**Comprehensive Benchmarking and Integration of Tumour Microenvironment Cell Estimation Methods**

Alejandro Jiménez-Sánchez[*,^,1], Oliver Cast[*,1], Martin L. Miller[^,1]

* Equal Contributions

^ Correspondence: ajs.scientia@google.com, +1 (646) 238-0035 (AJS), martin.miller@cruk.cam.ac.uk, +44 (0) 1223 769 657 (MLM).

1 - Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK.

Running title: ConsensusTME for tumour microenvironment cell estimation

The authors declare no potential conflicts of interest.

# Abstract

Various computational approaches have been developed for estimating the relative abundance of different cell types in the tumour microenvironment (TME) using bulk tumour RNA data. However, a comprehensive comparison across diverse data sets that objectively evaluates the performance of these approaches has not been conducted. Here we benchmarked seven widely used tools and gene sets and introduce ConsensusTME, a method that integrates gene sets from all the other methods for relative TME cell estimation of 18 cell types. We collected a comprehensive benchmark dataset consisting of pan-cancer data (DNA-derived purity, leukocyte methylation, and H&E-derived lymphocyte counts) and cell-specific benchmark data sets (peripheral blood cells and tumour tissues). Although none of the methods outperformed others in every benchmark, ConsensusTME ranked top three in all cancer-related benchmarks and was the best performing tool overall. We provide a web resource to interactively explore the benchmark results and an objective evaluation to help researchers select the most robust and accurate method to further investigate the role of the TME in cancer (www.consensusTME.org).


# Statement of Significance

This work shows an independent and comprehensive benchmarking of recently developed and widely used tumour microenvironment cell estimation methods based on bulk expression data and integrates the tools into a consensus approach

# Introduction

The tumour microenvironment (TME) plays an active role in tumour initiation, progression, metastasis, and treatment response (1). Thus, studying the TME is a central paradigm of cancer research. However, a great variety of stromal and immune cell types populate tumour tissues, and the complex interactions between these different components of the tumour microenvironment is still unclear. Traditionally, cells from the TME have been quantified using immunohistochemistry (IHC), immunofluorescence (IF), flow cytometry, and more recently using cytometry by time of flight (CyTOF) mass spectrometry. These methods, although accurate, are laborious, low throughput and require pre-selected cellular markers, making their application in large number of samples and measurements challenging. Thus, the systematic application of these assays for comprehensively investigating the various different cell types in the TME in an unbiased manner is limited. Single cell RNA sequencing (scRNA-seq) has begun to fill this gap, however, scRNA-seq is expensive to apply on large patient cohorts, requires specific sample preparation, and cannot be applied to existing data sets, such as The Cancer Genome Atlas (TCGA) which consists of thousands of genomically profiled and clinically well-annotated tumour samples. The study of the different cell subpopulations of the TME in TCGA has become an important goal, but also an important challenge for bioinformatics, since cell type information identity is mixed in bulk tumour transcriptomics data.

Estimation of non-cancerous cell proportions from bulk tumour samples can be performed using genomics data such as whole-exome sequencing, microarrays, RNA-seq, or DNA methylation data. During the last decade, multiple computational approaches have been developed intending to quantitatively or semi-quantitatively calculate distinct TME cell type population estimates (2). A variety of statistical frameworks and algorithmic procedures have been employed with each method using different benchmarking data sets (2). In general, two different algorithmic classes

exist into which most TME cell estimation methods can be classified: regression-based deconvolution algorithms and gene set enrichment-based methods. Importantly, both classes rely on cell type-specific markers that are selected according to prior knowledge. The deconvolution algorithms use signature matrices containing gene expression profiles of purified immune cells and impute to what degree each of the gene expression profile is represented in the bulk tumour gene expression profile. Gene set enrichment-based methods assign curated gene sets to represent cell types before computing enrichment scores as a function of the expression of the genes within each gene set. Both classes of computational approaches that intend to estimate TME cell content in tumours (i.e. TME cell estimation methods) require two components: a statistical framework (e.g. regression or gene set enrichment) and a signature for the specific cell type of interest (e.g. a signature matrix or signature gene set).

Cell type-specific estimation in the TME using bulk tumour data is a challenging task as certain stromal- and immune cell populations are lowly abundant cell populations and is further convoluted because expression of particular genes is rarely unique to any particular cell type. Thus, there is not a straightforward solution for accurate TME cell estimation with various different gene signatures and statistical frameworks suggested as the optimal solution (2). One of the problems in the field is that each method has claimed to outperform others in their own benchmarking experiments (3,4). Thus, the need for independent and more comprehensive benchmarks has become increasingly important (5,6).

Here, we developed a consensus approach (Consensus$^{TME}$) for 18 different cell types that compiles cell type-specific genes used by seven published gene sets or existing TME cell estimation methods: Bindea *et al*. gene sets (7), Davoli *et al*. gene sets (8), Danaher *et al*. gene sets (9), CIBERSORT (10), MCP-counter (11), TIMER (12), and xCell (13). We performed pan-cancer benchmarks using publicly available bulk genomic and transcriptomic data from TCGA

85 and cell type-specific benchmarks using “ground truth” data based on experiments carried out in each of the original manuscripts where available. The Consensus$^{TME}$ approach, available as an R package, is evolvable by design allowing new gene signatures and algorithms to be incorporated and their performance compared with continuously updated benchmark data sets. Overall, the Consensus$^{TME}$ approach provides a robust and improved method for the relative cell
90 type estimation using bulk expression data of human tumour samples. We make all results accessible to the public (www.consensusTME.org) as well as provide a framework for incorporating new TME estimation methods and data sets.

# Materials and Methods

## Contact for Resource Sharing

95 Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Martin L. Miller (martin.miller@cruk.cam.ac.uk).

