www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

Compendiums of cancer transcriptomes for machine learning applications

Received: 15 February 2019
Accepted: 25 July 2019
Published online: 08 October 2019

Su Bin Lim[1,2], Swee Jin Tan[3], Wan-Teck Lim[4,5,6] & Chwee Teck Lim[1,2,7,8]

There are massive transcriptome profiles in the form of microarray. The challenge is that they are processed using diverse platforms and preprocessing tools, requiring considerable time and informatics expertise for cross-dataset analyses. If there exists a single, integrated data source, data-reuse can be facilitated for discovery, analysis, and validation of biomarker-based clinical strategy. Here, we present merged microarray-acquired datasets (MMDs) across 11 major cancer types, curating 8,386 patient-derived tumor and tumor-free samples from 95 GEO datasets. Using machine learning algorithms, we show that diagnostic models trained from MMDs can be directly applied to RNA-seq-acquired TCGA data with high classification accuracy. Machine learning optimized MMD further aids to reveal immune landscape across various carcinomas critically needed in disease management and clinical interventions. This unified data source may serve as an excellent training or test set to apply, develop, and refine machine learning algorithms that can be tapped to better define genomic landscape of human cancers.

Background & summary

The Cancer Genome Atlas (TCGA) increasingly serves as a 'training' reference to apply machine learning algorithms, having comprehensive, well-curated genomic data of over 11,000 tumors across 33 major cancer types. In recent years, this rich resource combined with machine learning has facilitated the development of cancer classifier[1], markers predictive of drug sensitivity[2], histopathology image-based prognostic predictor[3], and novel indices associated with oncogenic dedifferentiation[4]. There also exist vast datasets deposited at the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) in the form of microarray. Applying machine learning to exploit them, however, is not straightforward; they are often generated using diverse platforms and normalization tools, and are annotated with non-standardized texts and definitions. All of these features add computational complexity to the existing high-dimensional data, necessitating multiple and intricate analytics tools for data integration and analysis.

To increase the reuse of such legacy data, we generated single, merged microarray-acquired datasets (MMD) for 11 major cancer types using a uniform R pipeline (Fig. 1). This approach has been used in our earlier work to generate merged transcriptome data of a specific cancer type, non-small cell lung cancer (NSCLC), comprising both non-tumor (NT) and tumor tissue (TT) samples[5]. The resulting MMD was used to develop a predictive multi-gene classifier, termed as tumor matrisome index (TMi), for prognosis and prediction of response to adjuvant chemotherapy among NSCLC patients[6].

Here, we extend the framework to include various carcinomas of epithelial origin. Consistent with prior works[7–11], comparably correlated patterns of genome-wide differential expression (DE) were observed between microarray (MMD) and RNA-seq (TCGA). Next, we demonstrate the potential application of MMD as training data to develop clinical predictive models that can be applied cross platform. By applying CIBERSORT[12], we

[1]NUS Graduate School for Integrative Sciences & Engineering, National University of Singapore, Singapore, Singapore. [2]Department of Biomedical Engineering, National University of Singapore, Singapore, Singapore. [3]Regional Scientific Affairs, Sysmex Asia Pacific, Singapore, Singapore. [4]Division of Medical Oncology, National Cancer Centre Singapore, Singapore, Singapore. [5]Office of Academic and Clinical Development, Duke-NUS Medical School, Singapore, Singapore. [6]IMCB NCC MPI Singapore Oncogenome Laboratory, Institute of Molecular and Cell Biology (IMCB), A*STAR, Singapore, Singapore. [7]Mechanobiology Institute, National University of Singapore, Singapore, Singapore. [8]Institute for Health Innovation and Technology (iHealthtech), National University of Singapore, Singapore, Singapore. Correspondence and requests for materials should be addressed to C.T.L. (email: ctlim@nus.edu.sg)

