Accurate pan-cancer tumor purity estimation from gene expression data

Egor Revkov[1,2], Tanmay Kulshrestha[2], Ken Wing-Kin Sung[1,2], Anders Jacobsen Skanderup[1,2,3*]

[1]National University of Singapore, School of Computing, 119077, Singapore

[2]Genome Institute of Singapore, 138672, Singapore

[3]National Cancer Centre Singapore, Division of Medical Oncology, 169610, Singapore

*skanderupamj@gis.a-star.edu.sg

Abstract

Tumors are complex masses composed of malignant and non-malignant cells. Variation in tumor purity (proportion of cancer cells in a sample) can both confound integrative analysis and enable studies of tumor heterogeneity. Here we developed PUREE, which uses a weakly supervised learning approach to infer tumor purity from a tumor gene expression profile. PUREE was trained on gene expression data and genomic consensus purity estimates from 7864 solid tumor samples. PUREE predicted purity with high accuracy across distinct solid tumor types and generalized to tumor samples from unseen tumor types and cohorts. Gene features of PUREE were further validated using single-cell RNA-seq data from distinct tumor types. In a comprehensive benchmark, PUREE outperformed existing transcriptome-based purity estimation approaches. Overall, PUREE is a highly accurate and versatile method for estimating tumor purity and interrogating tumor heterogeneity from bulk tumor gene expression data.

Introduction

Cancerous tumors are complex mixtures of malignant and non-malignant cells shaping the tumor microenvironment (TME). The composition and relative proportions of malignant cells and non-malignant components (comprising stromal, epithelial, and infiltrating immune cells) can display substantial variation across tumors (1–4). The composition of the TME is also associated with the disease stage and treatment response (5). The proportion of malignant cancer cells in the tumor mass, herein referred to as tumor purity, also impacts genomic analysis such as estimation of clonal composition (6) and tumor mutation burden (7), critical for predicting treatment outcomes and selecting patients for immunotherapy. Moreover, tumor purity can guide tumor transcriptome deconvolution and the estimation of gene expression profiles for malignant and non-malignant cell populations inside tumors (8,9), enabling new insights on TME biology (10) and its impact on clinical treatment response.

Traditionally, the cancer cell proportion has been estimated by pathologists inspecting nuclei in hematoxylin and eosin (H&E)-stained tissue slides. However, such estimates may often be imprecise, as demonstrated by the noticeable variation in estimates when the same sample is evaluated by different pathologists (11). More recent computational approaches to estimate tumor purity are based on DNA sequencing data where variation in allele frequencies of somatic DNA mutations, copy-number alterations (CNAs), or DNA methylation patterns are used to infer the malignant cell proportion (5,12–18). Genomics-based purity estimation methods, despite differences in underlying statistical models and input data, have been shown to produce concordant estimates of tumor purity (9,19).

Tumor purity can also be estimated from the tumor gene expression profile (20), which has been used to derive clinically relevant molecular subtypes (21–24), perform quality control of tumor samples (25,26), and analyze treatment responses after immunotherapy (27). Existing methods that estimate tumor purity from a tumor gene expression profile adopt different analytical strategies. ESTIMATE calculates a combined enrichment score for infiltrating immune and stromal cells followed by training of a supervised model (28). EPIC uses constrained least square optimization in combination with non-malignant cell-type reference profiles to perform cell type proportion deconvolution (29). DeMixT uses probabilistic modeling to infer proportions of stromal and cancer-cell components from a set of input samples, comprising both tumor and normal-tissue samples (8). LinSeed constructs an undirected weighted linearity network of genes to determine mutually linear features followed by simplex-based deconvolution (30). CIBERSORTx defines a cell-type signature matrix followed by support vector regression to infer the proportions of cell types in each sample (31,32). Similarly, DeconRNASeq solves a non-negative least squares problem using a pre-defined cell-type signature matrix to derive the cellular proportions (33). However, due to the inherent modeling assumptions, these methods might not always capture all the biological variation between stroma and cancer cells, required to predict require malignant cells proportions. Additionally, while these methods show convincing results in their own benchmarks, it is not clear how accurate they are when applied to distinct cancer types and compared with each other.

Our goal was to develop an accurate reference-free method for predicting tumor purity from a tumor gene expression profile. To reduce the modeling limitations of existing approaches, we utilized minimal prior modeling assumptions and instead relied on a statistical learning approach to infer gene expression patterns related to stroma and cancer components. We used a weakly supervised learning strategy, training a machine learning model using gene expression data from 7864 tumors and 20 solid cancer types (34) in combination with orthogonal consensus

genomics-based tumor purity estimates. The resulting method, PUREE, is able to robustly predict purity values with high correlation and low root mean squared error (RMSE) when compared to consensus genomics-based estimates from the same samples, outperforming existing deconvolution methods both on a TCGA test set (0.2 increase in Pearson's correlation and 0.17 decrease in RMSE compared to the respective second best approaches) and seven external validation datasets of lung, colorectal, uterine, paraganglioma, and testicular cancers.

Results

Overview of approach

Our goal was to develop an accurate method for estimating tumor purity from a tumor gene expression profile. Such a method should be able to generalize across different solid cancer types and exhibit high concordance with orthogonal purity estimates derived from tumor DNA data (Fig. 1a). We therefore assembled a training dataset comprising matched genomic and gene expression profiles from 7864 tumors spanning 20 solid cancer types from TCGA (34) (Suppl. Tab. S1). The orthogonal ("pseudo" ground truth) purity label of each tumor was estimated using the tumor genomic profile, using the consensus of four existing algorithms that generally displayed high concordance (mean Pearson r=0.85, Methods, Suppl. Fig. S1). Next, we adopted a weakly supervised learning strategy to train a model that could predict tumor purity labels from the matched gene expression profiles. Gene expression profiles were rank-percentile transformed to provide robustness to variation in scale and normalization of different gene expression datasets and platforms (e.g. FPKM, TPM; Methods). From the 60,000 transcripts profiled in TCGA, we further selected and focused on 9,554 (10K) highly expressed protein-coding autosomal genes for model development (Methods). We explored the performance of a range of machine learning methods (Suppl. Fig. S2). In particular, given the regression task of predicting the bounded continuous tumor purity value, we tested both a range of linear and non-linear machine learning architectures. This comparison showed that a simple linear regression model could achieve optimal accuracy using only a limited set of gene expression features (Suppl. Fig. S2). PUREE was therefore developed using a linear regression and weakly supervised learning strategy to enable accurate estimation of tumor purity from a solid tumor gene expression profile (Fig. 1b).

