformation in all samples for 94.42% of genes (Figure 7A). For the remaining 5.58% of genes, X-Plat still had the lowest array prediction error in at least 86.49% of these genes. HARMONY predicted array expression values with the lowest error in 2.7% of samples for six genes (*Cdr2*, *Cep350*, *Hsf4*, *Shisa3*, *Wdr91*, and *Xrcc6*). TDM had the lowest array prediction error in 13.51% of test samples for Clcf1, in 10.81% of samples for *Arl6ip6*, *Cep350*, *Rictor*, *Stk11*, *Sycp3*, *Tbc1d10b*, and *Tmem39a*, in 5–10% of samples for 96 genes, and in ≤5% of samples for 508 genes.

For the array → seq transformation in rat, X-Plat had the lowest prediction error in ≥80% of samples for 95.05% of genes and in all samples for 57.66% of genes (Figure 7A). Among the remaining genes, X-Plat had the lowest seq prediction error in 60–80% of samples for 442 genes (4.01%), in 40–60% of samples for 73 genes, in 20–40% of samples for 21 genes, and in ≤20% of samples for nine genes, including *Ank2*, for which X-Plat had the lowest seq prediction error in only 5.41% of samples.

For *Arabidopsis*, X-Plat had the lowest prediction error for the seq → array transformation in all samples for 91.54% of genes (Figure 7B). For the remaining 8.46% of genes, X-Plat still had the lowest array prediction error in at least 67.86% of samples. HARMONY did not have the lowest array prediction error for any sample for any gene. TDM had the lowest array prediction error in 32.14% of test samples for *at3g19780*, in 28.57% of samples for *at3g18140*, in 21.43% of samples for *at5g61450*, *at2g45220*, *at3g49780*, *at4g11850*, *at5g24530*, *at5g17050*, *at4g34180*, *at5g12050*, *at3g23030*, and *at4g27880*, in 10–20% of samples for 114 genes, and in ≤10% of samples for 616 genes.

For the array → seq transformation in *Arabidopsis*, X-Plat had the lowest prediction error in ≥80% of samples for 76.39% of genes and in all samples for 31.05% of genes (Figure 7B). Among the remaining genes, X-Plat had the lowest seq prediction error in 60–80% of samples for 1 735 genes (19.75%), in 40–60% of samples for 273 genes, and in 20–40% of samples for 66 genes.