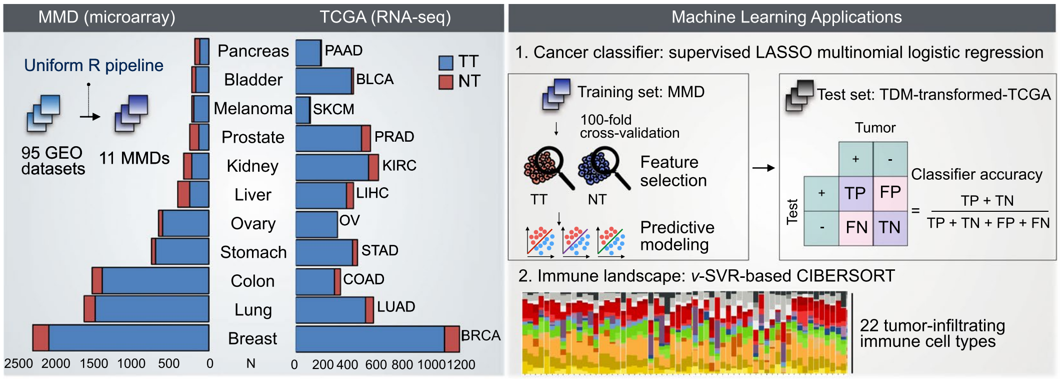

Fig. 1 MMD: development, validation, and potential applications in oncology. Microarray-based datasets containing raw transcriptome profiles of patient-derived tumor tissues (TT) and non-tumor (NT) tissues were processed, merged, and batch-effect corrected using an integrated R pipeline. Validation of each cancer type-specific MMD was performed using PCA and RRHO algorithms. Clinical models trained using MMD can be applied to TCGA, facilitating the discovery of new biomarkers, development of prognostic models, and parallel cross-platform analyses with TCGA.

further show how MMDs can be used to de-convolve tumor immune microenvironment by parsing specific sub-populations of infiltrating immune cell, comparatively with TCGA datasets of matching cancer types.

Through pan-cancer analysis of MMDs, we recently identified clinically significant matrisomal changes associated with immune response and targetable immune checkpoints for a subset of cancers across different malignancies[13]. The generated cancer type-specific MMDs, the associated clinical metadata and R codes are available at ArrayExpress and figshare (see Data Records and Code Availability). Our open resource of curated large-scale transcriptomic data may provide the basis for the analytical and computational techniques to derive unbiased and new information, enabling predictive modeling for precision oncology.

Methods

MMD generation. A careful GEO search (http://www.ncbi.nlm.nih.gov/geo) was done to ensure the selection of MIAME compliant datasets having the following attributes in the original GEO submission: (1) raw data in CEL files, (2) tissue origin annotation (i.e., NT or TT), and (3) Affymetrix platform annotation. Here, only datasets generated using the GPL570 platform (Affymetrix Human Genome U133 Plus 2.0 Array) were specifically selected to ensure uniform curation of the same probe-sets (i.e., 54,675 probes). Altogether, 95 independent GEO datasets comprising a total of 8,386 samples spanning over 11 cancer types were subjected to pre-processing, normalization, batch-effect correction, data integration and analyses (Table S1). The number of NT and TT samples in each GEO dataset is summarized in Table S2.

Raw expression data from each dataset was first imported and loaded into R Bioconductor[14] (RStudio version 1.1.447) using the affy package (version 1.48.0)[15]. The ReadAffy function was called with default parameters to read all CEL files, except for the function argument "cdfname" which was set to "hgu133plus2". The rma function was subsequently used to normalize and background correct all the annotated probe-sets-derived expression data. This preprocessing step was applied to all 95 datasets for uniform processing and feature annotation prior to merging based on cancer type. Batch effects were identified and removed using ComBat via the inSilicoMerging package (version 1.14.0)[16]. Probes having maximum mean expression values across samples in each MMD were collapsed to the genes, and were annotated using the hgu133plus2SYMBOL object in the hgu133plus2.db package (version 3.2.2)[17] for subsequent DE analysis.

TCGA datasets. The Cancer Genome Atlas (TCGA) data were retrieved and processed via the TCGA-Assembler package (version 2.0)[18] (Table S1). Normalized RPKM count values were extracted using the ProcessRNASeqData function via the TCGA-Assembler package (version 2.0)[18]. Only genes with at least 1 count per million (cpm) or RPMK value in at least 20% of total number of samples in each cohort were kept via the edgeR package (version 3.12.1)[19]. The number of genes filtered out in each TCGA dataset is summarized in Table S3. Selected genes were normalized by Trimmed Mean of M-values (TMM), and were subjected to DE analyses using the voom and lmFit functions in the limma package (version 3.26.9)[20]. Of note, ovarian (OV) and melanoma (SKCM) TCGA cohorts were excluded in DE and RRHO analyses due to lack of NT samples (Table S1). Clinical data including disease status (NT vs. TT) were downloaded via the DownloadBiospecimenClinicalData function in the TCGA-Assembler package (version 2.0)[18].

PCA, DE and RRHO analysis. Principal component analysis (PCA) was performed using the prcomp function in the built-in R stats package (version 3.2.2). The first two PCs were visualized using the ggbiplot package (version 0.55)[21]. The lmFit and eBayes functions in the limma package (version 3.26.9)[20] were used to perform DE analysis. All genes annotated in each MMD and TCGA dataset were ranked by log fold change (logFC) computed based on their DE between NT and TT samples. These ranked lists were further reconstructed to only include genes that were common to both MMD- and TCGA-derived lists[22] (Table S3). These files were loaded into a web-based executable simplified version of rank-rank hypergeometric overlap (RRHO) tool (http://systems.crump.ucla.edu/rankrank/rankranksimple.php). In all cases, the step size was set to 300 to generate Benjamin-Yekutieli corrected hypergeometric matrix and RRHO heatmaps.

