

1 Data-driven inference of crosstalk in the tumor microenvironment

2
3 Umesh Ghoshdastider [1], Marjan Mojtabavi Naeini [1], Neha Rohatgi [1], Egor Revkov [1], Angeline
4 Wong [1], Sundar Solai [1], Tin Trung Nguyen [1], Joe Yeong [2,3,4], Jaber Iqbal [2], Puay Hoon Tan
5 [2,5,6], Balram Chowbay [7,8,9], Ramanuj DasGupta [1], Anders Jacobsen Skanderup [1]*

6
7 [1] Genome Institute of Singapore (GIS), A*STAR, Singapore

8 [2] Division of Pathology, Singapore General Hospital, Singapore

9 [3] Singapore Immunology Network (SIgN), A*STAR, Singapore

10 [4] Institute of Molecular Cell Biology (IMCB), A*STAR, Singapore

11 [5] Duke-NUS Medical School, Singapore

12 [6] Department of Anatomy, Yong Loo Lin School of Medicine, National University of Singapore

13 [7] Clinical Pharmacology Laboratory, Division of Cellular & Molecular Research, National Cancer Centre, Singapore, Singapore

14 [8] Centre for Clinician-Scientist Development, Duke-NUS Medical School, Singapore, Singapore

15 [9] Clinical Pharmacology Core Laboratory, SingHealth, Singapore

16 * Correspondence: skanderupamj@gis.a-star.edu.sg

17 Abstract

18 Signaling between cancer and nonmalignant (stromal) cells in the tumor microenvironment
19 (TME) is key to tumorigenesis yet challenging to decipher from tumor transcriptomes. Here,
20 we report an unbiased, data-driven approach to deconvolute bulk tumor transcriptomes and
21 predict crosstalk between ligands and receptors on cancer and stromal cells in the TME of 20
22 solid tumor types. Our approach recovers known transcriptional hallmarks of cancer and
23 stromal cells and is concordant with single-cell and immunohistochemistry data, underlining
24 its robustness. Pan-cancer analysis reveals previously unrecognized features of cancer-
25 stromal crosstalk. We find that autocrine cancer cell cross-talk varied between tissues but
26 often converged on known cancer signaling pathways. In contrast, many stromal cross-talk
27 interactions were highly conserved across tumor types. Interestingly, the immune checkpoint
28 ligand PD-L1 was overexpressed in stromal rather than cancer cells across all tumor types.
29 Moreover, we predicted and experimentally validated aberrant ligand and receptor expression
30 in cancer cells of basal and luminal breast cancer, respectively. Collectively, our findings
31 validate a data-driven method for tumor transcriptome deconvolution and establishes a new
32 resource for hypothesis generation and downstream functional interrogation of the TME in
33 tumorigenesis and disease progression.

35 Conflicts of Interest

36 The authors declare no potential conflicts of interest.

39 Keywords

40 Cancer, Tumor Microenvironment, Systems Biology

1 The tumor microenvironment (TME) is a multi-faceted cellular environment that both
2 constrains the evolving tumor (Hanahan and Coussens, 2012) and plays a pivotal role in tumor
3 progression and therapeutic response (Junttila and de Sauvage, 2013). Existing experimental
4 methods to characterize the TME, such as imaging and single-cell based approaches, cannot
5 be applied retrospectively to existing large-scale bulk tumor datasets, representing a vast and
6 mostly unexplored resource for studying cancer cell ligand-receptor repertoires and cross-talk
7 in the. Existing approaches to deconvolve bulk tumor gene expression profiles have mainly
8 focused on estimation of cell type fractions (Aran et al., 2017; Newman et al., 2015) or have
9 been optimized or tested on a limited set of tumor types (Ahn et al. 2013; Moffitt et al., 2015;
10 Quon et al., 2013; Wang et al., 2018).

11 Previous studies have analysed tumor purity in a pan-cancer context (Aran et al., 2015)
12 and developed regression-based methods for accurate mixed-tissue transcriptome
13 deconvolution (Shen-Orr et al., 2010). Here we combined these approaches to estimate
14 cancer and stromal (comprising any non-cancer cell) compartment molecular profiles from
15 bulk tumor RNA-seq data and infer signaling interactions between average representative
16 cells in these two compartments. Uniquely, our approach (TUMERIC) avoids making
17 assumptions about the transcriptional profiles of cancer and stromal cells in a given tumor by
18 integrating independent estimates of tumor purity derived from DNA sequencing and copy
19 number profiling data. We validate the approach using public cancer genomics data, mass
20 spectrometry protein abundance data, single cell RNA-seq data, and immunohistochemistry
21 imaging data. We apply the method to ~8000 tumor samples across 20 solid tumor types from
22 The Cancer Genome Atlas (TCGA) to infer cancer and stromal cell crosstalk conserved across
23 tumor types as well as tissue specific interactions. Finally, we compare cross talk across the
24 molecular subtypes of breast cancer and infer signaling interactions specific to the aggressive
25 basal subtype.

26

27

1 **Results**

2 *Overview of approach*

3 We first estimate the cancer cell fraction (tumor purity) from somatic mutation allele
4 frequencies, DNA copy number profiles, and mRNA expression signatures of the tumors using
5 a robust consensus approach (Fig. 1 and Methods). The tumor transcriptome profiles are then
6 deconvolved into an average cancer and stromal cell profile using non-negative linear
7 regression. To infer cancer and stromal cell ligand and receptor repertoires, as well as
8 potential crosstalk between these compartments, we combine the inferred expression profiles
9 with a curated database of ligand-receptor (LR) interactions (Ramilowski et al., 2015) to
10 estimate relative LR complex concentrations under equilibrium (Methods).

11

12 *Estimation of tumor purity across 8000 tumor samples*

13 To analyze crosstalk in a pan-cancer context, we first estimated tumor purities of ~8000
14 samples across 20 solid tumor types from TCGA. Briefly, we obtained purity estimates based
15 on DNA somatic variant allele frequency (PurBayes and AbsCN-seq), copy number (ASCAT),
16 and mRNA expression (ESTIMATE) data using four existing algorithms, followed by
17 imputation and quantile normalization, to produce robust consensus tumor purity estimates
18 across all sample. Most tumors had a purity in the range 40-70%, but there was large variation
19 within and between tumor types (Fig. 2a, Suppl. Fig. 1). Pancreatic adenocarcinoma (PAAD)
20 tumors had very low purity (median ~39%), consistent with previous observations (Wood and
21 Hruban, 2012). The glioblastoma (GBM) and ovarian cancer (OV) samples had the highest
22 purity estimates, likely influenced by sample selection bias in the first phase of the TCGA
23 project.

24 We found that previously published consensus tumor purity estimates (Aran et al.,
25 2015), while positively correlated with TUMERIC estimates, were likely overestimating purity
26 by >30% as compared to genome-derived purity estimates (Suppl. Fig. 1). We then compared
27 TUMERIC consensus purity values with purity estimates from recently published
28 transcriptome deconvolution methods, using tumor purity estimates computed by the TCGA
29 pan-cancer consortium (ABSOLUTE method) as an independent benchmark (Hoadley et al.,
30 2018). One method (CiberSortX) could only be applied in three tumor types because it
31 required a tumor type specific signature matrix for deconvolution. Another method (DeMixT)
32 could only be applied in tumor types where matched normal tissue samples were available.
33 Overall, the concordance with ABSOLUTE purity was generally low for the three tested
34 transcriptome deconvolution methods (Pearson $r > 0.4$ in only 2/8 tested tumor types, Figure
35 2c). Furthermore, all methods failed ($r < 0.15$) to estimate tumor purity in at least two of the

1 tested tumor types. In contrast, TUMERIC consensus purity estimates were generally highly
2 correlated ($r > 0.75$ in 6/8 tumor types, lowest $r = 0.42$ for THCA) with TCGA ABSOLUTE
3 purity estimates across cancer types.

4

5 *Validation of approach*

6 We performed multiple analyses to evaluate the accuracy of TUMERIC in deconvolution of
7 cancer and stromal compartment transcriptomes. Firstly, known stromal (*FAP*, *CD3D*, *CD4*,
8 *CSF1R*) and epithelial-cell specific factors (*EPCAM*) showed expected strong and consistent
9 gene expression differences between stromal and cancer compartments across cancer types
10 (Suppl. Fig. 2-6). Next, since somatic copy number alterations (CNAs) are hallmarks of cancer
11 cell genomes, we reasoned that genes expressed exclusively in the stromal compartment
12 should not be affected by tumor CNAs. We used TUMERIC to infer the top-100 cancer and
13 stromal-cell specific genes in each tumor type. The cancer-specific genes in each tumor type
14 were distributed widely across the genome (Suppl. Fig. 17). We found strong correlation
15 between bulk tumor CNA and expression of cancer-specific genes, but no correlation between
16 tumor CNA and expression of stroma-specific genes (Fig. 3a). Variation in correlation between
17 tumor types could be explained by the overall prevalence of CNAs in a given tumor type,
18 where tumor types with higher levels of chromosomal instability showed higher correlation of
19 tumor CNA and expression of cancer-specific genes (Fig. 3a). Similarly, as a positive control,
20 we found that previously derived stromal and immune-cell specific genes (Yoshihara et al.,
21 2013) were inferred by TUMERIC to have markedly higher expression in the stroma
22 compartment of all tumor types (Fig 3b).