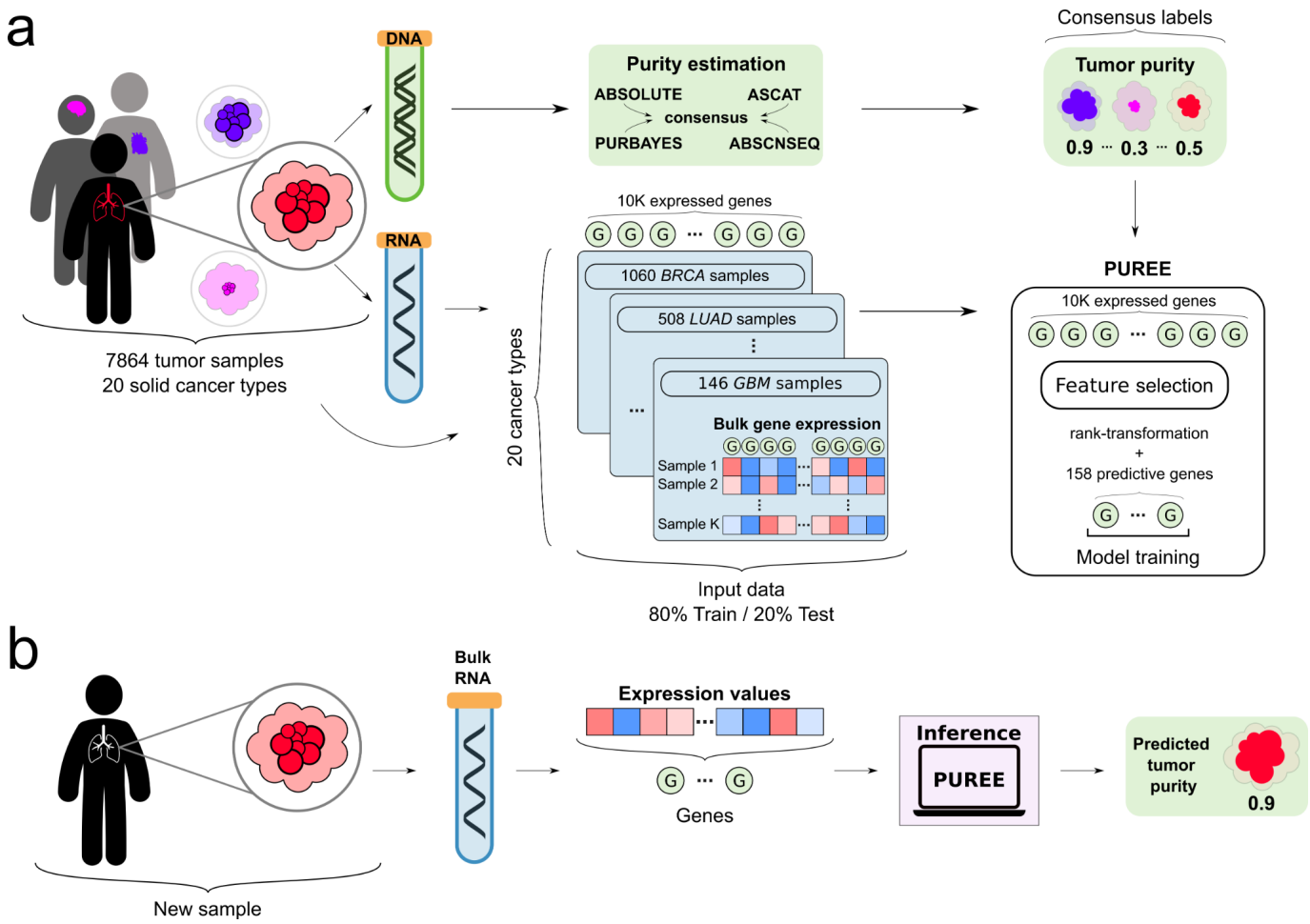


Figure 1: Overview of PUREE. a) PUREE is trained using a weakly supervised learning approach. Consensus genomics-based purity estimates are used as orthogonal (pseudo ground-truth) labels, and a predictive model is trained on rank-transformed gene expression profiles from 7864 tumor samples spanning 20 solid tumor types (80%/20% train/test split). **b)** For a new solid tumor sample, PUREE infers the purity from the corresponding tumor gene expression profile.

Feature selection to account for cancer type and tumor purity imbalance

The TCGA training dataset showed strong cancer type and purity range imbalance (Fig. 2a, Suppl. Tab. S1). To reduce the impact of this imbalance during model training, we adopted a two-step feature selection strategy (Methods). Briefly, the first step consisted of selecting features which could predict purity at both lower and higher purity ranges (Fig. 2b). The second step further filtered this feature set to identify the genes most predictive across the entire purity range, resulting in 158 features (Fig. 2c).

We further explored the properties of the resulting reduced feature set. Using gene set enrichment analysis, the 158 genes were enriched in pathways relating to angiogenesis, KRAS signaling and epithelial-mesenchymal transition (Methods, Suppl. Fig. S3). Genes positively correlated with purity showed enrichment in cancer-related pathways and processes such as epithelial-mesenchymal transition (BASP1, COL4A1, THBS2), genes involved in the TNFA signaling via NFKB (SPSB1, SMAD3), and genes upgregulated by KRAS activation (CFB, MAFB). Genes negatively correlated with purity showed enrichment in stroma-related processes

such as inflammatory response (IL1R1, STAB1, MSR1) and also genes involved in epithelial-mesenchymal transition (TGFBR3, CXCL12, CRLF1, PMP22, SDC1). Overall, this confirmed the original hypothesis of the PUREE model selecting cancer and stroma-related genes.