Multi-gene classifiers. Expression data of TMi and other gene signatures of commercially available or previously validated multi-gene tests (MGTs) were extracted from all TT samples across MMD and TCGA datasets, and were loaded into Morpheus (http://software.broadinstitute.org/morpheus/) for sample stratification. The list of MGT genes and the associated references are summarized in Table S4. K-means clustering was performed with "one minus pearson correlation" metric and 1,000 iterations.

CIBERSORT. Consisting of over 1,500 samples, breast, colon, and lung MMDs exceeded the load capacity (500MB) of the CIBERSORT analysis (http://cibersort.standford.edu/)[12]. 1,000 samples were thus randomly selected to generate the input "mixture" file for these MMDs. All samples in the rest of MMDs were included in the CIBERSORT analysis. Each run was performed with a default LM22 (22 immune cell types) gene signature using 100 permutations. The resulting immune cell profiles were used to compute the mean fractions of 22 immune cell types and the quantitative change between the two groups (NT vs. TT), denoted as delta (TT – NT, %), per dataset.

ROC analysis. A summary of four MGTs applied to MMDs, including gene signatures, the associated references, computation method for respective prognostic index, is provided in Table S5. Diagnostic accuracy of MGTs in classifying TT from NT samples was evaluated through the receiver operating characteristic (ROC) analysis. The area under the ROC curve (AUC), sensitivity, and specificity with the optimal cutoff for respective prognostic index were computed using the pROC package (version 1.10.0)[23].

Data Records

Our 11 MMDs are available at ArrayExpress for lung[24], pancreas[25], prostate[26], kidney[27], stomach[28], colon[29], ovary[30], breast[31], liver[32], bladder[33], and melanoma cancer[34].

Technical Validation

Principal component analysis (PCA). PCA was performed to assess the performance of ComBat in correcting batch effects, as previously described[6,35]. The first two PCs that capture the most variance are shown for both untransformed and ComBat-transformed datasets (Fig. 2). Batch-effect corrected MMDs exhibit an apparent overlay of PCs colored by the study (i.e., original dataset), and are separated by the disease status (i.e., NT vs. TT), demonstrating successful adjustment of batch effects arising from independent datasets of different sources. The PCA plots of MMD data exclusively comprising TT samples further distinguished the two risk groups (TMi_{high} and TMi_{low}) stratified by a pan-cancer multi-gene TMi classifier (Fig. S1; see Methods).

Differential expression (DE) analysis. Prior to in-depth genome-wide DE analysis, expression levels of cancer-related genes and three reference genes (i.e., GAPDH, UBB, and ACTB) were compared between the two groups (NT vs. TT) using MMDs. The selected housekeeping genes are stably expressed across tissues to maintain cellular function, and are commonly used for normalization in transcriptomics studies. While expression levels of cancer-associated gene were significantly different between NT and TT samples, that of all reference genes were almost the same in the two groups across all cancer types, validating the robustness of ComBat in adjusting technical batch effects while maintaining biological variation across samples (Fig. S2).

All MMDs were next subjected to genome-wide, limma-based DE analysis to rank all the genes by logFC based on DE between NT and TT samples (see Methods). These ranked lists were used to generate volcano plots visually depicting differentially expressed genes that met our statistical threshold (i.e., absolute value of $\text{logFC} > 1$ and adjusted $\text{P-value} < 0.001$) in TT relative to NT samples (Fig. S3 and Table S5). To validate these results in an independent cohort of patients, we processed TCGA data of matching cancer types (see Methods), and applied the same methods to construct the list of differentially expressed genes.

Rank-rank hypergeometric overlap (RRHO) analysis. RRHO algorithm[36] was used to assess the overlap intensity between MMD- and TCGA-derived lists of genes ranked by DE between NT and TT samples per cancer type (Fig. 3). As compared to conventional single arbitrary cut-off-based approaches, RRHO heatmaps have been widely used to visually compare genome-wide DE patterns across different species and profiling platforms, without having to correct for batch effects for the two distinct data files[36,37]. A significant overlap was observed for lung, prostate, kidney, colon, breast, and liver cancer, for which RRHO map max ranged from 1083 for kidney cancer to 1592 for colorectal cancer (Fig. 3, top row). The weak correlation observed across pancreas, stomach, and bladder cancers between MMD and TCGA datasets is likely due to a relatively small number of tumor-free tissues available in respective TCGA datasets (Table S1).

To test whether this would indeed be the case, we utilized the TMi annotation (TMi_{high} or TMi_{low}) previously derived from MMD data exclusively comprising TT samples (Fig. S1), and further classified TMi group for all TCGA TT samples using the same approaches (Table S3; see Methods). Except for bladder cancer, RRHO map max increased significantly from 135 to 1014 for pancreatic cancer and 437 to 1203 for gastric cancer (Fig. 3, middle row). Similarly, highly concordant RRHO results were derived from TT subgroups stratified by other commercially available or previously validated cancer type-specific multi-gene classifiers (Fig. 3, bottom row; see Methods). These QC steps altogether demonstrate the robustness of our uniform workflow for cross-cancer analysis (Fig. S4).

