biomedicines

Review

Transcriptomic Harmonization as the Way for Suppressing Cross-Platform Bias and Batch Effect

Nicolas Borisov [1,2,*] and Anton Buzdin [1,2,3,4]

[1] World-Class Research Center "Digital Biodesign and Personalized Healthcare",
Sechenov First Moscow State Medical University, 119435 Moscow, Russia
[2] Moscow Institute of Physics and Technology, 141701 Dolgoprudny, Russia
[3] Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, 117997 Moscow, Russia
[4] PathoBiology Group, European Organization for Research and Treatment of Cancer (EORTC),
1200 Brussels, Belgium
* Correspondence: borisov.nm@mipt.ru; Tel.: +7-9032187261

Abstract: (1) Background: Emergence of methods interrogating gene expression at high throughput gave birth to quantitative transcriptomics, but also posed a question of inter-comparison of expression profiles obtained using different equipment and protocols and/or in different series of experiments. Addressing this issue is challenging, because all of the above variables can dramatically influence gene expression signals and, therefore, cause a plethora of peculiar features in the transcriptomic profiles. Millions of transcriptomic profiles were obtained and deposited in public databases of which the usefulness is however strongly limited due to the inter-comparison issues; (2) Methods: Dozens of methods and software packages that can be generally classified as either flexible or predefined format harmonizers have been proposed, but none has become to the date the gold standard for unification of this type of Big Data; (3) Results: However, recent developments evidence that platform/protocol/batch bias can be efficiently reduced not only for the comparisons of limited transcriptomic datasets. Instead, instruments were proposed for transforming gene expression profiles into the universal, uniformly shaped format that can support multiple inter-comparisons for reasonable calculation costs. This forms a basement for universal indexing of all or most of all types of RNA sequencing and microarray hybridization profiles; (4) Conclusions: In this paper, we attempted to overview the landscape of modern approaches and methods in transcriptomic harmonization and focused on the practical aspects of their application.

Keywords: gene expression; transcriptional profiles; RNA sequencing; microarray hybridization; data normalization and harmonization; batch effect; machine learning; Big Data; universal data indexing

check for updates

Citation: Borisov, N.; Buzdin, A. Transcriptomic Harmonization as the Way for Suppressing Cross-Platform Bias and Batch Effect. *Biomedicines* **2022**, *10*, 2318. https://doi.org/10.3390/biomedicines10092318

Academic Editor: Christos K. Kontos

Received: 22 August 2022
Accepted: 16 September 2022
Published: 18 September 2022

1. The Problem of Transcriptomic Data Harmonization

The digital ocean of whole-transcriptome gene expression profiles has flooded since the early 2000s when the first generation of robust and reproducible mRNA microarray hybridization (MH) techniques was introduced into the routine laboratory practice [1–4]. The outstandingly high importance of the open-access gene expression data that could be accumulated and extracted from public databases was recognized immediately, thus leading to emergence of popular online repositories such as Gene Expression Omnibus (GEO) [5,6] or ArrayExpress [7,8]. Later on, this has also inspired many impactful large-scale integrative biomedical cooperative projects such as The Cancer Genome Atlas (TCGA) [9,10] for cancer genomics and transcriptomics, Gene-Tissue Expression (GTEx) [11,12], and Atlas of Normal Tissue Expression (ANTE) [13] for normal human tissue expression profiles, the CancerRxGene database [14] for genomes and transcriptomes of cell lines connected with their response to hundreds of drugs, and the Broad Institute deconvoluted profiles for gene expression changes in cells under the influence of gene constructs, drugs, and other chemicals [15,16].

Shortly after the critical mass of gene expression profiles has accumulated, the following two conceptual problems with the data analysis were recognized. First, poor technical compatibility of the expression profiles is obtained using different experimental platforms/equipment, protocols, and reagents [17–21]. Indeed, this can be readily explained by the different physico-chemical principles of gene detection and interrogation [22,23] and by specific library preparation enzymatic bias [24]. The second problem (so-called batch effect) dealt and still deals with the unclear compatibility of gene expression profiles obtained with the same equipment and reagents, but in different series of experiments, e.g., they are performed in different times or in different labs [25,26]. There is no clear explanation of the nature of the batch effect (e.g., it may be due to relatively different activities of enzymes and chemicals for library preparation and MH or RNA sequencing from batch to batch), but the effect itself is sound and frequently inevitable [25].

The compromised compatibility of gene expression profiles obtained using different platforms and protocols was experimentally explored in the international projects MAQC (for MH) and SEQC (for RNA sequencing). Both MAQC [17–19] and SEQC [27] projects investigated compatibilities of gene expression profiles obtained using various microarray and sequencing platforms for the same set of four sample types (named A, B, C, and D), each performed in multiple replicates. Type A samples were the commercially available Stratagene Universal Human Reference RNA specimens for all but brain human tissues; type B samples were also commercially available Ambion Human Brain Reference RNA. Types C and D samples were the mixtures of A and B with the A:B ratios of 3:1 and 1:3, respectively. In the MAQC project [17–19], the samples of types A, B, C, and D were profiled using the MH platforms Agilent-012391 Whole Human Genome Oligo Microarray G4112A (GPL1708), Affymetrix Human Genome U133 Plus 2.0 Array (GPL570) and Illumina Sentrix Human-6 Expression Beadchip (GPL2507). In the SEQC project [27], the same samples were profiled using the NGS platform Illumina HiSeq 2000 (GPL11154), as well as three MH platforms: Illumina HumanHT-12 V4.0 expression beadchip (GPL10558), Affymetrix Human Gene 2.0 ST Array (GPL17930), and Affymetrix GeneChip® PrimeView™ Human Gene Expression Array (GPL16043).

The MAQC and SEQC projects investigated the correlations between the transcriptome profiles of the same biological type, yet obtained using the different experimental platforms. Although these correlations were high [17–19,27], without the some special cross-platform normalization methods (quantile normalization [28] was not enough), the overall collections of profiles were grouped according to the experimental platforms, rather than to the biological type of samples, in terms of both clustering dendrograms and of principal component analysis (PCA) [29–34].

As the reaction of the scientific community, a bunch of first-generation harmonization/normalization methods was generated in the first decade of the 21st century, aimed at the standardization of multi-platform expression profiles using specific algorithms. These methods were mostly trained on the different types of MH gene expression data and could dynamically transform gene profiles into a flexible yet inter-comparable form [35]. The following alternative approaches that have different principles and different destinies could be mentioned in this review: Quantile Normalization (QN) [28], Quantile Discretization (QD) [36], Normalized Discretization (NorDi) [37], Distribution Transformation (DisTran) [38], Empirical Bayes (EB)/ComBat [39], Distance-Weighted Discrimination (DWD) [40–42], Cross-Platform Normalization (XPN) [29,31], Gene Quantiles (GQ) [43], and PLatform-Independent Latent Dirichlet Allocation (PLIDA) [30].

Further approaches were largely influenced by the coming era of routine next-generation sequencing (NGS) of mRNA (RNA sequencing or RNAseq) that has started roughly in the second decade of this century. Nowadays, RNAseq has become the gold standard and the basic tool for transcriptomic profiling [44–50]. In addition to measuring gene activities, RNAseq has also the potential of detecting mutations and overall tumor mutational burden [51], gene splice isoforms [52], and oncogenic fusion transcripts [53–56]. During the RNAseq era, a new group of cross-platform data comparison methods was developed [27]. However, the RNAseq gene expression profiles have outnumbered the MH counterparts relatively recently. It was only in 2019 [32] when the number of datasets for the most popular RNAseq platform (Illumina HiSeq 2000, GPL11154) exceeded the number of datasets for the most popular MH platform (Affymetrix U133, GPL570). Moreover, the total number of individual profiles for those two platforms is still comparable in 2022 as well. Many unique transcriptomic profiles exist only as the MH data, e.g., clinically annotated expression profiles for some pathological conditions including cancers [57].

