quanTIseq: quantifying immune contexture of human tumors

Francesca Finotello[1], Clemens Mayer[1], Christina Plattner[1], Gerhard Laschober[1], Dietmar Rieder[1], Hubert Hackl[1], Anne Krogsdam[1], Wilfried Posch[2], Doris Wilflingseder[2], Sieghart Sopper[3], Marieke IJsselsteijn[4], Douglas Johnson[5], Yaomin Xu[6], Yu Wang[6], Melinda E. Sanders[7], Monica V. Estrada[7], Paula Ericsson-Gonzalez[7], Justin Balko[5], Noel de Miranda[4], Zlatko Trajanoski[1]

[1]Biocenter, Division of Bioinformatics, Medical University of Innsbruck, Austria
[2]Division of Hygiene and Medical Microbiology, Medical University of Innsbruck, Austria
[3]Department of Haematology and Oncology, Medical University of Innsbruck, Austria
[4]Department of Pathology, Leiden University Medical Centre, The Netherlands
[5]Vanderbilt University, Nashville/TN, USA
[5]Department of Medicine, Vanderbilt University Medical Center, Nashville/TN, USA
[6]Department of Biostatistics, Vanderbilt University Medical Center, Nashville/TN, USA
[7]Department Pathology Microbiology and Immunology, Vanderbilt University Medical Center, Nashville/TN, USA

Keywords: cancer immunology, immunotherapy, deconvolution, RNA-seq, immune contexture

We introduce quanTIseq, a method to quantify the tumor immune contexture, determined by the type and density of tumor-infiltrating immune cells. quanTIseq is based on a novel deconvolution algorithm for RNA sequencing data that was validated with independent data sets. Complementing the deconvolution output with image data from tissue slides enables *in silico* multiplexed immunodetection and provides an efficient method for the immunophenotyping of a large number of tumor samples.

Cancer immunotherapy with antibodies targeting immune checkpoints has shown durable benefit or even curative potential in various cancers[1,2]. As only a fraction of patients are responsive to immune checkpoint blockers, efforts are underway to identify predictive markers as well as mechanistic rationale for combination therapies with synergistic potential. Thus, comprehensive and quantitative immunological characterization of tumors in a large number of clinical samples is of utmost importance, but it is currently hampered by the lack of simple and efficient methods. Cutting-edge technologies like single-cell RNA sequencing and multi-parametric flow or mass cytometry are technically and logistically challenging and cannot be applied to archived samples. Multiplexed

immunohistochemistry (IHC)[3] or immunofluorescence (IF) assays can be performed only in specialized labs and require sophisticated equipment and extensive optimization of protocols for specific cancer entities. Computational methods for quantitative immunophenotyping of tumors from RNA sequencing (RNA-seq) data hold potential for efficient and low-cost profiling of large number of samples, but currently suffer from several limitations. Methods based on enrichment analysis of immune gene sets compute only semi-quantitative scores[4]. Conversely, deconvolution algorithms (reviewed in [4–6]) can enable a quantitative estimation of the proportions of the cell types of interest and, hence, of the immune contexture - defined by the type and density of tumor-infiltrating immune cells. The immune contexture has not only major prognostic value in colorectal cancer (CRC)[7] and other cancer types[8], but can also provide information that is relevant for the prediction of treatment response. However, currently available deconvolution algorithms have important limitations and are not suitable for the quantification of immune contexture of human tumors. For instance, CIBERSORT, a method based on support-vector regression for deep-deconvolution of 22 immune-cell phenotypes, can only infer cell fractions relative to the total immune-cell population and has been developed and validated using microarray data[9]. TIMER performs deconvolution of six immune cell types, but the results cannot be interpreted directly as cell fractions, nor compared across different immune cell types and data sets[10]. EPIC, a deconvolution method recently developed using RNA-seq data, estimates relative fractions referred to the whole cell mixture, but does not consider immune cells relevant for cancer immunology like regulatory T cells (T_{reg}) cells, dendritic cells, and classically (M1) and alternatively (M2) activated macrophages[11]. Most importantly, these methods have not been validated using independent data sets comprising tumor RNA-seq and a gold standard method like IHC or IF of the same sample.