## Quantification and Statistical Analysis

### Single-sample gene set enrichment analysis

100 Single-sample gene set enrichment analysis (ssGSEA) (14), a modification of standard GSEA (15), was performed on RNA measurements for each sample using the GSVA package version 1.32.0 (16) in R version 3.6.0 with parameter: `method = “ssgsea”`. TME cell gene sets obtained from previous publications or described as below (7–9)

## Consensus$^{TME}$

105 To generate the Consensus$^{TME}$ gene sets we identified cell types for which there were signatures from at least two different sources were available, 18 cell types in total. To extract genes from the signature matrix "LM22" used by CIBERSORT, we first filtered out genes whose expression value was below 1.96 standard deviations of the mean for each cell type. In addition, we collapsed activated and resting states for corresponding cell types. Once we had collected signature genes
110 from Bindea et al., Danaher et al., Davoli et al. CIBERSORT, MCP-Counter and xCell we created a unique union of the genes for each cell type. From this union of genes a set of cell type-specific genes was curated for each of the TCGA cancer types. This was done using a similar approach to the TIMER algorithm where genes were only included if the expression of that gene has a negative correlation (pearson's correlation < 0.2, *p*-value $\leq$ 0.05) with tumour purity (ABSOLUTE
115 derived) for the corresponding cancer type (12,17). The assumption behind this filter is that it removes immune- and stromal cell genes which may be aberrantly expressed by cancer cells in some cancer types hereby ensuring that filtered the gene sets likely represent the presence of the implicated immune cell while also leading to gene sets that account for tissue specific variability in the immune response (18). The implementation of such an expression filter may on
120 the other hand unintentionally remove immune- and stromal-specific genes that may indeed be cancer specific and hereby exclude interesting genes that are interesting to investigate in other contexts. However, in the current version the implementation of the tumour-specific expression filter ensures that as additional signatures are incorporated into the Consensus$^{TME}$ supersets performance continues to increase. Finally, ssGSEA was employed to calculate NES for each cell
125 type as described above. General immune scores for each tumour types were generated by combining the genes of the different immune cells into one gene set for each TCGA cancer type. The Consensus$^{TME}$ method can be used through installation of the "ConsensusTME" R package via GitHub (https://github.com/cansysbio/ConsensusTME/).

## Comparison statistical metrics

130 Concordance between computational estimates and ground truth values was measured using either Kendall's rank correlation coefficient or the multiple linear regression goodness of fit metrics: adjusted R-squared, Akaike information criterion (AIC), and Bayesian information criterion (BIC). AIC and BIC z-score values were calculated to compare across different tumour types in the comparisons since AIC and BIC values are unitless. Differences between groups of
135 variables in the TIMER benchmark were identified using one-way ANOVA with Tukey honest significant differences post-hoc tests. All statistical tests were adjusted for multiple testing using the Benjamini-Hochberg procedure to control for false discovery rate (FDR).

## TCGA immune estimations

TCGA RNA-sequencing (RNA-seq) data was collected from cBioPortal (19). Batch normalisation
140 had been applied and gene expression values calculated using the "RSEM" pipeline (20). Four existing TME cell estimation methods and three published gene sets were used alongside Consensus$^{TME}$ to produce relative abundances of immune cell types per sample across 32 tumour types. For each method, a general immune score was also derived if it was not already provided, representing the total level of immune cell infiltration in each tumour sample, for the TME methods
145 that were gene sets this was done by collapsing the genes for each cell type together to form a new category and for the regression based methods this was done by summing the regression coefficients as per method used by CIBERSORT algorithm with parameters: `absolute = TRUE, abs_method = 'sig.score'`.

## TME cell estimation methods

150 TME estimation methods were used to estimate abundances of cell types from RNA expression profiles. The TME methods that were benchmarked were CIBERSORT (run in absolute mode)

(10), MCP-counter (11), TIMER (12), xCell (13), as well as gene sets collected from three previous publications (7–9) (Supplementary Table 1A).

## Bindea et al., Danaher et al. and Davoli et al. gene sets

Gene sets provided by Bindea et al., Danaher et al. and Davoli et al. were used with ssGSEA to provide enrichment scores for each of the immune signatures (7–9). To generate general immune scores, genes selected for immune cells where combined into one gene set for each method independently.

## xCell

The “xCell” R package (version 1.12) was used to generate immune estimates for the xCell method (13). A general immune estimation score is already generated by xCell.

## MCP-counter

Estimations for the MCP-counter method were produced using the “MCPcounter” R package (version 1.1.0) (11). Immune scores for this method were produced in a similar manner as the ssGSEA methods by creating a union of signature genes for each of the cell types. The “`MCPcounter.estimate`” function was altered to allow for the new signature.

## CIBERSORT

CIBERSORT estimations were produced using the R source code, provided on request from the web resource (10). CIBERSORT was run in “Absolute mode” (under beta development) using 100 permutations and quantile normalisation disabled as recommended for RNA-seq data.

175 Absolute scores representing the “overall immune content” is produced natively by the algorithm in absolute mode.

## TIMER

TIMER estimations were produced using R source code, available from the web resource (12).
180 Immune scores for TIMER were produced as a sum of the coefficients for each cell type.

## Purity score benchmark

Pan-cancer purity scores were downloaded from the NIH Genomic Data Commons (https://gdc.cancer.gov/about-data/publications/PanCan-CellOfOrigin) (Supplementary Table 1B)
185 (21). Purity scores were generated using ABSOLUTE (17) which uses copy number, variant allele frequency, and tumour specific karyotype data to calculate the cancer fraction of a tumour samples. To benchmark the immune estimation methodologies using purity of samples the immune scores were added to an independent stromal score; calculated through the use of ESTIMATE (version 1.0.13) (22). The stromal score was added as this could negatively affect the
190 performance of estimating RNA-based tumour purity for TME cell estimation methods which only estimate immune cells compared to those that estimate both immune- and stromal cells. ABSOLUTE’s derived tumour purity and the different TME methods tumour purity scores were correlated independently for each tumour type.

## 195 Leukocyte fraction benchmark

Methylation derived leukocyte fraction scores were downloaded from the NIH Genomic Data Commons (https://gdc.cancer.gov/about-data/publications/PanCan-CellOfOrigin) (Supplementary Table 1B) (21). Multiple linear regression was then employed using TME method cell type estimates that were in the category of being leukocytes (Supplementary Table 1C) as

200 explanatory variables and the methylation derived leukocyte fraction as the response variable. Leukocyte methylation data was log transformed to meet the normality and heteroscedasticity assumptions of the model. Adjusted R squared, AIC, and BIC metrics were calculated to compare the goodness of fit between the methods while taking into consideration the number of variables included in the model.

205

## Somatic single nucleotide mutation data

Somatic single nucleotide mutation data was downloaded from the Broad Institute GDAC Firehose (https://gdac.broadinstitute.org/).