23 To test the concordance of TUMERIC with tumor single-cell RNA-seq (scRNA-seq)
24 profiling, we analyzed TUMERIC expression estimates for cancer and stromal-cell specific
25 genes identified by scRNA-seq of melanoma tumors (Tirosh et al., 2016). TUMERIC inferred
26 significantly higher stroma-compartment expression for scRNA-seq stromal-cell specific
27 genes ($P=2e-55$, Mann Whitney, two-tailed), and significantly higher cancer-compartment
28 expression for cancer-cell specific genes ($P=3.6e-4$, Fig. 3c). Next, using gene set enrichment
29 analysis, we found consistent association of compartment-wise inferred gene expression and
30 known hallmarks of cancer (e.g. cell cycle and DNA repair) and stromal cells (e.g.
31 angiogenesis and immune response) across cancer types (Fig. 3d, Suppl. Fig. 7). We then
32 evaluated the extent that the deconvolved mRNA profiles represent an accurate proxy for
33 protein levels in the cancer and stromal cells. We used TUMERIC to deconvolve protein
34 abundance data from two TCGA tumor types (Edwards et al., 2015), and found strong
35 concordance between the mRNA and protein expression profiles (Fig 3e). Finally, to test the

1 concordance of TUMERIC with immunohistochemistry (IHC) data, we identified genes with
2 high variability of cancer and stromal cell differential expression across tumor types and used
3 IHC data from The Human Protein Atlas (Uhlen et al., 2017) to confirm that expression
4 patterns of two such genes were indeed variable and concordant across tumor types (Fig 3f
5 and Suppl. Fig 8). Although this IHC comparison is limited to two genes, the combined
6 concordance of transcriptome deconvolution data with both proteomic and IHC data supports
7 an overall concordance between transcriptome and protein-level data. Overall these diverse
8 comparisons with orthogonal types of data support that TUMERIC can robustly deconvolve
9 gene expression profiles of cancer and stromal cell compartments across the 20 tested tumor
10 types.

11

12 *Compartment specific expression of ligands and receptors*

13 To explore ligand-receptor (LR) signaling across cancer types, we focused on 263 ligands and
14 242 receptors (603 LR pairs) with detectable bulk tumor gene expression (>1 RPKM in >25%
15 of samples) in at least half of the 20 cancer types. We first analyzed for compartment
16 specificity of these ligands and receptors. Expectedly, ligands belonging to the complement
17 system (e.g. *C1QB* and *C3*) as well as leukocyte specific chemokines (e.g. *CCL5* and *CCL21*)
18 had high stroma specific expression across tumor types (Fig 4a). Cancer-specific ligand
19 expression across tumor types was much less frequent and pronounced. Among the most
20 common cancer-specific ligands were endothelial cell targeting factors (*PODXL2* and
21 *VEGFA*), consistent with the hypothesis that cancer cells interact with and induce tumor
22 vascularization. Immune (e.g. *CD2* and *CD4*) and macrophage cell (*CSF1R*) specific factors
23 were expectedly among the top stromal specific receptors across tumor types (Fig 4b). Similar
24 to ligands, receptors with cancer-cell specific expression across tumor types were less
25 common. The top common cancer specific receptors included known cancer associated
26 members of the EGF-family (*ERBB2* and *ERBB3*), Wnt-family (*LRP4*, *LRP5* and *LRP6*), and
27 FGF-family (*FGFR3*). In summary, these data underline that our approach is able to tease
28 apart cancer and stromal-compartment specific expression of ligands and receptors from bulk
29 tumor transcriptome profiles.

30

31 *Recurrence of ligand-receptor interactions across tumor types*

32 Next, we developed the Relative Crosstalk (RC) score to quantify and differentiate between
33 types of autocrine and paracrine (here denoting signaling within or between compartments,
34 respectively) ligand-receptor (LR) crosstalk (Fig 4c and Suppl. Table 1). We first evaluated the
35 extent that known LR pairs had consistent crosstalk directionality across tumors types and

1 found a striking difference between the cancer and stromal compartment. While only 3 LR
2 pairs had strong autocrine cancer signaling scores (median RC score > 40%) across tumor
3 types, 264 LR pairs showed strong autocrine stroma signaling RC scores across cancer types
4 (Fig 4d). This suggests that, in solid tumors, autocrine cancer signaling tends to be tumor type
5 specific and determined by the cancer cell-of-origin. In contrast, stromal autocrine signaling is
6 often conserved and independent of tumor type and tissue. Interestingly, the paracrine
7 signaling interface between cancer and stromal cell compartments also had a high number of
8 recurrent interactions (26-40 conserved interactions, respectively), highlighting the importance
9 of the tumor environment on cancer cell biology.

10 Inferred recurrent cancer-cell autocrine LR pairs included receptors such as *FGFR8*,
11 *LRP6* and *MST1R* (Fig. 4e). Signaling through *ACVR2B* was notably both among the top
12 inferred cancer autocrine and stroma-to-cancer signaling interactions across tumor types (Fig
13 4e,g), highlighting a common role for ACVR-coupled TGF-beta/SMAD signaling in
14 tumorigenesis. We analyzed recurrent cancer-to-stroma LR interactions to investigate how
15 cancer cells may engage and shape their microenvironment (Fig 4f). Top inferred cancer-to-
16 stroma interactions included the lymphocyte-specific selectin (*PODXL2-SELL*), highlighting a
17 possible conserved interaction between cancer cells and leukocytes or endothelial cells
18 (Fieger et al., 2003) (only the leukocyte-specific selectin (*SELL*) was abundantly expressed
19 across most tumors and included in our analysis).

20

21 *PD-1 and CTLA-4 immune checkpoint ligands are overexpressed in stromal cells*
22 The top recurrent stroma-stroma interactions included many chemokine signaling interactions,
23 especially involving *CXCR3* and *CXCR5* receptors, suggesting common stroma induced
24 recruitment of leukocytes and lymphocytes to solid tumors (Suppl. Fig 9). Interestingly, the top
25 recurrent stromal interactions also included the known immune checkpoint interactions
26 *CD86/CTLA-4* (*CD86/CTLA4*) and *PD-L2/PD-1* (*PDCD1LG2/PDCD1*). The known *PD-L1/PD-1*
27 (*CD274/PDCD1*) checkpoint interaction was not among our initial curated and analyzed
28 ligand-receptor interactions, but a follow-up analysis of this LR pair revealed a similar stroma-
29 stroma enriched cross-talk pattern (median stroma-stroma RC=87%). Interestingly,
30 expression levels of immune checkpoint ligands (PD-L1, PD-L2, CD86) were highest in stroma
31 across all tumor types. This suggests that the bulk of immune checkpoint inhibitory signals in
32 the tumor microenvironment may be mediated by non-cancer cells. However, some tumor
33 types showed moderate to high expression of especially PD-L1 and CD86 in cancer cells.
34 Glioblastoma (GBM) was the tumor type with highest expression of both CD86 and PD-L2 in
35 both stroma and cancer cells. In contrast, brain tumors (GBM and LGG) also had the lowest

1 stromal expression of both the CTLA-4 and PD-1 receptor as compared to all other tumor
2 types, highlighting potential differences in T-cell infiltration levels, activation states, or
3 dynamics of checkpoint inhibition in brain tumor tissue.

4

5 *Expression of immune checkpoints and response to immune checkpoint inhibition*

6 Expectedly, inferred gene expression of the two checkpoint receptors, PD-1 and CTLA-4, was
7 high in stroma and almost absent in cancer cells (Fig. 4h). Stromal expression of these two
8 receptors was highest in melanoma (SKCM), the solid tumor type where immune checkpoint
9 blockade was first clinically approved and where therapeutic blockade currently shows the
10 most profound response rates (Sharma and Allison, 2015; Yarchoan et al., 2017). PD-L1 was
11 inferred to have high expression in cancer cells of cervical squamous cell carcinoma (CESC),
12 lung squamous cell carcinoma (LUSC), and renal cell carcinoma (KIRC, KIRP) (Fig 4h, Suppl.
13 Fig 9). These tumor types all have high response rates to PD-L1/PD-1 checkpoint inhibition
14 (Yarchoan et al., 2017). In contrast, colorectal cancer (CRC) has one of the poorest response
15 rates to checkpoint inhibition (Yarchoan et al., 2017) and also showed the lowest (near-zero)
16 inferred cancer cell expression of PD-L1 as compared to other tumor types. Overall, these
17 results indicate that bulk tumor deconvolution of immune checkpoint cell-cell interactions in
18 cancer and stroma compartments may provide insights into deciphering the conditions for
19 effective immune checkpoint therapy.

20

21 *Convergence of autocrine cancer cell crosstalk across tumor types*

22 The pan-cancer analysis indicated that cancer-cell directed LR interactions tended to differ
23 across tumor types. We therefore searched for LR interactions with extreme autocrine RC
24 scores in individual tumor types. Indeed, we found that most (14/20) cancer types had multiple
25 LR interactions with high cancer autocrine scores (RC score > 60%, Fig 4i). Interestingly, the
26 top autocrine cancer-cancer LR interactions tended to converge on the same signaling
27 pathways. 9/20 solid tumor types had (among top-2, Fig 4i) autocrine interactions converging
28 on TGF-beta and SMAD signaling pathways (BMP and ACVR-receptors). 8 tumor types had
29 interactions involving different fibroblast growth factor ligands and receptors, and 6 tumor
30 types had interactions converging on Eph/ephrin signaling. These results suggest that cancer
31 cells of solid tumors may often be dependent on autocrine activation of these signaling
32 pathways, and our data highlight specific LR candidate interactions associated with this
33 activation in each tumor type.

34

35 *Validation of compartment-specific ligand-receptor expression using scRNA-seq data*

1 We next investigated compartment-specific expression of candidate ligand-receptor pairs
2 using a melanoma scRNA-seq dataset (Tirosh et al., 2016). Firstly, we analyzed the 3 pan-
3 cancer cancer-to-cancer pairs with highest inferred cancer-cancer signalling in SKCM (*BMP7*
4 > *ACVR2B*, *BMP8B* > *ACVR2B*, *WNT3* > *LRP6*, Fig 4e). Among these pairs, the two receptors
5 (*ACVR2B* and *LRP6*) were inferred to be overexpressed in cancer cells (~4-fold higher
6 expression) and ligands were inferred to have unaltered expression in SKCM (Suppl. Table
7 1). We then compared the expression of the 2 receptors in malignant and non-malignant cells
8 as inferred by the authors (Tirosh et al., 2016). Confirming our results, both receptors were
9 significantly overexpressed in the cancer compartment ($P < 1e-10$, Wilcoxon rank-sum, Suppl.
10 Fig. 16). Next, we analysed the top 2 inferred SKCM cancer autocrine pairs (*BMP7* >
11 *BMPR1B*, *SEMA6A* > *PLXNA2*, Fig. 4i). Within each of these two pairs, the receptor *BMPR1B*
12 and the ligand *SEMA6A* were inferred to be significantly overexpressed in cancer cells (4 and
13 7-fold, respectively, Suppl. Table 1). Confirming our results, both of these genes were also
14 significantly overexpressed in cancer cells of the melanoma RNA-seq dataset ($P < 1e-10$,
15 Suppl. Fig. 16).