Therefore, we developed quanTIseq, a computational pipeline for the **quan**tification of the **T**umor **I**mmune contexture using RNA-**seq** data and images of haematoxylin and eosin (H&E)-stained tissue slides (**Fig. 1a**). As part of quanTIseq, we first developed a deconvolution algorithm based on constrained least squares regression[12]. We then designed a signature matrix from a compendium of 51 RNA-seq data sets (**Supplementary Table 1**) from ten different immune cell types: B cells, M1 and M2 macrophages, monocytes (Mono), neutrophils (Neu), natural killer (NK) cells, $CD4^+$ and $CD8^+$ T cells, T_{reg} cells, and dendritic cells (DC) (**Fig. 1b and Supplementary Table 2**). Notably, as the preprocessing steps, including gene annotation and expression normalization, have a strong impact

on the final estimates and can lead to inconsistencies between the mixture and the signature matrix, we implemented a full analytical pipeline (available at: http://icbi.i-med.ac.at/software/quantiseq/doc/index.html), consisting of read pre-processing, quantification of gene expression, deconvolution of cell fractions, and computation of cell densities (**Fig. 1c**).

To validate quanTIseq we first used both simulated data and published data. We simulated 1,700 RNA-seq data sets from human breast tumors by mixing various numbers of reads from tumor and immune-cell RNA-seq data, considering different immune compositions and sequencing depths. quanTIseq obtained a high correlation between the true and the estimated fractions and accurately quantified tumor content (measured by the fraction of "other" cells) (**Supplementary Figure 1**). We then validated quanTIseq using experimental data from a previous study[13], in which peripheral blood mononuclear cell (PBMC) mixtures were subjected to both, RNA-seq and flow cytometry. A high accuracy of quanTIseq estimates was also observed with this data set (**Fig. 1d** and **Supplementary Figure 2**). Additionally, we successfully validated quanTIseq using two previous published data sets (**Supplementary Figures 3** and **4**).

As most of the validation data sets available in the literature are based on microarray data or consider a limited number of phenotypes, we generated RNA-seq and flow cytometry data from mixtures of peripheral-blood immune cells collected from nine healthy donors. Flow cytometry was used to quantify all the immune sub-populations considered by quanTIseq signature matrix except macrophages, which are not present in blood. Comparison between quanTIseq cell estimates and flow cytometry fractions showed a high correlation at a single and multiple cell-type level (**Fig. 1e** and **Supplementary Figure 5**).

We then validated quanTIseq using two independent data sets. The first data set was generated from samples from 31 melanoma patients (Vanderbilt cohort). We carried out RNA-seq and, wherever possible, IHC staining for $CD8^+$, $CD4^+$ or $FOXP3^+$ cells from consecutive whole-tissue slides. To quantify specific immune cells from the scanned images, we developed an analysis pipeline (available at https://github.com/mui-icbi/IHCount) to perform semi-automatic cell counting. The second data set was generated from samples from nine CRC patients (Leiden cohort). RNA-seq data and multiplexed IF stainings for $CD8^+$ T, $CD4^+$ T and T_{reg} cells were carried out. As it can be seen, cell fractions obtained with quanTIseq correlated with the respective image cell densities for both the Vanderbilt

(**Fig. 2a**) and the Leiden cohort (**Fig. 2b**). Thus, the results of our extensive validation using simulated data, published data, data from blood cell mixtures, and two independent data sets demonstrate that quanTIseq can faithfully and quantitatively decompose immune profiles in human tumors using RNA-seq data.