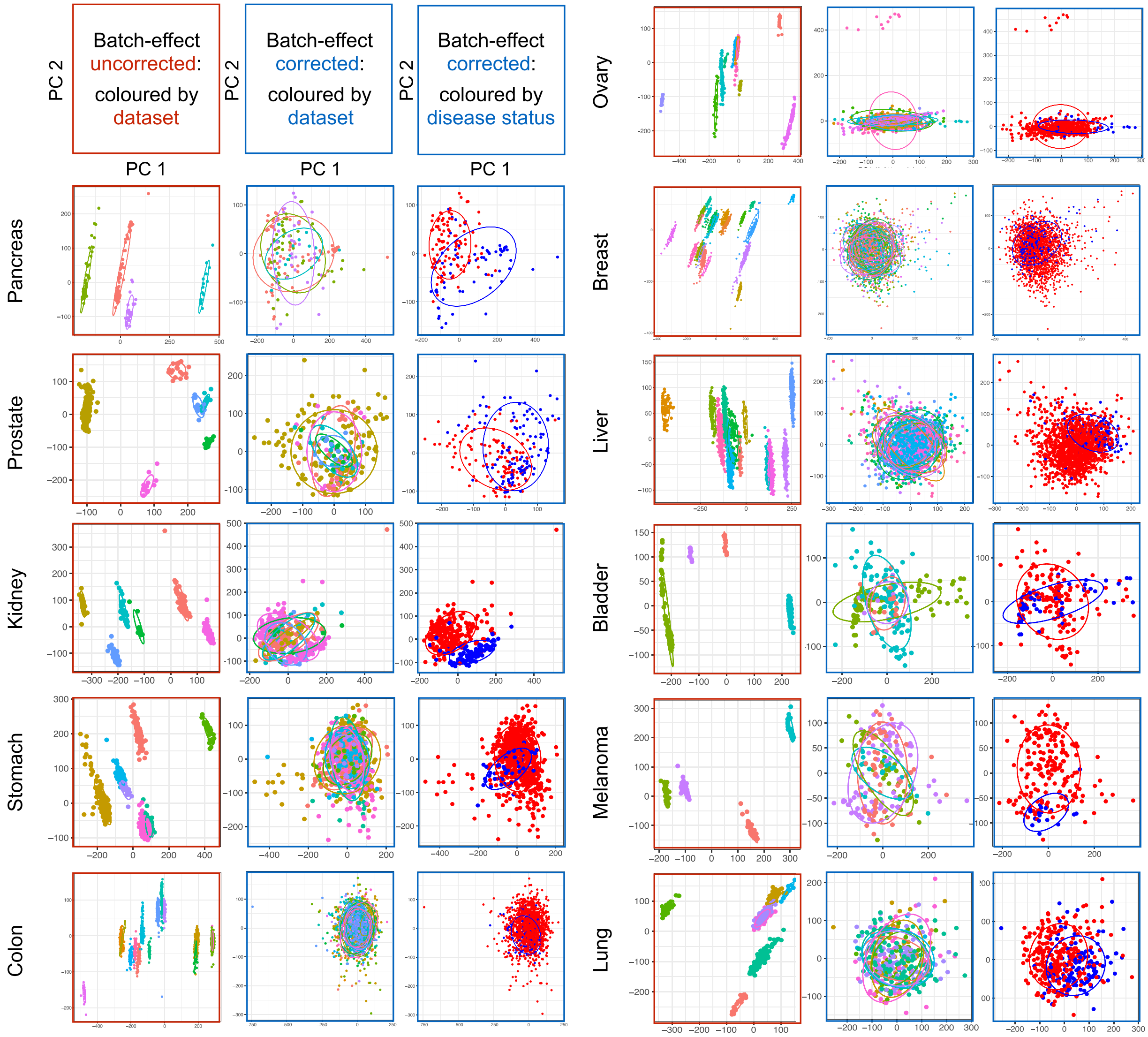

Fig. 2 QC metrics for MMDs. The first two PCs capturing the most variance are shown. PCA plots with red colored-border show PCs of merged data before batch-effect correction, which are colored by dataset (left). PCA plots with blue colored-border show PCs of merged data after Combat adjustment, which are colored by dataset (middle) and disease status (i.e., TT vs. NT; right). Ellipses are drawn one standard deviation away from the mean of the Gaussian fitted to each MMD.

