Cancer gene expression profiles associated with clinical outcomes to chemotherapy treatments

Nicolas Borisov1,2*, Maxim Sorokin1,3, Victor Tkachev1, Andrew Garazha1 and Anton Buzdin1,2,3,4

From 11th International Young Scientists School "Systems Biology and Bioinformatics" – SBB-2019
Novosibirsk, Russia. 24-28 June 2019

Abstract

Background: Machine learning (ML) methods still have limited applicability in personalized oncology due to low numbers of available clinically annotated molecular profiles. This doesn’t allow sufficient training of ML classifiers that could be used for improving molecular diagnostics.

Methods: We reviewed published datasets of high throughput gene expression profiles corresponding to cancer patients with known responses on chemotherapy treatments. We browsed Gene Expression Omnibus (GEO), The Cancer Genome Atlas (TCGA) and Tumor Alterations Relevant for GENomics-driven Therapy (TARGET) repositories.

Results: We identified data collections suitable to build ML models for predicting responses on certain chemotherapeutic schemes. We identified 26 datasets, ranging from 41 till 508 cases per dataset. All the datasets identified were checked for ML applicability and robustness with leave-one-out cross validation. Twenty-three datasets were found suitable for using ML that had balanced numbers of treatment responder and non-responder cases.

Conclusions: We collected a database of gene expression profiles associated with clinical responses on chemotherapy for 2786 individual cancer cases. Among them seven datasets included RNA sequencing data (for 645 cases) and the others – microarray expression profiles. The cases represented breast cancer, lung cancer, low-grade glioma, endothermal carcinoma, multiple myeloma, adult leukemia, pediatric leukemia and kidney tumors. Chemotherapeutics included taxanes, bortezomib, vincristine, trastuzumab, letrozole, tipifarnib, temozolomide, busulfan and cyclophosphamide.

Keywords: Machine learning, Transcriptomics, Gene expression, RNA sequencing, Microarrays, Molecular diagnostics, Biomarkers detection, Cancer, Clinical oncology, Personalized medicine, Chemotherapy
**Background**

Personalized approach provides important advantages in clinical oncology in terms of improved patient survival and lower drug toxicities [1, 2]. However, so far it can only cover a minor fraction of cancer patients [3, 4] due to lack of robust prognostic biomarkers for most of the treatments [5]. The proportion of patients eligible for personalized oncology slightly grows. For example, the percentage of US patients with cancer estimated to benefit from personalized prescriptions of targeted therapeutics was only 0.7% in 2006, and it had increased to ~5% in 2018 [4]. However, this progress could be more significant if more companion diagnostic tests would be available for the standardly used cancer drugs. In this regard, gene expression data, either obtained by RNA sequencing [1] or using microarrays [6], frequently provide an advantage over genomic tests. Several trials and clinical case reports were published recently evidencing high efficiency of gene expression-based prescriptions of cancer chemotherapeutics. Cancer gene expression data can be used per se or can be normalized on the available profiles of healthy human tissues [7].

Using transcriptomic data, bioinformatic models can be built for patient-oriented ranking of cancer drugs [8]. These models can be hypothesis-driven, e.g. based on the knowledge of the specific mechanisms of drugs anticancer activities [9–11]. Alternatively, hypothesis-free approaches like machine learning (ML) don’t need any theoretic background but instead strongly require sufficient training and validation datasets. Many ML methods may be used for such applications, e.g. decision trees [12, 13], random forests, RF [14, 15], linear [16], logistic [17], lasso [18, 19], ridge [15, 20] regressions, multi-layer perceptron, MLP [12, 15, 21, 22], support vectors machines [12, 13, 15, 23–25], adaptive boosting [26–28], as well as binomial naïve Bayesian [15] method.

High-quality training and validation datasets are required to run both types of the above models. Nowadays there is a shortage of clinically annotated molecular data that would help developing ML-assisted diagnostic tools. The datasets available are usually considered too small for applying ML [23, 25, 26, 29–33]. Indeed, the figure of dozens or hundreds of annotated biosamples is negligible in comparison with ~20,000 protein coding genes measured in transcriptomic assays. Intelligent data filtering is, therefore, needed to reduce dimensionality of data [8]. However, a recent approach using dynamic feature extraction, or flexible data trimming, can significantly improve performances of ML-based methods for the real-world datasets [15, 25].

This study was performed to review available clinically annotated datasets of cancer transcriptomic profiles that may be suitable for applications in ML models. To our knowledge, this is the largest published collection of processed gene expression data coupled with case history excerpts indicating positive or negative response to certain treatment protocols for cancer patients. This manually curated collection of molecular datasets will be helpful for those working with the ML or artificial intelligence applications in oncology, as well as for the fundamental research and development of cancer biomarkers.

**Methods**

We curated GEO [34], TARGET [35] and TCGA [36] repositories to extract cancer gene expression profiles associated with the clinical outcomes of chemotherapeutic treatments. We attempted to build a knowledgebase of molecular datasets suitable for building ML classifiers of clinical responses on chemotherapy treatments (Table 1, Additional file 1). Every included dataset met the following criteria:

- at least 40 gene expression profiles present;
- data obtained for the same cancer type and using the same experimental platform
- every profile is linked with the case clinical history
- all cancers treated with at least one common drug or chemotherapy regimen
- treatment outcomes are available enabling to classify every case as either responder or non-responder.