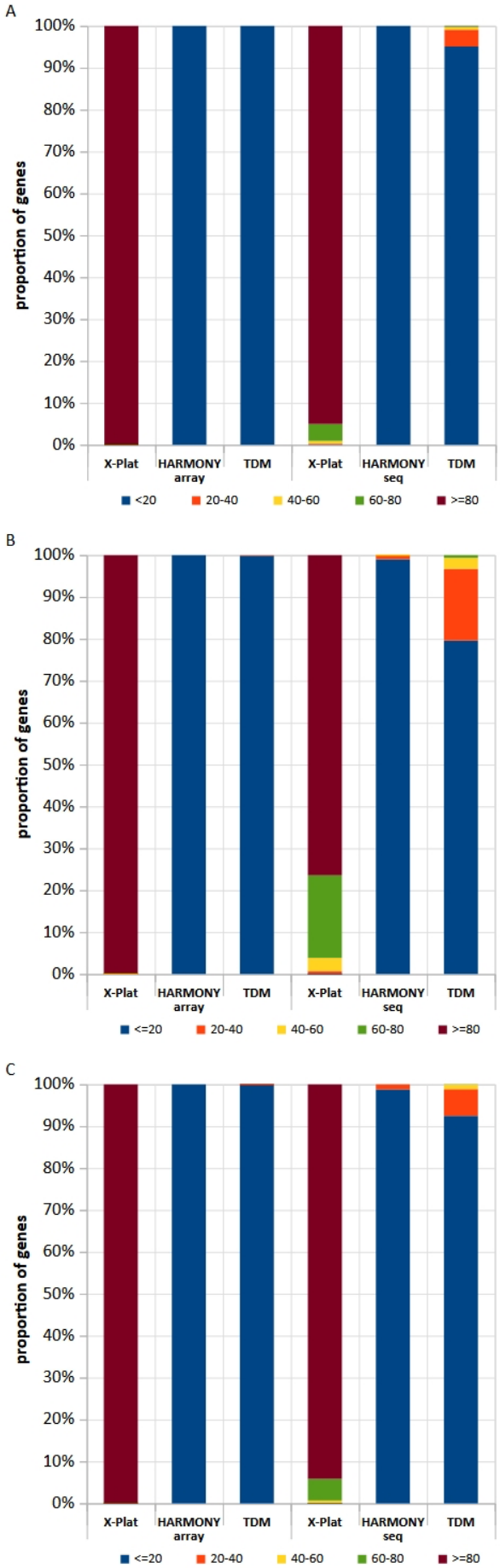


Figure 7: Comparison of results obtained using X-Plat, HARMONY, and TDM. Prediction errors and sample frequencies in five threshold categories (≤20%, 20–40%, 40–60%, 60–80%,

≥80%) are shown as histograms of the proportion of genes, separately for array and seq predictions, for rat (A), *Arabidopsis* (B), and human (C).

For human, X-Plat had the lowest prediction error for the seq → array transformation in ≥80% of samples for all genes and in all samples for 99.96% of genes (Figure 7C). HARMONY had the lowest array prediction error in 6.67% of samples for *CTAGE1* and *EFCAB5*, while TDM had the lowest error in 20% of samples for *ABCA12* and in 6.67% of samples for *ABCA17P*, *ABCA8*, and *ABCA9*. For the array → seq transformation, X-Plat had the lowest prediction error in all samples for 69.41% of genes and in ≥80% of samples for 94.19% of genes (Figure 7C). HARMONY had the lowest seq prediction error in 46.67% of samples for LIPE, in 40% of samples for *FAM107A* and *HLA*-C, in 20–40% of samples for 175 genes, and in ≤20% of samples for 2 478 genes. TDM had the lowest seq prediction error in 73.33% and 66.67% of samples for *HLA-DRB6* and *CLDN10*, respectively, and in 40–60%, 20–40%, and ≤ 20% of samples for 167, 908, and 1 523 genes, respectively.

Factors affecting differences in per-sample prediction error. We next explored factors that could explain differences in performance between the three tools, i.e. why X-Plat had the lowest predicted seq error for most genes with high sample frequency, but not for all genes, and to identify in which cases TDM and HARMONY instead had the lowest prediction error.

For each organism and each comparison category (ΔE_a(XT), ΔE_a(XH), ΔE_s(XT), and ΔE_s(XH)), we selected the top five genes with the largest positive Δ values and the top five with the largest negative Δ values (10 genes per category) for detailed inspection. These genes represent cases where X-Plat or the comparator tool (TDM or HARMONY) most strongly outperformed the other, allowing us to investigate expression-related factors associated with extreme differences in prediction error.