Machine learning applications for predictive medicine. *Cancer classifier*. Publicly-accessible data repositories, such as GTEx[38], TCGA[39], HPA[40], and ArrayExpress[41], host genome-wide expression profiles assayed with various profiling technologies. Having sufficient read depth[10], higher resolution[11], higher dynamic range[42], and lower technical variation[43], RNA-seq is increasingly the platform of choice in translational-biomarker studies. Paralleling this trend, cross-platform normalization tools continue to be developed, facilitating comparison of data from different platforms. PREBS[44], VOOM[45], and TDM[42] are examplary techniques that are specifically designed to transform RNA-seq data to make it compatible with microarray data. Other conventional methods also exist in dealing with such 'dataset shifts'[46], such as quantile normalization, $\log_2$ transformation, and nonparanormal transformation[42].

Using supervised machine learning, we developed new cancer classifiers trained on MMDs, and evaluated their classifying performance on their respective RNA-seq-acquired TCGA datasets (Fig. 4a). Among the existing transformation methods, TDM transformation best fitted the reference MMD data distribution (Fig. 4b). Using the glmnet package (version 2.0.13)[47], we performed LASSO multinomial logistic regression[48] with 100 fold cross-validation (CV) to build best predictive model in distinguishing TT from NT samples. Predictive model built from each MMD was then tested directly on TDM-transformed-TCGA dataset. Except for breast MMD, all MMDs achieved an average AUC of 0.96 (ranging from 0.913 to 0.997) in classifying TCGA cancers (Fig. 4c). Other commercially available MGTs, including the Myriad myplan^{TM} Lung Cancer, Pervenio^{TM}, Oncotype DX and MammaPrint, further achieved the AUC ranging from 0.714 to 0.862 (Table S6, Fig. S5; see Methods).

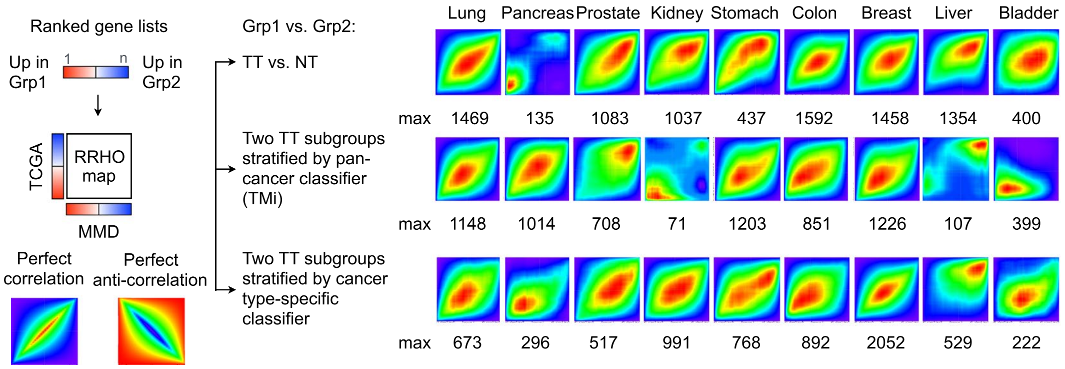

Fig. 3 Parallel genome-wide differential expression (DE) analyses with TCGA. Rank-rank hypergeometric overlap (RRHO) heatmaps are drawn to visualize the overlap intensity between MMD- and TCGA-derived lists of genes ranked by DE between the two groups: TT vs. NT group (top row), between the two TT subgroups classified by TMi (middle row) and by known cancer type-specific classifier (bottom row). RRHO map max values, denoted as max, are stated.

