CITEgeist: Accurate deconvolution of spatial transcriptomics with same-slide proteomics reveals midkine as a secreted microenvironment modulator in ESR1 mutant breast cancer.

Alexander Chih-Chieh Chang [1,2,*], Brent T. Schlegel [1,2,*], Neil Carleton [1,2], Priscilla F. McAuliffe [1,3], Hunter Waltermire [1,2], Steffi Oesterreich [1,4], Russell Schwartz [5,6], and Adrian V. Lee [1,4,7,#]

[1] Women's Cancer Research Center, UPMC Hillman Cancer Center, Magee-Womens Research Institute, Pittsburgh, PA, USA

[2] University of Pittsburgh, School of Medicine, Pittsburgh, PA, USA

[3] Department of Surgery, Division of Breast Surgical Oncology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

[4] Department of Pharmacology and Chemical Biology, University of Pittsburgh, Pittsburgh, PA, USA

[5] Ray and Stephanie Lane Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA, USA

[6] Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA, USA

[7] Institute of Precision Medicine, Pittsburgh, PA, USA

[*] This author contributed equally to this manuscript

[#] Corresponding Author.

Abstract

Background: Dysplastic tissue architecture in estrogen receptor-positive (ER+) breast cancer across therapy-naïve and therapy-exposed cancer tissues presents unique challenges in the analysis of spatial transcriptomics. Many tools for deconvolution are developed on well-structured tissue architectures such as the 10x Genomics mouse hippocampus dataset. Spatial transcriptomics analysis could offer valuable insights into treatment response, but faces limitations in cellular resolution.

Methods: To address this problem, we developed CITEgeist, a computational tool for spatial transcriptomic deconvolution using integrated proteomics data from the same slide. Visium Antibody Capture technology was applied alongside our novel algorithm to analyze the tumor microenvironment. We demonstrate the reliability of our method using pre- and post-treatment samples from six breast cancer cases.

Results: Our approach revealed previously undetectable cellular interactions within the tumor microenvironment. By taking an interoperable approach to software development and grounding our algorithm in interpretable variables, we demonstrate how CITEgeist deconvolution is not only accurate but robust enough to be directly used as input in external analytical tools developed by other research teams. We then applied this approach to a set of specimens from a prospective trial our group ran and further validated the findings in a series of in vitro experiments as a demonstrated use case of the utility, necessity, and flexibility of CITEgeist; and the potential of our method to rapidly translate novel clinical samples to new biological insights.

Conclusions: CITEgeist addresses a critical technical gap in spatial multi-omics analysis through an integrated, multi-disciplinary approach. This work demonstrates the value of combining clinical, translational, and computational expertise to identify novel mechanisms of treatment resistance, potentially transforming therapeutic strategies for resistant disease.

Keywords: Breast cancer, Spatial transcriptomics, Proteomics, Tumor microenvironment, ESR1 mutation, Elderly patients, Treatment resistance, Computational deconvolution, Midkine signaling, Multi-omics

Background

Estrogen receptor-positive (ER+) breast cancer accounts for approximately 85% of new breast cancer diagnoses [1], with surgery and anti-estrogen therapy established as the first-line treatment [2]. How breast cancer changes post-treatment is of vital importance to disentangling how treatment resistance develops, and spatial transcriptomics offer a unique opportunity to dissect the mechanisms of cancer evolution in native tissue architectural contexts. Additionally, many techniques allow for the joint collection of proteomic information from the same slide, allowing integration of multi-omic informational landscapes, and hence, deeper insights.

However, no tool is currently capable of integrating this proteomic information for analysis, and there are also limitations in various single-cell reference-dependent deconvolution tools that have been the dominant design choice in the field.

Single-cell reference-based deconvolution methods have several disadvantages. First, it is costly to generate a single cell reference to match each spatial dataset, at approximately double the cost [3]. Second, external references or atlases are not always available, especially for heterogeneous post-treatment cancer samples, which in themselves are rare and difficult to collect routinely [4]. Third, biosamples such as core biopsies often provide too little tissue for single-cell sequencing [8], and, even if available, the biopsy tissue used for the reference may not capture the heterogeneity of the cancer slice used in the spatial transcriptomics. [5,6]. Fourth, there is a significant risk of data hallucination, where gene expression or cell-type abundance, as estimated from single-cell data, is artificially imposed in a spatial context, erroneously granting spatial properties to transcriptomic data when they are actually artifacts [7]. Finally,

many methods suffer from a lack of interpretability, requiring complex assumptions and architectures to integrate the two informational landscapes of single-cell transcriptomics and spatial transcriptomics [8], which often is incompatible with other downstream analysis options and tools popular in the field, i.e., by failing to preserve raw counts needed in DeSeq [9].