16

17 *Autocrine and paracrine crosstalk in brain tumors*

18 The two brain tumor types (LGG and GBM) were both characterized by very high autocrine
19 Delta-Notch signaling scores (*DLL1-NOTCH1*, RC > 80%, Fig 4i). Both *DLL1* and *NOTCH1*
20 were inferred to have >4 fold higher expression in cancer compared to tumor stromal cells and
21 normal brain tissue in both tumor types. *DLL1* cancer cell expression was ~16-fold higher than
22 stroma and normal tissue in LGG (Suppl. Fig 10). These data are consistent with previous
23 studies showing that Notch autocrine/juxtacrine signaling is critical for glioma tumorigenesis
24 (Purow et al., 2005; Teodorczyk and Schmidt, 2015).

25 In contrast to autocrine cancer cell signaling, stroma-to-cancer crosstalk is lost when
26 cancer cells are cultured *in vitro*. Interestingly, the top-2 stroma-to-cancer specific interactions
27 for glioblastoma (GBM) converged on *EGFR* (Suppl. Fig 11). *EGFR* is a well-studied
28 oncogene in GBM where 40-50% of patient tumors have *EGFR* overexpression driven by gene
29 amplification (Brennan et al., 2013). We inferred >30-fold higher *EGFR* expression in GBM
30 cancer cells compared to the stromal compartment and normal brain tissue (Suppl. Fig 11). In
31 contrast, all canonical EGFR ligands were expressed at highest levels in the stroma (Suppl.
32 Fig 11). *SPINK1* and *AREG* ligands were inferred to be exclusively expressed in stroma, and
33 the most abundant ligand, *EFEMP1*, had ~8-fold higher expression in the stromal
34 compartment compared to cancer cells and normal brain tissue. A similar pattern was
35 observed in lower grade gliomas (LGG) (Suppl. Fig 11), suggesting that EGFR signaling in

1 glioma and glioblastoma cancer cells is driven mainly by ligands produced by the tumor
2 infiltrating stroma. These data are also consistent with the observation that glioblastoma
3 cancer cells rapidly loose EGFR amplifications during cell culture (Pandita et al., 2004; William
4 et al., 2017).

5

6 *Crosstalk associated with subtypes of breast cancer*

7 To investigate whether TUMERIC could identify differences in crosstalk between subtypes of
8 a given tumor type, we inferred cross-talk associated with basal-like triple-negative breast
9 cancers (TNBCs), a more aggressive subtype of breast cancer that generally do not
10 overexpress HER2 and estrogen receptors. Expectedly, cancer cells of basal tumors had ~60-
11 fold and 250-fold lower expression of HER2 (*ERBB2*) and estrogen (*ESR1*) receptors as
12 compared to cancer cells of HER2+ and Luminal subtypes, respectively (Suppl. Figure 12).
13 NOTCH1 receptor expression was elevated in cancer cells of basal tumors, with *MFAP2* and
14 *JAG1* as the top autocrine and paracrine specific ligands, respectively (Fig 5a-c, Suppl. Table
15 1). Consistent with this observation, cancer cells of basal tumors showed gene expression
16 signatures of notch pathway activation relative to non-basal tumors (Suppl. Figure 13). Cancer
17 cells of basal tumors also showed upregulated *FZD7* receptor expression and enrichment of
18 autocrine *WNT11* and autocrine/paracrine *WNT3* interactions. Gene set enrichment analysis
19 indicated an overall enrichment for activation of WNT/beta-catenin signaling in cancer cells of
20 basal tumors (Suppl. Figure 13). Interestingly, the Frizzled/Wnt ligand antagonist, *SFRP1*, was
21 inferred to have high expression in cancer cells of basal tumors while being nearly absent in
22 cancer cells of other breast cancer subtypes (Fig 5c). At the receptor tyrosine kinase interface,
23 cancer cells of basal tumors were characterized by increased *KIT* and decreased *RET*
24 signaling compared to non-basal tumor types. Our analysis also inferred strong up-regulation
25 of IL-6 signaling through GP130 (*IL6ST*) in cancer cells of non-basal tumors.

26 Next, we wanted to validate the TUMERIC predicted differential expression of *SFRP1*
27 and *IL6ST* in the cancer cells of basal versus luminal breast cancer subtypes. We performed
28 RNA in situ hybridization (ISH) using specific RNAscope probes generated against either
29 *SFRP1* or *IL6ST* in FFPE sections of breast tumors from each subtype (see Methods, Fig 5d).
30 Cancer cell mRNA expression in each subtype was quantified using two different approaches
31 yielding similar results (Fig 5e and Suppl. Fig 14). Remarkably, and consistent with the
32 TUMERIC prediction, we observed higher expression of *SFRP1* in the cancer cells of the basal
33 subtype, compared to the luminal tumors (Fig 5e). In contrast, *IL6ST* had lower expression in
34 cancer cells of the basal tumors as compared to the luminal subtype.

1 Overall, these data highlight putative differences in crosstalk of basal and non-basal
2 breast cancer tumors and demonstrates how our approach can be applied to study cell
3 signaling associated with specific molecular, genetic, or clinical subtypes of tumors.

4

5

1 **Discussion**

2 Tumors are heterogenous mixtures of cancer cells and infiltrating non-cancer (stromal) cells.
3 Molecular measurements obtained from a bulk tumor sample reflect an average across these
4 components. By chance, the composition of these components will vary across tumor
5 samples, introducing noise for downstream integrative analysis. However, in this study we
6 demonstrate how this variability also enables inference of (average) cancer and stromal cell
7 gene expression profiles across a collection of tumor samples. Our approach, TUMERIC, uses
8 genomic data to first produce unbiased tumor purity estimates followed by constrained
9 regression to deconvolve the matched bulk tumor transcriptome profiles. We show that this
10 approach consistently recovers known hallmarks of cancer and stromal cell transcriptomes
11 and is concordant with orthogonal single cell transcriptome and immunohistochemistry
12 imaging data. The general methodology is in theory not restricted to transcriptomic data and
13 might be used to deconvolve other types of bulk tumor molecular data such as proteomic (Fig
14 3e) or epigenetic profiles.

15 By applying the method to ~8000 tumor samples across 20 solid tumors, we inferred
16 ligand and receptor expression profiles in cancer and stromal cells across tumor types (Suppl.
17 Table 1). We then nominated potential crosstalk within and between cancer and stromal
18 compartments across tumor types. A large number of crosstalk interactions were inferred to
19 be highly stroma specific across all tumor types. These interactions included many chemokine
20 interactions as well as therapeutically important PD-1 and CTLA-4 immune checkpoint
21 interactions. Checkpoint receptors were expectedly exclusively expressed in stromal cells.
22 However, we inferred substantial expression of especially PD-L1 and CD86 ligands in the
23 cancer cells of many tumor types, consistent with the hypothesis that cancer cells can express
24 these ligands to attenuate T-cell responses (Sharma and Allison, 2015). Unexpectedly, most
25 tumor types showed even higher expression of these checkpoint ligands in the stromal
26 compartment, suggesting that the bulk of immune checkpoint inhibitory signals in the tumor
27 microenvironment may be mediated by non-cancer cells. Interestingly, the tumor types with
28 highest (e.g. lung squamous and kidney cancers) and lowest (e.g. colorectal cancer) inferred
29 PD-L1 cancer cell expression have markedly different response rates to PD-L1/PD-1
30 checkpoint inhibition in clinical trials (Yarchoan et al., 2017). Further studies are however
31 needed to explore the hypothesis that bulk tumor deconvolution of immune checkpoints can
32 provide novel insights into the conditions for effective immune checkpoint therapy.

33 Our inferred recurrent cancer-cell autocrine LR pairs included the *MST1R* (RON)
34 receptor, which acts upstream of the RAS-ERK and PI3K-AKT signaling pathways and is a
35 prognostic marker and candidate therapeutic target in many solid cancer types (Yao et al.,

1 2013). Additionally, LR interactions involving different low-density lipoprotein receptor-related
2 proteins (LRPs) were frequently recurrently directed towards cancer cells (e.g. *WNT3-LRP6*,
3 *APOE-LRP5*, *LTF-LRP11*), corroborating the importance of Wnt-signaling on cancer cell
4 biology. Furthermore, our analysis showed that a recently proposed non-canonical *CXCR4*
5 ligand, *CEL* (Panicot-Dubois et al., 2007), is frequently up-regulated in cancer cells of many
6 tumor types, indicating a mechanism by which cancer cells might perturb *CXCR4* signaling in
7 addition to the canonical *CXCL12-CXCR4* axis (Guo et al., 2015).