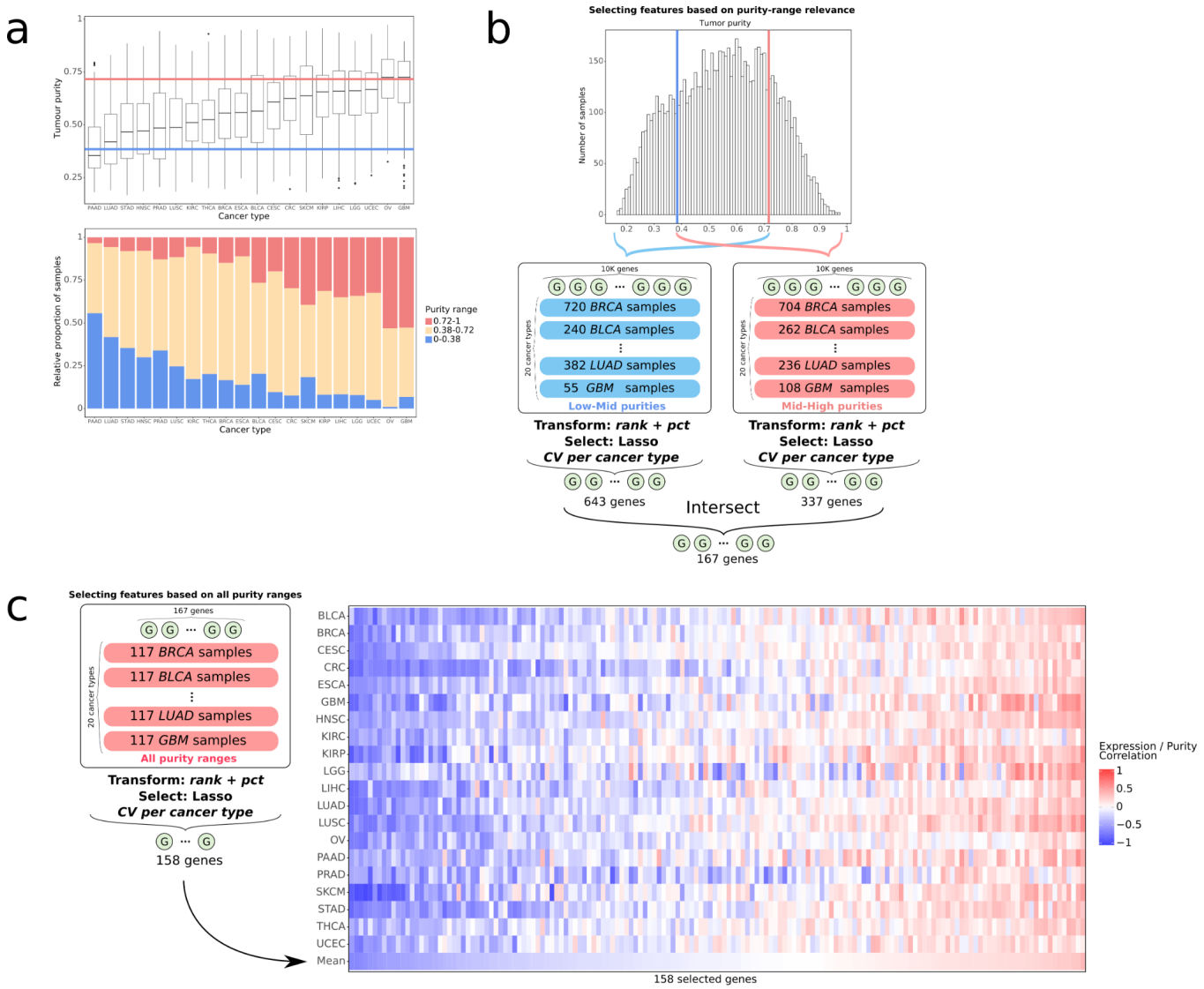


Figure 2: Feature selection to account for cancer type and tumor purity imbalance. a) top: genomics-based tumor purity across 20 cancer types from TCGA, blue and red lines marking the bottom/top 20% purity samples; bottom: genomics-based tumor purity separated into low, medium, and high purity ranges. **b)** the first step of the feature selection strategy: lasso feature selection, cross-validated on cancer types as folds, was performed separately on low-mid (0.17-0.72 purities) and mid-high (0.38-0.97) purity range samples, the two feature sets were intersected resulting in 167 genes. **c)** the second step of the feature selection strategy: using the features from the first step, lasso feature selection was iteratively performed across all training samples, cross-validated on cancer types as folds, resulting in 158 gene features used for the final model.

Comparing the performance of pan-cancer and cancer-type specific models

As an alternative to PUREE's pan-cancer tumor purity prediction model, we explored whether models trained for a specific cancer type could more accurately predict purity. To test this, we

trained cancer-type specific models (Methods) and compared their performance with PUREE across all cancer types. Interestingly, PUREE showed comparable and often improved performance, with comparable median correlation (0.784 vs 0.790, *P*=0.1, Wilcoxon signed-rank test, two-tailed) and lower median RMSE (0.094 vs 0.096, P=0.08) with the orthogonal genomics-based purity estimates (Fig. 3a, Suppl. Fig. S4, S5, S6). Overall, this confirmed that PUREE's pan-cancer feature selection and training approach provided a robust and accurate prediction across all individual cancer types.

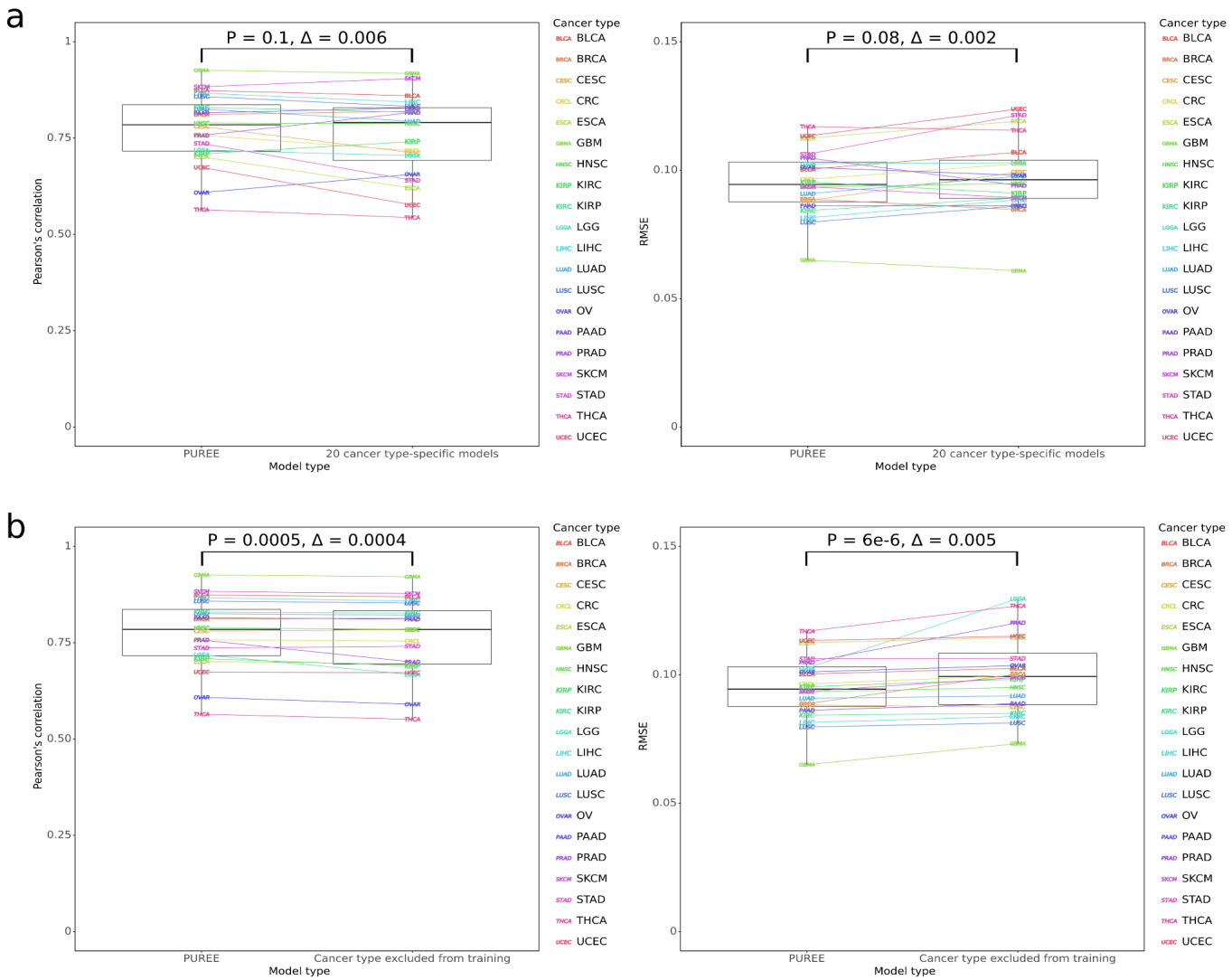


Figure 3: PUREE versus cancer type-specific architectures and unseen cancer types. a) Performance (Pearson's correlation and RMSE) of PUREE as compared to 20 cancer-type specific models on the test split of the TCGA dataset; **b)** Performance (Pearson's correlation and RMSE) of PUREE when predicting on a cancer type included in training data versus when the same cancer type was excluded from the training data. Differences were evaluated with the Wilcoxon signed-rank test (two-tailed), and delta is the difference between medians of metrics across all cancer types. P-values from Wilcoxon signed-rank test (two-tailed) shown.