We used different approaches to discriminate between the treatment responders and non-responders. Where available, e.g. for the datasets extracted from the GEO repository, we used the responder/non-responder marks assigned by the authors of the original communications publishing these data. In many instances, the number of response groups was more than two and included groups like “partial responders”. However, most frequently binary ML-assisted drug response classifiers are needed that classify patients in only two classes: either responders or non-responders [8, 23, 25, 29, 30].

If a binary classifier is needed, then the number of clinical response groups in the training/validation datasets must be also condensed to two, i.e. responders and non-responders. In such case, the groups identified by the authors as partial responders probably can be combined with the responders. This is the case for all current breast cancer datasets, namely GSE25066 [37, 38], GSE41998 [39], GSE20271 [40], GSE50948 [41], GS E18728 [48], GSE20181 [49, 59], GSE20194 [50], GSE23988 [51], GSE22358 [52], GSE32646 [53], GSE37946 [54], GSE42822 [55], GSE59515 [57] and GSE76360 [58].

For the TCGA profiles, namely for the low-grade glioma (TCGA-LGG), lung cancer (TCGA-LC), and uterine corpus endothelial carcinoma (TCGA-UEC) datasets, and for the acute myeloid leukemia dataset GSE5122
### Table 1: Overview of selected transcriptomic datasets of responders/non-responders to cancer chemotherapy, responders (R) vs non-responders (NR)

| Reference   | Dataset ID | Disease type                      | Therapy                                                                 | Experimental platform | Number N of cases (R vs NR) | Number of core marker genes (S) |
|-------------|------------|-----------------------------------|-------------------------------------------------------------------------|------------------------|-----------------------------|---------------------------------|
| [37, 38]    | GSE25066   | Breast cancer with different hormonal and HER2 status | Neoadjuvant taxane + anthracycline                                       | Affymetrix Human Genome U133A Array | 508 (118 R: complete response + partial response; 389 NR: residual disease + progressive disease) | 20                              |
| [39]        | GSE41998   | Breast cancer with different hormonal and HER2 status | Neoadjuvant doxorubicin + cyclophosphamide, followed by paclitaxel       | Affymetrix Human Genome U133A Array | 124 (90 R: complete response + partial response; 34 NR: residual disease + progressive disease) | 11                              |
| [40]        | GSE20271   | Breast cancer with different hormonal and HER2 status | Paclitaxel + fluorouracil + adriamycin + cyclophosphamide                | Affymetrix Human Genome U133A Array | 85 (18 R: complete response + partial response; 66 NR: residual disease + progressive disease) | 11                              |
| [41]        | GSE50948   | Breast cancer with different hormonal and HER2 status | Paclitaxel + doxorubicin followed by cyclophosphamide + methotrexate/ fluorouracil followed by trastuzumab | Affymetrix Human Genome U133A Plus 2.0 Array | 156 (53 R: complete response + partial response; 103 NR: residual disease + progressive disease) | 19                              |
| [42]        | GSE9782    | Multiple myeloma                   | Bortezomib monotherapy                                                   | Affymetrix Human Genome U133A Array | 169 (85 R: complete response + partial response; 84 NR: no change + progressive disease) | 18                              |
| [43]        | GSE39754   | Multiple myeloma                   | Vincristine + adriamycin + dexamethasone followed by autologous stem cell transplantation (ASCT) | Affymetrix Human Exon 1.0 ST Array | 136 (74 R: complete, near-complete and very good partial responders; 62 NR: partial, minor and worse) | 16                              |
| [44]        | GSE68871   | Multiple myeloma                   | Bortezomib-thalidomide-dexamethasone                                     | Affymetrix Human Genome U133A Plus 2.0 Array | 118 (69 R: complete, near-complete and very good partial responders; 49 NR: partial, minor and worse) | 12                              |
| [45]        | GSE55145   | Multiple myeloma                   | Bortezomib followed by ASCT                                               | Affymetrix Human Exon 1.0 ST Array | 61 (33 R: complete, near-complete and very good partial responders; 28 R: partial, minor and worse) | 14                              |
| [35, 46]    | TARGET-50  | Childhood kidney Wilms tumor       | Vincristine sulfate + cyclosporine, cytarabine, daunorubicin + conventional surgery + radiation therapy | Illumina HiSeq 2000 | 122 (36 R: complete, near-complete and very good partial responders; 86 NR: partial, minor and worse) | 14                              |
| [35, 47]    | TARGET-10  | Childhood 8 acute lymphoblastic leukemia | Vincristine sulfate + carboplatin, cyclophosphamide, doxorubicin         | Illumina HiSeq 2000 | 98 (30 R, 68 NR: see Fig. 1) | 14                              |
| [35]        | TARGET-20  | Childhood acute myeloid leukemia   | Non-target drugs (asparaginase, cyclosporine, cytarabine, daunorubicin, etoposide; methotrexate, mitoxantrone) including busulfan and cyclophosphamide | Illumina HiSeq 2000 | 54 (31 R, 23 NR: see Fig. 1) | 10                              |
| [35]        | TARGET-20  | Childhood acute myeloid leukemia   | Same non-target drugs, but excluding busulfan and cyclo- phosphamide | Illumina HiSeq 2000 | 142 (62 R, 80 NR: see Fig. 1) | 16                              |
| [48]        | GSE18728   | Breast cancer                      | Docetaxel, capecitabine                                                 | Affymetrix Human Genome U133A Plus 2.0 Array | 61 (23 R: complete response + partial response; 38 NR: residual disease + progressive disease) | 16                              |
| [49]        | GSE20181   | Breast cancer                      | Letrozole                                                              | Affymetrix Human Genome U133A Array | 52 (37 R: complete response + partial response; 15 NR: residual disease + progressive disease) | 11                              |
| [50]        | GSE20194   | Breast cancer                      | Paclitaxel; (tri) fluoroacetyl chloride; 5-fluorouracil, epirubicin, cyclophosphamide | Affymetrix Human Genome U133A Array | 52 (11 R: complete response + partial response; 41 NR: residual disease + progressive disease) | 10                              |
| [51]        | GSE23988   | Breast cancer                      | Docetaxel, capecitabine                                               | Affymetrix Human Genome U133A Array | 61 (20 R: complete response + partial response; 41 NR: residual disease + progressive disease) | 18                              |
| [52]        | GSE22358   | Breast cancer                      | Docetaxel, capecitabine                                               | Agilent UNC Perou Lab Homo sapiens IXX44K Custom Array | 122 (116 R: complete response + partial response; 6 NR: residual disease + progressive disease) | 2                               |
| [53]        | GSE32646   | Breast cancer                      | Paclitaxel, 5-fluorouracil, epirubicin, cyclophosphamide               | Affymetrix Human Genome U133A Plus 2.0 Array | 115 (27 R: complete response + partial response; 88 NR: residual disease + progressive disease) | 17                              |
Table 1 Overview of selected transcriptomic datasets of responders/non-responders to cancer chemotherapy, responders (R) vs non-responders (NR) (Continued)