To demonstrate the utility of quanTIseq, we then analyzed RNA-seq data from more than 8,000 TCGA samples across 20 TCGA solid cancers (**Fig. 2c**). We obtained high agreement between quanTIseq results and the lymphocytic infiltration[14] and tumor purity[15] estimates reported in two previous studies (**Supplementary Note 1**). The results of the survival analyses using the computed TCGA cell fractions (**Supplementary Figure 7**) show that the prognostic power for single cell types is highly context dependent. Moreover, within cancer entities the immune cell compositions were highly variable. As an example we present the immune fractions of the CRC patients stratified into four consensus molecular subtypes (CMS)[16]. The results revealed higher infiltration of M1 macrophages and $CD8^+$ T cells in the "CMS1 - MSI immune" group (p-values<0.02), which is has a good prognosis, and of B cells and M2 macrophages in the "CMS4 - mesenchymal" group (p-values<0.03) (**Fig. 2d)**.

We also show the value of quanTIseq for cancer immunotherapy and present the results of the quantification of immune contexture in pre-treatment samples from melanoma patients on anti-PD1 treatment (subset from the Vanderbilt cohort). We carried out deconvolution using RNA-seq data and scaled the fractions using cells densities extracted from images to perform *in silico* multiplexed immunodetection. The cell densities of ten immune cell types showed large heterogeneity across the patients and some differences between responders and non-responders, although not statistically significant (p>0.09) (**Fig. 2e**). However, due to the limited number of samples, further studies are necessary to determine which immune cell contextures have predictive power.

Finally, all quanTIseq results from the TCGA, the Vanderbilt cohort, and two additional cohorts of melanoma patients treated with immune checkpoint blockers[18,19], were deposited in The Cancer Immunome Atlas (https://tcia.at)[17] to make them available to the scientific community and facilitate the generation of testable hypothesis.

In conclusion, we developed quanTIseq, a computational pipeline for the analysis of raw RNA-seq and tissue imaging data that quantifies the fractions and densities of ten different immune cell types relevant for cancer immunology. Unlike previous approaches, quanTIseq is specifically designed for

RNA-seq, which is the current reference technology for high-throughput quantification of gene expression[20] and was extensively validated (**Supplementary Table 2**). Moreover, in order to avoid inconsistencies between the mixture and the signature matrix, we assembled and provide a complete analytical pipeline. As bulk RNA-seq is now widely applied to profile fresh-frozen and archived tumor specimens, quanTIseq can be applied to effectively mine these data[4]. Specifically, quanTIseq can be used to quantify the immune contexture in a large number of archived samples in order to identify immunogenic effects of conventional and targeted drugs and hereby gain mechanistic rationale for the design of combination therapies. Thus, quanTIseq represents an important enhancement to the computational toolbox for dissecting tumor-immune cell interactions, and can further be applied to autoimmune, inflammatory, or infectious diseases.

Acknowledgements

We would like to thank Dr. Paul Hoertnagl (Innsbruck Blood Bank, Austria) for the collection of blood samples from healthy donors and Dr. Kristen L. Hoek (Vanderbilt Institute for Infection, Immunology and Inflammation, US) for providing access to the flow cytometry data for algorithm validation. This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 633592 (project “APERIM: Advanced Bioinformatics Tools for Personalised Cancer Immunotherapy”) and by Austrian Cancer Aid/Tyrol (project No. 17003, “quanTIseq: dissecting the immune contexture of human cancers”). The Leiden cohort work was supported by the Fight Colorectal Cancer-Michael’s Mission-AACR Fellowship (2015) and Alpe d’HuZes/KWF Bas Mulder Award (UL2015-7664). The Vanderbilt cohort work was supported by the Vanderbilt-Incyte Research Alliance Program Grant (JMB, DBJ, YX), as well as R00CA181491 (JMB), K23CA204726 (DBJ), and the Breast Cancer Specialized Program of Research Excellence (SPORE) P50 CA098131.

Author contributions

FF developed quanTIseq and performed deconvolution analyses and method benchmarking. FF and CP implemented quanTIseq software. FF and HH performed the statistical analyses. CM developed IHCount. CM and DR analyzed the Vanderbilt images and updated the TCIA database. ZT, FF, AK, GL, WP, DW and SS designed the validation experiment on blood mixtures. GL isolated the blood cells and performed the flow cytometry experiment and data analysis. AK performed RNA extraction. DJ contributed to the collection of Vanderbilt patient specimens and clinical annotation. YX and YW oversaw pre-processing and quality control of Vanderbilt RNA-seq data. MES oversaw histology techniques, analysis, and scoring techniques for the Vanderbilt cohort. MVE and PEG performed IHC data pre-processing. JB performed Vanderbilt data analysis, integration of techniques, performed and oversaw molecular assays and work involved in sample selection and isolation of nucleic acids. MI and NdM generated RNA-seq and multiplexed IF detection for the Leiden Cohort. FF and ZT wrote the manuscript. All authors read and approved the final manuscript.