Next, we evaluated the ability of PUREE to predict purity in cancer types absent from the training data (Methods). We compared performance metrics for PUREE and versions of PUREE where one cancer type was removed from the training data. This comparison showed only a minor decrease in correlation (median 0.7847 vs 0.7843, P=0.0005, Wilcoxon signed-rank test, two-tailed) and increase in RMSE (0.094 vs 0.099, P=6e-6) when the cancer type was absent

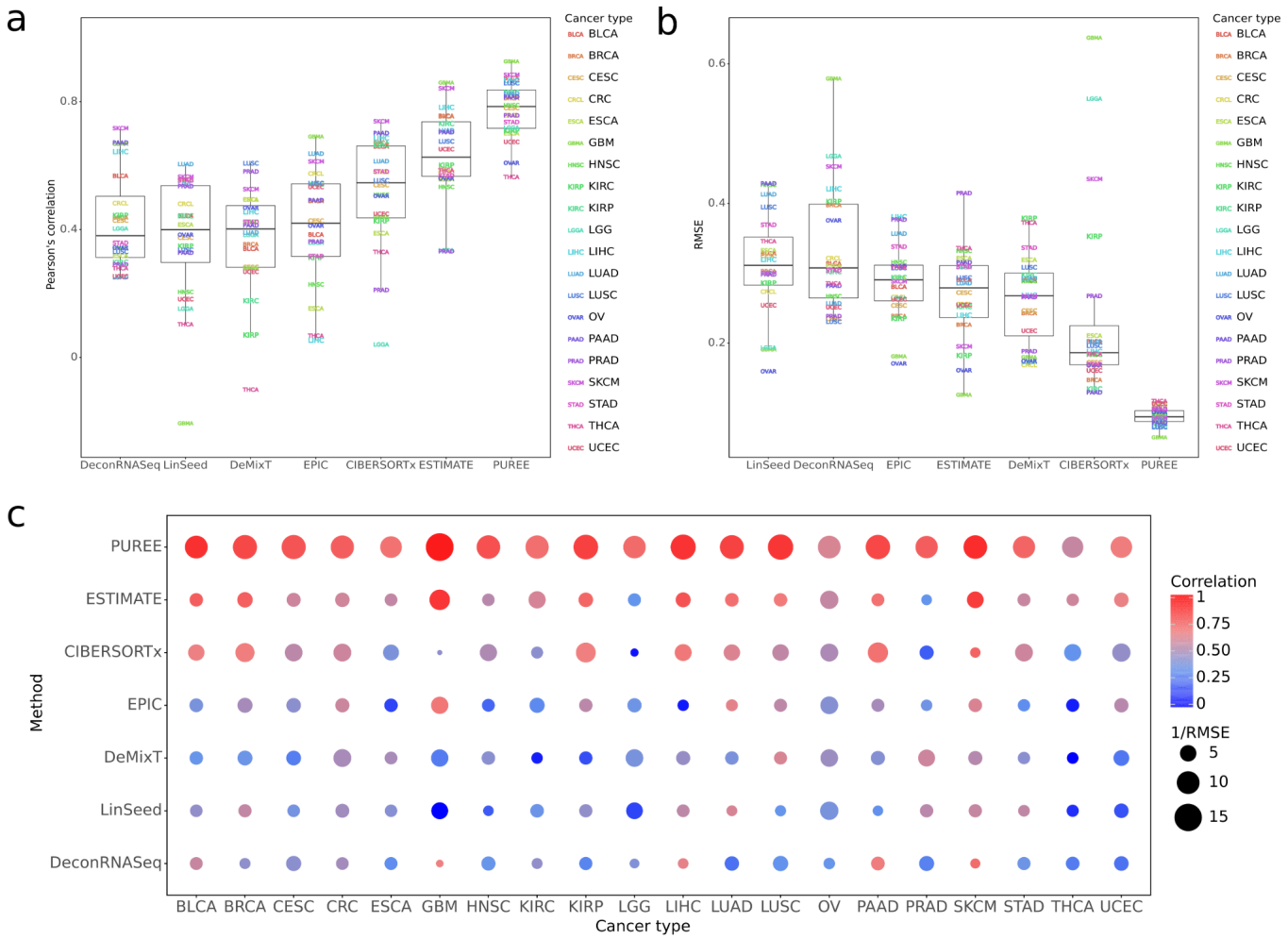
from the training data (Fig. 3b, Suppl. Fig. S7). This demonstrated that PUREE is robust and can generalize to solid tumor types not included in the training data, and that using the reduced feature set provides more robustness to the model.

Benchmarking of methods on independent datasets

We evaluated PUREE's performance on the withheld test sets from the TCGA dataset. We compared PUREE with six existing transcriptomics-based deconvolution and purity estimation methods (Methods).

Here, PUREE consistently demonstrated higher correlation and lower RMSE with consensus purity labels than the existing deconvolution methods (Fig. 4a,b, Suppl. Fig. S8). PUREE had the highest median correlation (r=0.78), followed by ESTIMATE (0.63) and CIBERSORTx (0.55). Similarly, PUREE had the lowest median RMSE of all methods (0.09), 53% lower than the next-best method (CIBERSORTx, 0.19), and PUREE displayed the lowest RMSE in each cancer type. PUREE also showed less variation in performance across cancer types as compared to the other methods, with an inter-quartile range for correlation and RMSE of 0.12 and 0.015, respectively. We additionally evaluated the performance of methods when tested across solid tumor types with likely distinct stromal composition (e.g. brain cancers and skin cancers). Consistent with our previous observations, PUREE outperformed other transcriptomics-based approaches, showing comparable high accuracy across cancer types with expected dissimilar stromal composition (Suppl. Fig. S9). A similar analysis showed that PUREE outperformed the other methods on the cancer types with extreme median tumor purities (Suppl. Fig. S10).

Next, we compared PUREE's and other methods' performance on two additional independent public lung cancer cohorts (35,36), a colorectal cancer cohort (37), and 4 TCGA cohorts of colorectal, uterine endometrial, pheochromocytoma and paraganglioma, and testicular cancer not present in the initial TCGA dataset used for model training and testing (Suppl. Tab. S2, Methods). Similar to the TCGA cohort, orthogonal genomics-based tumor purity estimates in these cohorts were estimated from tumor DNA sequencing data (Methods). Across all seven cohorts, PUREE demonstrated generally higher correlation and lower RMSE with the genomics-based tumor purity estimates (Fig. 5, Suppl. Fig. S11). We additionally compared PUREE's resource usage against the other methods in terms of memory (RAM) and compute time. This evaluation showed that PUREE uses less memory to run, in addition to consistently being the fastest method (Suppl. Fig. S12).



<u>Figure 4:</u> Comparing PUREE and existing methods on TCGA data. a,b,c) Comparison of PUREE and 6 existing transcriptome deconvolution methods. Mean Pearson's correlation (a) and RMSE (b) of methods per cancer type when compared with genomic tumor purity estimates on the TCGA test data split (1573 samples); (c) shows aggregated mean correlation and RMSE metrics per cancer type for each method.