| Reference | Platform | Classification | Chemotherapy | Array | R | NR |
|-----------|----------|----------------|--------------|-------|---|----|
| [54] GSE37946 | Breast cancer | Trastuzumab | Affymetrix Human Genome U133A Array | 50 (27 R: complete response + partial response; 23 NR: residual disease + progressive disease) | 14 |
| [55] GSE42822 | Breast cancer | Docetaxel, 5-fluorouracil, epirubicin, cyclophosphamide, capecitabine | Affymetrix Human Genome U133A Array | 91 (38 R: complete response + partial response; 53 NR: residual disease + progressive disease) | 13 |
| [56] GSE5122 | Acute myeloid leukemia | Tipifarnib | Affymetrix Human Genome U133A Array | 57 (13 R: complete response + partial response + stable disease; 44 NR: progressive disease) | 10 |
| [57] GSE59515 | Breast cancer | Letrozole | Illumina HumanHT-12 V4.0 expression beadchip | 75 (51 R: complete response + partial response; 24 NR: residual disease + progressive disease) | 15 |
| [58] GSE76360 | Breast cancer | Trastuzumab | Illumina HumanHT-12 V3.0 expression beadchip | 48 (42 R: complete response + partial response; 6 NR: residual disease + progressive disease) | 3 |
| [36] TCGA-LGG | Low-grade glioma | Temozolomide (optionally) mibebradil | Illumina HiSeq 2000 | 131 (100 R: complete response + partial response + stable disease; 31 NR: progressive disease) | 9 |
| [36] TCGA-LC | Lung cancer all types | Paclitaxel (optionally), cisplatin/carboplatin, reolysin | Illumina HiSeq 2000 | 41 (24 R: complete response + partial response + stable disease; 17 NR: progressive disease) | 7 |
| [36] TCGA-UC | Uterine corpus endometrial carcinoma | Paclitaxel (optionally) carboplatin, cisplatin, doxorubicin | Illumina HiSeq 2000 | 57 (57 R: complete response + partial response + stable disease; 7 NR: progressive disease) | 2 |

[56], stable disease cases can be most probably classified as the responders whereas progressive disease cases – as the non-responders. For the multiple myeloma dataset GSE9782 [42], the classification can be used as defined by the authors, where patents with complete and partial response were annotated as the responders, and with no change and progressive disease – as the non-responders. For three other multiple myeloma datasets, namely GSE39753 [43], GSE68871 [44], and GSE55145 [45], complete, near-complete and very good partial response groups can be most likely considered as the responders, whereas partial, minor and worse response groups – as the non-responders.