8 Cancer cell specific cross-talk interactions were often variable across tissues,
9 underscoring the notion that cancer cell signaling is often tissue specific and likely dependent
10 on the cell of origin. However, inferred cancer-cell autocrine cross talk often converged on
11 Wnt, TGF-beta, Ephrin, and FGFR-family signaling pathways. The importance of these
12 pathways in tumorigenesis is well established (Hanahan and Weinberg, 2011), underscoring
13 the validity and utility of our approach. Moreover, the top cancer cell-specific interactions in
14 individual tumor types highlight specific putative cancer cell signaling dependencies. Further
15 functional studies are needed to validate these interactions and test whether they represent
16 vulnerabilities that could be targeted therapeutically. Encouragingly, the inferred cancer cell
17 specific interactions in brain tumors were highly concordant with previous studies. Firstly, the
18 top inferred autocrine interaction was *DLL1-NOTCH1*, an autocrine/juxtacrine interaction
19 critical for glioma tumorigenesis (Purow et al., 2005; Teodorczyk and Schmidt, 2015).
20 Secondly, the top stroma-to-cancer specific interactions converged on *EGFR*, and all
21 canonical *EGFR* ligands were overexpressed in the stroma. Stroma-to-cancer crosstalk is
22 interesting because it comprises interactions and dependencies that are potentially lost when
23 cancer cells are cultured *in vitro* or engrafted in a non-human microenvironment. Indeed, GBM
24 cancer cells rapidly lose *EGFR* amplification and overexpression following *in vitro* cell culture
25 (Pandita et al., 2004; William et al., 2017), however, amplification and overexpression can be
26 maintained if the cancer cells are co-cultured with EGF ligands (William et al., 2017). Overall,
27 these observations are consistent with our data and suggest that our methodology and data
28 could help design *in vitro* assays and co-culture models that more accurately mimic the biology
29 of the human TME.

30 Furthermore, we demonstrate that TUMERIC can nominate differences in crosstalk
31 between molecular or clinical subtypes of a given tumor type. By inferring crosstalk specific to
32 basal breast cancers, we recovered the expected patterns for HER2 and estrogen receptor
33 expression across breast cancer subtypes. We inferred specific Notch and Wnt autocrine
34 cancer cell signaling interactions specific to basal tumors. Interestingly, the Wnt inhibitory
35 ligand, *SFRP1*, was inferred to have very high expression in cancer cells of basal tumors while

1 being nearly absent in cancer cells of other breast cancer subtypes (Fig 4d). Downregulation
2 of SFRP1 has been linked to epithelial to mesenchymal transition (EMT) in breast cancer cells
3 (Scheel et al., 2011), indicating potential differences in EMT pathways or states between
4 luminal and basal breast cancer cells. Interestingly, our analysis also inferred strong up-
5 regulation of IL-6 signaling through GP130 (*IL6ST*) in cancer cells of non-basal tumors,
6 consistent with previous studies highlighting a functional role for RET-IL6 crosstalk in ER-
7 positive breast tumors (Gattelli et al., 2013). Importantly, we could validate the predicted
8 subtype specific expression of *SFRP1* and *IL6ST* (GP130) using RNA-ISH in breast cancer
9 FFPE tissue sections. Overall, these data suggest that tumor subtypes can have profound
10 differences in crosstalk, and our analysis can provide a list of putative, testable interactions
11 underlying breast cancer subtypes.

12 We note that a key limitation of our approach is that local concentrations of ligands
13 and receptors may differ considerably from our average compartment-level estimates. While
14 this is especially important when considering between-compartment crosstalk, it likely matters
15 less for the inference of cancer-to-cancer autocrine signaling (where the same cell produces
16 the interacting ligand and receptor). We expect future high-throughput technologies such as
17 sample-level spatial transcriptomics could be combined with our cohort-scale approach to
18 provide a more accurate and comprehensive description of tumor crosstalk.

19 Overall, our study provides unique insights into the complex interactions shaping the
20 TME. Our approach is especially useful in settings where bulk tumor biopsy data is either
21 already abundant or the only feasible data source, which is the case in many clinical settings.
22 We anticipate that the method could complement existing biomarker and target discovery
23 approaches by providing computationally purified molecular profiles of cancer cells as they
24 exist inside human tumors.

25
26

27 **Methods**

28

29 *Tumor data sources*

30 We analyzed 20 solid tumor types with TCGA acronyms BLCA, BRCA, CESC, CRC (COAD
31 and READ combined), ESCA, GBM, HNSC, KIRC, KIRP, LGG, LIHC, LUAD, LUSC, OV,
32 PAAD, PRAD, SKCM, STAD, THCA and UCEC. The full names of cancer types and sample
33 size are indicated in the Suppl. Table 1. We obtained somatic mutation (SNV) and copy
34 number variation (CNV) data for 20 tumor types from the Broad Institute Firehose website

1 (See data availability section below). Uniformly processed TCGA RNA-seq (FPKM) data was
2 obtained from the UCSC Xena server.

3

4 *Tumor purity estimation*

5 We used 4 different published methods for consensus tumor purity estimation: AbsCNseq,
6 PurBayes, AscAt and ESTIMATE. AbsCNseq uses CNA segmentation and SNV variant allele
7 frequency (VAF) data of individual tumors(Bao et al.); PurBayes utilizes SNV VAF data of
8 diploid genes (inferred from CNA data)(Larson and Fridley); AscAt purity estimation is based
9 on CNA (SNP array) data, where tumor ploidy and purity are co-estimated to identify allele
10 specific CNA(Loo et al.), pre-computed AscAt tumor purity estimates for the TCGA cohort were
11 obtained from the COSMIC website (See data availability section). ESTIMATE uses mRNA
12 expression signatures of known immune and stromal gene signatures to infer tumor
13 purity(Yoshihara et al., 2013), and tumor purity values were obtained by applying ESTIMATE
14 to the TCGA RNA-seq (log2 FPKM) data. In order to derive consensus tumor purity estimates,
15 we carried out missing data imputation followed by quantile normalization separately for each
16 cancer type. Some tumor purity values were missing because the algorithms failed to
17 converge on certain input data. Additionally, we observed some instances of very high (>98%)
18 or low (<10%) purity estimates, but such cases were usually only found by a single method
19 for a given tumor and were therefore also assigned as missing data. Missing data was then
20 imputed using an iterative Principal Component Analysis of the incomplete algorithm-vs-
21 sample tumor purity matrix (using the missMDA R package(Josse and Husson, 2016)). We
22 used quantile normalization to further align and standardize the tumor purity distributions of
23 different algorithms per tumor type. Briefly, a mean reference purity distribution is computed
24 and these mean values are substituted back into the purity distributions of the individual
25 algorithms. Since ESTIMATE generated purity estimates with a large bias compared to the
26 other three methods (generally 30-50% higher, Suppl. Fig 1), we did not use ESTIMATE purity
27 values to compute the mean reference distribution (see code in Supplementary Data). The
28 final TUMERIC consensus tumor purity estimate was obtained as the mean of these
29 normalized purity values.

30

31 *Comparison of transcriptome deconvolution methods*

32 TCGA RNA-seq HTSeq count data for BRCA, GBM, HNSC, LGG, LUAD, LUSC, SKCM and
33 THCA tumor types (used for DeMixT (Wang et al., 2018), LinSeed (Zaitsev et al., 2019)) and
34 TPM data for HNSC, LUAD, LUSC tumor types (used for CiberSortX (Newman et al., 2019))
35 were downloaded from the UCSC Xenahub platform (<https://xenabrowser.net/>). All methods
36 were executed on 100 randomly selected tumor transcriptome samples from each tumor type.
37 The datasets were normalized and parameters were set according to authors instructions (see

1 Supplementary Methods). CiberSortX could only be run for tumor types where a compatible
2 signature matrix was available (HNSC, LUAD, and LUSC), and DeMixT could only be run for
3 tumor types with available normal samples (all selected tumor types except SKCM). DeMixT
4 and LinSeed were run using R v.3.6.1 and their respective R packages, CiberSortX was run
5 using its online web interface. Estimated tumor purity fractions were compared to values
6 computed by the TCGA pan-cancer consortium using the ABSOLUTE algorithm
7 (<https://gdc.cancer.gov/about-data/publications/pancanatlas>).

8

9 *Database of ligand-receptor interactions*

10 We obtained ~1400 ligand-receptors pairs supported by evidence in the literature as compiled
11 and curated by Ramilowski et al. (Ramilowski et al., 2015). 7 additional known immune
12 checkpoint interactions (Khalil et al., 2016) were manually added to the analysis: CD274-
13 PDCD1, CD80-CTLA4, CD80-CD28, NECTIN2-CD226, NECTIN2-TIGIT, PVR-TIGIT,
14 SIGLEC1-SPN. The complete list of ligand-receptor interactions can be found in Suppl. Table
15 2.

16

17 *Cancer-stroma gene expression deconvolution*

18 We assume tumors to be comprised of cancer and stromal (any non-cancer) cells. Measured
19 bulk tumor mRNA abundance is then given by the sum of mRNA molecules derived from these
20 two compartments. Tumor mRNA expression ($e_{tumor,i}$) measured for a given gene in sample
21 i can then be expressed as:

22

$$e_{tumor,i} = p_i \times \bar{e}_{cancer} + (1 - p_i) \times \bar{e}_{stroma}$$

23

24 Here p_i denotes the cancer cell proportion (tumor purity) in sample i , and \bar{e}_{cancer} and \bar{e}_{stroma}
25 are average expression levels for the gene in the cancer and stromal compartment (not
26 dependent on sample i), respectively. We make the simplifying assumption that these (non-
27 negative) average compartment expression levels are constant across the set of tumors and
28 estimate them using non-negative least squares regression. 95%-confidence intervals and
29 standard deviations for the cancer and stroma point estimates are estimated using
30 bootstrapping.

31 Bulk tumor RNA-seq FPKM data was log-transformed, $\log_2(X+1)$, before
32 deconvolution. It has previously been discussed whether mixed-tissue gene expression
33 deconvolution should be done using linear or log-transformed gene expression values(Shen-
34 Orr et al., 2012; Zhong and Liu, 2012). Firstly, we observed that the relationship between
35 tumor purity and raw bulk tumor gene expression was generally heteroscedastic: Across all

1 cancer types, tumor purity had overall stronger linear correlation with log-transformed RNA-
2 seq gene expression data, as can be observed in deconvolution of known stromal-specific
3 genes (Suppl. Fig 2-5). Secondly, we directly compared results for raw and log-transformed
4 data, and while the results were overall similar, we found that log-transformation provided
5 better separation between inferred cancer and stroma compartment gene expression for
6 known stromal genes (Suppl. Fig 15), which is consistent with previous empirical analysis
7 (Shen-Orr et al., 2012).