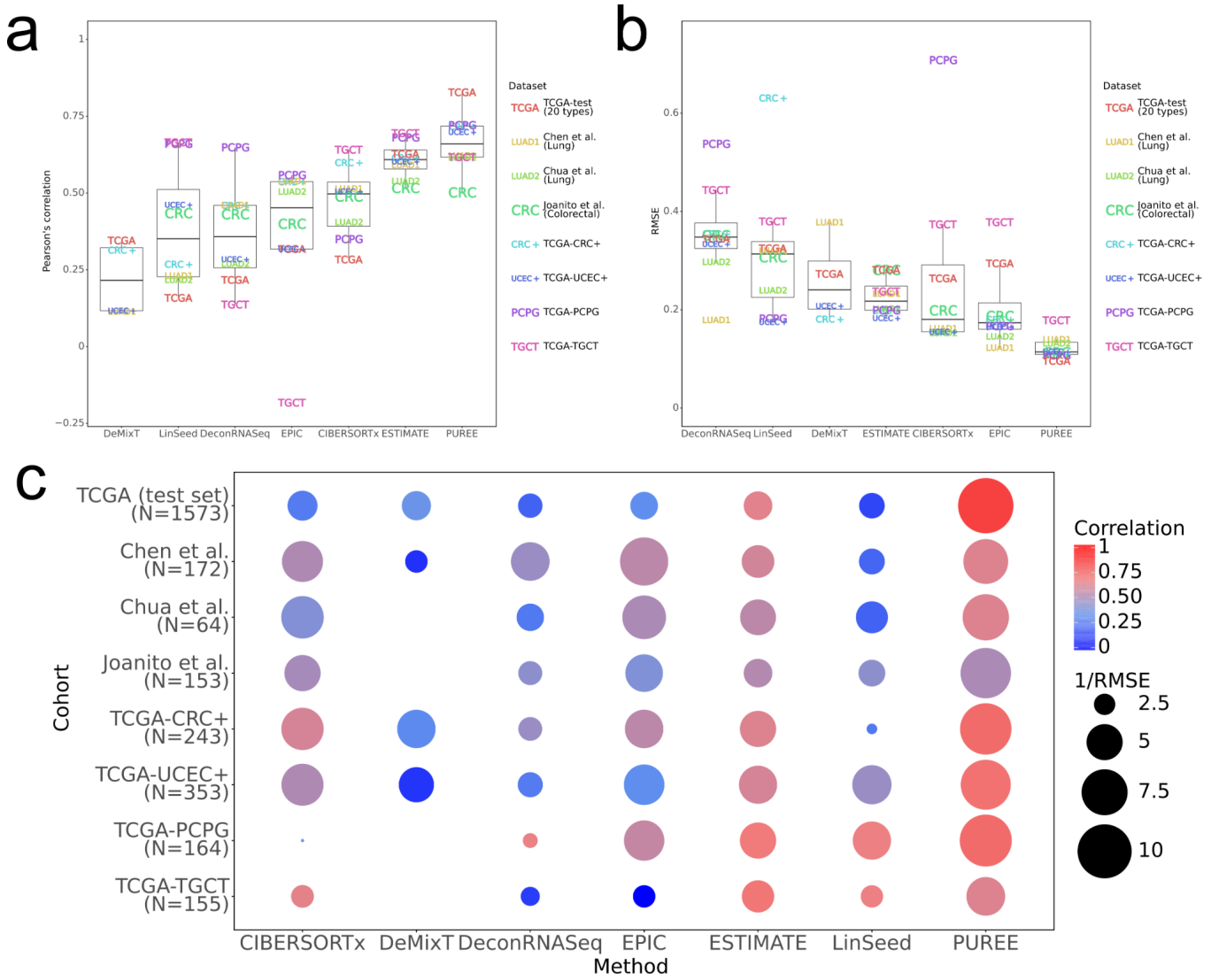


Figure 5: Benchmarking of methods on independent datasets. Comparing the purity estimates of PUREE and 6 existing transcriptome deconvolution methods. The predictions from each method were compared with genomic tumor purity estimates across 7 cohorts: the TCGA test set, lung cancer samples from Chen et. al (35), lung cancer samples from Chua et. al (36), a colorectal cancer cohort from Joanito et al. (37), and 4 TCGA cohorts of colorectal (CRC+), uterine endometrial (UCEC+), pheochromocytoma and paraganglioma (PCPG), and testicular cancer (TGCT) not present in the initial TCGA dataset used for training and testing. **a, b)** Mean Pearson's correlation (a) and RMSE (b) per cohort for all methods. **c)** Mean correlation and RMSE per cohort of every method shown together. DeMixT could not be run on the Chua et al., Joanito et al., PCPG and TGCT cohorts due to the absence of normal samples there.

Exploring the PUREE feature set using single-cell RNA-seq data

We performed orthogonal analysis and validation of the 158 gene-features in the PUREE model using single-cell RNA-seq data. We used published scRNA-seq data from head and neck cancer (38) (5902 cells total, 2539 classified as malignant, 3363 as non-malignant) and melanoma (39) (4513 cells total, 3256 classified as malignant, 1257 as non-malignant). We computed mean cell-wise z-scores of the expression of the genes with positive and negative

purity-expression correlation. Interestingly, these genes showed noticeable expression differences between malignant and non-malignant cells in both tumor types (Fig. 6, Mann-Whitney P<1e-90; Suppl. Fig. S13). Genes with positive purity-expression correlation had markedly higher expression in malignant cells as compared to non-malignant cells. In contrast, genes with negative purity-expression correlation were upregulated in non-malignant cells. This result further confirmed that the gene feature set used by PUREE has the ability to distinguish between and quantify the proportion of cancer and non-cancer cells in the tumors.