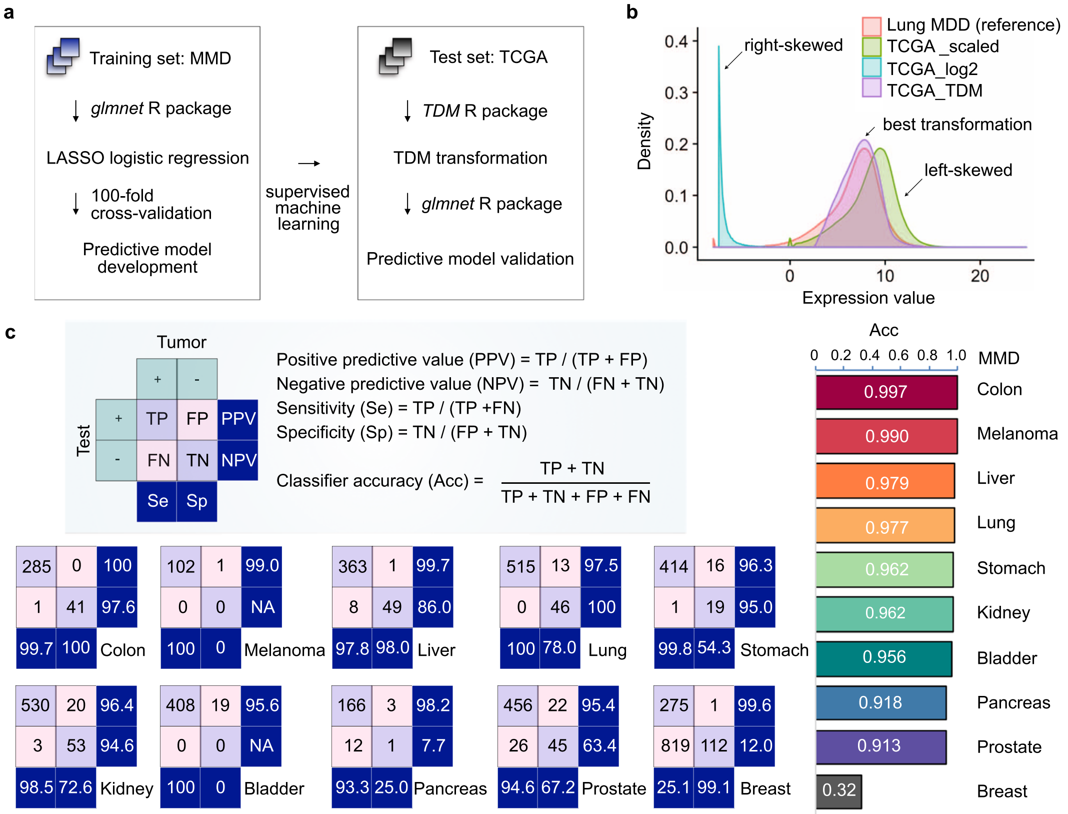

Fig. 4 Supervised machine learning classifies cancer. (**a**) Schematic workflow: cancer classifiers are built from MMDs, and are tested on TCGA of matching cancer types using LASSO logistic regression. (**b**) TDM-transformed testing data (TCGA LUAD) best fits the training data distribution (lung MDD). (**c**) Classifying accuracy of MMD-derived cancer classifier.

Pan-cancer immunogenomic analyses. TCGA data are increasingly being used to study the prognostic influence of the composition of tumor-infiltrating lymphocytes (TILs)[49,50], neoantigens[51,52] and immune cytolytic activity[53], all of which are putative markers predictive of clinical response to immune checkpoint inhibitor (ICI) treatments. The recent advancements in computational techniques have further facilitated high-resolution, large-scale immunogenomic analyses of the tumor-immune interface[54]. Of the developed analytical pipelines, CIBERSORT serves as an exemplary *in silico* deconvolution method to estimate the relative proportion of 22 immune cell populations from heterogeneous bulk tissues. By applying CIBERSORT to MMDs, we next tested if the generated

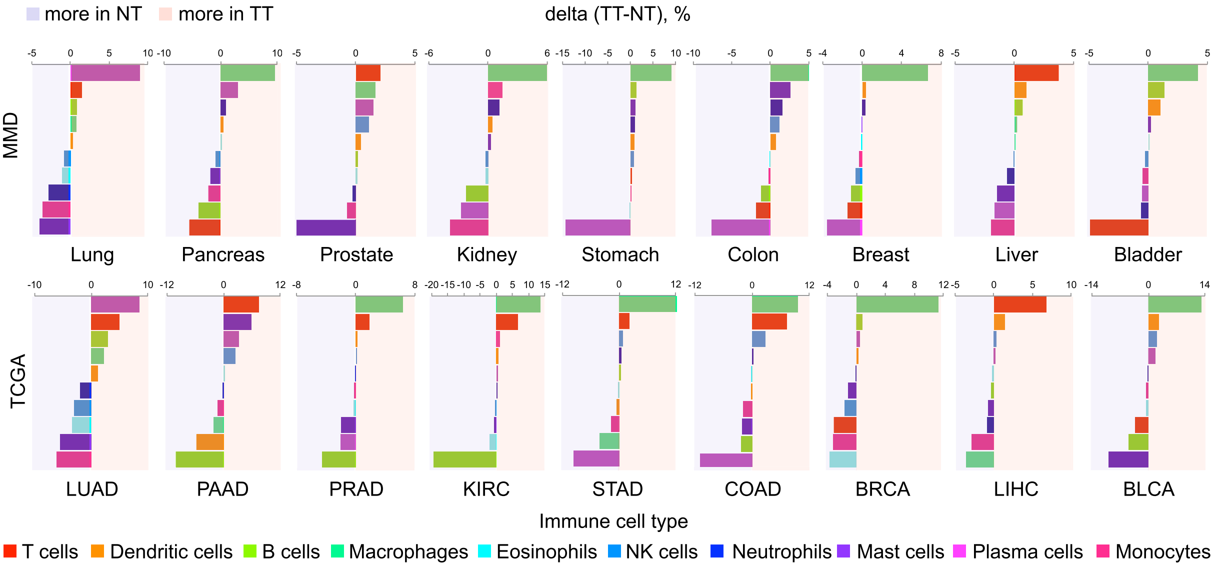

Fig. 5 Immune cell composition in NT and TT samples. Quantified changes of CIBERSORT-estimated fractions of immune cell populations between the two groups using MMD (top) and TCGA (bottom) datasets.

compendiums could further provide the basis for the developed computational infrastructure to reveal clinically significant immune landscape across multiple cancer types (see Methods).