## 210 H&E deep learning lymphocyte fractions benchmark

Lymphocyte fractions were generated by Saltz et al. for 13 TCGA cancer types using deep learning based image analysis (Supplementary Table 1B) (23). Multiple linear regression was applied in a similar manner as for the leukocyte methylation analysis, instead using a hyperbolic sine transformation of lymphocyte fraction as a response variable to meet normality and
215 heteroscedasticity assumptions of the model. Models for each method were fitted using only method estimates of lymphocytes as explanatory variables (Supplementary Table 1C).

## Independent cell type-specific benchmarks

The benchmarking validation analyses for each of the methods were replicated, where possible,
220 to match the parameters used in the original publications. Of the seven methods there were four benchmarking datasets available; either online or provided by the authors. Each of the datasets contained samples with bulk gene expression values along with matched “ground truth” values (Supplementary Table 1B). The CIBERSORT benchmarking dataset, provided by the authors on request (10), consisted of flow cytometry values of different immune cell types from PBMC

225 samples. The xCell benchmarking datasets, SDY311 and SDY420, were publicly available for download from ImmPort (24), and consisted of RNA-Seq and matching CyTOF quantification of immune cells from PBMC samples. The MCP-counter publication used gene expression profiles from GEO (accession number GSE39582) and IHC counts of CD3+, CD8+, and CD68+ cells (available on request from the authors) (11). The TIMER benchmark consisted of H&E stained
230 slides from TCGA Bladder urothelial carcinoma (BLCA) study. A pathologist manually reviewed each of these slides to categorise each sample into one of three categorical levels for neutrophil abundance: “Low”, “Medium” or “High”; estimations are available from the TIMER online resource (12). For all benchmarking experiments, except TIMER, concordance was measured using correlation between “ground truth” values and the immune estimations of each method. Due to
235 the variation in the degree of specificity to which cell subsets were defined, summations of subsets was required to allow accurate comparisons in some cases (Supplementary Table 1D). For the TIMER benchmark, samples were grouped by low, medium and high pathological estimation, then TME method estimates were compared by ANOVA with Tukey post hoc.

# Results

## 240 Consensus cell type-specific gene supersets for estimating tumour microenvironment cell populations

Following the generation of large data sets of tumour genomic profiles such as TCGA and the International Cancer Genome Consortium (ICGC), various approaches to assessing TME cell populations have been developed, each using different algorithms, gene markers, and validation
245 benchmarks (2). To build on the knowledge of cell type-specific gene sets represented in the diversity of these TME cell estimation methods, we sought an integrative strategy that incorporates knowledge from existing signatures and statistical approaches. The Consensus$^{TME}$

method integrates cell type-specific gene markers from independent TME cell estimation methods and uses single sample gene set enrichment analysis (ssGSEA) to compute relative TME cell type- and tumour-specific enrichment scores from bulk expression data (Fig. 1A). The first step of the Consensus$^{TME}$ approach is the selection of publicly available signatures or tools which for the current version included four tools: CIBERSORT, TIMER, MCP-counter, xCell, and three gene sets curated and used by Bindea et al., Danaher et al. and Davoli et al. Second, cell types are selected when at least two methods estimate their abundance. Third, we generated a gene set for each cell type by using the union of genes used by the methods to estimate that cell type. Fourth, we removed genes that correlate (rho > -0.2) with tumour purity in a tumour type-specific manner as applied in the TIMER method (12) (see Methods). Fifth, for statistical framework the ssGSEA approach was selected because gene set enrichment frameworks treat microarray and RNAseq values in the same way, being based on the ranked genes rather than the mRNA transcript abundance values. In addition, ssGSEA outperforms other gene set enrichment calculations with the Consensus$^{TME}$ gene sets (Supplementary Fig. S1A). Finally, the output from Consensus$^{TME}$ are normalised enrichment scores (NES), which accounts for variations in gene set size (14) and captures the relative level of estimated abundance of specific cell types across samples. For example, in a sample with a high NES for a specific cell type, the individual genes of the gene set would rank higher in a sorted list of the abundance level of all genes in the transcriptome compared to samples with low NES of the same cell type. In sum, to improve performance, Consensus$^{TME}$ aggregates cell type specific genes that have been independently considered relevant by different methods, and estimates their relative abundance in a tumour type specific manner (Supplementary Fig. S1B) .

## Pan-cancer leukocyte and lymphocyte TCGA benchmarks

To benchmark the different methods in an objective and systematic manner, we used publicly available data from multiple tumour types comprising 9,142 samples in total. All immune

estimation methods were utilised setting appropriate parameters but without optimisation requiring changes to source code. This included running CIBERSORT absolute mode and TIMER using the appropriate cancer type for the dataset. For each of the Bindea et al., Danaher et al. and Davoli et al. gene sets, ssGSEA was applied with the same parameters as Consensus$^{TME}$ (see methods). First, to evaluate the ability of each TME cell estimation method to capture the overall amount of immune component in the TME, we calculated total immune scores independently with each method. Briefly, when immune scores were not calculated by default by a method, we customly derived it by either creating a unique union of all genes for gene set enrichment methods and adding that as an additional signature or for regression based approaches summing the coefficients from all cell types (see Methods). Since tumour purity does not account only for immune cell infiltration, but also for other stromal cells (e.g. fibroblasts and endothelial cells) (25), we inferred stromal non-immune related content of all samples using ESTIMATE (22) and added this value to each of the TME estimation methods' immune scores to create a RNA-based immune- and stromal composition score (TME score). We found that all seven methods and Consensus$^{TME}$ perform very similar to each other as estimated by the correlation (Kendall's correlation coefficient, $\tau$) between the RNA-based TME score calculated from each of the methods and the DNA-derived tumour purity score based on ABSOLUTE (17,21) (Fig. 1B). As expected, in nearly all comparisons (all methods by all cancer types) there was a negative correlation between the TME score and the DNA-based tumour purity score. Across all cancer types (pan-cancer) we found that CIBERSORT, Consensus$^{TME}$, and Danaher performed as the top 3 TME estimation methods. Furthermore, using this specific correlation measure, all the methods performed well overall with many correlations being statistically significant, but there were some cancer types in which all methods performed poorly, e.g. pancreatic adenocarcinoma (PAAD). Variation in performance was largely independent of cancer cellularity, mutation load, leukocyte fraction, and sample size for any of the methods.