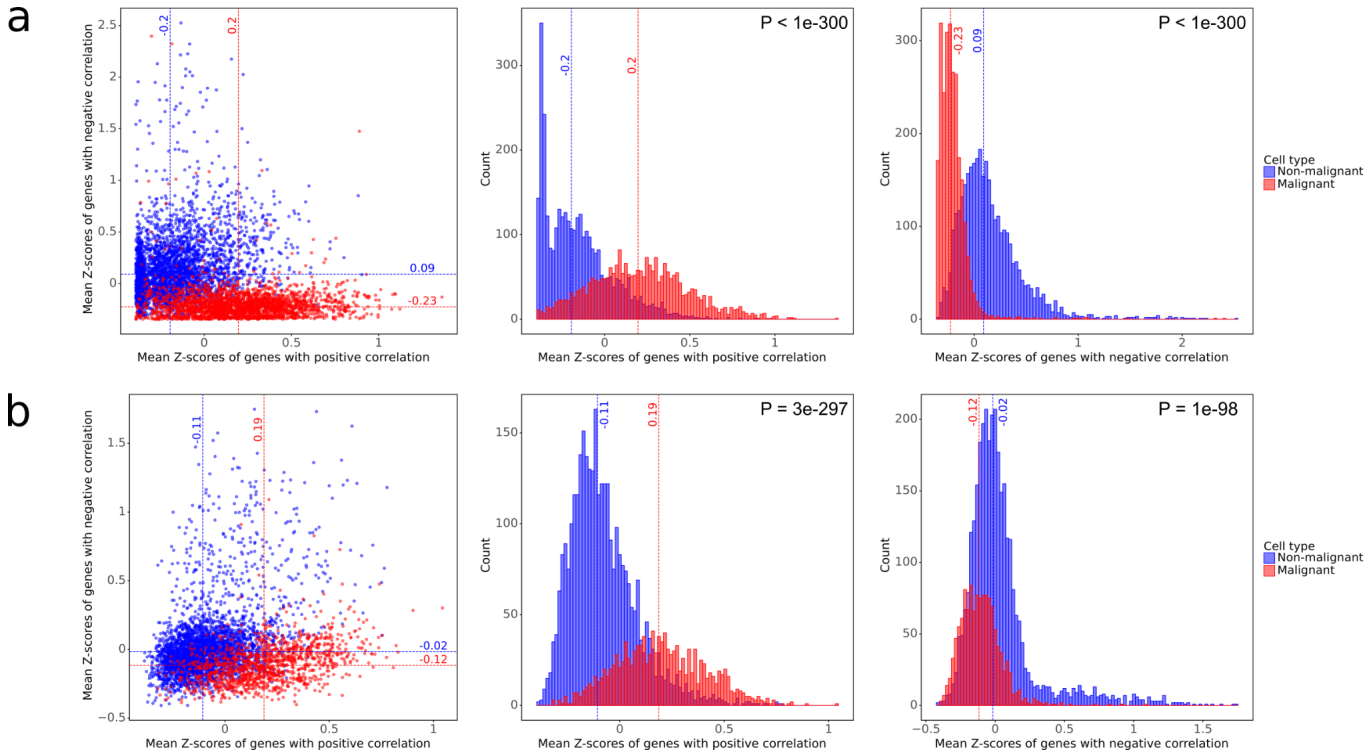


Figure 6: Validation of PUREE's features using single-cell RNA-seq data. Comparing the expression of PUREE gene features in malignant and non-malignant cells in **a)** head and neck cancer cells from Puram et al. (38) and **b)** melanoma cells from Tirosh et al. (39). Malignant and non-malignant cell labels were obtained from the original publications. Genes were separated into groups of positive (top 20%) and negative (bottom 20%) purity-vs-expression correlation in TCGA data averaged across cancer types. The dashed red and blue lines indicate medians of non-malignant and malignant cells' Z-scores distributions respectively. P-values are calculated using the Mann-Whitney U rank test (two-tailed).

Discussion and Conclusions

We developed a novel method, PUREE, that can predict the proportion of cancer cells in a sample (tumor purity) from a bulk tumor gene expression profile. The method adopts a machine learning-based feature selection strategy in combination with a linear regression architecture. The model is trained using weak supervision and consensus tumor purity labels obtained from tumor DNA sequencing data spanning 20 solid tumor types and a range of median genomics-based purities from 35% in pancreatic cancer to 72% in ovarian cancer. Compared with existing transcriptome deconvolution methods, we demonstrate that PUREE has superior accuracy across multiple independent test cohorts, spanning median purity ranges from 45% to 76%. Additionally, PUREE is fast and user friendly as the underlying model is pre-trained in advance.

PUREE adopts a stringent feature selection strategy, with which we were able to reduce the initial feature set of 9554 genes to 158 predictive genes. Combined with a pan-cancer training strategy, we demonstrate that this sparse model can generalize well to unseen tumor types and cohorts. We also show that PUREE's pan-cancer architecture has comparable accuracy to cancer-type specific models. This suggests that the model is able to capture expression signatures of cancer and stromal cells conserved across solid tumor types, and that cancer-type specific expression signatures do not provide additional discriminatory information for tumor purity estimation. We found that these pan-cancer conserved feature genes were enriched for known cancer and stromal cell specific processes such as epithelial-mesenchymal transition and immune cell activity.

The gene feature set used by PUREE also demonstrated a remarkable ability to distinguish between malignant and non-malignant cells in single-cell RNA-seq data. This provides orthogonal validation of our feature selection and pan-cancer training strategy as well as further confirming the predictive power of the selected gene set. This result also suggests that a modified version of our approach could potentially be repurposed to classify malignant and non-malignant cells in single-cell RNA-seq data. Finally, due to its supervised machine learning approach, PUREE has some limitations stemming from the composition of the training data. Specifically, the method has only been trained and tested on solid tumor samples and will therefore likely have suboptimal performance if applied to other non-solid cancer types.

In summary, we have shown that PUREE is a highly accurate and efficient method for purity estimation from a tumor gene expression profile, enabling robust and accurate interrogation of tumor purity and heterogeneity from bulk tumor gene expression data. We envision PUREE to be especially useful in settings where the DNA-seq data is either hard to obtain or absent. Even when tumor DNA-seq data is available, PUREE may provide an additional and orthogonal approach to tumor purity estimation. This may be especially relevant in cohorts and settings where the DNA and RNA are extracted from different aliquots of a tumor.

Methods

Availability of data and materials

TCGA gene expression data on 20 solid cancer types was downloaded through the UCSC Xena Hub (40). The gene expression data of the lung cancer validation cohorts used in the external benchmark were obtained from the publication by Chen et al. (35) and Chua et al. (36). The gene expression data of the colorectal cancer cohort used in the external benchmark was obtained from the publication by Joanito et al. (37). The preprocesed gene expression data of 4 TCGA cohorts of colorectal, uterine, paraganglioma and testicular cancers were also obtained from the UCSC Xena Hub. The gene expression data and the cell type labels of the head and neck cancer single-cell cohort used in the feature set validation was taken from the publication by Tirosh et al. (Gene Expression Omnibus ID GSE72056) (39). The gene expression data and the cell type labels melanoma single-cell cohort used in the feature set validation was obtained from the publication by Puram et al. (Gene Expression Omnibus ID GSE103322) (38).

In the first lung cancer validation dataset from Chen et. al (35), genomic purity estimates were originally computed as a mean of THetA2 (41), TitanCNA (42), AbsCNSeq (18) and PurBayes (16) and obtained from the respective publication. In the second lung cancer validation dataset, Chua et al (36), genomic purity estimates were originally based on ASCAT (15) and Sequenza (43) methods and taken from the respective publication. In the colorectal cancer validation cohort from Juanito et al., genomic consensus purity values were re-computed as a mean of THetA2 (41), TitanCNA (42), AbsCNSeq (18) and PurBayes (16) methods. ABSOLUTE tumor purity estimates for TCGA samples were obtained from the NCI Genomic Data Commons (GDC) database (34). Only samples that had genomic tumor purity available were used.

Code availability

PUREE is available as a web service (https://puree.genome.sg/) and the respective Python package (https://github.com/skandlab/PUREE). The Python code used to generate the figures, conduct the methods benchmark in R, as well as the other minor Python functions are available at https://github.com/erevkov/PUREE_supplementary.

Genomics-based consensus tumor purity estimates

For TCGA samples, genomic-based consensus tumor purities were computed as a mean of predictions from ABSOLUTE (17), AbsCNSeq (18), ASCAT (15) and PurBayes (16) following the approach reported in Ghoshdastider et al. (44). AbsCNSeq and PurBayes estimates are based on mutation variant allele frequency data, and ASCAT and ABSOLUTE on SNP-array data. Briefly, samples with extremely low (<0.1) and extremely high (>0.98) purity estimates from individual methods were flagged as missing data, as recommended by Ghoshdastider et al. These missing data values were instead imputed using an iterative principal component analysis approach (45). Quantile normalization was used to standardize and average the tumor purity distributions of different algorithms per cancer type. Finally, consensus purity estimates were estimated as the sample-wise medians of the normalized purity estimates from individual methods.