These technical limitations prompted the development of CITEgeist, a novel computational tool that enables spatial transcriptomic deconvolution using integrated proteomics data obtained from identical tissue sections via integer linear programming. We demonstrate the accuracy of CITEgeist using a simulated reference test. Additionally, we run this tool on post-treatment samples from a clinical trial examining endocrine therapy in older women who forego surgery to emphasize the rapid cutting-edge applicability of this method. Furthermore, we develop intrinsic tests on the non-simulated real tissue data to confirm the confidence of CITEgeist outputs on post-treatment tumor data. Lastly, we demonstrate how our output is fully compatible with tools developed by other groups, and use them to identify a novel environmental effect in ESR1 mutant breast cancer, which we then validate in cell culture in the laboratory.

91 **Methods**

92 **CITEgeist Model**

93 CITEgeist deconvolutes spatial transcriptomics data by leveraging antibody capture
94 from the same tissue section, eliminating dependency on external single-cell RNA
95 sequencing references. For the complete derivation, please see the Supplementary
96 Methods. The method consists of three key components: (1) protein-based
97 deconvolution using an Expectation-Maximization (EM) algorithm, (2) regional
98 refinement with neighborhood optimization, and (3) spatial gene deconvolution with
99 variable-sized neighborhoods.

100 **Software**

101 CITEgeist is implemented in Python 3.10 using Gurobi 11.0.2 for Integer Linear
102 Programming (ILP) optimization. Data preprocessing includes gene filtering (> 0 counts
103 in ≥ 1% spots, mean expression > 1.1) and normalization (target sum 10,000). Antibody
104 data undergoes winsorization for the top and bottom 5% of data points, followed by
105 global centered log ratio (CLR) normalization. Complete mathematical derivations and
106 implementation details are provided in Supplementary Methods.

107 **Single-Cell RNA-seq Reference**

108 To establish a baseline for gene expression profiles — both for deconvolution tasks and
109 the simulation of spatially resolved transcriptomic datasets — we utilized a publicly
110 available atlas of single-cell transcriptomic data from human breast cancers, explicitly
111 focusing on estrogen receptor-positive (ER+) tumors. This dataset served as the
112 reference for deconvolution methods like Seurat [10] and cell2location [8], as well as the

113 foundation for our simulations, enabling us to generate in-silico spatial datasets with a
114 quantifiable ground truth of cell type proportions and gene expression levels.

115 We selected 12,000 cells from ER+ tumor samples across 11 ER+/HER2- cases in the
116 Wu et al. breast cancer reference dataset [11], maintaining cell type proportions through
117 random down-sampling from the primary atlas according to the annotations of 9 distinct
118 cell types. Due to the computational constraints of the scCube autoencoder framework
119 [12], we used this down-sampled version of the atlas as our primary reference for
120 spatial transcriptomic simulations.

121 For the benchmarking of reference-based deconvolution methods, we utilized two
122 distinct permutations of our reference, the first of which retained the complete set of
123 ER+ tumor cells from the 11 patients from the Wu et al. atlas [11]. This ensured that the
124 reference expression profiles for cell types were as comprehensive as possible,
125 avoiding potential biases introduced by down-sampling. We then performed the same
126 analysis using our second permutation of the reference, down-sampled from ~30,000 to
127 8,000 cells to reflect both cancer heterogeneity and a realistic simulation of a single-
128 sample scRNA-seq reference. By using both approaches, we ensured that our
129 benchmarking fairly evaluated the deconvolution performance without artificial
130 constraints on the reference profiles, while simultaneously reflecting realistic conditions
131 in our simulated experiments.