We further evaluated the performance of the TME estimation methods by using leukocyte
300 fractions derived from DNA methylation data for 30 tumour types (see methods) (21). The leukocyte DNA methylation data, which is not a gold standard in itself, was used as an orthogonal inference for leukocyte infiltration across tumour types studied by the TCGA consortium. To assess the performance of each of the methods using leukocyte methylation data and to account for accuracy across multiple cell types, we fitted multiple linear regression models using the
305 leukocyte fraction as a response variable and only cell type estimates in the category of being leukocytes as explanatory variables for each method (Supplementary Table 1C). As there are many cell types classed as leukocytes, multiple linear regression was chosen as it allows us to assess how variation across all cell types in combination could explain the variation in the methylation derived leukocyte fraction. Since different methods estimate different number of
310 leukocytes, the coefficient of determination can be artificially increased by the number of variables in a model (i.e. overfitting). Thus, to more appropriately compare the models in an unbiased way we used adjusted coefficients of determination ($R^2$), the Akaike information criterion (AIC), and Bayesian information criterion (BIC) for model comparison. These penalise model complexity (i.e. number of cell types used in the models) to varying degrees with BIC putting the greatest
315 emphasis on finding a parsimonious model. When comparing the $R^2$ of the different models, the best performing methods were Bindea, Consensus$^{TME}$, and Danaher (Fig. 1C). Similarly, AIC and BIC scores showed that Consensus$^{TME}$, Davoli, Danaher and Bindea models perform better than the other methods (models with lower AIC and BIC values are preferred).

320 Similarly, we implemented multiple linear regression analysis using tumour-infiltrating lymphocyte counts derived from digitised H&E-stained images analysed through a deep-learning convolutional neural network approach (23). The methods that showed better fit metrics were Consensus$^{TME}$, Davoli, Bindea and Danaher (Fig. 1D). Together, these broad pan-cancer benchmarks show a variation in the performance of the different methods when compared to each

325 other, and no single method consistently outperforms the others. There was also cancer specific performance variation observed that should be taken into account when considering the appropriateness of using these TME estimation methods.

## Performance evaluation based on cell type-specific datasets

330 All TME estimation methods tested, except Bindea and Davoli, performed their own independent benchmarks using experimental data with cell type-specific measurements in the original publications. We collected benchmarking data for CIBERSORT, xCell, TIMER, and MCP-counter to carry out a side-by-side comparisons. We used the CIBERSORT benchmark data that consisted of peripheral blood mononuclear cells (PBMCs) of 20 healthy human subjects quantified
335 by flow cytometry (10). Correlations between estimated immune cell types and the flow cytometry fractions showed that the best performing methods were MCP-counter, CIBERSORT, and xCell (Supplementary Fig. 1C). However, most of the correlations lacked statistical significance, and due to the different cell types estimated by the different methods it is difficult to reach a conclusion. Similarly, the xCell benchmark data set consisted of 16 PBMC leukocyte subsets from two
340 different studies with 61 and 104 healthy human subjects each, where PBMCs fractions were measured using CyTOF (13). MCP-counter, CIBERSORT, and xCell were the methods that showed best performance in these PBMC benchmarks; however many cell types did not reach statistical significance (Supplementary Fig. 1C).

345 Finally, we used cancer-related benchmarks from TIMER (12) and MCP-counter (11). For TIMER's benchmark, 404 TCGA bladder cancer samples were analysed by a pathologist who categorised them as low, medium, or high according to their neutrophil counts using H&E stained slides. Performance was assessed by measuring the significance of difference between the computational estimates of samples in each category. Here, Consensus$^{TME}$, Bindea, and TIMER

350 obtained the best separation between categories, but only Consensus$^{TME}$ and Bindea separated significantly the three categories after multiple test correction (Fig. 2A). Interestingly, xCell, CIBERSORT, and MCP-counter did not differentiate the three categories significantly, while Davoli does not estimate neutrophils. For MCP-counter's benchmark, IHC digital quantification of CD3+ (T cells), CD8+ (CD8+ T cells), and CD68+ (Monocytic lineage) cell densities were
355 analysed from 38 colorectal cancer samples. Correlations between the TME methods' estimations and the cellular fractions were computed (Fig. 2B). Consensus$^{TME}$, MCP-counter, and Danaher provided the best correlations, with Consensus$^{TME}$ performing best on the three cell types evaluated.

360 When observing the rank of methods across all benchmarking experiments (Fig. 2C) no one method was shown to consistently outperform all others. However, the integrative Consensus$^{TME}$ approach was in the top three for all cancer-based benchmarks and achieved the best mean rank of all TME estimation methods.

# 365 Discussion

With the recent generation of large publicly available molecular profiling of cancer samples, a variety of computational methods for analysis of cell components of the TME have been generated. In principle, the method of choice should be based on performance, however, the popularity and ease of use can also be reasons behind the specific method that researchers tend
370 to select (5). In the case of TME cell estimation from bulk expression data, this problem is magnified by the lack of objective and independent benchmark analyses, since most methods use their own benchmarks which may introduce biases and reliance on one type of data. Here we performed an unbiased and objective benchmarking exercise comparing seven of the most widely

used and recent TME estimation methods, and also developed Consensus$^{TME}$: a gene set enrichment based method that integrates cells types and genes from these seven different methods in order to generate a consensus gene superset for each cell type. The novelty of Consensus$^{TME}$ is that it a) draws on the current knowledge of cell specific gene by integrating gene sets from publicly available methods, b) provides a new TME cell estimation method that consistently performs well with the best mean rank across all benchmarks, and c) will evolve and improve as new methods and data sets are published. We also provide a web resource (www.consensustme.org) for evaluating current and future TME cell estimation methods across a diverse range of datasets allowing researchers to make informed decisions about the strengths and weaknesses of the different approaches to immune estimation.