TCGA training and test set construction

The TCGA dataset consisted of 7864 samples from 20 solid cancer types. 80% of samples were selected for model training (TCGA train split, 6291 samples) and 20% for testing (TCGA test split, 1573 samples). The training and test sets were randomly sampled so they had comparable cancer-type and purity distributions. The initial gene expression feature matrix was filtered to only include autosomal and protein-coding genes. Genes with low expression (median TPM<1) and low variance (variance<1) in all cancer types in the TCGA train split were further filtered, leaving 9554 gene expression features for subsequent steps (referred to as the 10K features set).

Gene expression data rank-transformation

Gene expression data is rank-percentile normalized (sample-wise) when serving as input for PUREE. The initial rank-transformation allows for generalization across different gene expression platforms (e.g. RNA-seq, microarray) and measurement units (e.g. Transcripts Per Million (TPM), Fragments Per Kilobase Million (FPKM)). Briefly, gene expression values (e.g. TPMs [0, 0, 1, 5, 100]) are first ranked based on their position within a sample in ascending order, assigning the lowest possible rank for tied groups ([1, 1, 3, 4, 5]). The percentile of the resulting rank is then computed ([0.2, 0.2, 0.6, 0.8, 1]). The resulting percentiles computed relative to the ranking universe of the 10K feature set serve as input values to PUREE.

Construction of machine learning models

All machine learning models (Elastic Net, Gradient Boosting, Support Vector Regression, Lasso, Logit Regression and Linear Regression models, as well as Simple Imputer for missing values imputation) were trained using the Scikit-Learn Python package (46). A fully connected Neural Net consisting of a variable number of relu-activated fully-connected hidden layers, depending on the feature size of the input data, was constructed using the Keras submodule of the Tensorflow Python package (47) (Supplementary).

Feature selection

We used Lasso regression to reduce the number of input features in PUREE. During the first step, we selected features relevant to both low and high purity ranges. As feature selection models, two lasso regression models, cross-validated using cancer types as folds, were trained on all training data except the bottom and top-20% purity values, respectively. These two resulting feature sets were intersected, resulting in 167 genes. Next, we further iteratively selected features equally relevant to all cancer types by training a Lasso model (cross-validated with cancer type as folds) on a balanced subset of the full TCGA training set comprising all purity ranges, using the earlier selected 167 genes as initial features. Briefly, this is done by training on samples from N-1 cancer types, and testing on samples from the remaining withheld cancer type (leave-one-cancer-type out). We balance this step of feature selection by selecting only 117 samples from each cancer type per fold, as it is the least number of samples per cancer type (GBM, N=117) found in the TCGA train split. This resulted in 158 genes with

non-zero weights, which serve as predictive features in the final PUREE model. Correlations of the genes' expression with genomic-based tumor purity were computed on the train portion of TCGA as means per each cancer type averaged across all cancer types.

Predicting on unseen samples

The PUREE pipeline consists of three parts: rank-percentile normalization, missing values imputation and linear regression model inference. During the first step, rank-percentiles of the overlap of the genes in the input data and the 10K genes are computed. The data is further reduced to the 158 PUREE input genes. During the second step, if there are any missing values in the 158 selected genes, they are imputed based on the medians of the values in TCGA train set. During the third step, the pre-trained linear regression model is applied to the resulting data in order to predict purity values. The predicted values that fall out of the [0,1] range are rounded to the nearest in-range value.

Cancer-type specific models and test experiments

For experiments with cancer type-specific models, we constructed training/test sets comprising the 10K features for each individual cancer type. We used 5-fold cross-validation to train lasso models on these cancer-type specific training sets. For experiments where individual cancer types were excluded during model training, we constructed a pan-cancer training set (158 features) comprising all cancer types except the cancer type being withheld, followed by linear regression and 5-fold cross-validation as described for the main pan-cancer PUREE model.

Gene set enrichment analysis

Gene set enrichment analysis was evaluated using the GSEApy Python package, which is based on Enrichr (48) and Gene Set Enrichment Analysis (GSEA) (49) (Supplementary).

Running other transcriptomics-based methods for purity prediction

CibesortX was run using the web interface provided by the authors. DeMixT, EPIC, ESTIMATE, DeconRNASeq and LinSeed were run using their respective R packages (Supplementary).

Statistics and reproducibility

To calculate P-values, Mann-Whitney U rank test was used for non-paired data (single-cell z-scores) and Wilcoxon signed-rank test for paired samples (cancer types). Pearson's correlation and root mean squared error were used to calculate the mean statistics for each cancer type or cohort.

TCGA cohort of 20 cancer types includes 7864 samples with 6291 in the train set and 1573 in the test set. Chen et al. lung cancer cohort consists of 172 samples. Chua et al. lung cancer cohort consists of 64 samples. Joanito et al. colorectal cancer cohort consists of 153 samples.

TCGA-CRC+ colorectal cancer cohort consists of 243 samples. TCGA-UCEC+ uterine endometrial cancer cohort consists of 353 samples. TCGA-PCPG pheochromocytoma and paraganglioma cohort consists of 164 samples. TCGA-TGCT cohort consists of 155 samples.

The head and neck single-cell RNA-seq cohort from Puram et al. consists of 5902 cells, 3363 of which are malignant and 2539 non-malignant. The melanoma single-cell dataset from Tirosh et al. consists of 4513 cells, 3256 of which are malignant and 1257 non-malignant (132 cells with unresolved cell type assignment were dropped).

List of abbreviations

FPKM - Fragments Per Kilobase Million

GSEA - Gene Set Enrichment Analysis

H&E - Hematoxylin and Eosin

RMSE - Root Mean Square Error

TCGA - The Cancer Genome Atlas

TME - Tumor Microenvironment

TPM - Transcripts Per Million

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Funding

This research is supported by the Singapore Ministry of Health's National Medical Research Council under its OF-IRG program (OFIRG18may-0075), and Agency for Science, Technology and Research (A*STAR) under its CDAP program (grant no. 1727600057).

Author contributions

A.J.S. and E.R. conceived the idea and the experiments, E.R. conducted the experiments, developed the method, and wrote the manuscript. T.K. developed the web service and the Python API package. A.J.S. and K.S. supervised the method development and reviewed the manuscript.

Acknowledgments

During the method's development, E.R. was a student funded by the SINGA-A*STAR scholarship and is grateful for the financial support provided for the research above. The authors

would also like to express gratitude to Sim Ngak Leng for the help with relevant bioinformatics analysis, Kiran Krishnamachari for the extensive discussion on the manuscript, Sinem Kadioglu and Yu Amanda Guo for testing the method, and members of the Skanderup lab for other helpful inputs and overall support of the project.

Competing interests

The authors declare no competing interests.