8 We found that the equation above tended to misestimate stromal gene expression for
9 genes with somatic copy number alterations (CNA) affecting gene expression in a subset of
10 the samples (for example *ERBB2* in HER2-positive breast tumors). We therefore used a
11 modified approach for such genes. We first identified genes with correlation between CNA
12 and mRNA expression (comparing expression for samples with diploid and non-diploid CNA,
13 Mann-Whitney U-test, $P < 1e-6$, to account for multiple testing) in a given set of tumors, and
14 then estimated cancer and stromal compartment gene expression using a two-step approach.
15 Stromal compartment mRNA expression was first inferred using the above approach using
16 only samples with diploid copy number for the gene. We then used the inferred mean stroma
17 compartment expression, the measured mean tumor expression, and the mean purity of the
18 tumor samples to calculate the mean cancer compartment expression using the above
19 equation.

20 A limitation of TUMERIC is that the method estimates average cancer and stromal
21 compartment gene expression across a cohort of tumor samples. Since the stromal
22 compartment is a mix of many different types of cells, the averaged stromal expression
23 estimates therefore require careful interpretation. Furthermore, exploration of biologically
24 relevant between-sample variation in gene expression, not due to variation in tumor purity,
25 requires additional downstream analysis of the expression residuals.

26

27 *Deconvolution of iTRAQ tumor protein abundance data*

28 We obtained iTRAQ protein abundance data for BRCA and OV tumor types using CPTAC
29 consortium data available at cBioPortal (www.cbioportal.org). Bulk abundance profiles were
30 deconvolved into cancer and stroma compartment abundance profiles similar to the
31 procedures described for RNA-seq data above. Briefly, we have matched protein abundance
32 and consensus tumor purity data available for each tumor. We can therefore use the
33 consensus tumor purity estimates and NNLS to infer the (mean) abundance of individual
34 proteins in the cancer and stromal compartment, respectively.

35

36 *Ligand-Receptor Relative Crosstalk (RC) score*

1 Following the law of mass action, we can estimate the concentration of a ligand-receptor (LR)
2 complex from the molar concentrations of the individual ligand, L , receptor, R , and their
3 dissociation constant, K_D , under equilibrium:

4

5
$$[LR] = [L][R]k_D^{-1}$$

6

7 In our analysis we are only evaluating relative changes of individual LR complex
8 concentrations across samples or compartments. Since the dissociation constant cancels out
9 (present in numerator and denominator) when we compute relative concentrations (RC score
10 below) we can ignore the dissociation constant in downstream analysis. Our analysis depends
11 on multiple simplifying assumptions. We assume that compartment-level mRNA expression
12 estimates are reasonable proxies for ligand and receptor molar concentration at the site of
13 LR-complex formation. Additionally, we assume there is no external competition for individual
14 ligands and receptors of a given LR pair, or that these effects are constant across samples or
15 compartments.

16 To estimate the relative flow of signaling between cancer and stromal cell
17 compartments, we developed the Relative Crosstalk (RC) score. LR complex activity is first
18 approximated using the product of ligand and receptor gene expression inferred for the given
19 compartments (in linear scale). The RC score then estimates the relative complex
20 concentration given all four possible directions of signaling and a normal tissue state, e.g. for
21 cancer-cancer (CC) signaling:

22

23
$$RC_{CC} = \frac{e_{L,C} \times e_{R,C}}{e_{L,C} \times e_{R,C} + e_{L,C} \times e_{R,S} + e_{L,S} \times e_{R,C} + e_{L,S} \times e_{R,S} + e_{L,N} \times e_{R,N}}$$

24

25 The normal term in the denominator is included to normalize for complex activity in normal
26 tissue, and this term is calculated directly from the observed gene expression levels in normal
27 tissue samples available for each tumor type in TCGA.

28

29 *Gene-set enrichment analysis*

30 To study genes differentially expressed between cancer and stromal compartments, we
31 performed GSEA (Subramanian et al., 2005) pre-ranked analysis of genes sorted by
32 differential expression (log fold-change) in the two compartments. We analyzed all hallmark
33 gene sets and used a FDR cut-off of 0.25 to determine gene sets with differential enrichment.

34

35 *Immunohistochemistry (IHC) quantification analysis of Human Protein Atlas data*

1 In order to quantify cancer and stromal cells expression of genes, we performed color
2 deconvolution of IHC images obtained from the Human Protein Atlas (proteinatlas.org) (Uhlen
3 et al., 2017) using the ImageJ software package and standard protocols (Schindelin et al.,
4 2015). Following manual selection and segmentation of cancer and stromal cells (without
5 knowledge of antibody staining), color intensities were measured with ImageJ, and DAB
6 (target), hematoxylin (cells), and complementary components were estimated. Average
7 antibody intensities were then estimated for the cancer and stromal compartment of a given
8 slide. Further technical details are provided in the supplementary methods.

9

10 *RNA-ISH of breast cancer tissue*

11 We obtained Formalin-Fixed Paraffin-Embedded (FFPE) breast cancer tissue slides for 2
12 luminal (ER+/PR+) and 2 triple negative basal-like breast tumors. In situ hybridization (ISH)
13 was performed using specific RNAScope (<https://acdbio.com>) probes to evaluate IL6ST and
14 SFRP1 RNA level and localization. The FFPE slides were deparaffinised, rehydrated, and
15 pretreated using the RNAScope Sample Preparation kit according to the manufacturer's
16 recommendations. The slides were incubated in target retrieval solution (#322000, Advanced
17 Cell Diagnostics) for 15 minutes at 97°C followed by a protease solution (#322330, Advanced
18 Cell Diagnostics) for 30 minutes at 40°C. RNAScope IL6ST (#447251) and SFRP1 (#429381-
19 C2) target probe and RNAScope 2.5 HD Duplex Detection (Chromogenic) kit (#322430,
20 Advanced Cell Diagnostics) were applied to the slide according to the manufacturer's
21 instructions. RNA expression was quantified using two approaches. Firstly, according to the
22 manufacturer's instructions (TS 46-003), we segmented the Images using supervised machine
23 learning. Briefly, we used Fiji/ImageJ to train a 4-class random forest classifier to distinguish
24 regions with *SFRP1* staining, *IL6ST* staining, nuclear regions, and background signal. We
25 used this model to segment all 8 images. We selected 5 cancer cell regions in each image
26 and recorded the mean probability for the *SFRP1* and *IL6ST* staining classes in each region.
27 These staining probabilities were summarized (mean and standard deviation) across all
28 regions from the four images (2 patients) for the luminal and basal subtype, respectively. As
29 a second and different approach, we quantified *IL6ST* and *SFRP1* mRNA expression in cancer
30 cells of basal and luminal breast cancer tissues using 3-channel color deconvolution.
31 Channels were specified for probe 1 (*SFRP1*), probe 2 (*IL6ST*), and background. Multiple
32 areas enriched for cancer cells were selected in each tissue slide and the optical density (OD)
33 was calculated for each area. The average and standard deviation of the OD for each gene
34 and tissue type was calculated.

35

36 **Supplementary data**

1 All code used to generate the figures and statistics of the paper is included in Supplementary
2 Data 1. Source code for Tumeric is attached as Supplementary data 2. RC scores and other
3 summary metrics for all analyzed ligand-receptor pairs are summarized in Supplementary
4 Table 1. Supplementary data files can be accessed at: <https://bit.ly/2Ingp03>.

5

6 **Data availability**

7 TCGA SNV and CNV data: <https://gdac.broadinstitute.org/>. Data release version 2016/01/28.
8 TCGA RNA-seq data: https://toil.xenahubs.net/download/tcga_RSEM_gene_fpkm.gz
9 ASCAT purity estimates: <https://cancer.sanger.ac.uk/cosmic/download>
10 IHC data: <https://www.proteinatlas.org>.

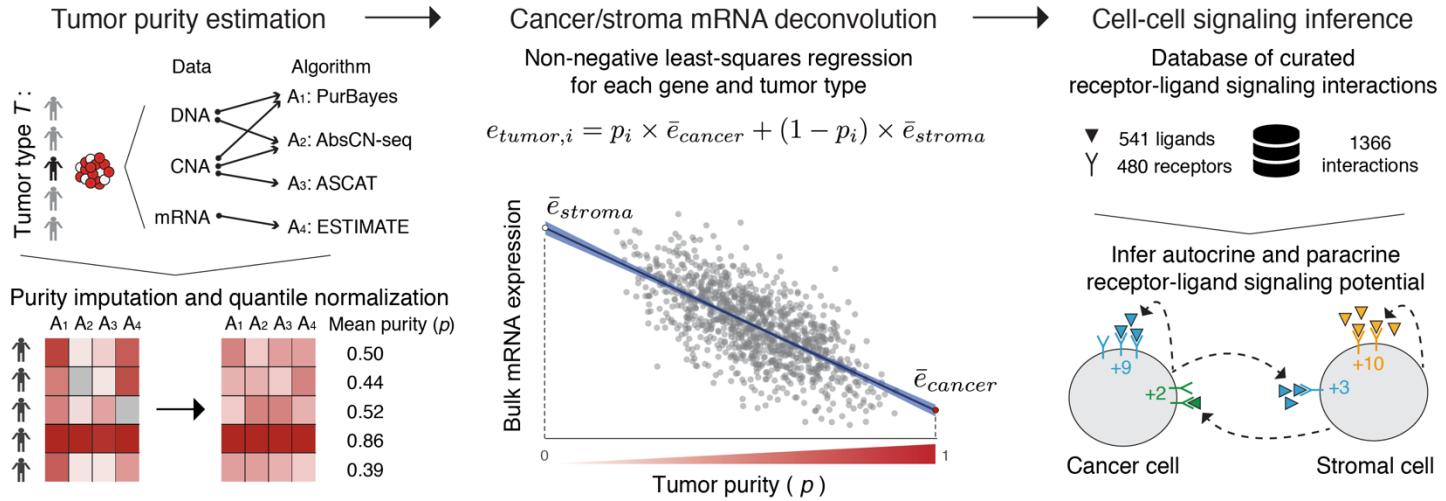
11

12 **Author contributions**

13 AJS conceived of the project. UG, MMN, NR, TTN, and AJS analyzed data. UG, SS, and AJS
14 designed and implemented methods and software. JY, PHT, JI and BC acquired and
15 annotated FFPE samples. AW, RD, and AJS performed RNAscope analysis. AJS, UG, and
16 RD wrote the manuscript.