Classification of the TARGET repository profiles was more sophisticated as no responder classification was given by the authors. This was the case for the datasets of pediatric Wilms kidney tumor (TARGET-50), acute myeloid leukemia (TARGET-20) and acute lymphoblastic leukemia (TARGET-10) extracted from the gene expression repository of National Cancer Institute [35]. However, these latter clinical cases were annotated by the time of event-free survival. Distributions of the event-free survival time enabled us to identify for every dataset two different modes of survival with different slopes (Fig. 1), that can be recognized as either responders or non-responders.

**Results**

For raw gene expression data, the number of features i.e. interrogated genes, usually exceeds the number of tumor cases by roughly two orders of magnitude. Therefore, for robust application of ML the dimensionality of data must be reduced to make the number of selected features lower than the number of tumor cases or at least comparable to it (Fig. 2a). To reduce dimensionality, the gene expression data can be aggregated into the higher-order molecular markers like activation profiles of molecular pathways [23, 29, 30, 60, 61]. Alternatively, the most informative fraction of the initial data can be selected that can distinguish between the responder and non-responder classes. For selection of such marker features, several approaches have been proposed, e.g. Pearson chi-squared test [62], correlation test [27, 62], variance thresholding, genetic algorithms [63], univariate feature selection, recursive feature elimination, principal component analysis [27], CUR matrix [64], decomposition [65] and covariate regression [66].

In the current research, we applied the following leave-one-out-based method for finding robust marker features [25] (Fig. 2c). Imagine that we have a gene expression dataset that embraces N clinical cases, each with corresponding expression profile. For each clinical case i = 1, … N, we determine the top Q marker genes that distinguish responding and non-responding cases in a sub-dataset that contains all samples but i. In other words, for all N sub-datasets each having N-1 cases, we interrogate each gene taken one by one and retrieve the top Q set of genes that showed the highest ROC AUC values for the difference between responder and non-responder profiles. The quality metric area under the ROC curve (AUC) is the universal metric of a biomarker robustness that depends on its sensitivity and specificity.
Fig. 1 Distribution of event-free survival time for the patients with (a) childhood kidney Wilms tumor from TARGET-50 dataset, (b) childhood ALL from TARGET-10 dataset and (c) childhood AML from TARGET-20 dataset [35]. Patients on the left from vertical threshold can be considered as the non-responders, and on the right – as the responders to the treatment.
It positively correlates with the quality of a biomarker and varies from 0.5 till 1. The standard discrimination threshold is 0.7 and the entries with higher AUC are considered high-quality biomarkers, and vice versa [68]. AUC is broadly used for detection of biomarkers in oncology [69–73].

To provide robust feature selection, the number $Q$ shouldn’t exceed the number of cases $N$. In the current application, we took $Q$ equal to 30 because all tdatasets under consideration had more than 40 cases. The final list of core marker genes was obtained by intersecting top $Q$ gene sets for all $N$ sub-datasets.

We applied this procedure to all the clinically annotated cancer transcriptomic datasets under consideration and identified for them core marker genes (Table 1). Twenty-three out of 26 datasets investigated provided 7–20 core marker gene features for further ML applications (Table 1). The remaining three datasets, namely GSE22358 [52], GSE76360 [58] and TCGA-UEC [36], were poorly balanced because the numbers of responders greatly exceeded the respective numbers of non-responders, or vice versa. For these three instances we were unable to generate robust core marker gene sets for ML applications because the number of such genes was too low (two-three per dataset, Table 1).
Discussion
By the current moment, ML has’t made a revolution in biomedicine [12]. This may be partly connected with the relatively recent emergence of experimental methods generating big amounts of biomedical data combined with the developed IT infrastructure. Among these game-changing methods the major role was played by the next-generation sequencing (NGS) and novel mass-spectrometry approaches which made whole genome-, transcriptome-, proteome- and metabolome analyses relatively fast and cheap [74–76].

Yet further development of ML methods in personalized oncology is still strongly limited by the low number of clinically annotated cancer patient molecular datasets. A dataset suitable for ML should have all together enough number of high-throughput molecular profiles and also the associated clinical case history records featuring success of the therapeutic regimen used.

In this paper we reviewed three major repositories of omics data for the available responder/non-responder datasets including more the 40 cancer cases treated with the same chemotherapeutics. We identified 26 datasets with totally 2786 cases, ranging from 41 till 508 cases per dataset (Table 1). We checked the robustness of these datasets and their suitability for ML applications using our previous method of core maker feature determination [25]. According to this test, 23/26 datasets were suitable for ML, each having 7–20 core marker genes/features for further ML applications. Contrarily, the remaining three datasets produced only two or three features, which may seem insufficient for the ML. Poor performance of these three datasets was most likely due to unbalanced numbers of clinical responder/non-responder cases included.

To increase the number of cases (Fig. 2b), the datasets for the same disease or drug treatment conditions can be merged using cross-dataset harmonization. Different methods can be used to harmonize data obtained using the same [77, 78] or two different experimental platforms [79, 80], or even using multiple platforms [81] (Fig. 2b).

In addition, when the cases are deficient, transfer learning methods may be used for a certain disease or drug condition. Using this approach, the ML training process may be preformed on the multiple available molecular profiles corresponding to cell culture treated with certain drugs [82], whereas the ML classifier validation may be done on more rare patient cancer cases [23, 29, 30].

