Newman et al. Genome Biology (2017) 18:128
DOI 10.1186/s13059-017-1257-4

Genome Biology

CORRESPONDENCE Open Access

Data normalization considerations for digital tumor dissection

Aaron M. Newman[1,2*], Andrew J. Gentles[3,4], Chih Long Liu[1,2], Maximilian Diehn[1,5,6] and Ash A. Alizadeh[1,2,3,6,7*]

Please see related Li et al correspondence: www.dx.doi.org/10.1186/s13059-017-1256-5
and Zheng correspondence: www.dx.doi.org/10.1186/s13059-017-1258-3

Abstract

In a recently published article in *Genome Biology*, Li and colleagues introduced TIMER, a gene expression deconvolution approach for studying tumor-infiltrating leukocytes (TILs) in 23 cancer types profiled by The Cancer Genome Atlas. Methods to characterize TIL biology are increasingly important, and the authors offer several arguments in favor of their strategy. Several of these claims warrant further discussion and highlight the critical importance of data normalization in gene expression deconvolution applications.

Computational approaches for enumerating cell subsets from bulk tissue expression profiles have significant potential for studying tumor cellular ecosystems, including tumor-infiltrating leukocytes (TILs) [1–3]. We therefore read with interest the recent *Genome Biology* article by Li and colleagues in which they introduce TIMER, an in silico method for TIL deconvolution [4]. TIMER relies on prior knowledge of immune signature genes as input and consists of three major steps: (1) gene expression normalization across platforms and sample types; (2) selection of immune signature genes that are negatively correlated with tumor purity; and (3) deconvolution of RNA admixtures using a previously described technique for iterative linear least squares regression (LLSR) [5]. They apply TIMER to the inference of six distinct immune subsets (B cells, CD4 T cells, CD8 T cells, neutrophils, macrophages, and dendritic cells) in The Cancer Genome Atlas (TCGA) bulk tumor expression profiles and investigate links between TIL heterogeneity, tumor genomic features, and survival in 23 cancer types.

Several groups, including ours, have also proposed methods for gene expression deconvolution [1, 3, 5, 6]. We recently described CIBERSORT, an in silico tissue dissection approach that is robust to noise, unknown mixture content, and closely related cell types (collinearity) [6]. Notably, in benchmarking experiments CIBERSORT outperformed other deconvolution methods, including LLSR, and revealed complex associations between 22 distinct immune subsets and outcomes in a pan-cancer meta-analysis [6, 7]. We were therefore surprised by several claims in relation to CIBERSORT.

First, the authors assert that CIBERSORT succumbs to statistical collinearity (i.e., cell subsets with highly correlated expression profiles), leading to biased estimations. Evidence for this argument is primarily based on a simple experiment in which inferred levels of each immune subset were compared by Pearson correlation. After aggregating CIBERSORT results from 22 phenotypes into the same six subsets, the authors compared cross-correlation matrices between TIMER and CIBERSORT on four cancer types. Leukocyte levels estimated by TIMER were almost always positively correlated. According to the authors, positive correlations make intuitive sense because "immune cells work in synergy." In contrast, the correlations among the six phenotypes estimated by CIBERSORT were largely negative. The authors also observed negative correlations when analyzing more than six cell types with TIMER (i.e., LLSR), stating that negative correlations indicate a technical artifact due to collinearity.

In fact, CIBERSORT mitigates such bias through regularization, as was rigorously demonstrated through a battery of validation experiments [6, 7]. These analyses included an assessment of "deep deconvolution" in which in silico predictions of closely related leukocyte subsets were

* Correspondence: amnewman@stanford.edu; arasha@stanford.edu
[1]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University, Stanford, California 94305, USA
Full list of author information is available at the end of the article

directly compared against flow cytometry [6]. Moreover, an independent study confirmed that regularization improves the performance of gene expression deconvolution [8].