Competing financial interests

The authors declare no competing financial interests.

Figures

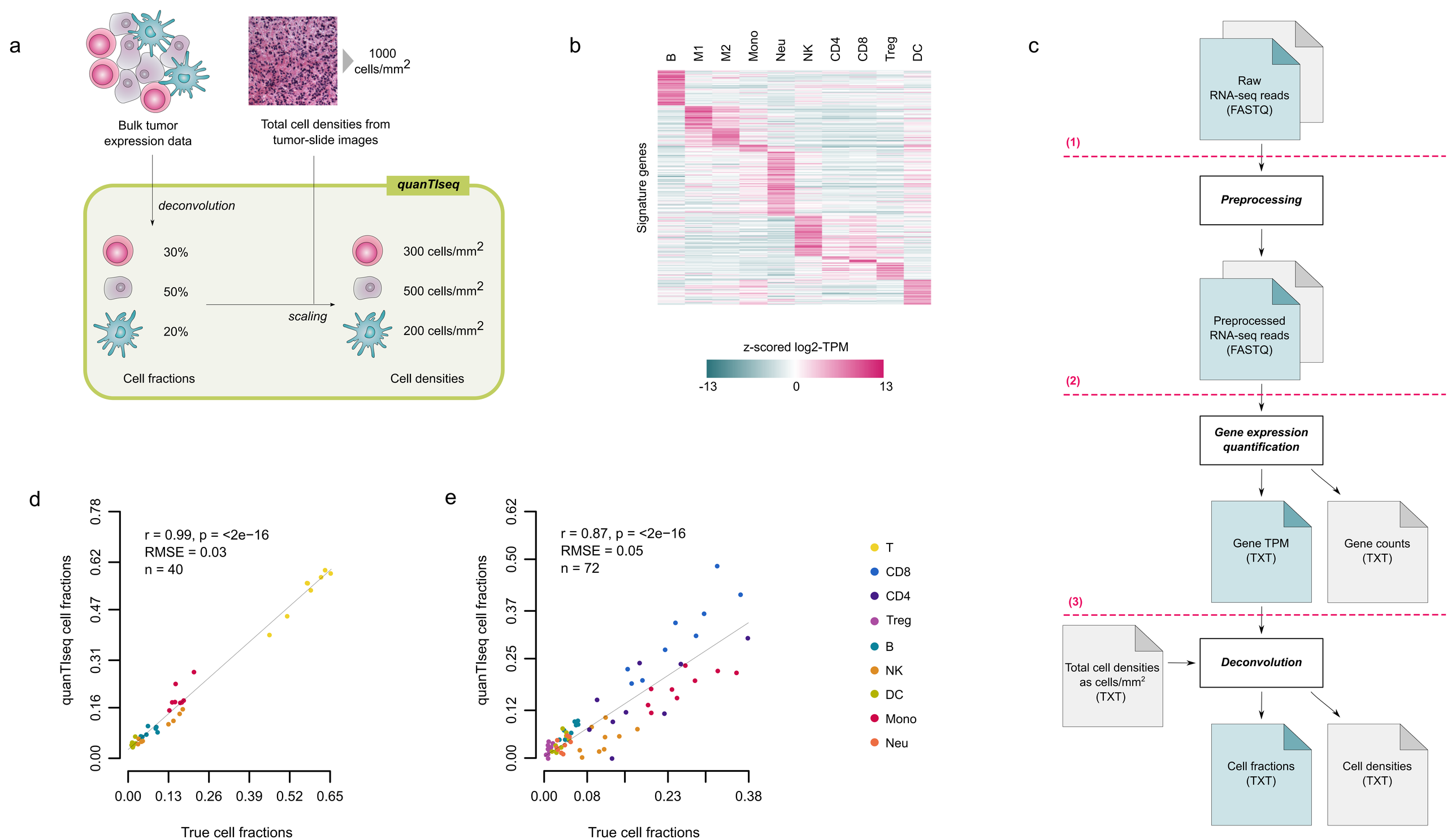
Figure 1: quanTIseq method and validation based on blood-cell mixtures. (a) quanTIseq characterizes the immune contexture of human tumors from expression and image data. Cell fractions are estimated from expression data and then scaled to cell densities (cells/mm^2) using total cell densities extracted from imaging data. **(b)** Heatmap of quanTIseq signature matrix, with z-scores computed from $\log_2$(TPM+1) expression values of the signature genes. **(c)** The quanTIseq pipeline consists of three modules that perform: (1) pre-processing of paired- or single-end RNA-seq reads in FASTQ format; (2) quantification of gene expression in transcripts-per-millions (TPM) and gene counts; (3) deconvolution of cell fractions and scaling to cell densities considering total cells per mm^2 derived from imaging data from H&E-stained slides. The analysis can be initiated at any step (e.g. pre-processed expression data can be analyzed starting from step 3). Optional files are shown in grey. Validation of quanTIseq with RNA-seq data from blood-derived immune-cell mixtures generated in [13] **(d)** and in this study **(e)**. Deconvolution performance was assessed with Pearson's correlation (r) and root-mean-square error (RMSE) using flow cytometry estimates as ground truth. The line represents the linear fit. B: B cells; CD4: CD4$^+$ T cells; CD8: CD8$^+$ T cells; DC: dendritic cells; M1: classically activated macrophages; M2: alternatively activated macrophages; Mono: monocytes; Neu: neutrophils; NK: natural killer cells; T: T cells; Treg: regulatory T cells. H&E: haematoxylin and eosin.