Bibliography

1. Fridman WH, Pagès F, Sautès-Fridman C, Galon J. The immune contexture in human tumours: impact on clinical outcome. Nat Rev Cancer. 2012 Apr;12(4):298–306.
2. Egeblad M, Nakasone ES, Werb Z. Tumors as Organs: Complex Tissues that Interface with the Entire Organism. Dev Cell. 2010 Jun;18(6):884–901.
3. Avila Cobos F, Vandesompele J, Mestdagh P, De Preter K. Computational deconvolution of transcriptomics data from mixed cell populations. Bioinformatics. 2018;34(11):1969–79.
4. Joyce JA, Pollard JW. Microenvironmental regulation of metastasis. Nat Rev Cancer. 2009;9(4):239–52.
5. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat Commun. 2015 Dec;6(1):8971.
6. Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, et al. PyClone: Statistical inference of clonal population structure in cancer. Nat Methods. 2014 Apr;11(4):396–8.
7. Anagnostou V, Niknafs N, Marrone K, Bruhm DC, White JR, Naidoo J, et al. Multimodal genomic features predict outcome of immune checkpoint blockade in non-small-cell lung cancer. Nat Cancer. 2020 Jan;1(1):99–111.
8. Wang Z, Cao S, Morris JS, Ahn J, Liu R, Tyekucheva S, et al. Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration. iScience. 2018 Nov;9:451–60.
9. Ghoshdastider U, Rohatgi N, Naeini MM, Baruah P, Revkov E, Guo YA, et al. Pan-Cancer Analysis of Ligand-Receptor Cross-talk in the Tumor Microenvironment. 2021; Available from: http://cancerres.aacrjournals.org/
10. Rohatgi N, Ghoshdastider U, Baruah P, Kulshrestha T, Skanderup AJ. A pan-cancer metabolic atlas of the tumor microenvironment. Cell Rep. 2022 May 10;39(6):110800.
11. Smits AJJ, Kummer JA, De Bruin PC, Bol M, Van Den Tweel JG, Seldenrijk KA, et al. The estimation of tumor cell percentage for molecular testing by pathologists is not accurate. Mod Pathol. 2014 Feb;27(2):168–74.
12. Dentro SC, Wedge DC, Van Loo P. Principles of Reconstructing the Subclonal Architecture of Cancers. Cold Spring Harb Perspect Med. 2017 Aug;7(8):a026625.
13. Chakravarthy A, Furness A, Joshi K, Ghorani E, Ford K, Ward MJ, et al. Pan-cancer deconvolution of tumour composition using DNA methylation. Nat Commun. 2018 Aug 13;9(1):3220.
14. Qin Y, Feng H, Chen M, Wu H, Zheng X. InfiniumPurify: An R package for estimating and accounting for tumor purity in cancer methylation research. Genes Dis. 2018 Mar 1;5(1):43–5.
15. Van Loo P, Naume B, Sun W, Weigman VJ, Marynen P, Borresen-Dale AL, et al. Allele-specific copy number analysis of tumors. Proc Natl Acad Sci. 2010;107(39):16910–5.
16. Larson NB, Fridley BL. PurBayes: Estimating tumor cellularity and subclonality in next-generation sequencing data. Bioinformatics. 2013;29(15):1888–9.
17. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012 May;30(5):413–21.
18. Bao L, Pu M, Messer K. AbsCN-seq: A statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data. Bioinformatics. 2014;30(8):1056–63.
19. Haider S, Tyekucheva S, Prandi D, Fox NS, Ahn J, Xu AW, et al. Systematic Assessment of Tumor Purity and Its Clinical Implications. JCO Precis Oncol. 2020 Nov;(4):995–1005.
20. Wang Y, Mashock M, Tong Z, Mu X, Chen H, Zhou X, et al. Changing Technologies of RNA Sequencing and Their Applications in Clinical Oncology. Front Oncol. 2020 Apr;10:447.

21. Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA, et al. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. Cell. 2016 Jan 28;164(3):550–63.
22. Bailey P, Chang DK, Nones K, Johns AL, Patch AM, Gingras MC, et al. Genomic analyses identify molecular subtypes of pancreatic cancer. Nature. 2016 Mar;531(7592):47–52.
23. Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. Nat Med. 2015 Nov;21(11):1350–6.
24. Moffitt RA, Marayati R, Flate EL, Volmar KE, Loeza SGH, Hoadley KA, et al. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. Nat Genet. 2015 Oct;47(10):1168–78.
25. Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature. 2016 Jun;534(7605):55–62.
26. Dudley JC, Gurda GT, Tseng LH, Anderson DA, Chen G, Taube JM, et al. Tumor Cellularity as a Quality Assurance Measure for Accurate Clinical Detection of BRAF Mutations in Melanoma. Mol Diagn Ther. 2014 Aug;18(4):409–18.
27. Riaz N, Havel JJ, Makarov V, Desrichard A, Urba WJ, Sims JS, et al. Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. Cell. 2017 Nov 2;171(4):934-949.e16.
28. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013 Dec;4(1):2612.
29. Racle J, Gfeller D. EPIC: A tool to estimate the proportions of different cell types from bulk gene expression data. In: Methods in Molecular Biology [Internet]. Humana Press Inc.; 2020. p. 233–48. Available from: https://pubmed-ncbi-nlm-nih-gov.ejproxy.a-star.edu.sg/32124324/
30. Zaitsev K, Bambouskova M, Swain A, Artyomov MN. Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. Nat Commun. 2019;10(1):1–16.
31. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol. 2019 Jul;37(7):773–82.
32. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015 May;12(5):453–7.
33. Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics. 2013 Apr 15;29(8):1083–5.
34. The Cancer Genome Atlas Program - National Cancer Institute [Internet]. [cited 2021 Jun 23]. Available from: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
35. Chen J, Yang H, Teo ASM, Amer LB, Sherbaf FG, Tan CQ, et al. Genomic landscape of lung adenocarcinoma in East Asians. Nat Genet. 2020 Feb;52(2):177–86.
36. Chua KP, Teng YHF, Tan AC, Takano A, Alvarez JJS, Nahar R, et al. Integrative Profiling of T790M-Negative EGFR-Mutated NSCLC Reveals Pervasive Lineage Transition and Therapeutic Opportunities. Clin Cancer Res. 2021 Nov 1;27(21):5939–50.
37. Joanito I, Wirapati P, Zhao N, Nawaz Z, Yeo G, Lee F, et al. Single-cell and bulk transcriptome sequencing identifies two epithelial tumor cell states and refines the consensus molecular classification of colorectal cancer. Nat Genet. 2022 Jul;54(7):963–75.
38. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell. 2017 Dec 14;171(7):1611-1624.e24.
39. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the

multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016 Apr 8;352(6282):189–96.
40. Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol. 2020 Jun;38(6):675–8.
41. Oesper L, Mahmoody A, Raphael BJ. THetA : Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. Genome Biol. 2013;14(7):1–41.
42. Ha G, Roth A, Khattra J, Ho J, Yap D, Prentice LM, et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. Genome Res. 2014 Nov;24(11):1881–93.
43. Favero F, Joshi T, Marquard AM, Birkbak NJ, Krzystanek M, Li Q, et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Ann Oncol. 2015 Jan;26(1):64–70.
44. Ghoshdastider U, Rohatgi N, Mojtabavi Naeini M, Baruah P, Revkov E, Guo YA, et al. Pan-cancer analysis of ligand-receptor crosstalk in the tumor microenvironment. Cancer Res. 2021 Feb;canres.2352.2020.
45. Josse J, Husson F. missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. J Stat Softw. 2016 Apr 4;70:1–31.
46. Pedregosa F, Michel V, Grisel O, Blondel M, Prettenhofer P, Weiss R, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.
47. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems [Internet]. 2015. Available from: https://www.tensorflow.org/
48. Xie Z, Bailey A, Kuleshov MV, Clarke DJB, Evangelista JE, Jenkins SL, et al. Gene Set Knowledge Discovery with Enrichr. Curr Protoc. 2021;1(3):e90.
49. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature. 2009 Nov;462(7269):108–12.