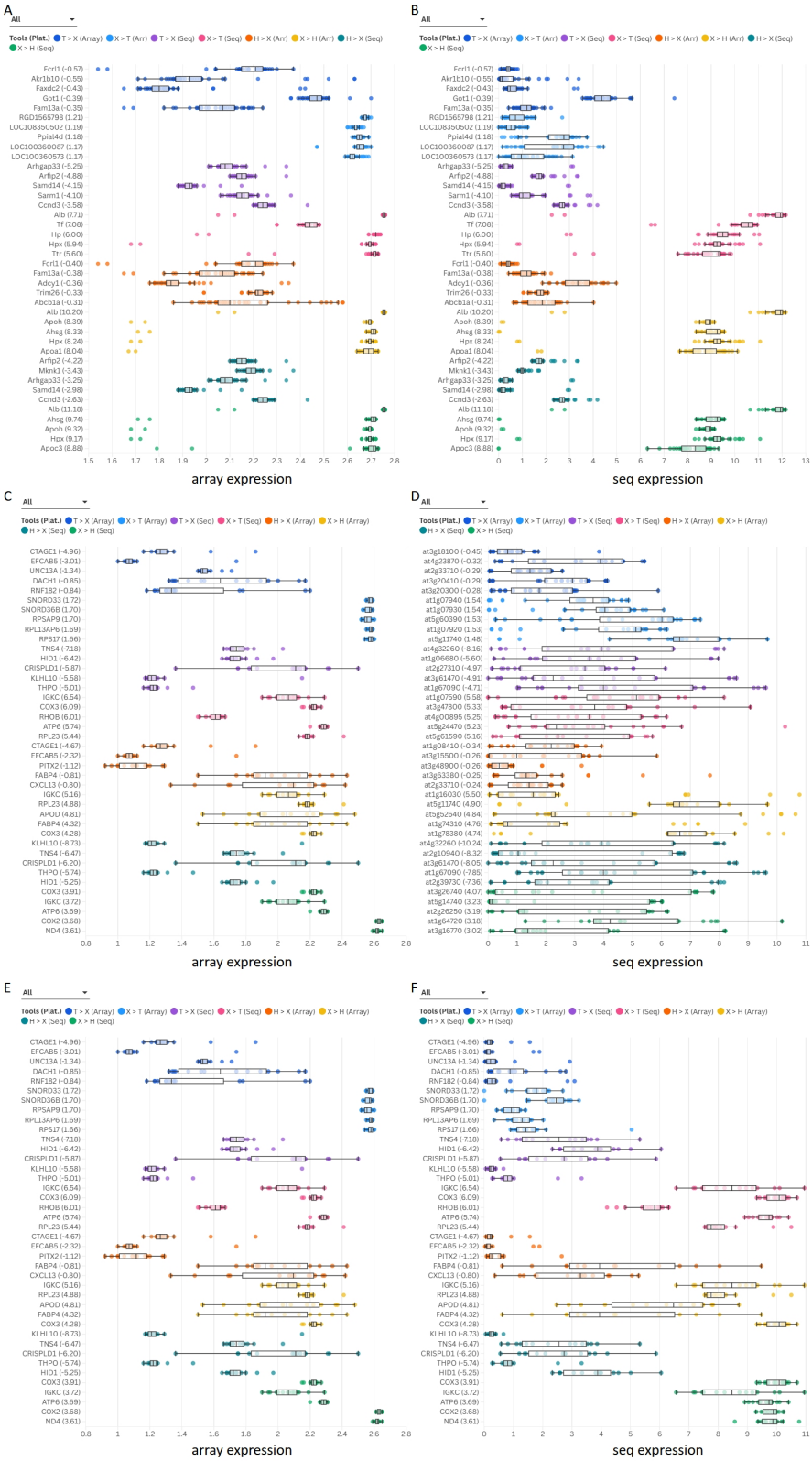

A
All
Tools (Plat.) T > X (Array) X > T (Arr) T > X (Seq) X > T (Seq) H > X (Arr) X > H (Arr) H > X (Seq) X > H (Seq)
array expression
B
All
Tools (Plat.) T > X (Array) X > T (Arr) T > X (Seq) X > T (Seq) H > X (Arr) X > H (Arr) H > X (Seq) X > H (Seq)
seq expression
C
All
Tools (Plat.) T > X (Array) X > T (Array) T > X (Seq) X > T (Seq) H > X (Array) X > H (Array) H > X (Seq) X > H (Seq)
array expression
D
All
Tools (Plat.) T > X (Array) X > T (Array) T > X (Seq) X > T (Seq) H > X (Array) X > H (Array) H > X (Seq) X > H (Seq)
seq expression
E
All
Tools (Plat.) T > X (Array) X > T (Array) T > X (Seq) X > T (Seq) H > X (Array) X > H (Array) H > X (Seq) X > H (Seq)
array expression
F
All
Tools (Plat.) T > X (Array) X > T (Array) T > X (Seq) X > T (Seq) H > X (Array) X > H (Array) H > X (Seq) X > H (Seq)
seq expression