We were also surprised by the authors' claim that hematopoietic cell types should generally track together, especially in light of their diverse functions (innate or adaptive, stimulatory or suppressive, etc.) and migration patterns (circulating, tissue-infiltrating, or tissue-resident) [9, 10]. For instance, while specific hematopoietic subsets infiltrating tumors can be positively correlated in a given tumor type [11], the expectation of universally positive correlations does not extend to all tumor types, or to all infiltrating immune cells [12]. Separately, age-related lymphomyeloid lineage skewing of hematopoiesis [13] would be expected to further confound this assumption. Finally, while acute and chronic inflammation can be substrates for tumor initiation, a number of distinct tumor-infiltrating immune cells are known to have either tumor-promoting or anti-tumor properties, and are associated with inverse prognostic correlations with cancer outcomes [9, 14].

We therefore reanalyzed previously published flow cytometry data of leukocyte subsets directly enumerated in peripheral blood mononuclear cells (PBMCs) from healthy donors and in tumor biopsies obtained from patients with lung squamous cell carcinoma (LUSC) [6, 7]. When we quantified each immune subset as a fraction of total leukocyte content, many of the pairwise correlations were negative, as were the mean correlation coefficients (Fig. 1a and b), consistent with CIBERSORT. However, when we instead considered absolute TIL levels in the same lung tumors, most of the correlations were positive (Fig. 1c), likely reflecting differences in tumor purity. Thus, data normalization in solid tumors significantly impacts the assessment of TIL heterogeneity and composition.

Given these results, we suspected that tumor purity would explain the discrepancy between TIMER and CIBERSORT. Indeed, after examining the TIMER source code, we found that, unlike most previous deconvolution methods including CIBERSORT, TIMER solves the regression problem without normalizing inferred cell subset frequencies to 1. TIMER results are therefore directly influenced by total leukocyte content, which is inversely correlated with tumor purity across TCGA (Fig. 2a). As a result, all six cell types strongly correlate with total leukocyte abundance in nearly every analyzed tumor type (Fig. 2b), making it difficult to discern the intercellular heterogeneity among the leukocyte subsets that variably infiltrate these tumors (Fig. 2c). When TIMER results were instead normalized in relative space (i.e., summing to 1) for each sample, all mean cross-correlation coefficients were negative (Fig. 2d). The inverse held true for CIBERSORT: mean cross-correlation coefficients became positive when we either (1) omitted the sum-to-1 normalization step, or (2) multiplied the