For the RNAseq data, a method called The Differential gene Expression in Sequencing, DESeq [58]/DESeq2 [59–61] based on the negative binomial distribution law, has rapidly become the standard in the field for the intra-platform normalization. However, effective method for the cross-platform RNAseq, or for RNAseq vs. MH harmonization was missing until recently, although several attempts for simultaneous normalization of MH and RNAseq data must be mentioned, such as Training Distribution Machine (TDM) [62], Universal exPression Code (UPC) [63], Feature-specific QN (FCQN) [64,65], MatchMixeR (MM) [66], Integrative Bayesian Network (IBN) [67], Rank-in [68], and Elastic Shared LASSO Regularization (ESLR) [69] methods. The divergence analysis method is another interesting attempt to compare the MH and NGS mRNA expression profiles, as well as microRNA and DNA methylation data [70]. The authors of the divergence analysis first applied conditional probability (Bayesian) models to mimic the unspecified (generalized-type) distribution that describes the gene expression/methylation data. This reconstruction was followed by the divergence analysis of one biological sample type from another. Although Dinalankara at al. [70] have managed to distinguish different samples after their divergence analysis, the applicability of their approach to a wide range of popular MH and NGS platforms remains unexplored [70].

In this review, we classified available intra- and cross-platform harmonization methods of transcriptomic profiles and compared their performance characteristics. Finally, we also included practical recommendations that may guide the reader to select optimal method depending on a specific task.

2. Principles of Harmonization Algorithms

Different harmonization methods are based on different algorithms aimed to suppress the platform bias and the batch effect. These algorithms may utilize different approaches to gene expression data processing and produce output data in different formats. Considering the mathematical apparatus, we proposed the following classification:

(1) Methods based on statistical transformations (considering quantiles, ranks, means, medians of gene expression levels, etc.):

   (a) Those using ranking of expression levels and setting the output levels according to the averaged values, such as QN [28], Feature-Specific QN (FCQN) [64], Quantile Discretization (QD) [36], Gene Quantiles (GQ) [43], Normalized Discretization (NorDi) [37], Distribution Transformation (DisTran) [36,38], Median Rank Scores (MRS) [36], YuGene [71], and Rank-in [68];

   (b) Those using piecewise rescaling of log-expression levels according to the mean/median values over distinct genes and samples, such as Column Sample (CS), Median-Centered (MC) [29], and Analysis of Variance (ANNOVA) [72] method;

(2) Methods using regression and/or maximum likelihood models for validation of predefined statistical hypotheses:

   (a) Those using negative binomial distribution, such as the DESeq [58]/DESeq2 [59–61];

   (b) Those using log-normal distribution with either covariance analysis [73], or with conditional/Bayesian models, as for the methods Universal exPression Code (UPC) [63,74], Empirical Bayes (ComBat) [39], Robust Microarray Analysis (RMA) [75], GeneChip Robust Multiarray Analysis (gcRMA) [76], Model-Based Expression Indices (MBEI) [77], Probe Logarithmic Intensity ERror

(PLIER) estimation [78], frozen Robust Microarray Analysis (fRMA) [79–82], MatchMixeR (MM) [66], Cross-Platform Comparison (XPC) [83];

(c) Those using Dirichlet and gamma distributions as for the method PLatform-Independent Latent Dirichlet Allocation (PLIDA) [30];

(d) Those using the empirical superposition of conditional probabilistic (Bayesian) models that describe the generalized-type distribution as for the method applied for the comparison of the MH, NGS, microRNA, and DNA methylation data [67,70];

(e) Those using the Least Absolute Shrinkage and Selection Operator (LASSO) regression models [69];

(3) Methods finding similar clusters in gene expression matrices of the datasets under normalization and then using iterative corrections to fit each cluster as close as possible to the target model:

(a) Those using piecewise linear interpolations in the log-expression space, such as Cross-Platform Normalization (XPN) [29];

(b) Those using piecewise cubic interpolations in the log-expression space, such CuBlock [34].

(4) Methods utilizing machine learning (ML) to find and artificially remove dissimilarities between datasets to be normalized:

(a) Those using the linear support vector machine (SVM) ML method, such as Distance-Weighted Discrimination (DWD) [40–42];

(b) Those using quantile-based regression models for data transfer from source to target datasets, such as Training Distribution Machine (TDM) [62].

Another important aspect that must be considered in this review is the format of output gene expression data generated by the harmonization techniques. Most of currently existing methods return the results in the flexible format. For the flexible normalization, the shape of the output transformed gene expression profiles is a variable that depends on all the profiles under harmonization. This has an important limitation that one cannot combine the output datasets generated after two or more acts of such harmonization. Even adding as few as just one transcriptional profile would require a new harmonization of the entire dataset. This clearly increases the calculation costs for large datasets that are being routinely updated.

Taken together, these factors complicate the analysis of not only single gene expression levels, but also of higher order gene-based biomarkers such as gene signatures [84], molecular pathway activation levels [85], algorithmically deduced cancer drug efficiency scores [86,87], and different ML models [88–90].

To overcome these limitations, an alternative concept was formulated comprising conversion of a whole set of profiles under harmonization into a pre-defined output shape, e.g., into a shape of a preferred gene interrogating experimental platform. In such a paradigm, the harmonized output should look as if it would be obtained using a predefined gene expression platform. The examples of predefined-shape harmonization methods include Frozen Robust Microarray Analysis (fRMA) [79–82], robust Quantile Normalization [91], Training Distribution Machine (TDM) [62], and Universal exPression Code (UPC) [63].

More recently, we proposed a new family of uniformly shaped cross-platform harmonizers termed Shambhala [32,33]. Harmonization here is performed not simultaneously for all the profiles under harmonization, but for the gene expression profiles taken one by one, when each individual profile is merged and quantile-normalized [28] with an auxiliary calibration dataset that is pre-defined by the method developers. Then, the resulting dataset is converted into the shape of the so-called reference definitive dataset. This creates an additional advantage of co-harmonizing datasets of different, even non-comparable, sizes.

Furthermore, such harmonization may use different mathematical transforms as the engine to reshape the transcriptional profiles. The first version of Shambhala used the piecewise linear method XPN [29,31] for profile reshaping [32], whereas the latest version [33] utilized the piecewise cubic transformation method CuBlock [34].

3. Evaluation of the Quality of Harmonization

Harmonization of transcriptional profiles is a complex process that can distort functionally relevant features such as clustering and neighborhood on a dendrogram and fold-change of gene expression with relation to control samples. We listed in Table 1 some of the quality assessment metrics and the abilities of different methods to retain the initial functional characteristics in the output profiles after harmonization.

The following quantitative metrics and methods may be applied to estimate the effect of harmonization:

(1) First, different statistical criteria may be used to estimate the following endpoints:
   - (a) Correlation analysis for the gene expression profiles before and after harmonization [29–31,33,34];
   - (b) Comparison of between- and within-class distances before and after harmonization [29];

(2) Alternatively, one may classify the samples according to gene expression data after normalization, involving various machine learning (ML) methods:
   - (a) Logistic regression [92], used in [30];
   - (b) SVM [93], used in [29,31];
   - (c) Nearest shrunken centroids Prediction Analysis for Microarrays (PAM) [94], used in [29].

As a typical material for such normalization quality benchmarks, in many studies, the investigators used standardized reference samples, whose gene expression was interrogated with different equipment using different experimental protocols. Probably, the most important series of such cross-comparisons was performed within the Microarray Quality Control (MAQC) [17–19] and Sequencing Quality Control (SEQC) [27] projects mentioned above in this article.