1 Figures

2



3

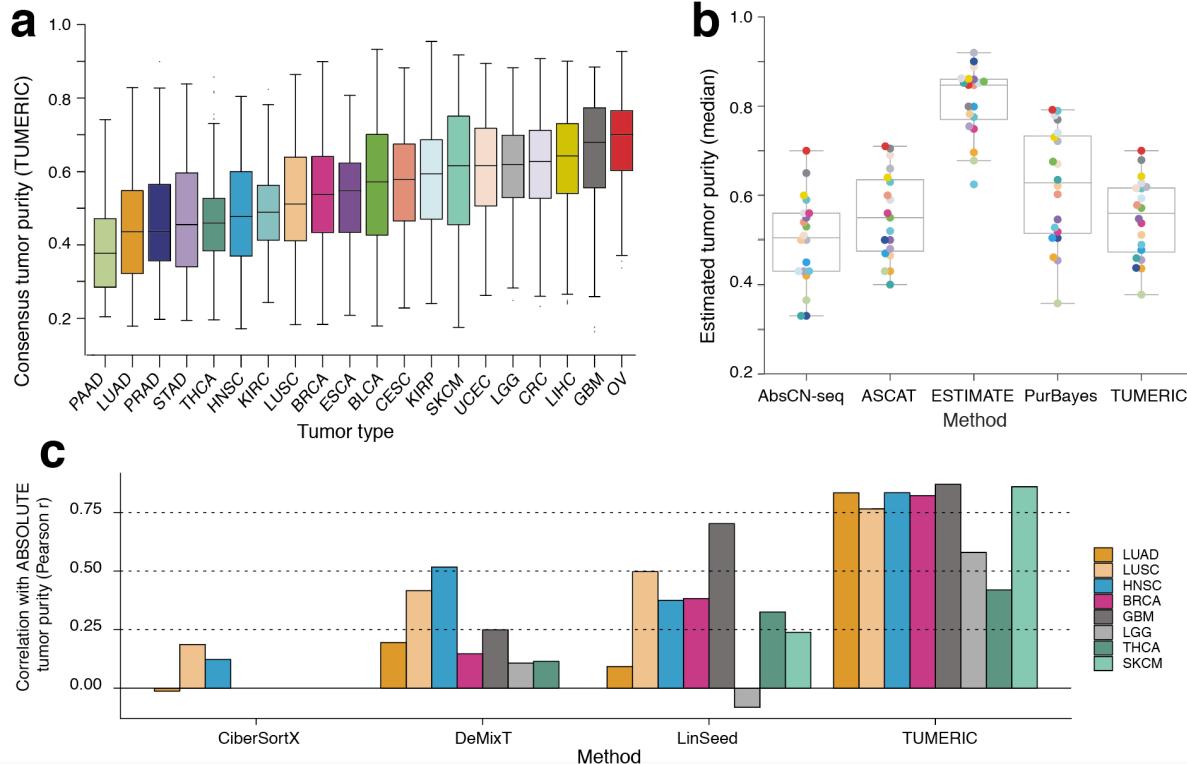
4

5 **Figure 1: Overview of approach.** The purity of each bulk tumor sample is first estimated
 6 using a consensus approach. mRNA expression levels in “average” cancer and stromal cells
 7 are inferred for a set of tumors (e.g. tumor type) using non-negative least-squares regression;
 8 figure shows data for *CD4* in breast cancer. Candidate autocrine and paracrine signaling
 9 interactions are inferred using a database of curated receptor-ligand signaling interactions.

10

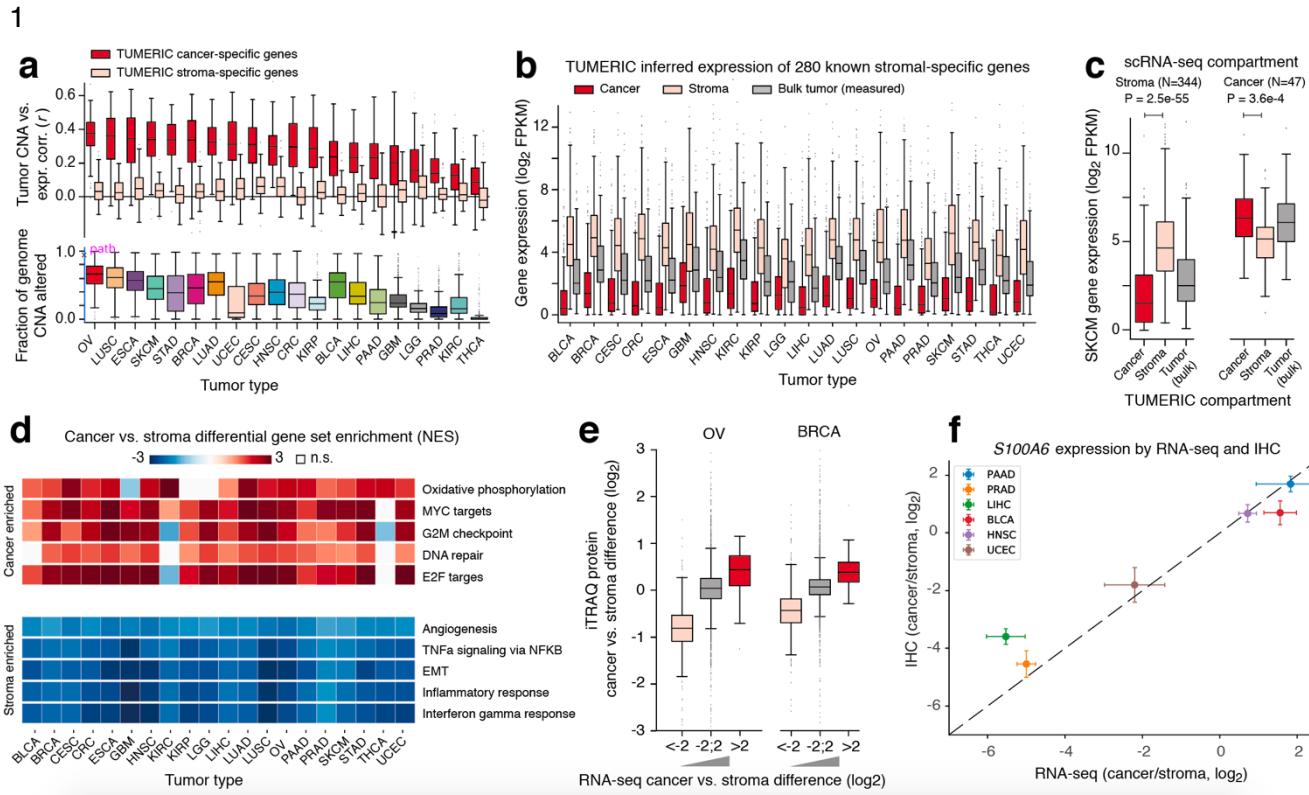
11

1
2



3
4

5 **Figure 2: Consensus tumor purity estimation.** **a)** TUMERIC consensus tumor purity
6 estimates for ~8000 bulk tumor samples across 20 solid tumor types. **b)** Median tumor purity
7 of each tumor type as estimated by 4 different methods, including TUMERIC consensus
8 estimates. Points (representing median purity across tumor type) are color coded according
9 to tumor types in panel a). **c)** Concordance of tumor purity estimates from existing
10 transcriptome deconvolution methods with ABSOLUTE (DNA-derived) tumor purity estimates;
11 TUMERIC consensus purity estimates included for comparison.