The Consensus$^{TME}$ approach is an evolvable method by design, meaning as new gene sets and signature matrices become available, they can be added to existing supersets and tested with already established benchmarks, thus potentially improving its performance as new methods and gene sets are developed. The “wisdom of the crowd” phenomenon, that the collective knowledge of a group supersedes that of individuals, is an approach that has previously been used to address complex problems in computational biology (26). With the complexity and diversity of the TME that also varies across different anatomical locations, the derivation of gene sets to capture this is a challenge that is well suited to this collaborative approach. While this may lead to the inclusion of spurious genes in the supersets, the collection of multiple gene signatures can better capture the diversity in the transcriptome of immune cells in across different biological contexts. To combat this and to make gene sets specific for each cancer type, we employed a purity-based gene selection in Consensus$^{TME}$, similar to that used by TIMER (see Methods) (12). Future versions may deploy alternative approaches such as gene exclusion or inclusion lists or weighting schemes, particularly as single cell sequencing data of tumours becomes available which will

make it possible to derive cancer-specific immune- and stromal gene signatures directly. To allow
400 dissemination of Consensus$^{TME}$ we developed an easy to implement R package.

Currently Consensus$^{TME}$ leverages a ssGSEA statistical framework, but the compartmentalised approach allows for gene sets to be used with other statistical frameworks where appropriate. While our benchmark of other available frameworks showed ssGSEA to be the most robust and
405 accurate, future development on get set enrichment algorithms may produce a new gold standard in the future (Supplementary Fig. S1A). Different frameworks can be especially important when considering the intended use of the output. Consensus$^{TME}$ NES's are imputed primarily for comparison of scores relatively across samples. Similar to another ssGSEA based method xCell, Consensus$^{TME}$ NES's also resemble absolute scores, which would allow comparison across cell
410 types, however, currently this has not been benchmarked (13). Regression based methods including CIBERSORT are better suited for solving this problem of comparing across cell types, but require more careful interpretation when comparing across samples since outputs are fractional.

415 We performed pan-cancer benchmarks using orthogonal data types generated for TCGA samples. While DNA-derived tumour purity scores correlated negatively with RNA-derived TME estimations for all methods, leukocyte methylation scores showed some variation across methods, and a lack of correlation for some methods. Also, different tumour types showed varying levels of concordance with leukocyte estimations that was irrespective of features of the cancer
420 type, such as average tumour cancer cell fraction, median mutation load, or average leukocyte fraction. The methylation derived leukocyte fraction was generated by Hoadley et al. who compared the methylation patterns of pure leukocyte cells and normal tissue methylation patterns before selecting regions of differential methylation between healthy or cancer cells and immune cells (21). This approach has the potential to be complementary to bulk RNA based deconvolution

425 (21,27,28). Lymphocyte deep learning H&E quantifications provided a lower association with RNA-derived lymphocyte estimations, an observation that has been reported previously and considered to be in part due to the RNA-derived estimates reflect more cell counts, while spatial image-derived estimates reflect the fraction of lymphocytes per area (23,27). Thus, this benchmark should be used with caution due to the uncertainty of both RNA-derived and imaged-
430 based derived lymphocyte estimates. However, similar performance results to the leukocyte methylation benchmark were observed.

Cell type-specific benchmarking on the PBMC data sets showed that while MCP-counter and CIBERSORT performed best on PBMCs, in general a low number of significant correlations were
435 obtained across methods. Due to the diversity of cells tested and estimated by the different methods, the introduction of potential biases was unavoidable, for example: a method predicting only highly abundant or transcriptionally distinct cell types would likely show good correlations but may not be as informative as a method providing estimates for a wider range of cell types. Thus, obtaining a concluding result out of the PBMC benchmarks was challenging, although the
440 interactive portal can be used to compare how methods performed on a cell-by-cell basis. Moreover, given the gene expression of PBMCs in the circulation of healthy individuals is different to that of tumour infiltrating lymphocytes, for the application of TME cell estimation using bulk RNA tumour data, PBMC benchmarks may be less informative (29,30). In contrast, benchmarks using bulk tumour gene expression from both bladder and colorectal cancer tumours for
445 neutrophils, CD3+, CD68+, and CD8+ cells showed significant associations for the majority of TME estimation methods, particularly Consensus$^{TME}$. Together, these benchmarks showed that while no one method consistently outperforms others, Consensus$^{TME}$ was unique in ranking among the top three best performing methods in all cancer-related benchmarks. Our comprehensive benchmarking on diverse pan-cancer and cell type-specific data complements

recently published efforts in evaluating TME cell estimation methods focussed mainly on specific cancer types (melanoma and ovarian) and scRNA-seq data (6).

In this work, we have provided a framework for a set of objective and unbiased benchmarks for immune- and stromal cell estimation and deconvolution methods. While each benchmark individually suffers from limitations, the combination of experiments allows a broad and informed assessment of the accuracy of current methods. The constant evolution within the field of deconvolution led us to develop a shiny app for exploration of results (www.consensustme.org). Rather than these benchmarks providing a temporal snapshot of performance of TME cell estimation methods, this resource will ensure that new methods and datasets will be benchmarked and integrated into a common framework and the results can be explored in an interactive environment. Through use of this benchmarking methodology, we show that the Consensus$^{TME}$ approach provides more accurate and robust cell type estimates than any of the existing methods individually for the estimation of immune cell subtypes in the setting of bulk tumours.

# Acknowledgments

A. Jiménez-Sánchez was supported by a doctoral fellowship from the Cancer Research UK Cambridge Institute and the Mexican National Council of Science and Technology (CONACyT). O. Cast and M.L. Miller were supported by the Brown Performance Group, Innovation in Cancer Informatics Discovery Grant (BD523775). M.L. Miller was supported by Cancer Research UK core grant (C14303/A17197) and the Target Ovarian Cancer Translational Project Grant (Cambridge-1320 MM18). We would also like to acknowledge the publishing authors of each of the TME methods for sharing their benchmarking datasets.

# References


1. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Cell. 2011;144:646–74.

2. Finotello F, Trajanoski Z. Quantifying tumor-infiltrating immune cells from transcriptomics data. Cancer Immunol Immunother. 2018;67:1031–40.

3. Li B, Liu JS, Liu XS. Revisit linear regression-based deconvolution methods for tumor gene expression data. Genome Biol. 2017. page 127.

4. Newman AM, Gentles AJ, Liu CL, Diehn M, Alizadeh AA. Data normalization considerations for digital tumor dissection. Genome Biol. 2017. page 128.