The MAQC and SEQC projects were focused on profiling the specific model human mRNA sample types. One was the commercial Stratagene universal human reference RNA mixture for all but brain tissues; another one was the commercial Ambion human brain reference mRNA, and the two remaining types were the mixtures of the Stratagene/Ambion samples in the ratios of 3:1 and 1:3, respectively.

The quality assessment is based on the expectation that a perfect harmonization must support the similarity of gene expression profiles according to the biological nature of the sample rather than depending on the equipment and reagents used to interrogate gene expression. Thus, early approaches used visual inspection of the principal component analysis (PCA) plots and/or cluster dendrograms to assess the cross-platform harmonization benchmarks [30–34]. However, this could only support a manual qualitative assessment without precise quantitative interrogation of the complex class distribution profiles.

We recently proposed a new metric for the algorithmic cluster analysis of dendrograms [33,95] called Watermelon Multisection (WM). WM measures the strength of data matching with the trait of interest. When moving from the root of the dendrogram to its distal branches, one can calculate general decrease of entropy and, therefore, information gain (IG) at each node of the dendrogram, i.e., its split into two shoulders [95]. This accumulated and normalized IG constitutes the WM metric for a given dendrogram, and a given set of classes under analysis. Consequently, the ratio $= \frac{WM_S}{WM_P}$, where WM_S is WM metric for clustering according to classes corresponding to biological nature and WM_P, according to the experimental platform used, may be used as a facile yet robust estimate of the harmonization quality. A higher R corresponds to a better quality, and vice versa [33].

Table 1. Selected benchmarks of harmonization methods.

Reference for Comparison	Methods	Materials	Experimental Platform	Qualitative Criteria	Quantitative Criteria	Best Methods
[29]	Cross-Platform Normalization, XPN [29]; Column Sample (CS); Median Center (MC); Empirical Bayes (EB) [39]; Distance-Weighted Discrimination (DWD) [41,42]	Three breast cancer datasets [96–98]	Affymetrix GeneChip U95Av2 arrays [96]; 25K Agilent oligonucleotide arrays [97,98]	—	Average distance to nearest sample in another platform; correlation with column standardization data; global integrative correlation; preservation of significantly differential genes	XPN

[31]	XPN; DWD; EB (ComBat) [39]; Median Rank Scores (MRS) [36]; Quantile Discretization (QD) [36]; Normalized Discretization (NorDi) [37]; Distribution Transformation (DisTran) [36,38]; Gene Quantiles (GQ) [43]; Quantile Normalization (QN) [28]	MAQC dataset [17–19]	Human Genome Survey Microarray v2.0;
			Agilent-012391 Whole Human Genome Oligo Microarray G4112A;
			Affymetrix Human Genome U133 Plus 2.0 Array;
			Illumina Sentrix Human-6 Expression Beadchip
				—	Mean-mean regression; cross-dataset data transfer for linear SVM [94] and nearest shrunken centroids [95] classification	XPN (for datasets of comparable size);
						DWD (for datasets of non-comparable size)

[30]	XPN; DWD; platform-independent latent Dirichlet allocation (PLIDA) [30]	Prostate cancer datasets [99,100];
		Breast cancer datasets [97,101];
		MAQC
			Affymetrix Human Genome U133 Array;
			Agilent Human 1A (V2);
			Human Genome Survey Microarray v2.0;
			Agilent-012391 Whole Human Genome Oligo Microarray G4112A;
			Affymetrix Human Genome U133 Plus 2.0 Array;
			Illumina Sentrix Human-6 Expression Beadchip
				Visual inspection of PCA plots.	Correlation analysis between the profiles before and after normalization;
					cross-dataset data transfer for logistic regression classification [92]
						PLIDA

[67]	MatchMixeR (MM) [66];
	DWD; XPM;
	ComBat
		NCI60 cell lines (dataset 1:58 lines; dataset 2:59 lines)	Affymetrix Human Genome U133A array;
			Human Genome U133 Plus 2.0 Array;
			Agilent Human Genome Whole Microarray;
			Illumina HiSeq 2000
				—	R^2 score (R^2 is the proportion of the variation in the dependent variable that is predictable from the independent variable [102] analysis;
					F1 score (F1 score is the harmonic mean of precision and recall [103,104]) analysis
						MM

[32]	Shambhala-1; QN; Differential Gene Expression in Sequencing 2 (DESeq2) [59–61]	MAQC; SEQC datasets [27]	Agilent-012391 Whole Human Genome Oligo Microarray G4112A;
			Affymetrix Human Genome U133 Plus 2.0 Array;
			Illumina Sentrix Human-6 Expression Beadchip;
			Illumina HiSeq 2000;
			Illumina HumanHT-12 V4.0 expression beadchip;
			Affymetrix Human Gene 2.0 ST Array;
			Affymetrix GeneChip® PrimeView™ Human Gene Expression Array
				Visual inspection of cluster dendrograms	—	Shambhala-1 (linear Shambhala)

Table 1. *Cont.*

Reference for Comparison	Methods	Materials	Experimental Platform	Qualitative Criteria	Quantitative Criteria	Best Methods

[34]	CuBlock [34];
	ComBat [39] YuGene [71];
	DBNorm [105]; Shambhala-1 [32];
	Universal exPression Code (UPC) [63]
		MAQC	Agilent-012391 Whole Human Genome Oligo Microarray G4112A;
			Affymetrix Human Genome U133 Plus 2.0 Array;
			Illumina Sentrix Human-6 Expression Beadchip
				Visual inspection of cluster dendrograms and PCA plots	Cross-dataset data transfer for support vector machine (SVM) classification [93]	CuBlock

[33]	Shambhala-2; Shambhala-1; QN; DESeq2; CuBlock; robust QN (QNR) [91]; Training Distribution Machine (TDM) [62]; UPC	GTEx [11], The Cancer Genome Atlas (TCGA) [10];
		Oncobox Atlas of Normal Tissue Expression (ANTE) [13];
		MAQC; SEQC
			Illumina HiSeq 2000;
			Illumina HiSeq 3000;
			Agilent-012391 Whole Human Genome Oligo Microarray G4112A;
			Affymetrix Human Genome U133 Plus 2.0 Array;
			Illumina Sentrix Human-6 Expression Beadchip;
			Illumina HumanHT-12 V4.0 expression beadchip;
			Affymetrix Human Gene 2.0 ST Array;
			Affymetrix GeneChip® PrimeView™ Human Gene Expression Array
				Visual inspection of PCA plots	Watermelon Multisection metric for quantitative assessment of clustering on dendrograms [95]	Shambhala-2 (cubic Shambhala)

4. Application Notes

During the last two decades, quantile normalization (QN) [28] and differential gene expression in sequencing 2, DESeq2 [59–61], have become methods of choice for intra-platform normalization of the MH and RNAseq gene expression data, respectively.

However, for the cross-platform harmonization, dozens of methods were developed for both MH and NGS types of gene expression data, but none of them was so far recognized as the gold standard. In fact, many, if not most, aspects of intra- and cross-platform normalization, such as incomparability of profiles obtained using different platforms, numerous methods for cross-platform normalization, and performance benchmarks for them, were studied in the first decade of the XXI century, in the co-called MH era. The advent of NGS, however, did not make this problem unimportant, at least because there is still a problem regarding how to harmonize old MH and new NGS data.