The extent of difference in immune cell composition between the two groups (NT vs. TT) varied depending on cancer type (Fig. S6), where the estimated fractions were generally comparable (<5% difference). Specific immune cell types particularly enriched in either NT or TT group were identified, including plasma cells in lung cancer, T cells in liver cancer, and B cells in kidney, stomach, colon, breast, and bladder cancers (Fig. 5). Their enrichment was further observed in respective TCGA datasets, demonstrating the potential use of MMDs to reveal the degree and distribution of TIL density, which might be a clinically relevant prognostic and predictive indicator across various carcinomas[55,56].

Code Availability

The R codes used to preprocess, merge, and correct for batch-effects for generation of all 11 cancer type-specific MMDs can be found in figshare (https://doi.org/10.6084/m9.figshare.7878086)[22]. The exemplary R codes and metadata used to develop clinical predictive models using lung MMD[57] are described in our earlier works[5,6,58].

References

1. Yuan, Y. *et al*. DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations. *BMC bioinformatics* **17**, 476 (2016).
2. Lee, S. I. *et al*. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nat Commun* **9**, 42 (2018).
3. Yu, K. H. *et al*. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat Commun* **7**, 12474 (2016).
4. Malta, T. M. *et al*. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell* **173**, 338–354 e315 (2018).
5. Lim, S. B., Tan, S. J., Lim, W.-T. & Lim, C. T. A merged lung cancer transcriptome dataset for clinical predictive modeling. *Sci Data* **5**, 180136 (2018).
6. Lim, S. B., Tan, S. J., Lim, W. T. & Lim, C. T. An extracellular matrix-related prognostic and predictive indicator for early-stage non-small cell lung cancer. *Nat Commun* **8**, 1734 (2017).
7. Wang, C. *et al*. The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nat Biotechnol* **32**, 926–932 (2014).
8. Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One* **9**, e78644 (2014).
9. Mooney, M. *et al*. Comparative RNA-Seq and microarray analysis of gene expression changes in B-cell lymphomas of Canis familiaris. *PLoS One* **8**, e61088 (2013).
10. Consortium, S. M.-I. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol* **32**, 903–914 (2014).
11. Nookaew, I. *et al*. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Res* **40**, 10084–10097 (2012).
12. Newman, A. M. *et al*. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12**, 453–457 (2015).
13. Lim, S. B. *et al*. Pan-cancer analysis connects tumor matrisome to immune response. *npj Precision. Oncology* **3**, 15 (2019).
14. Gentleman, R. C. *et al*. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80 (2004).
15. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307–315 (2004).
16. Taminau, J. *et al*. Unlocking the potential of publicly available microarray data using inSilicoDb and inSilicoMerging R/Bioconductor packages. *BMC Bioinformatics* **13**, 335 (2012).