Figure 8: Array (A, C, E) and sequencing (B, D, F) expression values for the genes with five largest positive and negative Δ values in each of ΔE_a(XT), ΔE_a(XH), ΔE_s(XT) and ΔE_s(XH) categories for rat (A–B), *Arabidopsis* (C–D), and human (E–F) datasets (see Methods for details on estimation of Δ*E*). T > X indicates cases where TDM predictions outperform those of X-Plat. The y-axes list gene names with the corresponding Δ value in parentheses.

For the seq → array direction, array expression values showed a much tighter spread with higher median expression (2.6–2.7 in rat, 2.5–2.62 in *Arabidopsis*, and 2.55–2.6 in human; Figure 8A, C, E) for genes with the top five maximum positive Δ array values (ΔE_a(XT), where X-Plat outperformed TDM in at least one sample) than for genes with the top five minimum negative Δ array values, where TDM outperformed X-Plat in at least one sample (1.7–2.5 in rat, 1.5–2.2 in *Arabidopsis*, and 1–2.2 in human; Figure 8A, C, E). This pattern suggests better performance of X-Plat for more highly expressed genes across organisms. A similar trend was observed when comparing X-Plat with HARMONY for predicted array expression in rat, and to a lesser extent in *Arabidopsis*, but not in human.

For the array → seq direction, sequencing expression values also showed a relatively higher spread (-8.5–12 in rat and -5.5–-10.3 in human; Figure 8B, F) for genes with the top five maximum positive Δ seq values (ΔE_s(XT), where X-Plat outperformed TDM in at least one sample) than for genes with the top five minimum negative Δ seq values, where TDM outperformed X-Plat (0–-3 in rat and -0–4.5 in human; Figure 8B, F). This again suggests that X-Plat performs better for genes with higher sequencing expression. In *Arabidopsis*, this trend was not seen with ΔE_s(XT) versus sequencing expression levels; instead, an inverse relationship was observed between ΔE_s(XT) and array expression (except for *atg27310*), suggesting that in this organism X-Plat can perform relatively better for less highly expressed genes (1.6–-2.2; Figure 8D). Similar expression-level trends were observed for the top five genes with maximum positive and negative ΔE_s(XH) values when comparing X-Plat and HARMONY in rat and human, but not in *Arabidopsis*.

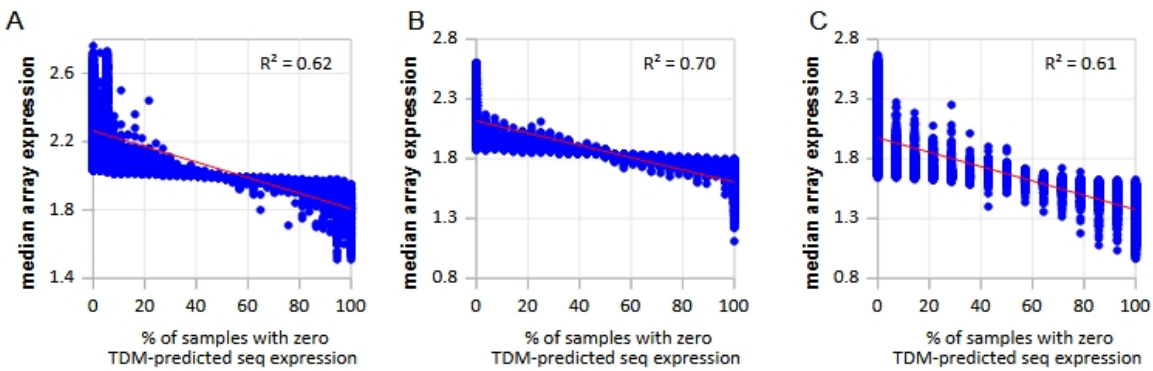


Figure 9: Percentage of samples with zero predicted seq expression values from TDM for rat (A), *Arabidopsis* (B), and human (C) datasets.