Most of cross-platform normalization methods return the output data in the flexible format, which requires recalculation of all previously processed profiles when adding new data to the analysis. This may dramatically increase calculation time and costs which can grow exponentially with the increase of the sample size. Furthermore, some methods which show the best performance in cross-platform normalization tests [31], such as XPN [29], have serious limitations. For instance, XPN allows normalization of only two datasets at once, with no subsequent application to other datasets [29]. In addition, an unbalanced size of groups of samples under harmonization may create obstacle to the analysis of the whole groups. The latter may force researchers to arbitrarily decrease samplings and not to include all available data into the analysis.

Thus, the need for the predefined, uniformly-shaped output for data harmonization was formulated about a decade ago [62,63,79–82]. Recently, we proposed a concept of uniformly shaped cross-platform harmonization of gene expression profiles [32,33]. The key feature of such harmonization is that each profile is converted into the shape of the reference definitive dataset independently from other profiles under harmonization. In such a way, the unlimited number of datasets of any size each can be harmonized. Furthermore, adding new data to the analysis does not require recalculation for the previously harmonized profiles, which spares time and reduces costs.

With such a concept in mind, we obtained the best results with the cubic data transformation algorithm adopted from the CuBlock method [34] and built Shambhala-2 package [33]. Shambhala-2 showed a strong capacity to restore the correct order of clusters on dendrograms, when the samples were grouped according to their biological nature, not the technical platform used to profile gene expression. This was effective for both MH and RNAseq types of data, including mixed MH-RNAseq datasets under harmonization. We hope, therefore, that this generation or next generations of Shambhala harmonizer will find their niche in the analysis of big transcriptomic data in the future.

5. Conclusions

We summarized our experience of using various harmonization methods for gene expression profiles in Table 2.

For the intra-platform harmonization of the MH data, QN [28] may seem the method of choice; however, the “robust QN” (QNR) [91] showed generally worse performance than the ordinary QN [33]. In turn, for the intra-platform harmonization of the NGS data, DESeq2 method could be recommended [59–61].

In case of cross-platform harmonization of two datasets with a comparable number of gene expression profiles, the best choice could be the XPN method [29,31]. When the data are MH profiles and there are more than two datasets under analysis, the method CuBlock [34] is preferred.

Finally, in the case of merging the MH and NGS expression datasets, or when merging of data from various platforms is needed, and the uniformly shaped (suitable for further intercomparisons) output format is required, then Shambhala-1 [32] or Shahmhala-2 [33] technique can be the best option.

To our knowledge, Shambhala methods were the first gene expression harmonizers with a uniformly shaped output, which were applied to merge the RNAseq and MH profiles [32,33]. Thus, these methods may become useful for the broad spectrum of applications.

However, one should keep in mind that Shambhala-1/2 approaches are algorithmically complex and, therefore, computational resource-demanding. Thus, parallel execution of the program code may be advantageous [33,89].

Table 2. Recommendations for the use of selected intra- and cross-platform harmonization methods.

Reference	Method	Mathematical Principle	Algorithmic Complexity	Advantages	Shortcomings
[28]	Quantile normalization (QN)	Ranking the expression levels of different genes within each profile and setting the expression level of each gene to the mean value (over all profiles) for the respective rank	Relatively simple	Gold standard method for intra-platform normalization of the MH data	Avoiding being used for cross-platform harmonization of the MH data; requiring recalculation of all gene expression-based values after addition of new samples
[59–61]	Differential Gene Expression in Sequencing 2 (DESeq2)	Transform based on the negative binomial distribution	Moderately complex	Gold standard for intra-platform normalization of RNAseq data	Requiring recalculation of all gene expression-based values after addition of new samples
[29]	Cross-Platform Normalization (XPN)	Piecewise linear iterative transform	Relatively complex	The method of choice for harmonization of two datasets of comparable size	Allowing normalization of more than two datasets; not recommending subsequent application to other datasets; requiring recalculation of all gene expression-based values after addition of new samples
[34]	CuBlock	Piecewise cubic iterative transform	Relatively complex	The method of choice for cross-platform normalization of more than two MH datasets	Requiring recalculation of all gene expression-based values after addition of new samples
[32]	Shambhala-1 (linear Shambhala)	Uniformly shaped harmonization based on the XPN method.	Complex	Working for harmonization of unlimited number of datasets of any size, for both MH and RNAseq data or their combinations; not requiring recalculation of gene expression-based values after addition of new samples	Resource-demanding
[33]	Shambhala-2 (cubic Shambhala)	Uniformly shaped harmonization based on the CuBlock method.	Complex	Working for harmonization of the unlimited number of datasets of any size, for both MH and RNAseq data or their combinations; not requiring recalculation of gene expression-based values after addition of new samples	Resource-demanding

Author Contributions: Conceptualization, N.B. and A.B.; investigation, N.B.; writing—original draft preparation, N.B.; writing—review and editing, A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financed by the Ministry of Science and Higher Education of the Russian Federation within the framework of state support for the creation and development of World-Class Research Centers "Digital biodesign and personalized healthcare" (No. 075-15-2022-304).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We thank Marianna Zolotovskaia and Ivan Petrov from Moscow Institute of Physics and Technology, Ilya Muchnik from Rutgers University, and Donald Geman from Johns Hopkins University for insightful comments.

Conflicts of Interest: The authors declare no conflict of interest.

List of Acronyms:

ANTE	Atlas of Normal Tissue Expression
ComBat	COMpensation of BATch effects
CS	Column Sample
CuBlock	Cubic Blocks
DBNorm	Distribution-Based Normalization
DisTran	Distribution Transformation
DESeq(2)	Differential Gene Expression in Sequencing (2)
DWD	Distance-Weighted Distribution
EB	Empirical Bayes
ESLR	Elastic Shared LASSO Regularization
FCQN	Feature-Specific QN
fRMA	Frozen Robust Microarray Analysis
gcRMA	GeneChip Robust Microarray Analysis
GEO	Gene Expression Omnibus
GQ	Gene Quantiles
GTEx	Genotype Tissue Expression
IBN	Integrative Bayesian Network
IG	Information Gain
LASSO	Least Absolute Shrinkage and Selection Operator
MAQC	Microarray Quality Control
MBEI	Model-Based Expression Indices
MC	Median Center
MH	Microarray Hybridization
ML	Machine Learning
MM	MatchMixeR
MRS	Median Rank Score
NGS	Next-Generation Sequencing
NorDi	Normalized Discretization
PAM	Prediction Analysis for Microarrays
PCA	Principal Component Analysis
PILER	Probe Logarithmic Intensity ERror
PLIDA	PLatform-Independent Latent Dirichlet Allocation
PRIDE	PRoteomics Identification DatabasE
QD	Quantile Discretization
QN	Quantile Normalization
QNR	Qunatile Normalization (Robust)
RMA	Robust Microarray Analysis
SEQC	Sequencing Quality Control
SVM	support vector machine
TCGA	The Cancer Genome Atlas
TDM	Training Distribution Machine
UPC	Universal exPression Code
WM	Watermelon Multisection
XPC	Cross-Platform Comparison
XPN	Cross-Platform Normalization

References

1. Lashkari, D.A.; DeRisi, J.L.; McCusker, J.H.; Namath, A.F.; Gentile, C.; Hwang, S.Y.; Brown, P.O.; Davis, R.W. Yeast Microarrays for Genome Wide Parallel Genetic and Gene Expression Analysis. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 13057–13062. [CrossRef] [PubMed]
2. King, H.C.; Sinha, A.A. Gene Expression Profile Analysis by DNA Microarrays: Promise and Pitfalls. *JAMA* **2001**, *286*, 2280. [CrossRef]
3. Bednár, M. DNA Microarray Technology and Application. *Med. Sci. Monit.* **2000**, *6*, 796–800. [PubMed]
4. Rew, D.A. DNA Microarray Technology in Cancer Research. *Eur. J. Surg. Oncol.* **2001**, *27*, 504–508. [CrossRef] [PubMed]
5. Edgar, R.; Domrachev, M.; Lash, A.E. Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository. *Nucleic Acids Res.* **2002**, *30*, 207–210. [CrossRef] [PubMed]