Conclusions
We identified 26 clinically annotated gene expression datasets ranging from 41 till 508 cases per dataset (Table 1). Collectively, they covered 2786 individual cancer cases. Among them seven datasets included RNA sequencing data (for 645 cases) and the others – microarray expression profiles. The datasets represented breast cancer, lung cancer, low-grade glioma, endothelial carcinoma, multiple myeloma, adult leukemia, pediatric leukemia and kidney tumors. Chemotherapeutics used included taxanes, bortezomib, vincristine, trastuzumab, letrozole, tipifarnib, temozolomide, busulfan and cyclophosphamide.

We hope that presented collection of clinically annotated transcriptomic profiles will be useful to those working with data analysis in oncology, as well as for the fundamental research and development of next-generation cancer biomarkers.

Supplementary information
Supplementary information accompanies this paper at https://doi.org/10.1186/s12920-020-00759-0.

Additional file 1. Clinically annotated datasets and samples they contain.

Abbreviations
ALL: Acute lymphoblastic leukemia; AML: Acute myelogenous leukemia; ASCT: Autologous stem cell transplantation; GEO: Gene expression omnibus; GSE: GEO series; LOO: Leave-one-out; ML: Machine learning; MLP: Multi-layer perceptron; PM: Personalized medicine; RF: Random forest; ROC: Receiver operating characteristic; SVM: Support vector machines; TARGET: Tumor Alterations Relevant for GEnomics-driven Therapy; TCGA: The Cancer Genome Atlas

Acknowledgements
Not applicable.

About this supplement
This article has been published as part of BMC Medical Genomics Volume 13 Supplement 8: Selected Topics in Systems Biology and Bioinformatics – 2019: medical genomics. The full contents of the supplement are available online at https://bmcmedgenomics.biomedcentral.com/articles/supplements/volume-13-supplement-8.

Authors’ contributions
NB, MS and AB contributed conception and design of the study. NB, MS, VT and AG analyzed the data. NB and AB wrote the manuscript. All authors read and approved the final manuscript.

Funding
The work was funded by the Russian Science Foundation grant no. 18-15-00061.

Availability of data and materials
All the data, including IDs of expression profiles, treatment methods and clinical response assessment, both done by the teams, who had worked with them, and our binary (“P-vs-N”) response classifications, are provided in Additional file 1.

Ethics approval and consent to participate
Current research did not involve any new human material. All the gene expression data that were used for research were taken from publicly available repositories GEO, TARGET and TCGA and had been previously anonymized by the teams, who had worked with them.

Consent for publication
Not applicable.
Competing interests
NB, MS, VT, AG and AB have a financial relationship with OmicsWay Corp.

Author details
1Department of Bioinformatics and Molecular Networks, OmicsWay Corporation, Walnut, CA 91788, USA. 2Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Oblast 141701, Russia. 3IM. Sechenov First Moscow State Medical University (Sechenov University), Moscow 119991, Russia. 4Stemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Moscow 117997, Russia.