5. Zheng S. Benchmarking: contexts and details matter. Genome Biol. 2017;18:129.

6. Sturm G, Finotello F, Petitprez F, Zhang JD, Baumbach J, Fridman WH, et al. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. Bioinformatics. Narnia; 2019;35:i436–45.

7. Bindea G, Mlecnik B, Tosolini M, Kirilovsky A, Waldner M, Obenauf AC, et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. Immunity. 2013;39:782–95.

8. Davoli T, Uno H, Wooten EC, Elledge SJ. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. Science. 2017;355.

9. Danaher P, Warren S, Dennis L, D'Amico L, White A, Disis ML, et al. Gene expression markers of Tumor Infiltrating Leukocytes. J Immunother Cancer. 2017;5:18.

10. 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;12:453–7.

11. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. 2016;17:218.

12. Li B, Severson E, Pignon J-C, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol. 2016;17:174.

13. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017;18:220.

14. 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;462:108–12.

15. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.

16. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7.

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;30:413–21.

18. Hu W, Pasare C. Location, location, location: tissue-specific regulation of immune responses. Journal of Leukocyte Biology. 2013;94:409–21.

19. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2:401–4.

20. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.

21. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell. 2018;173:291–304.e6.

22. 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;4:2612.

23. Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, et al. Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. Cell Rep. 2018;23:181–93.e7.

24. Bhattacharya S, Andorf S, Gomes L, Dunn P, Schaefer H, Pontius J, et al. ImmPort: disseminating data to the public for the future of immunology. Immunol Res. 2014;58:234–9.

25. Binnewies M, Roberts EW, Kersten K, Chan V, Fearon DF, Merad M, et al. Understanding the tumor immune microenvironment (TIME) for effective therapy. Nat Med. 2018;24:541–50.

26. Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, et al. Wisdom of crowds for robust gene network inference. Nat Methods. 2012;9:796–804.

27. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang T-H, et al. The Immune Landscape of Cancer. Immunity. 2018;48:812–30.e14.

28. 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;9:3220.

29. Baine MJ, Chakraborty S, Smith LM, Mallya K, Sasson AR, Brand RE, et al. Transcriptional profiling of peripheral blood mononuclear cells in pancreatic cancer patients identifies novel genes with potential diagnostic utility. PLoS One. 2011;6:e17014.

30. Sakai Y, Honda M, Fujinaga H, Tatsumi I, Mizukoshi E, Nakamoto Y, et al. Common transcriptional signature of tumor-infiltrating mononuclear inflammatory cells and peripheral blood mononuclear cells in hepatocellular carcinoma patients. Cancer Res.

2008;68:10267–79.

550 # Figure Legends

# Figure 1

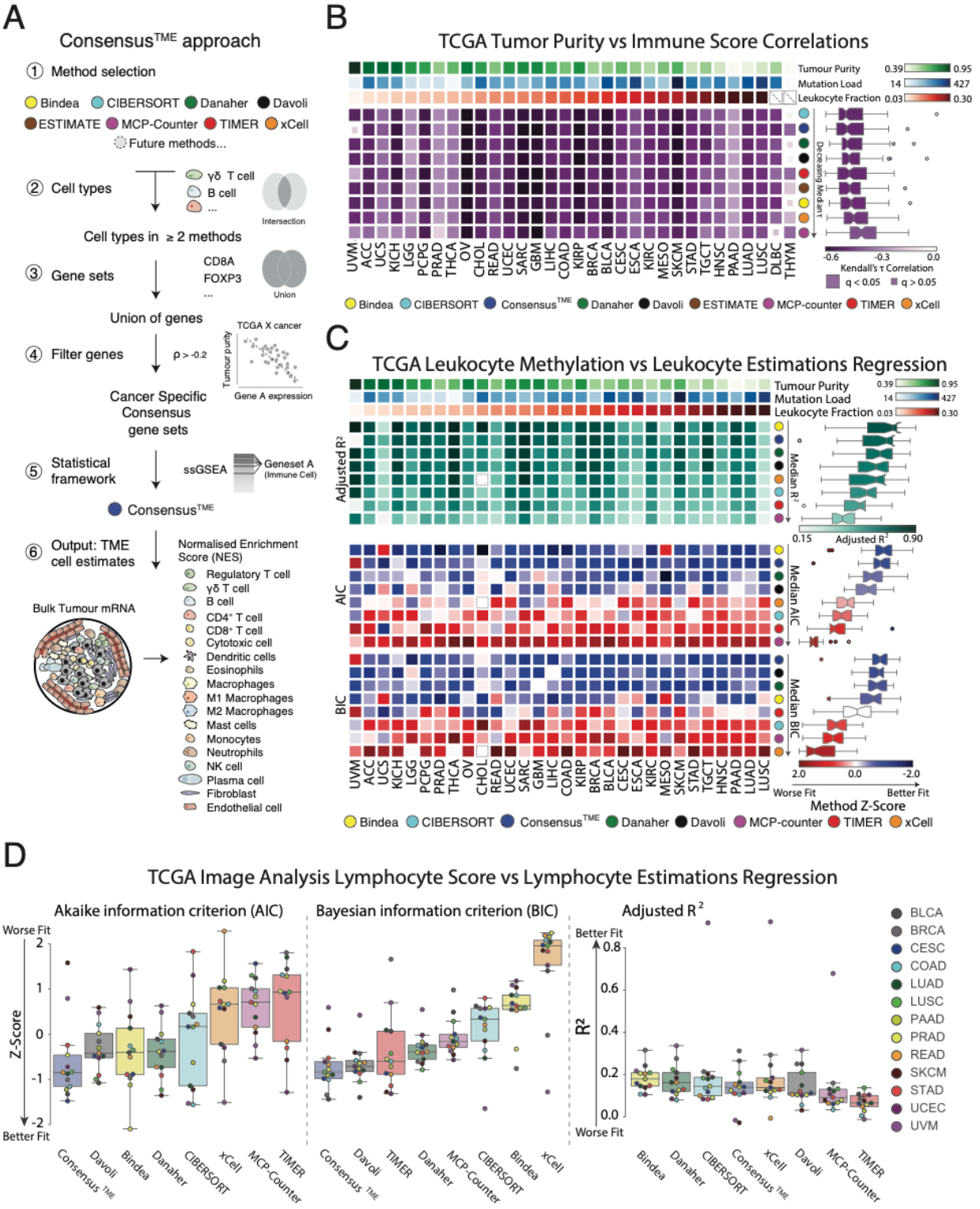
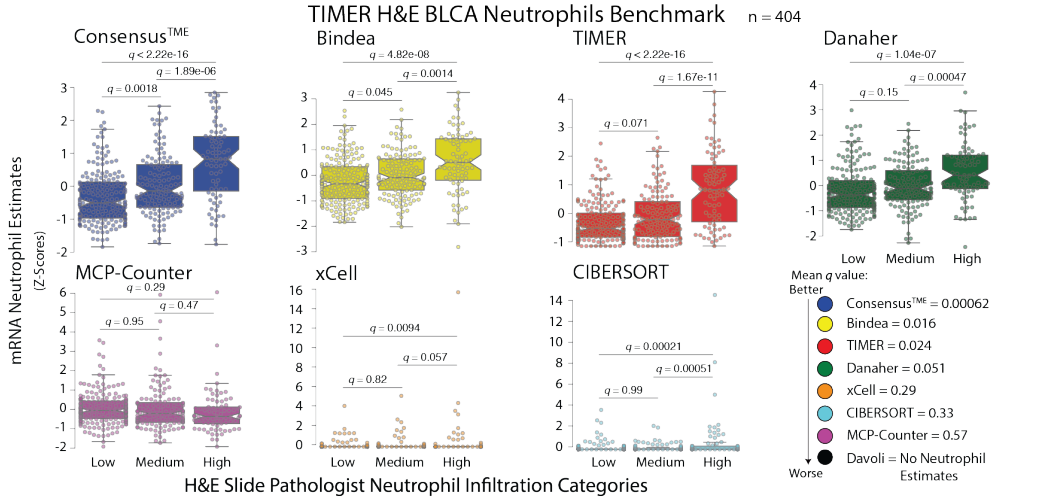