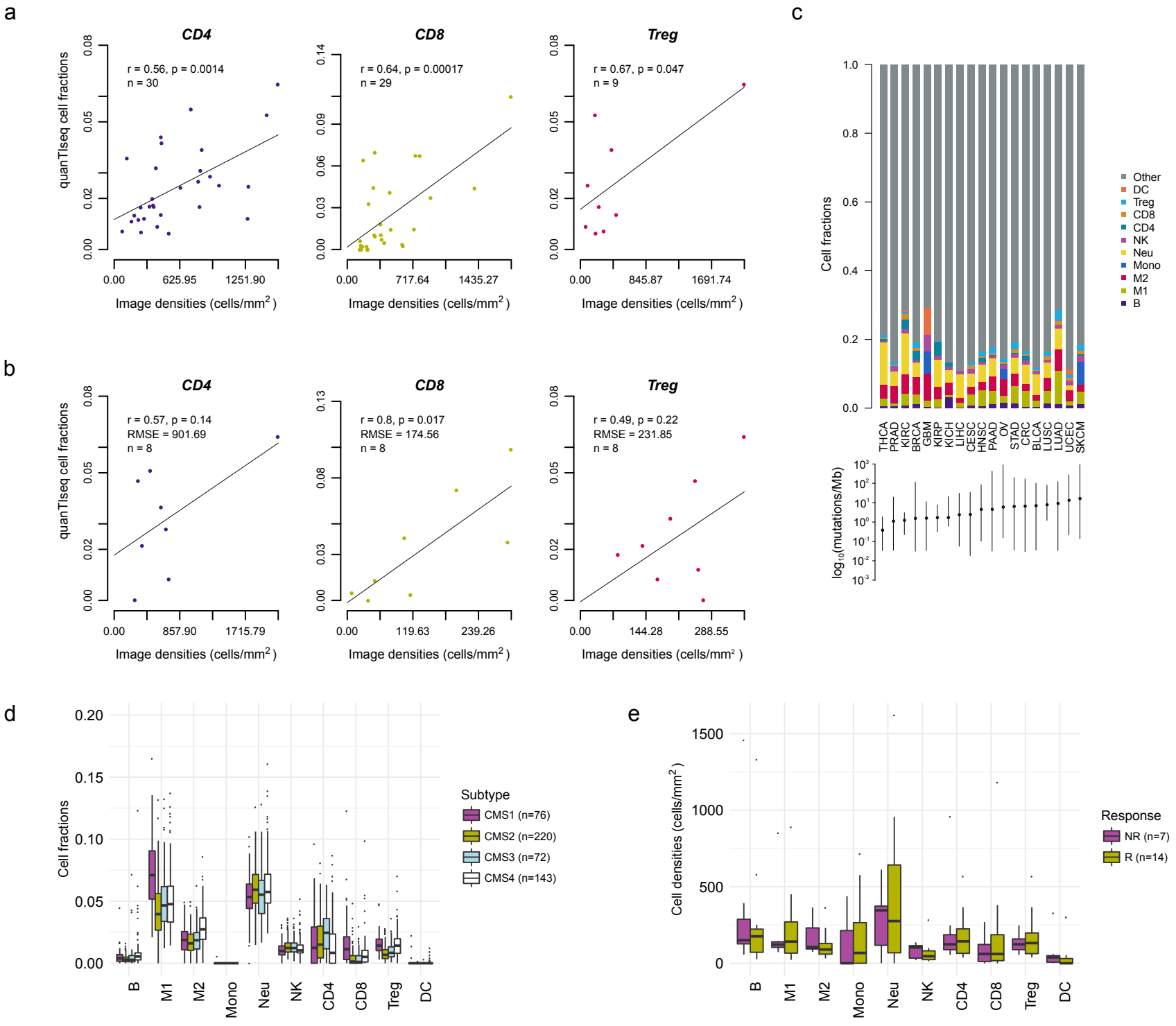
Figure 2: Deconvolution of tumor RNA-seq data with quanTIseq. Comparison of the cell fractions inferred for Vanderbilt melanoma patients **(a)** and Leiden colorectal cancer patients **(b)** with cells per mm^2 computed with immunofluorescence and immunohistochemistry, respectively. Deconvolution performance was assessed with Pearson's correlation (r) and root-mean-square error (RMSE). The line represents the linear fit. **(c)** Median cell fractions per cancer type across 8,243 TCGA samples, sorted according to the mutational load. The range and mean of the mutational loads, computed as the number of non-synonymous mutation per mega base (on $\log_{10}$ scale), are shown for each cancer type. **(d)** Immune cell fractions of TCGA colorectal cancer patients stratified according to consensus molecular subtypes (CMS). **(e)** Immune cell fractions from Vanderbilt melanoma patients stratified as responders (R) and non-responders (NR). B: B cells; CD4: CD4$^+$ T cells (including also CD4$^+$ regulatory T cells); CD8: CD8$^+$ T cells; DC: dendritic cells; M1: classically activated macrophages; M2: alternatively activated macrophages; Mono: monocytes; Neu: neutrophils; NK: natural killer cells; Treg: regulatory T cells; Other: other uncharacterized cells.