17. Carlson, M. hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2). *R package version 3.2.3* (2016).
18. Zhu, Y., Qiu, P. & Ji, Y. TCGA-assembler: open-source software for retrieving and processing TCGA data. *Nat Methods* **11**, 599–600 (2014).
19. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
20. Ritchie, M. E. *et al*. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
21. Vu, V. Q. ggbiplot: A ggplot2 based biplot. *R package version 0.55* (2011).
22. Lim, S. B. Compendiums of cancer transcriptome for machine learning applications. *figshare*. https://doi.org/10.6084/m9.figshare.7878086 (2019).
23. Robin, X. *et al*. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
24. Lim, S. B. A microarray meta-dataset of lung cancer. *ArrayExpress*, https://identifiers.org/arrayexpress:E-MTAB-6699 (2019).
25. Lim, S. B. A microarray meta-dataset of pancreatic cancer. *ArrayExpress*, https://identifiers.org/arrayexpress:E-MTAB-6690 (2019).
26. Lim, S. B. A microarray meta-dataset of prostate cancer. *ArrayExpress*, https://identifiers.org/arrayexpress:E-MTAB-6694 (2019).
27. Lim, S. B. A microarray meta-dataset of renal cancer. *ArrayExpress*, https://identifiers.org/arrayexpress:E-MTAB-6692 (2019).
28. Lim, S. B. A microarray meta-dataset of gastric cancer. *ArrayExpress*, https://identifiers.org/arrayexpress:E-MTAB-6693 (2019).
29. Lim, S. B. A microarray meta-dataset of colorectal cancer. *ArrayExpress*, https://identifiers.org/arrayexpress:E-MTAB-6698 (2019).
30. Lim, S. B. A microarray meta-dataset of ovarian cancer. *ArrayExpress*, https://identifiers.org/arrayexpress:E-MTAB-6691 (2019).
31. Lim, S. B. A microarray meta-dataset of breast cancer. *ArrayExpress*, https://identifiers.org/arrayexpress:E-MTAB-6703 (2019).
32. Lim, S. B. A microarray meta-dataset of liver cancer. *ArrayExpress*, https://identifiers.org/arrayexpress:E-MTAB-6695 (2019).
33. Lim, S. B. A microarray meta-dataset of bladder cancer. *ArrayExpress*, https://identifiers.org/arrayexpress:E-MTAB-6696 (2019).
34. Lim, S. B. A microarray meta-dataset of melanoma cancer. *ArrayExpress*, https://identifiers.org/arrayexpress:E-MTAB-6697 (2019).
35. Lim, S. B. *et al*. Addressing cellular heterogeneity in tumor and circulation for refined prognostication. *Proc. Natl Acad. Sci. USA* **116**, 17957–17962 (2019).
36. Plaisier, S. B., Taschereau, R., Wong, J. A. & Graeber, T. G. Rank-rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures. *Nucleic Acids Res* **38**, e169 (2010).
37. Cahill, K. M., Huo, Z., Tseng, G. C., Logan, R. W. & Seney, M. L. Improved identification of concordant and discordant gene expression signatures using an updated rank-rank hypergeometric overlap approach. *Sci Rep* **8**, 9588 (2018).
38. Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–585 (2013).
39. Atlas Research, CancerGenome N. *et al*. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113–1120 (2013).
40. Ting, D. T. *et al*. Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Rep* **8**, 1905–1918 (2014).
41. Rustici, G. *et al*. ArrayExpress update–trends in database growth and links to data analysis tools. *Nucleic Acids Res* **41**, D987–990 (2013).
42. Thompson, J. A., Tan, J. & Greene, C. S. Cross-platform normalization of microarray and RNA-seq data for machine learning applications. *PeerJ* **4**, e1621 (2016).
43. Wilhelm, B. T. & Landry, J. R. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* **48**, 249–257 (2009).
44. Uziela, K. & Honkela, A. Probe Region Expression Estimation for RNA-Seq Data for Improved Microarray Comparability. *PloS one* **10**, e0126545 (2015).
45. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**, R29 (2014).
46. 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* **19**, 185–193 (2003).
47. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **33**, 1–22 (2010).
48. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267–288 (1996).
49. Gentles, A. J. *et al*. The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat Med* **21**, 938–945 (2015).
50. Iglesia, M. D. *et al*. Genomic Analysis of Immune Cell Infiltrates Across 11 Tumor Types. *J Natl Cancer Inst* **108**, djw144 (2016).
51. Brown, S. D. *et al*. Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Res* **24**, 743–750 (2014).
52. Charoentong, P. *et al*. Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep* **18**, 248–262 (2017).
53. Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G. & Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160**, 48–61 (2015).
54. Hackl, H., Charoentong, P., Finotello, F. & Trajanoski, Z. Computational genomics tools for dissecting tumour-immune cell interactions. *Nat Rev Genet* **17**, 441–458 (2016).
55. Gnjatic, S. *et al*. Identifying baseline immune-related biomarkers to predict clinical outcome of immunotherapy. *J Immunother Cancer* **5**, 44 (2017).
56. Gibney, G. T., Weiner, L. M. & Atkins, M. B. Predictive biomarkers for checkpoint inhibitor-based immunotherapy. *Lancet Oncol* **17**, e542–e551 (2016).
57. Lim, S. B. A microarray meta-dataset of non-small cell lung cancer. *ArrayExpress*, https://identifiers.org/arrayexpress:E-MTAB-6043 (2018).
58. Lim, S. B. An extracellular matrix-related prognostic and predictive indicator for early-stage non-small cell lung cancer. *figshare*. https://doi.org/10.6084/m9.figshare.5350321 (2018).

Acknowledgements

This work was conceived and carried out at the MechanoBioEngineering laboratory at the Department of Biomedical Engineering, National University of Singapore (NUS). We acknowledge support provided by the National Research Foundation, Prime Minister's Office, Singapore under its Research Centre for Excellence, Mechanobiology Institute at NUS. W.-T.L. is supported by the National Medical Research Council (NMRC/CSA/040/2012 and NMRC/CSA-INV/0025/2017). S.B.L. is supported by NUS Graduate School for Integrative Sciences and Engineering (NGS), Mogam Science Scholarship Foundation, and Daewoong Foundation.

Author Contributions

S.B.L., S.J.T., W.-T.L. and C.T.L. conceptualized and designed the study. S.B.L. developed the R pipeline to generate MMDs. S.B.L., S.J.T., W.-T.L. and C.T.L. analyzed and interpreted the data. S.B.L., S.J.T., W.-T.L. and C.T.L. reviewed and contributed to the manuscript.

Additional Information

Supplementary Information is available for this paper at https://doi.org/10.1038/s41597-019-0207-2.

Competing Interests: The authors declare no competing interests.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.