A
Consensus$^{TME}$ approach
① Method selection
Bindea CIBERSORT Danaher Davoli
ESTIMATE MCP-Counter TIMER xCell
Future methods...
② Cell types
γδ T cell
B cell
...
Intersection
Cell types in ≥ 2 methods
③ Gene sets
CD8A
FOXP3
...
Union
Union of genes
④ Filter genes
ρ > -0.2
TCGA X cancer
Tumour purity
Gene A expression
Cancer Specific Consensus gene sets
⑤ Statistical framework
ssGSEA
Geneset A (Immune Cell)
Consensus$^{TME}$
⑥ Output: TME cell estimates
Normalised Enrichment Score (NES)
Regulatory T cell
γδ T cell
B cell
CD4+ T cell
CD8+ T cell
Cytotoxic cell
Dendritic cells
Eosinophils
Macrophages
M1 Macrophages
M2 Macrophages
Mast cells
Monocytes
Neutrophils
NK cell
Plasma cell
Fibroblast
Endothelial cell
Bulk Tumour mRNA
B
TCGA Tumor Purity vs Immune Score Correlations
Tumour Purity 0.39 0.95
Mutation Load 14 427
Leukocyte Fraction 0.03 0.30
Decreasing Median τ
Kendall's τ Correlation
q < 0.05 q > 0.05
UVM ACC UCS KICH LGG PCPG PRAD THCA OV CHOL READ UCEC SARC GBM LIHC COAD KIRP BRCA BLCA CESC ESCA KIRC MESO SKCM STAD TGCT HNSC PAAD LUAD LUSC DLBC THYM
Bindea CIBERSORT Consensus$^{TME}$ Danaher Davoli ESTIMATE MCP-counter TIMER xCell
C
TCGA Leukocyte Methylation vs Leukocyte Estimations Regression
Tumour Purity 0.39 0.95
Mutation Load 14 427
Leukocyte Fraction 0.03 0.30
Adjusted R²
Median R²
0.15 Adjusted R² 0.90
AIC
Median AIC
BIC
Median BIC
2.0 0.0 -2.0
Worse Fit Better Fit
Method Z-Score
UVM ACC UCS KICH LGG PCPG PRAD THCA OV CHOL READ UCEC SARC GBM LIHC COAD KIRP BRCA BLCA CESC ESCA KIRC MESO SKCM STAD TGCT HNSC PAAD LUAD LUSC
Bindea CIBERSORT Consensus$^{TME}$ Danaher Davoli MCP-counter TIMER xCell
D
TCGA Image Analysis Lymphocyte Score vs Lymphocyte Estimations Regression
Akaike information criterion (AIC)
Bayesian information criterion (BIC)
Adjusted R²
Worse Fit
Better Fit
Z-Score
R²
Consensus TME Davoli Bindea Danaher CIBERSORT xCell MCP-Counter TIMER
BLCA BRCA CESC COAD LUAD LUSC PAAD PRAD READ SKCM STAD UCEC UVM