References

1. Topalian, S. L., Taube, J. M., Anders, R. A. & Pardoll, D. M. Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy. *Nat. Rev. Cancer* **16,** 275–287 (2016).
2. Chen, D. S. & Mellman, I. Elements of cancer immunity and the cancer-immune set point. *Nature* **541,** 321–330 (2017).
3. Tsujikawa, T. *et al.* Quantitative Multiplex Immunohistochemistry Reveals Myeloid-Inflamed Tumor-Immune Complexity Associated with Poor Prognosis. *Cell Rep.* **19,** 203–217 (2017).
4. Hackl, H., Charoentong, P., Finotello, F. & Trajanoski, Z. Computational genomics tools for dissecting tumour-immune cell interactions. *Nat. Rev. Genet.* **17,** 441–458 (2016).
5. Newman, A. M. & Alizadeh, A. A. High-throughput genomic profiling of tumor-infiltrating leukocytes. *Curr. Opin. Immunol.* **41,** 77–84 (2016).
6. Aran, D. & Butte, A. J. Digitally deconvolving the tumor microenvironment. *Genome Biol.* **17,** 175 (2016).
7. Galon, J. *et al.* Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science* **313,** 1960–1964 (2006).
8. Fridman, W. H., Pagès, F., Sautès-Fridman, C. & Galon, J. The immune contexture in human tumours: impact on clinical outcome. *Nat. Rev. Cancer* **12,** 298–306 (2012).
9. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12,** 453–457 (2015).
10. Li, B. *et al.* Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.* **17,** 174 (2016).
11. Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous Enumeration Of Cancer And Immune Cell Types From Bulk Tumor Gene Expression Data. *eLIFE* **6,** e26476 (2017).
12. Haskell, K. H. & Hanson, R. J. An algorithm for linear least squares problems with equality and nonnegativity constraints. *Math. Program.* **21,** 98–118 (1981).
13. Hoek, K. L. *et al.* A cell-based systems biology assessment of human blood to monitor immune responses after influenza vaccination. *PloS One* **10,** e0118528 (2015).
14. Cancer Genome Atlas Network. Genomic Classification of Cutaneous Melanoma. *Cell* **161,** 1681–1696 (2015).

15. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6,** 8971 (2015).
16. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21,** 1350–1356 (2015).
17. 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).
18. Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350,** 207–211 (2015).
19. Hugo, W. *et al.* Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell* **165,** 35–44 (2016).
20. Finotello, F. & Di Camillo, B. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Brief. Funct. Genomics* **14,** 130–142 (2015).


a
Bulk tumor expression data
1000 cells/mm2
Total cell densities from tumor-slide images
quanTIseq
deconvolution
30%
50%
20%
scaling
300 cells/mm2
500 cells/mm2
200 cells/mm2
Cell fractions
Cell densities
b
B
M1
M2
Mono
Neu
NK
CD4
CD8
Treg
DC
Signature genes
z-scored log2-TPM
-13
0
13
c
Raw RNA-seq reads (FASTQ)
(1)
Preprocessing
Preprocessed RNA-seq reads (FASTQ)
(2)
Gene expression quantification
Gene TPM (TXT)
Gene counts (TXT)
(3)
Total cell densities as cells/mm2 (TXT)
Deconvolution
Cell fractions (TXT)
Cell densities (TXT)
d
r = 0.99, p = <2e−16
RMSE = 0.03
n = 40
quanTIseq cell fractions
True cell fractions
0.00
0.16
0.31
0.47
0.62
0.78
0.13
0.26
0.39
0.52
0.65
e
r = 0.87, p = <2e−16
RMSE = 0.05
n = 72
quanTIseq cell fractions
True cell fractions
0.00
0.12
0.25
0.37
0.50
0.62
0.08
0.23
0.38
T
CD8
CD4
Treg
B
NK
DC
Mono
Neu


a
CD4
r = 0.56, p = 0.0014
n = 30
CD8
r = 0.64, p = 0.00017
n = 29
Treg
r = 0.67, p = 0.047
n = 9
quanTIseq cell fractions
Image densities (cells/mm2)
b
CD4
r = 0.57, p = 0.14
RMSE = 901.69
n = 8
CD8
r = 0.8, p = 0.017
RMSE = 174.56
n = 8
Treg
r = 0.49, p = 0.22
RMSE = 231.85
n = 8
c
Cell fractions
Other
DC
Treg
CD8
CD4
NK
Neu
Mono
M2
M1
B
THCA PRAD KIRC BRCA GBM KIRP KICH LIHC CESC HNSC PAAD OV STAD CRC BLCA LUSC LUAD UCEC SKCM
log10(mutations/Mb)
d
Cell fractions
Subtype
CMS1 (n=76)
CMS2 (n=220)
CMS3 (n=72)
CMS4 (n=143)
B M1 M2 Mono Neu NK CD4 CD8 Treg DC
e
Cell densities (cells/mm2)
Response
NR (n=7)
R (n=14)
B M1 M2 Mono Neu NK CD4 CD8 Treg DC