6. Brazma, A.; Hingamp, P.; Quackenbush, J.; Sherlock, G.; Spellman, P.; Stoeckert, C.; Aach, J.; Ansorge, W.; Ball, C.A.; Causton, H.C.; et al. Minimum Information about a Microarray Experiment (MIAME)-toward Standards for Microarray Data. *Nat. Genet.* **2001**, *29*, 365–371. [CrossRef]
7. Rocca-Serra, P.; Brazma, A.; Parkinson, H.; Sarkans, U.; Shojatalab, M.; Contrino, S.; Vilo, J.; Abeygunawardena, N.; Mukherjee, G.; Holloway, E.; et al. ArrayExpress: A Public Database of Gene Expression Data at EBI. *Comptes Rendus Biol.* **2003**, *326*, 1075–1078. [CrossRef] [PubMed]
8. Parkinson, H.; Kapushesky, M.; Shojatalab, M.; Abeygunawardena, N.; Coulson, R.; Farne, A.; Holloway, E.; Kolesnykov, N.; Lilja, P.; Lukk, M.; et al. ArrayExpress—a Public Database of Microarray Experiments and Gene Expression Profiles. *Nucleic Acids Res.* **2007**, *35*, D747–D750. [CrossRef]
9. The Cancer Genome Atlas Research Network. Comprehensive Genomic Characterization Defines Human Glioblastoma Genes and Core Pathways. *Nature* **2008**, *455*, 1061–1068. [CrossRef] [PubMed]
10. Tomczak, K.; Czerwińska, P.; Wiznerowicz, M. The Cancer Genome Atlas (TCGA): An Immeasurable Source of Knowledge. *Contemp. Oncol.* **2015**, *19*, A68–A77. [CrossRef] [PubMed]
11. Lonsdale, J.; Thomas, J.; Salvatore, M.; Phillips, R.; Lo, E.; Shad, S.; Hasz, R.; Walters, G.; Garcia, F.; Young, N. The Genotype-Tissue Expression (GTEx) Project. *Nature Genetics* **2013**, *45*, 580–585. [CrossRef] [PubMed]
12. The GTEx Consortium; Ardlie, K.G.; Deluca, D.S.; Segrè, A.V.; Sullivan, T.J.; Young, T.R.; Gelfand, E.T.; Trowbridge, C.A.; Maller, J.B.; Tukiainen, T.; et al. The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans. *Science* **2015**, *348*, 648–660. [CrossRef]
13. Suntsova, M.; Gaifullin, N.; Allina, D.; Reshetun, A.; Li, X.; Mendeleeva, L.; Surin, V.; Sergeeva, A.; Spirin, P.; Prassolov, V.; et al. Atlas of RNA Sequencing Profiles for Normal Human Tissues. *Sci. Data* **2019**, *6*, 36. [CrossRef]
14. Yang, W.; Soares, J.; Greninger, P.; Edelman, E.J.; Lightfoot, H.; Forbes, S.; Bindal, N.; Beare, D.; Smith, J.A.; Thompson, I.R.; et al. Genomics of Drug Sensitivity in Cancer (GDSC): A Resource for Therapeutic Biomarker Discovery in Cancer Cells. *Nucleic Acids Res.* **2013**, *41*, D955–D961. [CrossRef] [PubMed]
15. Chen, Y.; Li, Y.; Narayan, R.; Subramanian, A.; Xie, X. Gene Expression Inference with Deep Learning. *Bioinformatics* **2016**, *32*, 1832–1839. [CrossRef] [PubMed]
16. Subramanian, A.; Kuehn, H.; Gould, J.; Tamayo, P.; Mesirov, J.P. GSEA-P: A Desktop Application for Gene Set Enrichment Analysis. *Bioinformatics* **2007**, *23*, 3251–3253. [CrossRef]
17. Liang, P. MAQC Papers over the Cracks. *Nat. Biotechnol.* **2007**, *25*, 27–28, author reply 28–29. [CrossRef] [PubMed]
18. Chen, J.J.; Hsueh, H.-M.; Delongchamp, R.R.; Lin, C.-J.; Tsai, C.-A. Reproducibility of Microarray Data: A Further Analysis of Microarray Quality Control (MAQC) Data. *BMC Bioinform.* **2007**, *8*, 412. [CrossRef] [PubMed]
19. Shi, L.; Shi, L.; Reid, L.H.; Jones, W.D.; Shippy, R.; Warrington, J.A.; Baker, S.C.; Collins, P.J.; de Longueville, F.; Kawasaki, E.S.; et al. The MicroArray Quality Control (MAQC) Project Shows Inter- and Intraplatform Reproducibility of Gene Expression Measurements. *Nature Biotechnol.* **2006**, *24*, 1151–1161. [CrossRef]
20. Mane, S.P.; Evans, C.; Cooper, K.L.; Crasta, O.R.; Folkerts, O.; Hutchison, S.K.; Harkins, T.T.; Thierry-Mieg, D.; Thierry-Mieg, J.; Jensen, R.V. Transcriptome Sequencing of the Microarray Quality Control (MAQC) RNA Reference Samples Using next Generation Sequencing. *BMC Genom.* **2009**, *10*, 264. [CrossRef] [PubMed]
21. Wen, Z.; Wang, C.; Shi, Q.; Huang, Y.; Su, Z.; Hong, H.; Tong, W.; Shi, L. Evaluation of Gene Expression Data Generated from Expired Affymetrix GeneChip® Microarrays Using MAQC Reference RNA Samples. *BMC Bioinform.* **2010**, *11*, S10. [CrossRef]
22. Stelpflug, S.C.; Sekhon, R.S.; Vaillancourt, B.; Hirsch, C.N.; Buell, C.R.; Leon, N.; Kaeppler, S.M. An Expanded Maize Gene Expression Atlas Based on RNA Sequencing and Its Use to Explore Root Development. *Plant Genome* **2016**, *9*, 27898762. [CrossRef] [PubMed]
23. Han, S.; Van Treuren, W.; Fischer, C.R.; Merrill, B.D.; DeFelice, B.C.; Sanchez, J.M.; Higginbottom, S.K.; Guthrie, L.; Fall, L.A.; Dodd, D.; et al. A Metabolomics Pipeline for the Mechanistic Interrogation of the Gut Microbiome. *Nature* **2021**, *595*, 415–420. [CrossRef] [PubMed]
24. Tanaka, N.; Takahara, A.; Hagio, T.; Nishiko, R.; Kanayama, J.; Gotoh, O.; Mori, S. Sequencing Artifacts Derived from a Library Preparation Method Using Enzymatic Fragmentation. *PLoS ONE* **2020**, *15*, e0227427. [CrossRef]
25. Demetrashvili, N.; Kron, K.; Pethe, V.; Bapat, B.; Briollais, L. How to Deal with Batch Effect in Sequential Microarray Experiments? *Mol. Inform.* **2010**, *29*, 387–393. [CrossRef]
26. Lazar, C.; Meganck, S.; Taminau, J.; Steenhoff, D.; Coletta, A.; Molter, C.; Weiss-Solís, D.Y.; Duque, R.; Bersini, H.; Nowé, A. Batch Effect Removal Methods for Microarray Gene Expression Data Integration: A Survey. *Brief. Bioinform.* **2013**, *14*, 469–490. [CrossRef] [PubMed]
27. Xu, J.; Gong, B.; Wu, L.; Thakkar, S.; Hong, H.; Tong, W. Comprehensive Assessments of RNA-Seq by the SEQC Consortium: FDA-Led Efforts Advance Precision Medicine. *Pharmaceutics* **2016**, *8*, 8. [CrossRef] [PubMed]
28. Bolstad, B.M.; Irizarry, R.A.; Astrand, M.; Speed, T.P. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias. *Bioinformatics* **2003**, *19*, 185–193. [CrossRef]
29. Shabalin, A.A.; Tjelmeland, H.; Fan, C.; Perou, C.M.; Nobel, A.B. Merging Two Gene-Expression Studies via Cross-Platform Normalization. *Bioinformatics* **2008**, *24*, 1154–1160. [CrossRef]
30. Deshwar, A.G.; Morris, Q. PLIDA: Cross-Platform Gene Expression Normalization Using Perturbed Topic Models. *Bioinformatics* **2014**, *30*, 956–961. [CrossRef]