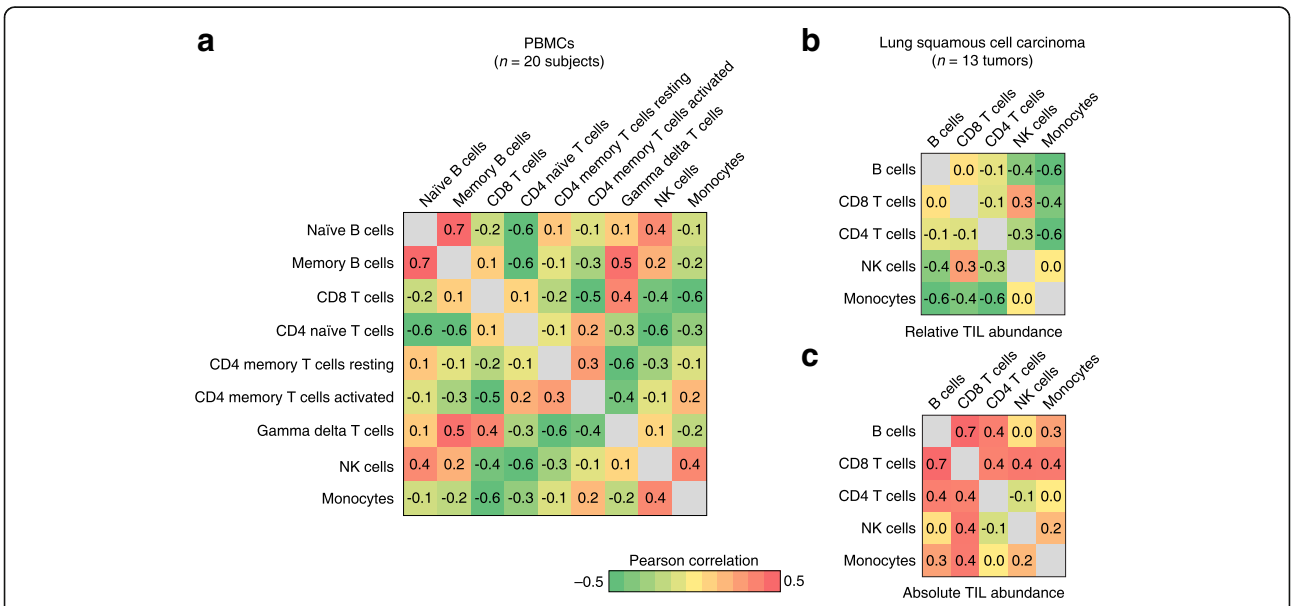

Fig. 1 Variable correlation of directly enumerated leukocyte frequencies in lung tumors and blood. **a** All-versus-all Pearson correlation matrix of nine distinct leukocyte subsets profiled by flow cytometry in peripheral blood mononuclear cells (*PBMCs*) from 20 healthy donors [6]. **b** Same as panel **a**, but for five immune subsets profiled by flow cytometry in 13 lung squamous cell carcinoma tumor biopsies [7]. **a**, **b** Leukocyte frequencies were normalized to sum to 1 prior to correlation analysis. **c** Same as panel **b**, except the frequency of each leukocyte subset was expressed as a percentage of viable singlets prior to correlation assessment. *NK* natural killer, *TIL* tumor-infiltrating leukocyte