Received: 14 July 2020 Accepted: 27 July 2020
Published: 18 September 2020

References
1. Buzdin A, Sorokin M, Garazha A, Glusker A, Aleshin A, Poddubskaya E, et al. RNA sequencing for research and diagnostics in clinical oncology. Semin Cancer Biol. 2020;60:311–23.
2. Zhukov NV, Tjulandin SA. Targeted therapy in the treatment of solid tumors: practice contradicts theory. Biochem Biophys. 2008;73:605–18.
3. Katz SJ, Ward KC, Hamilton AS, Mcleod WC, Wallner LP, Morrow M, et al. Gaps in receipt of clinically indicated genetic counseling after diagnosis of breast cancer. J Clin Oncol. 2018;36:1218–24.
4. Marquard J, Chen EY, Prasad V, et al. Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial. Nat Med. 2019;25:751–8.
5. Buzdin A, Sorokin M, Garazha A, Sekacheva M, Kim E, Zhukov N, et al. Molecular pathway activation - new type of biomarkers for tumor morphology and personalized selection of target drugs. Semin Cancer Biol. 2018;53:110–24.
6. Rodon J, Soria J-C, Berger R, Miller WH, Rubin E, Kugel A, et al. Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial. Nat Med. 2019;25:751–8.
7. Devizorova Z, Mironov S, Buzdin A. Theory of magnetic domain phases in ferromagnetic superconductors. Phys Rev Lett. 2019;123:117002.
8. Borisov N, Buzdin A. New paradigm of machine learning (ML) in personalized oncology: data trimming for squeezing more biomarkers from clinical datasets. Front Oncol. 2019;9:658.
9. Artemov A, Aliper A, Korzinkin M, Lezhina K, Jellen L, Zhukov N, et al. A method for predicting target drug efficacy in cancer based on the analysis of signaling pathway activation. Oncotarget. 2015;6:29347–56.
10. Shepelin D, Korzinkin M, Vanyushina A, Aliper A, Borisov N, Vasilev R, et al. Molecular pathway activation features linked with transition from normal skin to primary and metastatic melanomas in human. Oncotarget. 2016;7:655–70.
11. Zolotovskia MA, Sorokin MI, Emelianova AA, Borisov NM, Kuzmin DV, Borger P, et al. Pathway based analysis of mutation data is efficient for scoring target cancer drugs. Front Pharmacol. 2019;10:https://doi.org/10.3389/fphar.2019.00001.
12. Robin X, Turck N, Hainard A, Lisacek F, Sanchez J-C, Müller M. Bioinformatics for protein biomarker panel classification: what is needed to bring biomarker panels into in vitro diagnostics? Expert Rev Proteomics. 2009;6:675–89.
13. Bartlett P, Shawe-Taylor J. Generalization performance of support vector machines and other pattern classifiers. In: Advances in Kernel Methods: Support Vector Learning. Cambridge: MIT Press; 1999. p. 43–54. ISBN 0262194163.
14. Toljic I, Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics. 2011;27:1986–94.
15. 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.
16. Stigler SM. The history of statistics: the measurement of uncertainty before 1900. Cambridge: Belknap Press of Harvard University Press; 1986.
17. Cramer JS. The origins of logistic regression. SSRN Electron J. 2003. https://doi.org/10.2139/ssrn.360300.
18. Santos F, Symes WW. Linear inversion of band-limited reflection seismograms. SIAM J Sci Stat Comput. 1986;7:1307–30.
19. Tilbhirani R. The lasso method for variable selection in the Cox model. Stat Med. 1997;16:385–95.
20. Tikhonov AN, Arsenin V. Solutions of ill-posed problems. Washington, New York, Winston: distributed solely by Halsted Press; 1977.
21. Minsky ML, Papert SA. Perceptrons - expanded edition: an introduction to computational geometry. Boston: MIT Press; 1987.
22. Prados J, Kalousis A, Sanchez J-C, Alfar I, Carrette O, Hilario M. Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents. Proteomecs. 2004;4:2320–32.
23. Borisov N, Tkachev V, Suntsova M, Kovalchuk O, Zhavoronkov A, Muchnik I, et al. A method of gene expression data transfer from cell lines to cancer patients for machine-learning prediction of drug efficiency. Cell Cycle. 2018;17:486–91.
24. Osuna E, Freund R, Girosi F. An improved training algorithm for support vector machines: IEEE; 1997. p. 276–85. https://doi.org/10.1109/NNSP.1997.622408.
25. Tkachev V, Sorokin M, Mescheryakov A, Simonov A, Garazha A, Buzdin A, et al. 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. https://doi.org/10.3389/fgen.2019.00717.
26. Turk T, Wang JTL. Clinical intelligence: new machine learning techniques for predicting clinical drug response. Comput Biol Med. 2019;107:302–22.
27. Wang Z, Yang H, Wu Z, Wang T, Li W, Tang Y, et al. In silico prediction of blood-brain barrier permeability of compounds by machine learning and resampling methods. ChemMedChem. 2018;13:2189–201.
28. Yosipof A, Guedes RC, Garcia-Sosa AT. Data mining and machine learning models for predicting drug likeness and their disease or organ category. Front Chem. 2018;6:https://doi.org/10.3389/fchem.2018.00162.
29. Borisov N, Tkachev V, Muchnik I, Buzdin A. Individual drug treatment prediction in oncology based on machine learning using cell culture gene expression data: ACM Press; 2017. p. 1–6. https://doi.org/10.1145/3155077.3155078.