31. Rudy, J.; Valafar, F. Empirical Comparison of Cross-Platform Normalization Methods for Gene Expression Data. *BMC Bioinform.* **2011**, *12*, 467. [CrossRef] [PubMed]
32. Borisov, N.; Shabalina, I.; Tkachev, V.; Sorokin, M.; Garazha, A.; Pulin, A.; Eremin, I.I.; Buzdin, A. Shambhala: A Platform-Agnostic Data Harmonizer for Gene Expression Data. *BMC Bioinform.* **2019**, *20*, 66. [CrossRef] [PubMed]
33. Borisov, N.; Sorokin, M.; Zolotovskaya, M.; Borisov, C.; Buzdin, A. Shambhala-2: A Protocol for Uniformly Shaped Harmonization of Gene Expression Profiles of Various Formats. *Current Protocols* **2022**, *2*, e444. [CrossRef] [PubMed]
34. Junet, V.; Farrés, J.; Mas, J.M.; Daura, X. CuBlock: A Cross-Platform Normalization Method for Gene-Expression Microarrays. *Bioinformatics* **2021**, *37*, 2365–2373. [CrossRef] [PubMed]
35. Carter, S.L.; Eklund, A.C.; Mecham, B.H.; Kohane, I.S.; Szallasi, Z. Redefinition of Affymetrix Probe Sets by Sequence Overlap with CDNA Microarray Probes Reduces Cross-Platform Inconsistencies in Cancer-Associated Gene Expression Measurements. *BMC Bioinform.* **2005**, *6*, 107. [CrossRef]
36. Warnat, P.; Eils, R.; Brors, B. Cross-Platform Analysis of Cancer Microarray Data Improves Gene Expression Based Classification of Phenotypes. *BMC Bioinform.* **2005**, *6*, 265. [CrossRef]
37. Martinez, R.; Pasquier, N.; Pasquier, C. GenMiner: Mining Non-Redundant Association Rules from Integrated Gene Expression Data and Annotations. *Bioinformatics* **2008**, *24*, 2643–2644. [CrossRef]
38. Jiang, H.; Deng, Y.; Chen, H.-S.; Tao, L.; Sha, Q.; Chen, J.; Tsai, C.-J.; Zhang, S. Joint Analysis of Two Microarray Gene-Expression Data Sets to Select Lung Adenocarcinoma Marker Genes. *BMC Bioinform.* **2004**, *5*, 81. [CrossRef]
39. Johnson, W.E.; Li, C.; Rabinovic, A. Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods. *Biostatistics* **2007**, *8*, 118–127. [CrossRef]
40. Huang, H.; Lu, X.; Liu, Y.; Haaland, P.; Marron, J.S. R/DWD: Distance-Weighted Discrimination for Classification, Visualization and Batch Adjustment. *Bioinformatics* **2012**, *28*, 1182–1183. [CrossRef]
41. Marron, J.S.; Todd, M.J.; Ahn, J. Distance-Weighted Discrimination. *J. Am. Stat. Assoc.* **2007**, *102*, 1267–1271. [CrossRef]
42. Benito, M.; Parker, J.; Du, Q.; Wu, J.; Xiang, D.; Perou, C.M.; Marron, J.S. Adjustment of Systematic Microarray Data Biases. *Bioinformatics* **2004**, *20*, 105–114. [CrossRef] [PubMed]
43. Xia, X.-Q.; McClelland, M.; Porwollik, S.; Song, W.; Cong, X.; Wang, Y. WebArrayDB: Cross-Platform Microarray Data Analysis and Public Data Repository. *Bioinformatics* **2009**, *25*, 2425–2429. [CrossRef] [PubMed]
44. Chu, Y.; Corey, D.R. RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation. *Nucleic Acid. Ther.* **2012**, *22*, 271–274. [CrossRef] [PubMed]
45. Nagalakshmi, U.; Wang, Z.; Waern, K.; Shou, C.; Raha, D.; Gerstein, M.; Snyder, M. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* **2008**, *320*, 1344–1349. [CrossRef]
46. Maher, C.A.; Kumar-Sinha, C.; Cao, X.; Kalyana-Sundaram, S.; Han, B.; Jing, X.; Sam, L.; Barrette, T.; Palanisamy, N.; Chinnaiyan, A.M. Transcriptome Sequencing to Detect Gene Fusions in Cancer. *Nature* **2009**, *458*, 97–101. [CrossRef]
47. Ingolia, N.T.; Brar, G.A.; Rouskin, S.; McGeachy, A.M.; Weissman, J.S. The Ribosome Profiling Strategy for Monitoring Translation in Vivo by Deep Sequencing of Ribosome-Protected MRNA Fragments. *Nat. Protoc.* **2012**, *7*, 1534–1550. [CrossRef]
48. Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: A Revolutionary Tool for Transcriptomics. *Nat. Rev. Genet.* **2009**, *10*, 57–63. [CrossRef]
49. Korir, P.K.; Geeleher, P.; Seoighe, C. Seq-Ing Improved Gene Expression Estimates from Microarrays Using Machine Learning. *BMC Bioinform.* **2015**, *16*, 286. [CrossRef]
50. Taylor, K.C.; Evans, D.S.; Edwards, D.R.V.; Edwards, T.L.; Sofer, T.; Li, G.; Liu, Y.; Franceschini, N.; Jackson, R.D.; Giri, A.; et al. A Genome-Wide Association Study Meta-Analysis of Clinical Fracture in 10,012 African American Women. *Bone Rep.* **2016**, *5*, 233–242. [CrossRef]
51. Hollern, D.P.; Xu, N.; Thennavan, A.; Glodowski, C.; Garcia-Recio, S.; Mott, K.R.; He, X.; Garay, J.P.; Carey-Ewend, K.; Marron, D.; et al. B Cells and T Follicular Helper Cells Mediate Response to Checkpoint Inhibitors in High Mutation Burden Mouse Models of Breast Cancer. *Cell* **2019**, *179*, 1191–1206.e21. [CrossRef] [PubMed]
52. Thind, A.S.; Monga, I.; Thakur, P.K.; Kumari, P.; Dindhoria, K.; Krzak, M.; Ranson, M.; Ashford, B. Demystifying Emerging Bulk RNA-Seq Applications: The Application and Utility of Bioinformatic Methodology. *Brief. Bioinform.* **2021**, *22*, bbab259. [CrossRef]
53. Vellichirammal, N.N.; Albahrani, A.; Li, Y.; Guda, C. Identification of Fusion Transcripts from Unaligned RNA-Seq Reads Using ChimeRScope. In *Chimeric RNA*; Li, H., Elfman, J., Eds.; Methods in Molecular Biology; Springer: New York, NY, USA, 2020; Volume 2079, pp. 13–25. ISBN 978-1-4939-9903-3.
54. Kekeeva, T.; Tanas, A.; Kanygina, A.; Alexeev, D.; Shikeeva, A.; Zavalishina, L.; Andreeva, Y.; Frank, G.A.; Zaletaev, D. Novel Fusion Transcripts in Bladder Cancer Identified by RNA-Seq. *Cancer Lett.* **2016**, *374*, 224–228. [CrossRef]
55. Gu, J.; Chukhman, M.; Lu, Y.; Liu, C.; Liu, S.; Lu, H. RNA-Seq Based Transcription Characterization of Fusion Breakpoints as a Potential Estimator for Its Oncogenic Potential. *BioMed. Res. Int.* **2017**, *2017*, 9829175. [CrossRef] [PubMed]
56. Schmidt, B.M.; Davidson, N.M.; Hawkins, A.D.K.; Bartolo, R.; Majewski, I.J.; Ekert, P.G.; Oshlack, A. Clinker: Visualizing Fusion Genes Detected in RNA-Seq Data. *GigaScience* **2018**, *7*, giy079. [CrossRef] [PubMed]
57. Borisov, N.; Sorokin, M.; Tkachev, V.; Garazha, A.; Buzdin, A. Cancer Gene Expression Profiles Associated with Clinical Outcomes to Chemotherapy Treatments. *BMC Med. Genom.* **2020**, *13*, 111. [CrossRef]
58. Anders, S.; Huber, W. Differential Expression Analysis for Sequence Count Data. *Genome Biol.* **2010**, *11*, R106. [CrossRef]