We found that TDM predictions for seq expression values contained many zeros. This apparent artefact was observed across all three organisms, with the proportion of samples receiving zero predictions ranging from 0% to 100% in a gene-dependent manner (Figure 9). The sample frequencies of zero seq value predictions by TDM were inversely correlated with the median array expression value in rat, *Arabidopsis*, and human, indicating that lower-expression genes were more likely to be driven to zero by TDM.

DISCUSSION

We developed X-Plat as a cross-platform conversion tool that enables effective use of legacy gene expression and methylation data for model building, thereby expanding the available sample pool for training and validation. While the transition to next-generation sequencing–based assays is inevitable as microarray platforms become obsolete and reagent supplies are discontinued, X-Plat ensures that the vast amount of data generated using microarrays and single-gene experiments remains valuable and usable. Although several tools have been proposed to transform data between platforms (5, 8, 14–16), to our knowledge, no existing method provides a practically applicable, gene-specific and bidirectional conversion framework that spans both expression and methylation data, and some explicitly restrict sequencing reads to regions where microarray probes were designed (15, 16, 27). By contrast, X-Plat estimates gene-wise mappings that can be applied to new samples without requiring probe-level sequencing read information, making it suitable for large-scale public compendia and routine model deployment.

Following the development of X-Plat, we compared its performance with two widely used conversion tools, TDM (8) and HARMONY (26). Across platforms and organisms, X-Plat consistently outperformed both TDM and HARMONY. In rat, *Arabidopsis*, and human expression datasets, X-Plat achieved lower average cross-validated RMSE than HARMONY and TDM for ≥95% of genes in all sequencing-to-array transformations and in most array-to-sequencing transformations, with only the *Arabidopsis* array-to-sequencing setting showing a slightly lower proportion (~82%). X-Plat also achieved low RMSE (≤0.2) for the majority of CpG regions in paired human array–sequencing methylation datasets. These findings are consistent with prior work showing that careful cross-platform normalization can enable robust model transfer between microarray and RNA-seq data, but indicate that a gene-wise, bidirectional mapping such as X-Plat can further reduce prediction error across diverse organisms and assays (8, 28). We observed discrepancies in RMSE values for rat compared with *Arabidopsis* and human; this is likely contributed by larger annotation gaps and

less mature genome builds for rat, whereas *Arabidopsis* and human benefit from extensive, community-driven and long-term annotation refinement, respectively, which improves mapping and quantification and thus supports more accurate cross-platform modeling (29, 30).

We further investigated factors underlying the observed performance differences among the three tools and found that X-Plat performed particularly well for highly expressed genes across organisms, for both microarray- and sequencing-based data. Expression-range analyses showed that genes with larger dynamic range and higher median expression tended to have lower RMSE under X-Plat than under TDM or HARMONY, whereas lowly expressed or low-range genes were more likely to show smaller or reversed differences, consistent with the fact that RMSE is dominated by large-magnitude errors. Down-sampling of the training data (Supplementary Figure S2) resulted in only a marginal increase in RMSE for X-Plat predictions of both array and sequencing expression values, indicating that the method is relatively robust to reductions in training sample size. Although TDM occasionally exhibited lower sequencing RMSE than X-Plat, this improvement often—but not always—coincided with its tendency to predict zero expression values for sequencing data; in such cases, the apparent superior performance of TDM reflects an artefact of zero-inflation rather than a genuine gain in predictive accuracy (31). In practical terms, X-Plat provides more stable performance across genes and organisms, with fewer extreme failure modes, whereas TDM and HARMONY tend to underperform except in a small subset of genes where their assumptions happen to align with the data (28).