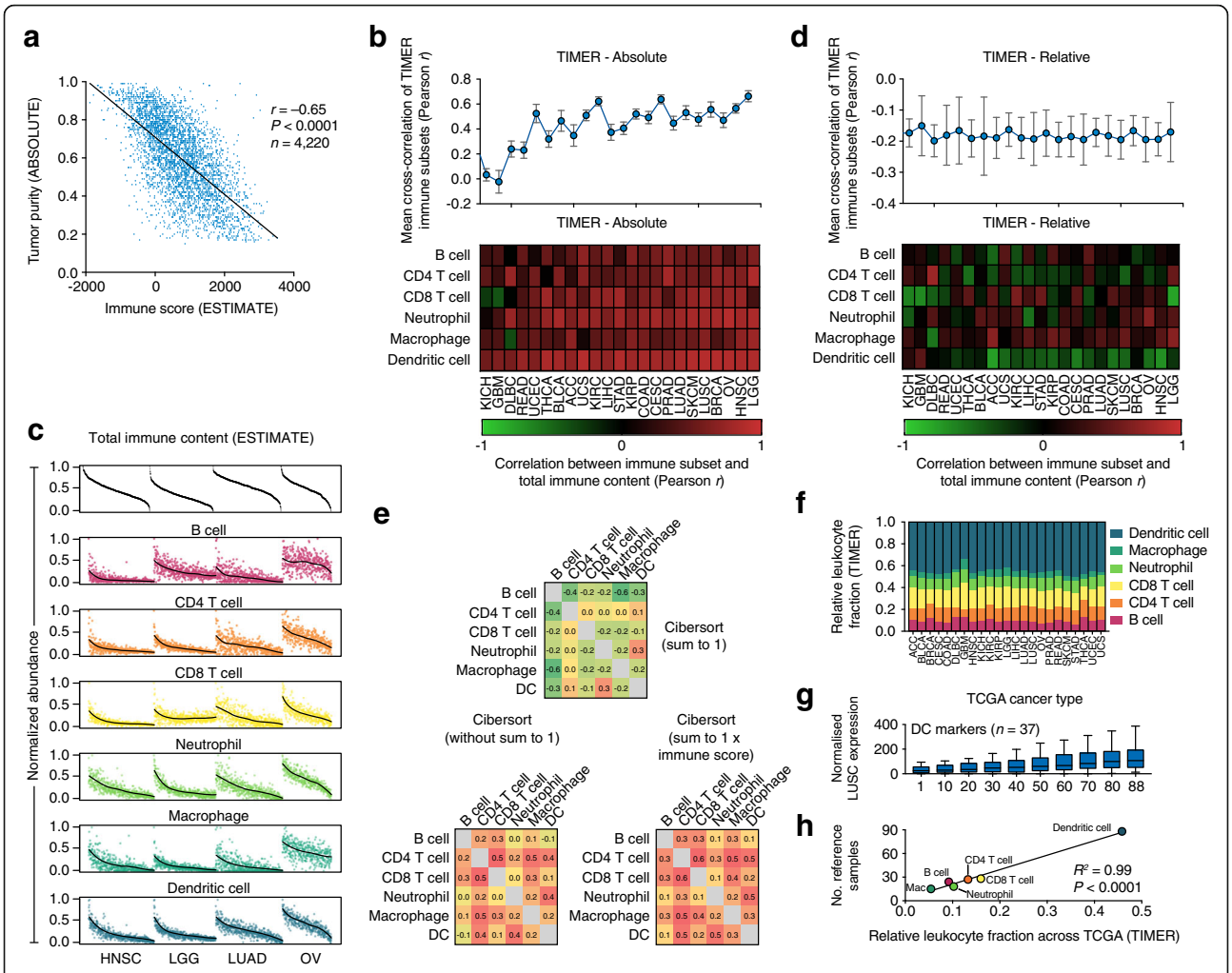

Fig. 2 Impact of data normalization on in silico tumor-infiltrating leukocyte profiling. **a** Tumor purity inferred by ABSOLUTE [29] versus immune content inferred by ESTIMATE [30], compared across 11 TCGA (The Cancer Genome Atlas) cancer types (ABSOLUTE data were obtained from [26]). **b** *Bottom* heat map showing Pearson correlations comparing overall leukocyte content, inferred by ESTIMATE, with immune subset abundance, inferred by TIMER, across 23 TCGA tumor types. Cancers are ordered from *left* to *right* by the mean correlation coefficient calculated across the six immune cell types. *Top* mean cross-correlation coefficient of the six immune subsets compared with each other, omitting self-comparisons. Cancer types are vertically aligned, and correlation coefficients are expressed as mean ± SEM. **c** TIMER results are shown for four representative TCGA cancer types, along with immune content inferred by ESTIMATE. Overall leukocyte content and estimates of individual tumor-infiltrating leukocyte (TIL) subsets are normalized from 0 to 1 within each cancer type, and ordered from *left* to *right* by decreasing immune content. *Regression lines* (shown in *black*) were calculated by cubic splines. **d** Same as panel **b**, but after normalizing inferred levels of the six leukocyte subsets to one in each patient. **e** Cross-correlation matrix of CIBERSORT results before and after adjustment by total leukocyte content. Results are shown for lung squamous cell carcinoma (*LUSC*) microarrays profiled by TCGA ($n = 130$ tumor samples). ESTIMATE was used to infer total leukocyte content, denoted *immune score*. **f** Average representation of the six immune subsets inferred by TIMER across 23 TCGA cancer types. **g** Impact of source datasets on tumor gene expression levels following batch correction. Li et al. applied ComBat [17] to merge expression profiles of bulk tumors with a reference database containing six immune cell types with variable representation. Here, the number of dendritic cell (*DC*) samples in the authors' reference database ($n = 88$) was randomly sampled from 1 to 88 while the remaining immune subsets were left unchanged. For each iteration, ComBat was applied to merge the reference immune profiles with RNA-Seq data from LUSC, which we used as a representative TCGA cancer type ($n = 555$ tumors). The median expression level of each DC marker gene (used in Li et al. and originally obtained from [31]) was determined across the LUSC cohort; markers are represented as medians, quartiles, and 10th and 90th percentiles. **h** Analysis of the number of immune reference samples versus the relative fraction of each immune subset inferred by TIMER across TCGA (colored as in panel **f**)

normalized results by a separate estimate of overall immune content (Fig. 2e). While we acknowledge that TIMER estimates were not intended to be analyzed in relative space (as described below), the same reasoning should have been applied by Li et al. to CIBERSORT; that is, CIBERSORT relative abundance estimates should not have been directly compared with absolute leukocyte abundance (as in Table S6 from [4]). Collectively, these

data highlight the importance of data normalization in comparing gene expression deconvolution methods.