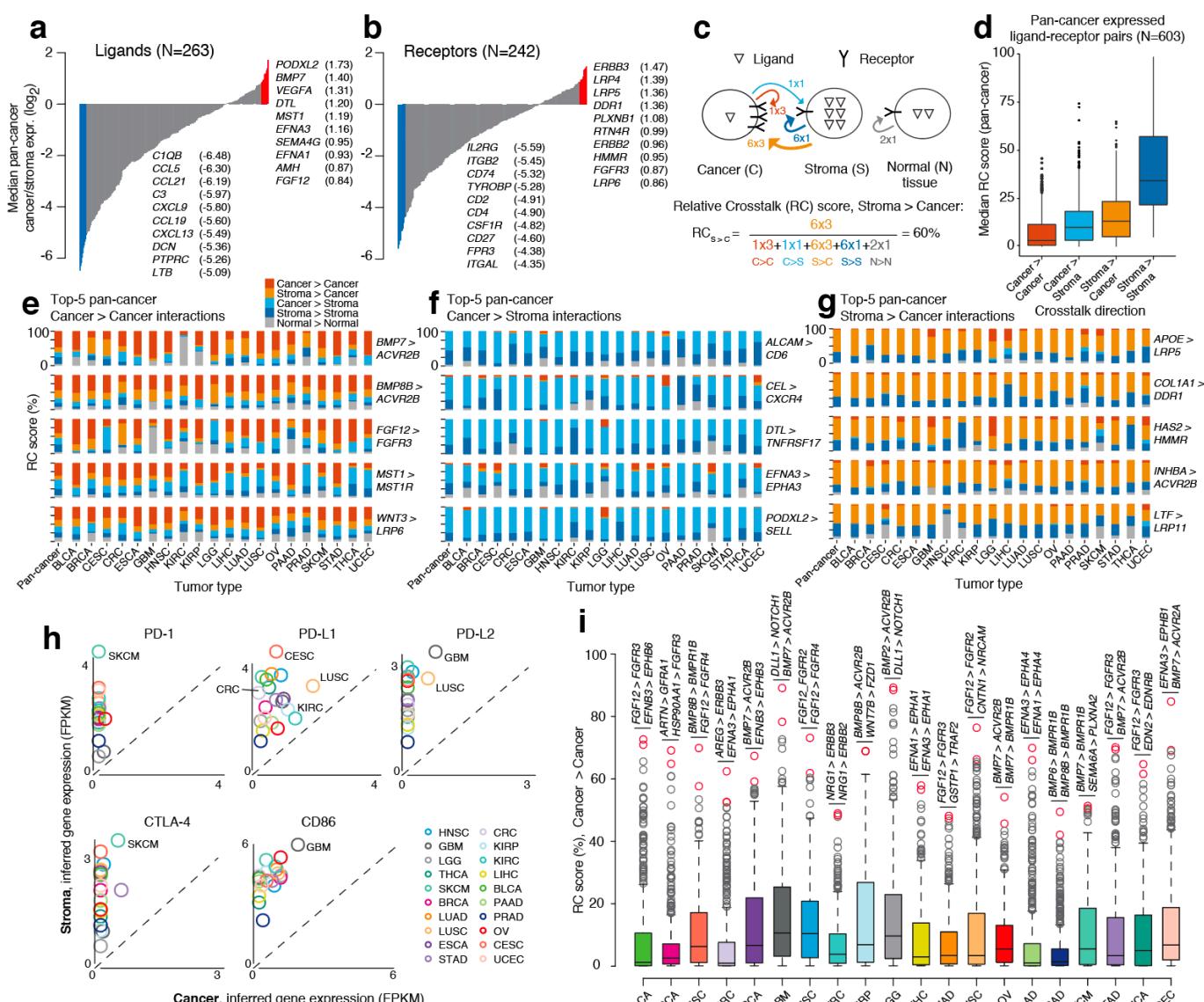


2

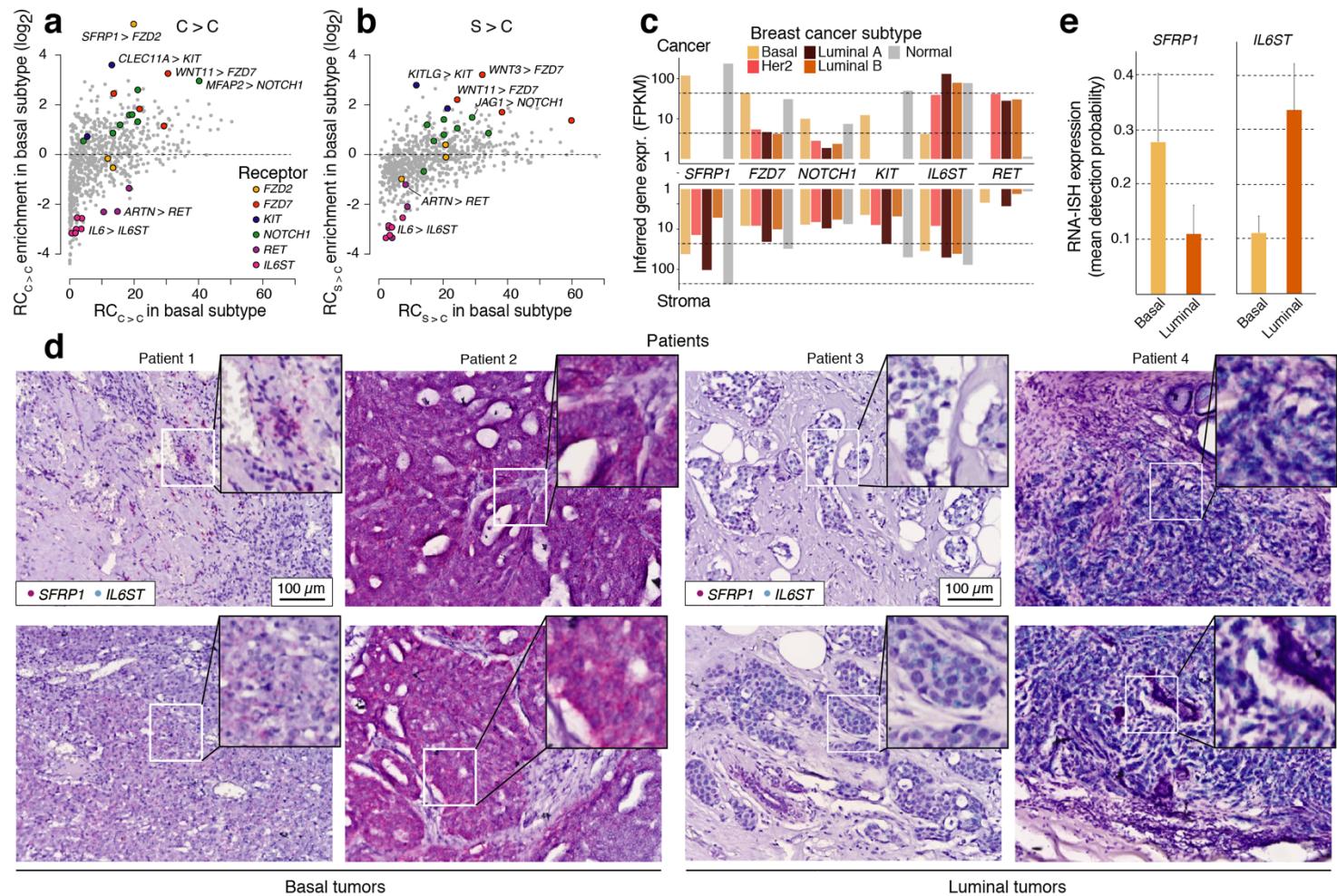
3 **Figure 3: Validation of tumor gene expression deconvolution.** **a)** Genes specifically expressed in cancer and stromal cells were inferred for each tumor type, and the correlation (Pearson r) between mRNA expression and somatic copy number alterations (CNA) at each gene locus was evaluated (top panel). Tumor types are ordered by the difference in means of cancer and stromal gene correlations. The fraction of genome altered by CNA was determined for each tumor sample (bottom panel). **b)** Inferred cancer and stroma compartment expression levels for 280 known stromal-specific genes. **c)** Inferred cancer and stroma compartment expression levels in melanoma (SKCM) for cancer and stroma specific genes previously identified with melanoma scRNA-seq. **d)** Genes were ordered by inferred expression difference between cancer and stroma compartments in each tumor type, and gene set enrichment analysis (GSEA) was used to identify cancer and stroma enriched gene sets. **e)** Protein expression was inferred for cancer and stroma compartments in (OV) and breast (BRCA) cancer cohorts using iTRAQ protein quantification data and compared to deconvolved RNA-seq expression data. **f)** Comparison of immunohistochemistry (IHC) and deconvolved RNA-seq expression for a gene (*S100A6*) with highly variable cancer/stroma expression across cancer types; error bars reflect standard deviations of the point estimates.

19

20



1
2 **Figure 4: Pan-cancer inference of crosstalk.** **a, b)** Differential expression (median) between
3 cancer and stroma compartments of ligands **a)** and receptors **b)** across cancer types. **c)** The
4 Relative Crosstalk (RC) score approximates the relative flow of signaling in the four possible
5 directions between cancer and stromal cell compartments, including a bulk (non-deconvolved)
6 matched normal tissue reference. In the concrete example, a given ligand is expressed with 1, 6,
7 and 2 copies (receptor in 3, 1, and 1 copies) in cancer, stroma, normal cells, respectively. This
8 yields a stroma-to-cancer (S>C) RC score of 60%. **d)** Median RC score across the 20 solid tumor
9 types was estimated and plotted for each direction of signaling. **e, f, g)** RC scores for the 5 ligand-
10 receptor pairs with highest median pan-cancer cancer-cancer autocrine **e)**, cancer-stroma **f**, and
11 stroma-cancer **g**) paracrine signaling scores. **h)** Inferred cancer and stromal compartment gene
12 expression across tumor types for immune checkpoint ligands (PD-L1, PD-L2, CD86) and
13 receptors (PD-1, CTLA-4). **i)** Distributions of RC scores for cancer-cancer autocrine signaling
14 across tumor types; the top-2 ligand-receptor pairs are highlighted for each tumor type.



1
2
3 **Figure 5: Inference of crosstalk in basal breast cancer.** **a, b**) Enrichment of RC scores in
4 breast cancer tumors of basal subtype for **a**) autocrine cancer-to-cancer and **b**) paracrine
5 stroma-cancer crosstalk. **c**) Inferred expression of selected receptors and ligands in cancer
6 and stromal compartment across breast cancer subtypes, expression in normal tissue (non-
7 deconvolved) included for comparison. **d**) RNA ISH of *SFRP1* and *IL6ST* in 2 basal and 2
8 luminal FFPE breast tumor samples (2 slides per tumor). **e**) Quantification of *SFRP1* and
9 *IL6ST* expression (mean detection probability across cancer cell regions) in the tissue sections
10 using a supervised classification approach.

11
12
13

1

2 References

3

4

5

6 Ahn, J., Yuan, Y., Parmigiani, G., Suraokar, M.B., Diao, L., Wistuba, I.I., and Wang, W.
7 DeMix: deconvolution for mixed cancer transcriptomes using raw measured data.
8 Bioinformatics 29, 1865–1871.

9

10 Aran, D., Sirota, M., and Butte, A.J. (2015). Systematic pan-cancer analysis of tumour purity.
11 Nat Commun 6, 8971.

12

13 Aran, D., Hu, Z., and Butte, A.J. (2017). xCell: digitally portraying the tissue cellular
14 heterogeneity landscape. Genome Biology 220.