30. Borisov N, Tkachev V, Buzdin A, Muchnik I. Prediction of drug efficiency by transferring gene expression data from cell lines to cancer patients. In: Rozonov L, Mirkin B, Muchnik I, editors. Braverman readings in machine learning. Key ideas from inception to current state. Cham: Springer International Publishing; 2018. p. 201–12. https://doi.org/10.1007/978-3-319-99492-5_9.
31. Turk T, Wei Z. A link prediction approach to cancer drug sensitivity prediction. BMC Syst Biol. 2017;11. https://doi.org/10.1186/s12918-017-0463-8.
32. Turk T, Wei Z, Wang JTL. Transfer learning approaches to improve drug sensitivity prediction in multiple myeloma patients. IEEE Access. 2017;5:7381–93.
33. Turk T, Wei Z, Wang JTL. A transfer learning approach via procurates analysis and mean shift for cancer drug sensitivity prediction. J Bioinforma Comput Biol. 2018;16:1840014.
34. Edgar D, Domachew M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30:2007–10.
35. Goldman M, Craft B, Swatloski T, Cline M, Morozova O, diekhans M, et al. The UCSC cancer genomics browser: update 2015. Nucleic Acids Res. 2015;43:D812–7.
36. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol Poznan Pol. 2015;19:669–77.
37. Hatzis C, Puztas L, Valero V, Boozer DJ, Esseran L, Lluch A, et al. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. JAMA. 2011;305:1873–81.
38. Itzh M, Iwamoto T, Matsuka J, nogami T, Motoki T, Shien T, et al. Estrogen receptor (ER) mRNA expression and molecular subtype distribution in ER-negative/progesterone receptor-positive breast cancers. Breast Cancer Res Treat. 2014;143:403–9.
39. Horak CE, Puztas L, Xing G, Tiffany OC, Saura C, Tseng L-M, et al. Biomarker analysis of neoadjuvant doxorubicin/cyclophosphamide followed by ixabepilone or paclitaxel in early-stage breast cancer. Clin Cancer Res Off J Am Assoc Cancer Res. 2013;19:1587–95.
40. Tabchy A, Valero V, vidaurre T, lluch A, gomez H, Martin M, et al. Evaluation of a 30-gene paclitaxel, fluorouracil, doxorubicin, and cyclophosphamide chemotherapy response predictor in a multicenter randomized trial in breast cancer. Clin Cancer Res Off J Am Assoc Cancer Res. 2010;16:5351–61.
41. Prat A, Bianchini G, Thomas M, Belousova A, Cheang MCU, koehler A, et al. Research-based PAM50 subtype predictor identifies higher responses and
improved survival outcomes in HER2-positive breast cancer in the NOAH study. Clin Cancer Res Off J Am Assoc Cancer Res. 2014;20:511–21.
42. Mulligan G, Mitsiades C, Bryant B, Zhan F, Chng WI, Roels S, et al. Gene expression profiling and correlation with outcome in clinical trials of the proteasome inhibitor bortezomb. Blood. 2007;109:3177–88.
43. Chauhan D, Tian Z, Nicholson B, Kumar KGS, Zhou B, Carrasco R, et al. A small molecule inhibitor of ubiquitin-specific protease-7 induces apoptosis in multiple myeloma cells and overcomes bortezomib resistance. Cancer Cell. 2012;22:345–58.
44. Terragna C, Remondini D, Martello M, Zamagni E, Pantani L, Patriarca F, et al. The genetic and genomic background of multiple myeloma patients achieving complete response after induction therapy with bortezomb, thalidomide and dexamethasone (VTD). Oncotarget. 2016;7:9666–79.
45. Amin SB, Yip W-K, Minvielle S, Broyl A, Li Y, Hanlon B, et al. Gene expression profile alone is inadequate in predicting complete response in multiple myeloma. Leukemia. 2014;28:2229–34.
46. Walz AL, Ooms A, Gadd S, Gerhard DS, Smith MA, Guidy Avill JM, et al. Recurrent DGR9B, DROSHA, and SIX Homedomain mutations in favorable histology Wilms tumors. Cancer Cell. 2015;27:286–97.
47. Tricoli JV, Blair DG, Anders CK, Bleyer WA, Boardman LA, Khan J, et al. Biologic and clinical characteristics of adolescent and young adult cancers: acute lymphoblastic leukemia, colorectal cancer, breast cancer, melanoma, and sarcoma: biology of AYA cancers. Cancer. 2016;12:1017–28.
48. Korde LA, Lusa L, McShane L, Lebowitz PF, Lukes L, Camphausen K, et al. Gene expression pathway analysis to predict response to neoadjuvant doctetaxel and capcitabine for breast cancer. Breast Cancer Res Treat. 2010;119:685–99.
49. Miller WR, Larionov A. Changes in expression of oestrogen regulated and proliferation genes with neoadjuvant treatment highlight heterogeneity of clinical resistance to the aromatase inhibitor, letrozole. Breast Cancer Res. 2010;12:R52.
50. Popovic V, Chen W, Gallas BG, Hatzi C, Shi W, Samuelson FW, et al. Effect of training-sample size and classificaiton difficulty on the accuracy of genomic predictors. Breast Cancer Res. 2010;12:R5.
51. Iwamoto T, Bianchini G, Booser D, Qi Y, Coutant C, Shiang CY-H, et al. Gene identification of molecular predictors of response in a study of tipifarnib treatment in relapsed and refractory acute myelogenous leukemia. Clin Cancer Res Off J Am Assoc Cancer Res. 2017;13:2074–2084.
52. Glück S, Ross JS, Royce M, McKenna EF, Perou CM, Avisar E, et al. TP53 achieves complete response after induction therapy with bortezomb, thalidomide and dexamethasone (VTD). Oncotarget. 2016;7:9666–79.