We and others have previously shown that regression-based gene expression deconvolution can robustly quantify cell-type proportions [5–7, 15, 16]. Therefore, we were surprised by the claim that "levels of different cell types are not comparable" in the output of TIMER [4]. Upon further examination of this output, we found disproportionately high levels of rare dendritic cells (DCs) across all 23 cancer types (approximately 50% by inferred fractional abundance; Fig. 2f), suggesting problems with marker gene selection and/or data normalization. We hypothesized that this result might be due to the authors' use of ComBat [17] to purge batch effects between two highly distinct sample types: bulk tumors profiled by TCGA and a knowledgebase of purified leukocytes used for signature genes. In support of this hypothesis, we found that the number of DC reference profiles in the knowledge base was strongly correlated with the expression of DC marker genes in normalized tumors (Fig. 2g). Further analysis revealed a strong association between predicted abundance in TCGA (Fig. 2f) and representation in the knowledge base for all six leukocyte subsets (Fig. 2h). Thus, misapplication of ComBat distorted important biological signals that correlate with experimental batches [18], preventing TIMER from estimating cell type proportions.

Separately, we wish to address the claim that CIBERSORT is only applicable to microarray data [2, 4]. While microarray datasets were indeed the focus of our previous studies, this is not an inherent restriction of the deconvolution algorithm itself, which is platform agnostic. In fact, the analytical assumptions made by CIBERSORT are likely to hold for any mixture that can be modeled as a linear sum of its parts and for which an appropriate signature matrix exists. Such mixtures include RNA-Seq datasets, as others have already shown for bulk tumor profiling [19–21], and for single-cell RNA-Seq profiling [22], as well as other genomic features associated with cell lineage [23]. For example, CIBERSORT was recently used to enumerate hematopoietic subsets in bone marrow biopsies from healthy and diseased patients based on genomic patterns of nucleosome accessibility profiled by ATAC-Seq [23], demonstrating its broad applicability.

Finally, in response to this correspondence, Li et al. [24] have made a number of new claims that warrant clarification. In order to comprehensively address these claims, we have included a detailed point-by-point response, including new analyses, in Additional file 1: Figures S1 and S2). We summarize three key points:

1. The authors continue to ignore the significant impact of data normalization on deconvolution results, stating that CIBERSORT produces nonbiological negative correlations mainly due to collinearity. They dismiss the notion that regularization can help combat collinearity (despite significant literature on the topic [25], e.g. ridge regression, and [8]), and offer a flawed analysis to support their claim consisting of synthetic mixture datasets that are improperly defined since the mixed populations do not sum to 100% and are therefore unsuitable for addressing this topic (Additional file 1: Figure S1a and b).
2. Furthermore, Li et al. use a single flow cytometry experiment (Fig. 1a) to argue that closely related immune cell types should be positively correlated in abundance, whether in blood or in tumors. Since the default version of CIBERSORT produced negative correlations for the same cell types when enumerated in solid tumor biopsies, Li et al. claim these results contradict our own experimental data in Fig. 1a and are likely due to collinearity. In making these arguments, Li et al. disregard some fundamental immunological principles governing leukocyte migration patterns (see above and Additional file 1), relevant prior literature (e.g., Fig. 3a in [6]), and the main point of this correspondence (e.g., Fig. 2e). To further illustrate the impact of data normalization on deconvolution results, we extended our CIBERSORT analysis in Fig. 2e to 22 immune subsets (i.e., LM22 [6]) in TCGA LUSC tumors. As expected, the majority of pairwise correlations were positive when relative abundance estimates were scaled by total immune content, including correlations between closely related cell types (e.g., naïve versus memory B cells; Additional file 1: Figure S2). Moreover, we observed no significant association between pairwise correlations of leukocyte estimates in tumors and pairwise correlations of corresponding expression profiles in LM22 (Additional file 1: Figure S2). Therefore, leukocyte behavior is highly complex and unlikely to be distilled into simplistic comigration patterns without significant further investigation, especially without consideration of data normalization.
3. Finally, Li et al. claim that up to 25% of LM22 genes are positively correlated with tumor purity and, as a result, they contend that CIBERSORT's model is "frequently violated" when applied to tumors. Unfortunately, the authors ignore critical details of the algorithm and the LM22 signature matrix design. They also fail to consider many important factors in the interpretation of their own analyses, including the statistical significance, magnitude, and distribution of correlation coefficients, and the impact of positively correlated LM22 genes