59. Love, M.I.; Huber, W.; Anders, S. Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [CrossRef]
60. Varet, H.; Brillet-Guéguen, L.; Coppée, J.-Y.; Dillies, M.-A. SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data. *PLoS ONE* **2016**, *11*, e0157022. [CrossRef]
61. Maza, E. In Papyro Comparison of TMM (EdgeR), RLE (DESeq2), and MRN Normalization Methods for a Simple Two-Conditions-Without-Replicates RNA-Seq Experimental Design. *Front. Genet.* **2016**, *7*, 164. [CrossRef]
62. Thompson, J.A.; Tan, J.; Greene, C.S. Cross-Platform Normalization of Microarray and RNA-Seq Data for Machine Learning Applications. *PeerJ* **2016**, *4*, e1621. [CrossRef]
63. Piccolo, S.R.; Withers, M.R.; Francis, O.E.; Bild, A.H.; Johnson, W.E. Multiplatform Single-Sample Estimates of Transcriptional Activation. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 17778–17783. [CrossRef] [PubMed]
64. Franks, J.M.; Cai, G.; Whitfield, M.L. Feature Specific Quantile Normalization Enables Cross-Platform Classification of Molecular Subtypes Using Gene Expression Data. *Bioinformatics* **2018**, *34*, 1868–1874. [CrossRef] [PubMed]
65. Fauteux, F.; Surendra, A.; McComb, S.; Pan, Y.; Hill, J.J. Identification of Transcriptional Subtypes in Lung Adenocarcinoma and Squamous Cell Carcinoma through Integrative Analysis of Microarray and RNA Sequencing Data. *Sci. Rep.* **2021**, *11*, 8709. [CrossRef] [PubMed]
66. Zhang, S.; Shao, J.; Yu, D.; Qiu, X.; Zhang, J. MatchMixeR: A Cross-Platform Normalization Method for Gene Expression Data Integration. *Bioinformatics* **2020**, *36*, 2486–2491. [CrossRef] [PubMed]
67. Maleknia, S.; Salehi, Z.; Rezaei Tabar, V.; Sharifi-Zarchi, A.; Kavousi, K. An Integrative Bayesian Network Approach to Highlight Key Drivers in Systemic Lupus Erythematosus. *Arthritis Res. Ther.* **2020**, *22*, 156. [CrossRef]
68. Tang, K.; Ji, X.; Zhou, M.; Deng, Z.; Huang, Y.; Zheng, G.; Cao, Z. Rank-in: Enabling Integrative Analysis across Microarray and RNA-Seq for Cancer. *Nucleic Acids Res.* **2021**, *49*, e99. [CrossRef] [PubMed]
69. Huang, H.-H.; Rao, H.; Miao, R.; Liang, Y. A Novel Meta-Analysis Based on Data Augmentation and Elastic Data Shared Lasso Regularization for Gene Expression. *BMC Bioinform.* **2022**, *23*, 353. [CrossRef]
70. Dinalankara, W.; Ke, Q.; Xu, Y.; Ji, L.; Pagane, N.; Lien, A.; Matam, T.; Fertig, E.J.; Price, N.D.; Younes, L.; et al. Digitizing Omics Profiles by Divergence from a Baseline. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 4545–4552. [CrossRef] [PubMed]
71. Lê Cao, K.-A.; Rohart, F.; McHugh, L.; Korn, O.; Wells, C.A. YuGene: A Simple Approach to Scale Gene Expression Data Derived from Different Platforms for Integrated Analyses. *Genomics* **2014**, *103*, 239–251. [CrossRef] [PubMed]
72. Nguyen, T.N.; Nguyen, H.Q.; Le, D.-H. Unveiling Prognostics Biomarkers of Tyrosine Metabolism Reprogramming in Liver Cancer by Cross-Platform Gene Expression Analyses. *PLoS ONE* **2020**, *15*, e0229276. [CrossRef] [PubMed]
73. Ou-Yang, L.; Zhang, X.-F.; Wu, M.; Li, X.-L. Node-Based Learning of Differential Networks from Multi-Platform Gene Expression Data. *Methods* **2017**, *129*, 41–49. [CrossRef]
74. Piccolo, S.R.; Sun, Y.; Campbell, J.D.; Lenburg, M.E.; Bild, A.H.; Johnson, W.E. A Single-Sample Microarray Normalization Method to Facilitate Personalized-Medicine Workflows. *Genomics* **2012**, *100*, 337–344. [CrossRef] [PubMed]
75. Irizarry, R.A.; Hobbs, B.; Collin, F.; Beazer-Barclay, Y.D.; Antonellis, K.J.; Scherf, U.; Speed, T.P. Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics* **2003**, *4*, 249–264. [CrossRef]
76. Wu, Z.; Irizarry, R.A.; Gentleman, R.; Martinez-Murillo, F.; Spencer, F. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *J. Am. Stat. Assoc.* **2004**, *99*, 909–917. [CrossRef]
77. Li, C.; Wong, W.H. Model-Based Analysis of Oligonucleotide Arrays: Expression Index Computation and Outlier Detection. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 31–36. [CrossRef]
78. Therneau, T.M.; Ballman, K.V. What Does PLIER Really Do? *Cancer Inform* **2008**, *6*, 117693510800600. [CrossRef]
79. McCall, M.N.; Bolstad, B.M.; Irizarry, R.A. Frozen Robust Multiarray Analysis (FRMA). *Biostatistics* **2010**, *11*, 242–253. [CrossRef] [PubMed]
80. McCall, M.N.; Uppal, K.; Jaffee, H.A.; Zilliox, M.J.; Irizarry, R.A. The Gene Expression Barcode: Leveraging Public Data Repositories to Begin Cataloging the Human and Murine Transcriptomes. *Nucleic Acids Res.* **2011**, *39*, D1011–D1015. [CrossRef] [PubMed]
81. McCall, M.N.; Murakami, P.N.; Lukk, M.; Huber, W.; Irizarry, R.A. Assessing Affymetrix GeneChip Microarray Quality. *BMC Bioinform.* **2011**, *12*, 137. [CrossRef] [PubMed]
82. McCall, M.N.; Jaffee, H.A.; Irizarry, R.A. FRMA ST: Frozen Robust Multiarray Analysis for Affymetrix Exon and Gene ST Arrays. *Bioinformatics* **2012**, *28*, 3153–3154. [CrossRef] [PubMed]
83. Zhang, L.; Cham, J.; Cooley, J.; He, T.; Hagihara, K.; Yang, H.; Fan, F.; Cheung, A.; Thompson, D.; Kerns, B.J.; et al. Cross-Platform Comparison of Immune-Related Gene Expression to Assess Intratumor Immune Responses Following Cancer Immunotherapy. *J. Immunol. Methods* **2021**, *494*, 113041. [CrossRef] [PubMed]
84. Lee, J.S.; Nair, N.U.; Dinstag, G.; Chapman, L.; Chung, Y.; Wang, K.; Sinha, S.; Cha, H.; Kim, D.; Schperberg, A.V.; et al. Synthetic Lethality-Mediated Precision Oncology via the Tumor Transcriptome. *Cell* **2021**, *184*, 2487–2502.e13. [CrossRef] [PubMed]
85. Borisov, N.; Sorokin, M.; Garazha, A.; Buzdin, A. Quantitation of Molecular Pathway Activation Using RNA Sequencing Data. In *Nucleic Acid Detection and Structural Investigations*; Astakhova, K., Bukhari, S.A., Eds.; Springer: New York, NY, USA, 2020; Volume 2063, pp. 189–206. ISBN 978-1-07-160137-2.