In summary, our study contributes to broader efforts in cross-platform conversion and integration, addressing a key barrier to reusing legacy datasets (6, 8, 28). Because it is difficult to draw robust conclusions from data generated in any single laboratory—often due to limited sample sizes—it is increasingly important to correlate and integrate data across experiments, laboratories, platforms, assays, and technologies (32, 33). A platform that can comparably analyze data across assays and technologies, such as X-Plat, will help integrate useful legacy data with data generated using modern high-throughput sequencing instruments (28). This can enhance basic hypothesis generation and strengthen conclusions across domains, including disease research, where leveraging both historical microarray cohorts and contemporary sequencing datasets is particularly valuable (6, 34).

AUTHOR CONTRIBUTIONS

Neeraja M Krishnan: Conceptualization, Formal analysis, Methodology, Visualization, Validation, Writing—original draft, review & editing. Sarah I Rahman: Formal analysis. Lars Røn Olsen: Data generation. Binay Panda: Conceptualization, Methodology, Writing—review & editing.

SUPPLEMENTARY DATA

Supplementary Data are provided in a separate file.

CONFLICT OF INTEREST

None

FUNDING

This work was supported by an extramural grant from the Department of Biotechnology (DBT), Government of India to BP (BT/PR32828/BID/7/866/2019).

DATA AVAILABILITY

All data derived from sources in the public domain and are detailed in Supplementary Table S1 with Accession Numbers.

REFERENCES

1. Zhao S., Fung-Leung W.P., Bittner A., Ngo K. and Liu X. (2014) Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One, **9**, e78644.
2. Conesa A., Madrigal P., Tarazona S., Gomez-Cabrero D., Cervera A., McPherson A., Szcześniak M.W., Gaffney D.J., Elo L.L., Zhang X. and Mortazavi A. (2016) A survey of best practices for RNA-seq data analysis. Genome Biology, **17**, 13.
3. Rao M.S., Van Vleet T.R., Ciurlionis R., Buck W.R., Mittelstadt S.W., Blomme E.A.G. and Liguori M.J. (2019) Comparison of RNA-seq and microarray gene expression platforms for the toxicogenomic evaluation of liver from short-term rat studies. Frontiers in Genetics, **9**, 636.
4. Antoszewski K. and Chmielewska K. (2026) Designing RNA sequencing experiments: a practical guide to reproducible gene expression analysis. Computational and Structural Biotechnology Journal, **31**, 101–119.
5. Su Z., Łabaj P.P., Li S., Thierry-Mieg J., Thierry-Mieg D., Shi W., Wang C., Schroth G.P., Setterquist R.A., Thompson J.F. et al. (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nature Biotechnology, **32**, 903–914.
6. Wang C., Gong B., Bushel P.R., Thierry-Mieg J., Thierry-Mieg D., Xu J., Fang H., Hong H., Shen J., Su Z. et al. (2014) The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. Nature Biotechnology, **32**, 926–932.
7. Johnson W.E., Li C. and Rabinovic A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, **8**, 118–127.

8. Thompson J.A., Tan J. and Greene C.S. (2016) Cross-platform normalization of microarray and RNA-seq data for machine learning applications. PeerJ, **4**, e1621.
9. Titus A.J., Gallimore R.M., Salas L.A. and Christensen B.C. (2016) methyLiftover: cross-platform DNA methylation data integration. Bioinformatics, **32**, 2486–2488.
10. Tarazona S., Balzano-Nogueira L., Furió-Tarí P., Ferrer A., Conesa A. (2015) Data integration in genomics and transcriptomics: a review of methods and applications. Briefings in Bioinformatics, **16**, 977–991.
11. Shen R., Ghosh D. and Chinnaiyan A.M. (2004) Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. BMC Genomics, **5**, 94.
12. Wu C. and Lu Y. (2017) Empirical Bayes batch effect adjustment for microarray data with batch-specific parameters. Statistical Applications in Genetics and Molecular Biology, **16**, 343–360.
13. Allen J.D., Xie Y., Chen M. and Girard L. (2012) Probe mapping across multiple microarray platforms. Nucleic Acids Research, **40**, e147.
14. Frolova A., Lykhenko O. and Obolenska M. (2016) A comparison of approaches to integration of microarray expression data from multiple platforms. Biopolymers and Cell, **32**, 173–187.
15. Uziela K. and Honkela A. (2013) Probe-level expression change averaging improves cross-platform analysis of gene expression data. BMC Bioinformatics, **14**, 334.
16. Uziela K. and Honkela A. (2015) Improved probe-level RNA-seq expression estimates by modelling read counts. BMC Bioinformatics, **16**, 268.
17. Zhang W., Zhan X., Rosenberg N.A. and Guan Y. (2020a) MatchMixeR: a cross-platform normalization method for gene expression microarray data. Bioinformatics, **36**, 2486–2493.
18. Afgan E., Baker D., Batut B., van den Beek M., Bouvier D., Cech M., Chilton J., Clements D., Coraor N., Grüning B.A. et al. (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research, **46**, W537–W544.
19. Gautier L., Cope L., Bolstad B.M. and Irizarry R.A. (2004) affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics, **20**, 307–315.
20. Andrews S. (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
21. Bolger A.M., Lohse M. and Usadel B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, **30**, 2114–2120.
22. Kim D., Paggi J.M., Park C., Bennett C. and Salzberg S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology, **37**, 907–915.