on CIBERSORT results. When considering these variables, most of the significant positive correlations are of modest magnitude (e.g., 70% with $r < 0.2$) and only a small minority of LM22 genes are significantly positively correlated with tumor purity (3% with $r > 0.2$, approximately 0% with $r > 0.4$; Additional file 1: Figure S1c). Furthermore, since exclusion of all significantly positively correlated genes from LM22 had virtually no impact on tumor deconvolution performance (Additional file 1), we observed no empirical evidence consistent with the above claim.

In summary, our results address key conclusions in Li et al. [4, 24] and emphasize the importance of data normalization in deconvolution analyses. In particular, deconvolution methods cannot be meaningfully compared without taking normalization differences into account. By focusing on relative measures of TIL content in previous work [7], we avoided the confounding impact of tumor purity [26]. This approach has precedence in prior literature, particularly since many prognostic associations are more robust when defined as ratios of functionally distinct TILs (e.g., CD8 T cells versus Tregs, lymphocytes versus neutrophils, etc.) [7, 27, 28]. Whether absolute or relative measures of TIL abundance better capture tumor immunology in clinical settings remains an important consideration for future studies.

Additional file

Additional file 1: Detailed response to Li et al., [24]. (DOCX 561 kb)

Abbreviations

DC: Dendritic cell; LLSR: Linear least squares regression; LUSC: Lung squamous cell carcinoma; PBMC: Peripheral blood mononuclear cell; TCGA: The Cancer Genome Atlas; TIL: Tumor-infiltrating leukocyte

Authors' contributions

AMN and AAA wrote the manuscript with input and edits from all authors. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University, Stanford, California 94305, USA. [2]Division of Oncology, Department of Medicine, Stanford Cancer Institute, Stanford University, Stanford, California 94305, USA. [3]Center for Cancer Systems Biology, Stanford University, Stanford, California 94305, USA. [4]Department of Radiology, Stanford University, Stanford, California 94305, USA. [5]Department of Radiation Oncology, Stanford University, Stanford, California 94305, USA. [6]Stanford Cancer Institute, Stanford University, Stanford, California 94305, USA. [7]Division of Hematology, Department of Medicine, Stanford Cancer Institute, Stanford University, Stanford, California 94305, USA.