86. Poddubskaya, E.; Buzdin, A.; Garazha, A.; Sorokin, M.; Glusker, A.; Aleshin, A.; Allina, D.; Moiseev, A.; Sekacheva, M.; Suntsova, M.; et al. Oncobox, Gene Expression-Based Second Opinion System for Predicting Response to Treatment in Advanced Solid Tumors. *J. Clin. Oncol.* **2019**, *37*, e13143. [CrossRef]
87. Tkachev, V.; Sorokin, M.; Garazha, A.; Borisov, N.; Buzdin, A. Oncobox Method for Scoring Efficiencies of Anticancer Drugs Based on Gene Expression Data. In *Nucleic Acid Detection and Structural Investigations*; Astakhova, K., Bukhari, S.A., Eds.; Springer US: New York, NY, USA, 2020; Volume 2063, pp. 235–255. ISBN 978-1-07-160137-2.
88. Tkachev, V.; Sorokin, M.; Mescheryakov, A.; Simonov, A.; Garazha, A.; Buzdin, A.; Muchnik, I.; Borisov, N. FLOating-Window Projective Separator (FloWPS): A Data Trimming Tool for Support Vector Machines (SVM) to Improve Robustness of the Classifier. *Front. Genet.* **2019**, *9*, 717. [CrossRef]
89. Tkachev, V.; Sorokin, M.; Borisov, C.; Garazha, A.; Buzdin, A.; Borisov, N. Flexible Data Trimming Improves Performance of Global Machine Learning Methods in Omics-Based Personalized Oncology. *Int. J. Mol. Sci.* **2020**, *21*, 713. [CrossRef] [PubMed]
90. Turki, T.; Wang, J.T.L. Clinical Intelligence: New Machine Learning Techniques for Predicting Clinical Drug Response. *Comput. Biol. Med.* **2019**, *107*, 302–322. [CrossRef]
91. Bolstad, B. Preprocessing and Normalization for Affymetrix GeneChip Expression Microarrays. In *Methods in Microarray Normalization*; Stafford, P., Ed.; Drug Discovery Series; CRC Press: Boca Raton, FL, USA, 2008; Volume 0, pp. 41–59. ISBN 978-1-4200-5278-7.
92. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1–22. [CrossRef]
93. Vapnik, V.; Chapelle, O. Bounds on Error Expectation for Support Vector Machines. *Neural Comput.* **2000**, *12*, 2013–2036. [CrossRef]
94. Tibshirani, R.; Hastie, T.; Narasimhan, B.; Chu, G. Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 6567–6572. [CrossRef] [PubMed]
95. Zolotovskaia, M.A.; Sorokin, M.I.; Petrov, I.V.; Poddubskaya, E.V.; Moiseev, A.A.; Sekacheva, M.I.; Borisov, N.M.; Tkachev, V.S.; Garazha, A.V.; Kaprin, A.D.; et al. Disparity between Inter-Patient Molecular Heterogeneity and Repertoires of Target Drugs Used for Different Types of Cancer in Clinical Oncology. *Int. J. Mol. Sci.* **2020**, *21*, 1580. [CrossRef] [PubMed]
96. Huang, E.; Cheng, S.H.; Dressman, H.; Pittman, J.; Tsou, M.H.; Horng, C.F.; Bild, A.; Iversen, E.S.; Liao, M.; Chen, C.M.; et al. Gene Expression Predictors of Breast Cancer Outcomes. *Lancet* **2003**, *361*, 1590–1596. [CrossRef]
97. Hu, Z.; Fan, C.; Oh, D.S.; Marron, J.; He, X.; Qaqish, B.F.; Livasy, C.; Carey, L.A.; Reynolds, E.; Dressler, L.; et al. The Molecular Portraits of Breast Tumors Are Conserved across Microarray Platforms. *BMC Genom.* **2006**, *7*, 96. [CrossRef] [PubMed]
98. van't Veer, L.J.; Dai, H.; van de Vijver, M.J.; He, Y.D.; Hart, A.A.M.; Mao, M.; Peterse, H.L.; van der Kooy, K.; Marton, M.J.; Witteveen, A.T.; et al. Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer. *Nature* **2002**, *415*, 530–536. [CrossRef] [PubMed]
99. Wang, Y.; Xia, X.-Q.; Jia, Z.; Sawyers, A.; Yao, H.; Wang-Rodriquez, J.; Mercola, D.; McClelland, M. In Silico Estimates of Tissue Components in Surgical Samples Based on Expression Profiling Data. *Cancer Res.* **2010**, *70*, 6448–6455. [CrossRef] [PubMed]
100. Jia, Z.; Wang, Y.; Sawyers, A.; Yao, H.; Rahmatpanah, F.; Xia, X.-Q.; Xu, Q.; Pio, R.; Turan, T.; Koziol, J.A.; et al. Diagnosis of Prostate Cancer Using Differentially Expressed Genes in Stroma. *Cancer Res.* **2011**, *71*, 2476–2487. [CrossRef]
101. Desmedt, C.; Piette, F.; Loi, S.; Wang, Y.; Lallemand, F.; Haibe-Kains, B.; Viale, G.; Delorenzi, M.; Zhang, Y.; d'Assignies, M.S.; et al. Strong Time Dependence of the 76-Gene Prognostic Signature for Node-Negative Breast Cancer Patients in the TRANSBIG Multicenter Independent Validation Series. *Clin. Cancer Res.* **2007**, *13*, 3207–3214. [CrossRef] [PubMed]
102. Chicco, D.; Warrens, M.J.; Jurman, G. The Coefficient of Determination R-Squared Is More Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation. *PeerJ Comput. Sci.* **2021**, *7*, e623. [CrossRef] [PubMed]
103. Chicco, D. Ten Quick Tips for Machine Learning in Computational Biology. *BioData Min.* **2017**, *10*, 35. [CrossRef]
104. Chicco, D.; Jurman, G. The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef] [PubMed]
105. Meng, Q.; Catchpoole, D.; Skillicorn, D.; Kennedy, P.J. DBNorm: Normalizing High-Density Oligonucleotide Microarray Data Based on Distributions. *BMC Bioinform.* **2017**, *18*, 527. [CrossRef] [PubMed]