Transcriptomic Harmonization

Subjects: Biochemistry & Molecular Biology

Contributor: Nicolas Borisov , Anton Buzdin

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. 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.

gene expression transcriptional profiles RNA sequencing microarray hybridization

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][2][3][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][18][19][20][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][18][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][18][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][18][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][30][31][32][33][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 entry: 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][41][42], Cross-Platform Normalization (XPN) [29][31], Gene Quantiles (GQ) [43], and PLatform-Independent Latent Dirichlet Allocation (PLIDA) [30].

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, researchers 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) [44], Quantile Discretization (QD) [36], Gene Quantiles (GQ) [43], Normalized Discretization (NorDi) [37], Distribution Transformation (DisTran) [36][38], Median Rank Scores (MRS) [36], YuGene [45], and Rank-in [46];

   (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) [47] 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 [48]/DESeq2 [49][50][51];

   (b) Those using log-normal distribution with either covariance analysis [52], or with conditional/Bayesian models, as for the methods Universal exPression Code (UPC) [53][54], Empirical Bayes (ComBat) [39], Robust Microarray Analysis (RMA) [55], GeneChip Robust Multiarray Analysis (gcRMA) [56], Model-Based Expression Indices (MBEI) [57], Probe Logarithmic Intensity ERror (PLIER) estimation [58], frozen Robust Microarray Analysis (fRMA) [59][60][61][62], MatchMixeR (MM) [63], Cross-Platform Comparison (XPC) [64];

   (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 [65][66];

   (e) Those using the Least Absolute Shrinkage and Selection Operator (LASSO) regression models [67];

(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][41][42];

(b) Those using quantile-based regression models for data transfer from source to target datasets, such as Training Distribution Machine (TDM) [68].

Another important aspect that must be considered in this entry 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 [69], molecular pathway activation levels [70], algorithmically deduced cancer drug efficiency scores [71][72], and different ML models [73][74][75].

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) [59][60][61][62], robust Quantile Normalization [76], Training Distribution Machine (TDM) [68], and Universal exPression Code (UPC) [53].

More recently, researchers 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.

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][30][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 [77], used in [30];

   (b) SVM [78], used in [29][31];

   (c) Nearest shrunken centroids Prediction Analysis for Microarrays (PAM) [79], 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][18][19] and Sequencing Quality Control (SEQC) [27] projects mentioned above in this entry.

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][31][32][33][34]. However, this could only support a manual qualitative assessment without precise quantitative interrogation of the complex class distribution profiles.

Researchers recently proposed a new metric for the algorithmic cluster analysis of dendrograms [33][80] 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 [80]. This accumulated and normalized IG constitutes the WM metric for a given dendrogram, and a given set of classes under analysis. Consequently, the ratio R =W_{MS}/W_{MP}, 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 .

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.
2. King, H.C.; Sinha, A.A. Gene Expression Profile Analysis by DNA Microarrays: Promise and Pitfalls. JAMA 2001, 286, 2280.
3. Bednár, M. DNA Microarray Technology and Application. Med. Sci. Monit. 2000, 6, 796–800.
4. Rew, D.A. DNA Microarray Technology in Cancer Research. Eur. J. Surg. Oncol. 2001, 27, 504–508.
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.
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.
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.
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.
9. The Cancer Genome Atlas Research Network. Comprehensive Genomic Characterization Defines Human Glioblastoma Genes and Core Pathways. Nature 2008, 455, 1061–1068.
10. Tomczak, K.; Czerwińska, P.; Wiznerowicz, M. The Cancer Genome Atlas (TCGA): An Immeasurable Source of Knowledge. Contemp. Oncol. 2015, 19, A68–A77.
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.

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.
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.
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.
15. Chen, Y.; Li, Y.; Narayan, R.; Subramanian, A.; Xie, X. Gene Expression Inference with Deep Learning. Bioinformatics 2016, 32, 1832–1839.
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.
17. Liang, P. MAQC Papers over the Cracks. Nat. Biotechnol. 2007, 25, 27–28, author reply 28–29.
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.
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.
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.
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.
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.
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.

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.
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.
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.
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.
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.
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.
30. Deshwar, A.G.; Morris, Q. PLIDA: Cross-Platform Gene Expression Normalization Using Perturbed Topic Models. Bioinformatics 2014, 30, 956–961.
31. Rudy, J.; Valafar, F. Empirical Comparison of Cross-Platform Normalization Methods for Gene Expression Data. BMC Bioinform. 2011, 12, 467.
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.
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.
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.
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.

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.
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.
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.
39. Johnson, W.E.; Li, C.; Rabinovic, A. Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods. Biostatistics 2007, 8, 118–127.
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.
41. Marron, J.S.; Todd, M.J.; Ahn, J. Distance-Weighted Discrimination. J. Am. Stat. Assoc. 2007, 102, 1267–1271.
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.
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.
44. 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.
45. 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.
46. 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.
47. 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.
48. Anders, S.; Huber, W. Differential Expression Analysis for Sequence Count Data. Genome Biol. 2010, 11, R106.
49. 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.

50. 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.
51. 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.
52. 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.
53. 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.
54. 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.
55. 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.
56. 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.
57. 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.
58. Therneau, T.M.; Ballman, K.V. What Does PLIER Really Do? Cancer Inform 2008, 6, 117693510800600.
59. McCall, M.N.; Bolstad, B.M.; Irizarry, R.A. Frozen Robust Multiarray Analysis (FRMA). Biostatistics 2010, 11, 242–253.
60. 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.
61. McCall, M.N.; Murakami, P.N.; Lukk, M.; Huber, W.; Irizarry, R.A. Assessing Affymetrix GeneChip Microarray Quality. BMC Bioinform. 2011, 12, 137.
62. 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.
63. 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.

64. 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.

65. 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.

66. 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.

67. 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.

68. 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.

69. 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.

70. 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.

71. 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.

72. 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.

73. 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.

74. 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.

75. Turki, T.; Wang, J.T.L. Clinical Intelligence: New Machine Learning Techniques for Predicting Clinical Drug Response. Comput. Biol. Med. 2019, 107, 302–322.
76. 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.
77. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. 2010, 33, 1–22.
78. Vapnik, V.; Chapelle, O. Bounds on Error Expectation for Support Vector Machines. Neural Comput. 2000, 12, 2013–2036.
79. 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.
80. 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.