Received: 31 May 2017 Accepted: 12 June 2017
Published online: 05 July 2017

References

1. Newman AM, Alizadeh AA. High-throughput genomic profiling of tumor-infiltrating leukocytes. Curr Opin Immunol. 2016;41:77–84.
2. Aran D, Butte AJ. Digitally deconvolving the tumor microenvironment. Genome Biol. 2016;17:175.
3. Shen-Orr SS, Gaujoux R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. Curr Opin Immunol. 2013;25:571–8.
4. Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol. 2016;17:174.
5. Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PLoS One. 2009;4, e6098.
6. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12:453–7.
7. Gentles AJ, Newman AM, Liu CL, Bratman SV, Feng W, Kim D, et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. Nat Med. 2015;21:938–45.
8. Mohammadi S, Zuckerman N, Goldsmith A, Grama A. A critical survey of deconvolution methods for separating cell-types in complex tissues. arXiv. 2015: 1510.04583. https://arxiv.org/abs/1510.04583.
9. Fridman WH, Pages F, Sautes-Fridman C, Galon J. The immune contexture in human tumours: impact on clinical outcome. Nat Rev Cancer. 2012;12:298–306.
10. Shiao SL, Ganesan AP, Rugo HS, Coussens LM. Immune microenvironments in solid tumors: new targets for therapy. Genes Dev. 2011;25:2559–72.
11. Gao Q, Qiu SJ, Fan J, Zhou J, Wang XY, Xiao YS, et al. Intratumoral balance of regulatory and cytotoxic T cells is associated with prognosis of hepatocellular carcinoma after resection. J Clin Oncol. 2007;25:2586–93.
12. Stoll G, Bindea G, Mlecnik B, Galon J, Zitvogel L, Kroemer G. Meta-analysis of organ-specific differences in the structure of the immune infiltrate in major malignancies. Oncotarget. 2015;6:11894–909.
13. Rossi DJ, Bryder D, Zahn JM, Ahlenius H, Sonu R, Wagers AJ, Weissman IL. Cell intrinsic alterations underlie hematopoietic stem cell aging. Proc Natl Acad Sci U S A. 2005;102:9194–9.
14. Terzić J, Grivennikov S, Karin E, Karin M. Inflammation and colon cancer. Gastroenterology. 2010;138:2101–14.e2105.
15. Gong T, Hartmann N, Kohane IS, Brinkmann V, Staedtler F, Letzkus M, et al. Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. PLoS One. 2011;6, e27156.
16. Zhong Y, Wan YW, Pang K, Chow LM, Liu Z. Digital sorting of complex tissues for cell type-specific gene expression profiles. BMC Bioinformatics. 2013;14:89.
17. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27.
18. Jaffe AE, Hyde T, Kleinman J, Weinbergern DR, Chenoweth JG, McKay RD, et al. Practical impacts of genomic data "cleaning" on biological discovery using surrogate variable analysis. BMC Bioinformatics. 2015;16:372.
19. Tuong ZK, Fitzsimmons R, Wang SM, Oh TG, Lau P, Steyn F, et al. Transgenic adipose-specific expression of the nuclear receptor RORα drives a striking shift in fat distribution and impairs glycemic control. EBioMedicine. 2016;11: 101–17.
20. Mehnert JM, Panda A, Zhong H, Hirshfield K, Damare S, Lane K, et al. Immune activation and response to pembrolizumab in POLE-mutant endometrial cancer. J Clin Invest. 2016;126:2334–40.
21. Srinivasan S, Su M, Ravishankar S, Moore J, Head PE, Dixon JB, et al. TLR-exosomes exhibit distinct kinetics and effector function. arXiv. 2016 :1608. 08565v1. https://arxiv.org/abs/1608.08565.
22. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. Cell Syst. 2016;3:346–60.e4.
23. Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet. 2016;48:1193–203.

24. Li B, Liu JS, Liu XS. Revisit linear regression based deconvolution methods for tumor gene expression data. Genome Biol. 2017. doi:10.1186/s13059-017-1256-5.
25. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference and prediction. 2nd ed. New York: Springer; 2009.
26. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat Commun. 2015;6:8971.
27. Sato E, Olson SH, Ahn J, Bundy B, Nishikawa H, Qian F, et al. Intraepithelial CD8+ tumor-infiltrating lymphocytes and a high CD8+/regulatory T cell ratio are associated with favorable prognosis in ovarian cancer. Proc Natl Acad Sci U S A. 2005;102:18538–43.
28. Templeton AJ, McNamara MG, Seruga B, Vera-Badillo FE, Aneja P, Ocaña A, et al. Prognostic role of neutrophil-to-lymphocyte ratio in solid tumors: a systematic review and meta-analysis. J Natl Cancer Inst. 2014;106:dju124.
29. 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;30:413–21.
30. Yoshihara K, Shahmoradgoli M, Martinez 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;4:2612.
31. Abbas AR, Baldwin D, Ma Y, Ouyang W, Gurney A, Martin F, et al. Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. Genes Immun. 2005;6:319–3.