53. Taricha F, Schuster T, Müller E, Pellegatta S, Krämer M, Thurner B, et al. The genetic and genomic background of multiple myeloma patients achieving complete response after induction therapy with bortezomb, thalidomide and dexamethasone (VTD). Oncotarget. 2016;7:9666–79.
54. Liu JC, Voisin V, Bader GD, Deng T, Pusztai L, Symmans WF, et al. Seventeen-gene signature from enriched Her2/Neu mammary tumor-initiating cells predicts clinical outcome for human HER2+.ER α - breast cancer. Proc Natl Acad Sci U S A. 2012;109:5832–7.
55. Shabalin AA, Tjelmeland H, Fan C, Perou CM, Nobel AB. Merging two gene-expression data sets: bias, variance, and correction. Bioinformatics. 2012;28:1182–92.
56. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19:185–93.
57. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. The genetic and genomic background of multiple myeloma patients achieving complete response after induction therapy with bortezomb, thalidomide and dexamethasone (VTD). Oncotarget. 2016;7:9666–79.
58. Zolotovskiaia MA, Sorokin MI, Roumiantsev SA, Borisov NM, Budzin AA. Pathway instability is an effective new mutation-based type of cancer biomarkers. Front Oncol. 2019;9:https://doi.org/10.3389/fonc.2019.00658.
59. Chu Y, Corey DR. RNA sequencing: platform selection, experimental design, and data interpretation. Nucleic Acid Ther. 2012;22:271–4.
60. Cox J, Mann M. Quantitative, high-resolution proteomics for data-driven biomarker discovery in cancer cells. Nucleic Acids Res. 2013;41(Database Suppl 8):111.
61. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim J-S, et al. A novel signaling pathway impact analysis. Bioinformatics Oxf Engl. 2009;25:75–82.
62. Cho H-J, Lee S, Ji YG, Lee DH. Association of specific gene mutations derived from machine learning with survival in lung adenocarcinoma. PLoS One. 2018;13:e0207204.
63. Sofian O, Riefgoannis D, Kalnis P, Bajic V. DWFS: a wrapper feature selection tool based on a parallel genetic algorithm. PLoS One. 2015;10:1:e0117988.
64. Mahoney MW, Drineas P. CUR matrix decompositions for improved data analysis. Proc Natl Acad Sci. 2009;106:697–702.
65. Turk I, Welz I. Learning approaches to improve prediction of drug sensitivity in breast cancer patients. In: 2016 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC). Orlando: IEEE; 2016. p. 3314–20. https://doi.org/10.1109/EMBC.2016.7591437.
66. Menden MP, Casale FP, Stephan J, Bignell GR, Iorio F, McDermott U, et al. The germline genetic component of drug sensitivity in cancer cell lines. Nat Commun. 2018.https://doi.org/10.1038/s41467-018-05811-3.
67. Green DM, Sweats JA. Signal detection theory and psychophysics. Repr. ed edn. New York: Altas Hills: Peninsula Publ. 2000.
68. Boyd XC. Mathematical tools for demonstrating the clinical usefulness of biochemical markers. Scand J Clin Lab Invest Suppl. 1997;227:46–63.
69. Borisov NM, Terekhanova NV, Aliper AM, Venkova LS, Smirnov PY, Roumiantsev SA, et al. Signaling pathways activation profiles make better markers of cancer than expression of individual genes. Oncotarget. 2014;5:10198–205.
70. Chen L, Zhou Y, Tang X, Yang C, Tian Y, Xie R, et al. EGFR mutation decreases FDG uptake in non-small cell lung cancer via the NDA4/ROS/Glut1 pathway. Int J Oncol. 2018.https://doi.org/10.3892/ijo.2018.4626.
71. Liu T, Cheng G, Kang X, Xi Y, Zhu Y, Wang K, et al. Noninvasively evaluating the grading and IDH1 mutation status of diffuse gliomas by three-dimensional pseudo-continuous arterial spin labeling and diffusion-weighted imaging. Neuroradiology. 2018;60:593–702.
72. Tanjica M, Fan C, Parker JS, Hoadley KA, Hu Z, Li Y, et al. Integrated analysis of RNA and DNA from the phase III trial CALGB 40601 identifies predictors of response to trastuzumab-based neoadjuvant chemotherapy in HER2-positive breast cancer. Clin Cancer Res. 2018;24:5292–304.
73. Zolotovskiaia MA, Sorokin MI, Roumiantsev SA, Borisov NM, Budzin AA. Pathway instability is an effective new mutation-based type of cancer biomarkers. Front Oncol. 2019;9:https://doi.org/10.3389/fonc.2019.00658.
74. Chu Y, Corey DR. RNA sequencing: platform selection, experimental design, and data interpretation. Nucleic Acid Ther. 2012;22:271–4.
75. Cox J, Mann M. Quantitative, high-resolution proteomics for data-driven systems biology. Annu Rev Biochem. 2011;80:273–99.
76. Petterson E, Lundberg J, Ahmadian A. Generations of sequencing technologies. Genomics. 2009;93:105–11.
77. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014:15:550.
78. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics Oxf Engl. 2003;19:185–93.
79. Huang H, Lu X, Liu Y, Hiaauland P, Maron JS. R/WD: distance-weighted discrimination for classification, visualization and batch adjustment. Bioinformatics. 2012;28:1182–3.
80. Shabalin AA, Tjelmeland H, Fan C, Perou CM, Nobel AB. Merging two gene-expression studies via cross-platform normalization. Bioinformatics. 2008;24:1154–60.
81. Borisov N, Shabalina I, Tkachev V, Sorokin M, Garazha A, Pulin A, et al. Shambhala: a platform-agnostic data harmonizer for gene expression data. BMC Bioinformatics. 2019;20:666.
82. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013;41(Database issue):D655–61.