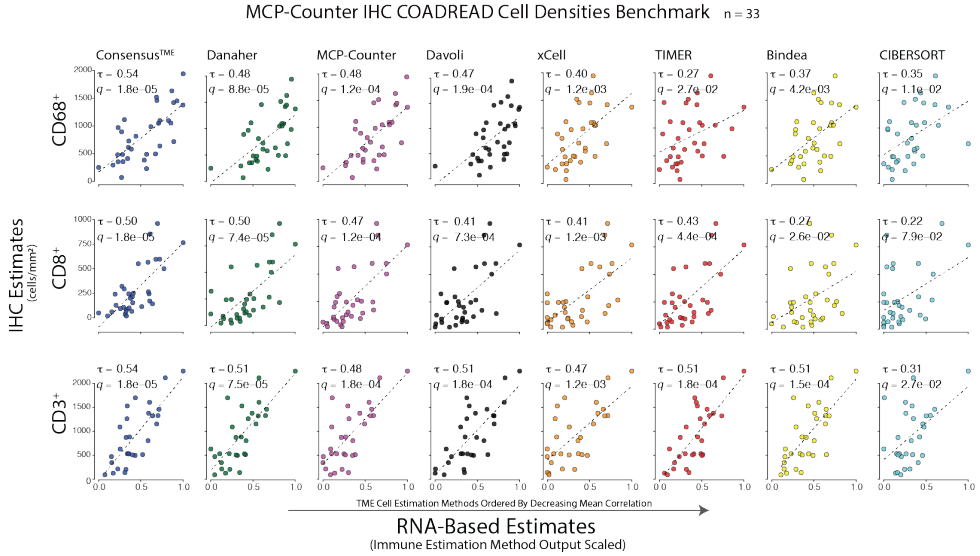
**Figure 1: Benchmark of methods for estimating TME cell components using purity and leukocyte DNA methylation data from TCGA data.** A) Consensus$^{TME}$ development strategy (see Methods). B) Each heatmap square represents Kendall's correlation coefficients (τ) of DNA-derived ABSOLUTE tumour purity scores (17,21) and RNA-derived TME score estimated by the different methods for the named TCGA cancer type. Box plots represent each methods performance across cancer types. C) Multiple linear regression models of leukocyte methylation scores as response variable (21) and RNA-derived leukocyte estimations as explanatory variables (Supplementary Table 1C). Column heatmaps are sorted according to methylation derived leukocyte fraction (Left: Low, Right: High), rows are sorted according to median performance (Top to Bottom: Decreasing performance). D) Multiple linear regression models of deep learning H&E-derived lymphocyte counts as response variable (23) and RNA-derived lymphocyte estimations as explanatory variables (Supplementary Table 1C). Adjusted $R^2$, Akaike Information Criterion (AIC) z-score, and Bayesian Information Criterion (BIC) z-score were compared across models generated by each methods cell type estimation. Lower AIC and BIC values represent a better goodness-of-fit penalising the number of variables. Mutation load: Median number of single nucleotide variants (SNVs). Tumour purity: DNA-derived ABSOLUTE cancer cell fraction. Leukocyte fraction: leukocyte cell fraction based on methylation data. Bar plots are sorted according to median correlation coefficient (Left to Right: Decreasing performance).

# Figure 2

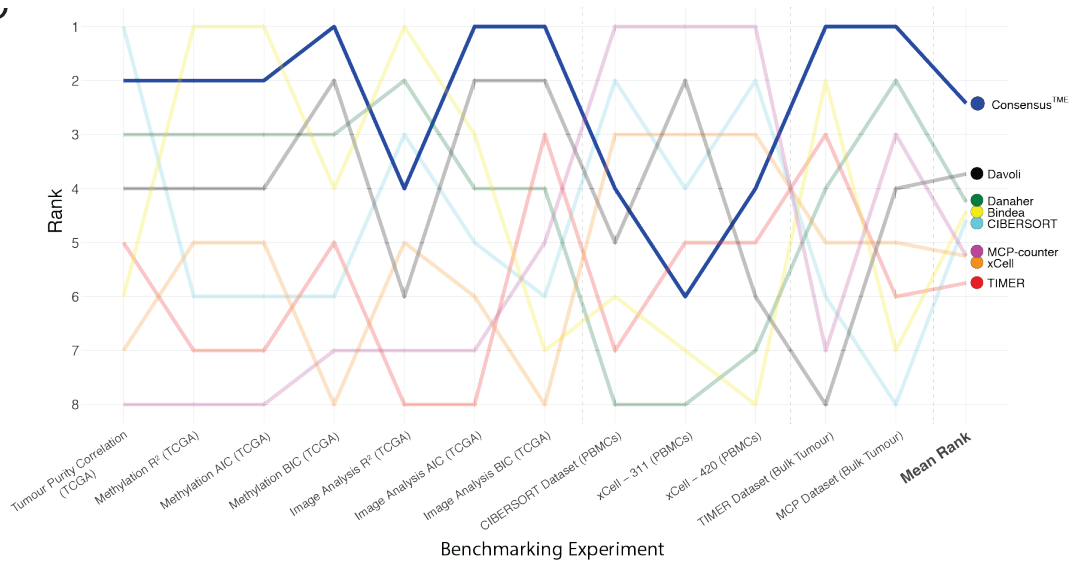
A


TIMER H&E BLCA Neutrophils Benchmark n = 404
Consensus^TME
Bindea
TIMER
Danaher
MCP-Counter
xCell
CIBERSORT
mRNA Neutrophil Estimates (Z-Scores)
Low
Medium
High
H&E Slide Pathologist Neutrophil Infiltration Categories
Mean q value:
Better
Worse
Consensus^TME = 0.00062
Bindea = 0.016
TIMER = 0.024
Danaher = 0.051
xCell = 0.29
CIBERSORT = 0.33
MCP-Counter = 0.57
Davoli = No Neutrophil Estimates

B


MCP-Counter IHC COADREAD Cell Densities Benchmark n = 33
Consensus^TME
Danaher
MCP-Counter
Davoli
xCell
TIMER
Bindea
CIBERSORT
CD68+
CD8+
CD3+
IHC Estimates (cells/mm²)
TME Cell Estimation Methods Ordered By Decreasing Mean Correlation
RNA-Based Estimates
(Immune Estimation Method Output Scaled)

C


Rank
Consensus^TME
Davoli
Danaher
Bindea
CIBERSORT
MCP-counter
xCell
TIMER
Tumour Purity Correlation (TCGA)
Methylation R² (TCGA)
Methylation AIC (TCGA)
Methylation BIC (TCGA)
Image Analysis R² (TCGA)
Image Analysis AIC (TCGA)
Image Analysis BIC (TCGA)
CIBERSORT Dataset (PBMCs)
xCell – 311 (PBMCs)
xCell – 420 (PBMCs)
TIMER Dataset (Bulk Tumour)
MCP Dataset (Bulk Tumour)
Mean Rank
Benchmarking Experiment

570 **Figure 2: Cell type-specific benchmark using original benchmarking datasets published by the individual methods.** A) Comparison between low, medium, and high categories of BLCA (n=404 samples) neutrophil H&E pathology counts from the TIMER benchmark data. One-way ANOVA with Tukey HSD post hoc tests were employed to calculate *q*-values. B) Kendall's correlation coefficients (τ) of MCP-counter COADREAD IHC (n=38 samples) cell densities (cell/$mm^2$) against TME methods estimates for CD8+ T

575 Cells (CD8+), T Cells (CD3+) and Macrophage/Monocytes (CD68+). Plots are sorted according to median correlation coefficient (left to right: decreasing performance). C) Overview of TME cell estimation methods across all benchmarking experiments, ConsensusTME method highlighted. Mean rank of cell estimations across all benchmarking experiments shown in final column.

580