15

16 Bao, L., Pu, M., and Messer, K. AbsCN-seq: a statistical method to estimate tumor purity,
17 ploidy and absolute copy numbers from next-generation sequencing data. Bioinformatics 30,
18 1056–1063.

19

20 Brennan, C.W., Verhaak, R.G., McKenna, A., Campos, B., Noushmehr, H., Salama, S.R.,
21 Zheng, S., Chakravarty, D., Sanborn, Z.J., Berman, S.H., et al. (2013). The Somatic
22 Genomic Landscape of Glioblastoma. Cell 155, 462–477.

23

24 Edwards, N.J., Oberti, M., Thangudu, R.R., Cai, S., McGarvey, P.B., Jacob, S., Madhavan,
25 S., and Ketchum, K.A. (2015). The CPTAC Data Portal: A Resource for Cancer Proteomics
26 Research. J Proteome Res 14, 2707–2713.

27

28 Fieger, C.B., ssetti, C., and Rosen, S.D. (2003). Endoglycan, a Member of the CD34 Family,
29 Functions as an L-selectin Ligand through Modification with Tyrosine Sulfation and Sialyl
30 Lewis x. Journal of Biological Chemistry 278, 27390–27398.

31

32 Gattelli, A., Nalvarte, I., Boulay, A., Roloff, T.C., Schreiber, M., Carragher, N., Macleod, K.K.,
33 Schleuderer, M., Lienhard, S., Kenner, L., et al. (2013). Ret inhibition decreases growth and
34 metastatic potential of estrogen receptor positive breast cancer cells. EMBO Molecular
35 Medicine 5, 1335–1350.

1

2 Guo, F., Wang, Y., Liu, J., Mok, S., Xue, F., and Zhang, W. (2015). CXCL12/CXCR4: a
3 symbiotic bridge linking cancer cells and their stromal neighbors in oncogenic
4 communication networks. *Oncogene* *35*, 816–826.

5

6 Hanahan, D., and Coussens, L. (2012). Accessories to the Crime: Functions of Cells
7 Recruited to the Tumor Microenvironment. *Cancer Cell* *21*.

8

9 Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of Cancer: The Next Generation. *Cell*
10 *144*, 646–674.

11

12 Hoadley, K.A., Yau, C., Hinoue, T., Wolf, D.M., Lazar, A.J., Drill, E., Shen, R., Taylor, A.M.,
13 Cherniack, A.D., Thorsson, V., et al. (2018). Cell-of-Origin Patterns Dominate the Molecular
14 Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* *173*, 291-304.e6.

15

16 Josse, J., and Husson, F. (2016). missMDA: A Package for Handling Missing Values in
17 Multivariate Data Analysis. *Journal of Statistical Software*.

18

19 Junntila, M.R., and de Sauvage, F.J. (2013). Influence of tumour micro-environment
20 heterogeneity on therapeutic response. *Nature* *501*, 346–354.

21

22 Khalil, D.N., Smith, E.L., Brentjens, R.J., and Wolchok, J.D. (2016). The future of cancer
23 treatment: immunomodulation, CARs and combination immunotherapy. *Nat Rev Clin Oncol*
24 *13*, nrclinonc.2016.25.

25

26 Larson, N., and Fridley, B. PurBayes: estimating tumor cellularity and subclonality in next-
27 generation sequencing data. *Bioinformatics* *29*, 1888–1889.

28

29 Loo, P., Nordgard, S.H., Lingjærde, O., Russnes, H.G., Rye, I.H., Sun, W., Weigman, V.J.,
30 Marynen, P., Zetterberg, A., Naume, B., et al. Allele-specific copy number analysis of
31 tumors. *pnas.org*.

32

33 Moffitt, R.A., Marayati, R., Flate, E.L., Volmar, K.E., Loeza, G.S., Hoadley, K.A., Rashid,
34 N.U., Williams, L.A., Eaton, S.C., Chung, A.H., et al. (2015). Virtual microdissection identifies
35 distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nature*

1 Genetics 47, 1168–1178.

2

3 Newman, A.M., Liu, C., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn,
4 M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression
5 profiles. *Nature Methods* 12, 453–457.

6

7 Newman, A.M., Steen, C.B., Liu, C., Gentles, A.J., Chaudhuri, A.A., Scherer, F.,
8 Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., et al. (2019). Determining cell
9 type abundance and expression from bulk tissues with digital cytometry. *Nature
10 Biotechnology* 1–10.

11

12 Pandita, A., Aldape, K.D., Zadeh, G., Guha, A., and James, D.C. (2004). Contrasting in vivo
13 and in vitro fates of glioblastoma cell subpopulations with amplified EGFR. *Genes,
14 Chromosomes and Cancer* 39, 29–36.

15

16 Panicot-Dubois, L., Thomas, G.M., Furie, B.C., Furie, B., Lombardo, D., and Dubois, C.
17 (2007). Bile salt–dependent lipase interacts with platelet CXCR4 and modulates thrombus
18 formation in mice and humans. *Journal of Clinical Investigation* 117, 3708–3719.

19

20 Purow, B.W., Haque, R.M., Noel, M.W., Su, Q., Burdick, M.J., Lee, J., Sundaresan, T.,
21 Pastorino, S., Park, J.K., Mikolaenko, I., et al. (2005). Expression of Notch-1 and Its Ligands,
22 Delta-Like-1 and Jagged-1, Is Critical for Glioma Cell Survival and Proliferation. *Cancer
23 Research* 65, 2353–2363.

24

25 Quon, G., Haider, S., Deshwar, A.G., Cui, A., Boutros, P.C., and Morris, Q. (2013).
26 Computational purification of individual tumor gene expression profiles leads to significant
27 improvements in prognostic prediction. *Genome Medicine* 5, 29.

28

29 Ramilowski, J.A., Goldberg, T., Harshbarger, J., Kloppman, E., Lizio, M., Satagopam, V.P.,
30 Itoh, M., Kawaji, H., Carninci, P., Rost, B., et al. (2015). A draft network of ligand–receptor-
31 mediated multicellular signalling in human. *Nature Communications* 6, 7866.

32

33 Scheel, C., Eaton, E., Li, S., Chaffer, C.L., Reinhardt, F., Kah, K.-J., Bell, G., Guo, W.,
34 Rubin, J., Richardson, A.L., et al. (2011). Paracrine and Autocrine Signals Induce and
35 Maintain Mesenchymal and Stem Cell States in the Breast. *Cell* 145, 926–940.

1

2 Schindelin, J., Rueden, C.T., Hiner, M.C., and Eliceiri, K.W. (2015). The ImageJ ecosystem:

3 An open platform for biomedical image analysis. *Molecular Reproduction and Development*

4 82, 518–529.

5

6 Sharma, P., and Allison, J.P. (2015). The future of immune checkpoint therapy. *Science* 348,

7 56–61.

8

9 Shen-Orr, S.S., Tibshirani, R., Khatri, P., Bodian, D.L., Staedtler, F., Perry, N.M., Hastie, T.,

10 rwal, M., vis, M., and Butte, A.J. (2010). Cell type–specific gene expression differences in

11 complex tissues. *Nat Methods* 7, 287.

12

13 Shen-Orr, S.S., Tibshirani, R., and Butte, A.J. (2012). Gene expression deconvolution in

14 linear space. *Nat Methods* 9, nmeth.1831.

15

16 Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A.,

17 Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment

18 analysis: A knowledge-based approach for interpreting genome-wide expression profiles.

19 *Proceedings of the National Academy of Sciences of the United States of America* 15545–

20 15550.

21

22 Teodorczyk, M., and Schmidt, M.H. (2015). Notching on Cancer’s Door: Notch Signaling in

23 Brain Tumors. *Frontiers in Oncology* 4, 341.

24

25 Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., Treacy, D., Trombetta, J.J., Rotem, A.,

26 Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of

27 metastatic melanoma by single-cell RNA-seq. *Science* 189–196.

28

29 Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhor, G., Benfeitas, R., Arif,

30 M., Liu, Z., Edfors, F., et al. (2017). A pathology atlas of the human cancer transcriptome.

31 *Science* 357, eaan2507.

32

33 Wang, Z., Cao, S., Morris, J.S., Ahn, J., Liu, R., Tyekucheva, S., Gao, F., Li, B., Lu, W.,

34 Tang, X., et al. (2018). Transcriptome Deconvolution of Heterogeneous Tumor Samples with

35 Immune Infiltration. *IScience*.

1

2 William, D., Mokri, P., Lamp, N., Linnebacher, M., Classen, C., Erbersdobler, A., and

3 Schneider, B. (2017). Amplification of the EGFR gene can be maintained and modulated by

4 variation of EGF concentrations in in vitro models of glioblastoma multiforme. PLOS ONE

5 12, e0185208.

6

7 Wood, L.D., and Hruban, R.H. (2012). Pathology and Molecular Genetics of Pancreatic

8 Neoplasms. *The Cancer Journal* 492–501.

9

10 Yao, H.-P., Zhou, Y.-Q., Zhang, R., and Wang, M.-H. (2013). MSP–RON signalling in

11 cancer: pathogenesis and therapeutic potential. *Nature Reviews Cancer* 13, 466–481.

12

13 Yarchoan, M., Hopkins, A., and Jaffee, E.M. (2017). Tumor Mutational Burden and

14 Response Rate to PD-1 Inhibition. *The New England Journal of Medicine* 377, 2500–2501.

15

16 Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W.,

17 Treviño, V., Shen, H., Laird, P.W., Levine, D.A., et al. (2013). Inferring tumour purity and

18 stromal and immune cell admixture from expression data. *Nature Communications* 4, 2612.

19

20 Zaitsev, K., Bambouskova, M., Swain, A., and Artyomov, M.N. (2019). Complete

21 deconvolution of cellular mixtures based on linearity of transcriptional signatures. *Nat*

22 *Commun* 10, 2209.

23

24 Zhong, Y., and Liu, Z. (2012). Gene expression deconvolution in linear space. *Nature*

25 *Methods* 9, 8–9.

26