23. Zhang Y., Parmigiani G. and Johnson W.E. (2020b) ComBat-seq: batch effect adjustment for RNA-seq count data. Nucleic Acids Research, **48**, e17.
24. Thompson J.R. and Foltz S.M. (2022) DGEobj.utils: Differential Gene Expression (DGE) Analysis Utility Toolkit. R package version 1.0.6. Available at: https://CRAN.R-project.org/package=DGEobj.utils.
25. 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. (2020) Array programming with NumPy. Nature, **585**, 357–362.
26. Borisov N., Shabalina I., Tkachev V., Sorokin M., Garazha A., Pulin A., Eremin I. and Buzdin A. (2019) Shambhala: a platform-agnostic data harmonizer for gene expression data. BMC Bioinformatics, **20**, 66.
27. Pedersen C.B., Nielsen F.C., Rossing M. and Olsen L.R. (2018) Using microarray-based subtyping methods for breast cancer in the era of high-throughput RNA sequencing. Molecular Oncology, **12**, 2136–2146.
28. Foltz S.M., Sattler M.C., Wang S., Cheng W.Y. and Greene C.S. (2023) Methodology to identify a gene expression signature by merging microarray datasets. Computers in Biology and Medicine, **160**, 106867.
29. Pruitt K.D., Brown G.R., Hiatt S.M., Thibaud-Nissen F., Astashyn A., Ermolaeva O., Farrell C.M., Hart J., Landrum M.J., McGarvey K.M. et al. (2014) RefSeq: an update on mammalian reference sequences. Nucleic Acids Research, **42**, D756–D763.
30. Atanur S.S., Birol I., Guryev V., Hirst M., Hummel O., Morrissey C., Behmoaras J., Fernandez-Suarez X.M., Johnson M.D., McLaren W.M. et al. (2010) The genome sequence of the spontaneously hypertensive rat: analysis and functional significance. Genome Research, **20**, 791–803.
31. Silverman J.D., Roche K., Mukherjee S. and David L.A. (2020) Naught all zeros in sequence count data are the same. Nature Communications, **11**, 1–13.
32. Qin, X., Huang, H., Li, Y. and Liu, Y. (2012) Multiple-platform data integration method with application to combined analysis of microarray and proteomic data. BMC Bioinformatics, **13**, 320.
33. Jiang Y., Redmond D., Nie K., Ghaffari S., Hsu Y.H., Thompson A., Sinha R. and Elemento O. (2021) Rank-in: enabling integrative analysis across microarray and RNA-seq for gene expression. Nucleic Acids Research, **49**, e99.
34. Wang M., Chen X., Zhang M., Klebanov L., Yakovlev A.Y. and others (2015) An investigation of biomarkers derived from legacy microarray data for their utility in the RNA-seq era. BMC Genomics, **16**, 1–12.