132 **Spatial Transcriptomic Simulation**

133 We utilized the scCube (v2.0.0) framework [12] to simulate Visium-like spatial
134 transcriptomic data from the downsampled ER+ scRNA-seq atlas. A Variational

135 Autoencoder (VAE) model with a hidden layer size of 128, a learning rate of 0.0001, and
136 10,000 epochs generated single-cell gene expression profiles for the nine major
137 reference cell types. Cells were generated based on the proportions of the annotated
138 cell types in the Wu atlas, with a batch size of 512, executed on a CUDA-enabled A100
139 GPU. The trained model was saved for future use.

140 To capture the spatial distribution of cells, each sample was modeled on a 50 × 50
141 spatial unit grid, divided into hexagonal Visium-like spots, with an average of five cells
142 per spot. This structure enabled the modeling of clustered and infiltrative patterns typical
143 of the breast tumor microenvironment. Clusters of cells representing various types were
144 assigned specific spatial parameters such as shape, twist, and scale, while infiltration
145 patterns were intentionally varied within and across clusters. Background cell types
146 were also incorporated, with proportions reflecting approximate tumor heterogeneity.

147 **Spatial Antibody Capture Simulation**

148 Spatial antibody capture data were simulated for each replicate by first defining two
149 simulated protein markers for each unique cell type in the dataset, resulting in a set of
150 cell-type-specific markers. For each simulated spatial spot, the proportion of each cell
151 type was retrieved, and the expected expression for each marker was calculated by
152 scaling the cell type proportion by a random factor drawn from a uniform distribution
153 between 20 and 50. Expression values for each marker were then simulated using a
154 negative binomial distribution with a mean proportional to the expected expression and
155 a dispersion parameter of 0.5. A 5% dropout rate was applied to the simulated
156 expression values, which set the expression to zero to simulate additional technical
157 noise. Low-level background expression was drawn from a negative binomial

distribution with lower-scale parameters for spots where a given cell type was absent. In addition to the cell type-specific markers, nonspecific proteins were simulated across all spots, drawing from a negative binomial distribution with a constant mean expression of 10. The final output was a matrix of simulated protein expression values for each marker across all spatial spots.

To comprehensively evaluate deconvolution methods, we simulated distinct sets of mixed and highly segmented populations, acknowledging that actual cancer data spans a continuum between these two states. By showcasing performance across these diverse tissue architectures, we offer a thorough assessment encompassing the full spectrum of tissue heterogeneity. Our analysis emphasizes both population types: the mixed populations, which are designed to more closely resemble complex tumor microenvironments, and highly segmented populations, which have been foundational in prior deconvolution method development.

In order to account for spatial variability, we generated five distinct spatial replicates for each sample type ("high_seg" and "mixed"), each exhibiting similar yet unique spatial patterns. This approach yielded spatial gene expression matrices at single-cell and spot-level resolutions, where each spot's gene expression values reflect the aggregated profiles of the cells within it. Additionally, we recorded the cell-type proportions per spot to provide a ground truth for evaluating spatial deconvolution accuracy. These datasets, featuring diverse spatial configurations and infiltration patterns, aim to offer a more realistic foundation for assessing different deconvolution methods on tumor-like spatial data.

181 **Cell Culture**

182 Establishment of MCF-7 and T47D mutant *ESR1* cell lines was previously reported [13].
183 Individual MCF-7 and T47D wild-type and CRISPR-edited D538G ESR1-mutant cell
184 clones were evenly pooled and maintained in Dulbecco's modified Eagle medium
185 (ThermoFisher 11965118) or RPMI (ThermoFisher 11875119), respectively, plus 10%
186 fetal bovine serum, and 1x penicillin-streptomycin (Sigma-Aldrich P4333) at 37°C in a
187 humidified incubator with 5% CO_2. For hormone treatment experiments, cells were
188 deprived of steroid hormones by placement in phenol red-free Improved Minimum
189 Essential Medium (Richter's Mod.) (Fisher Scientific MT10026CV) plus 10% charcoal
190 stripped serum (CSS). 17β-Estradiol (E2) (Sigma-Aldrich E2758-1G) was used at a final
191 concentration of 1nM for +E2 conditions. ELISA was performed using the Invitrogen
192 Human MDK ELISA Kit (EH319RB) according to manufacturer instructions.

193 **Immunofluorescence**

194 MCF-7 and T47D cells were plated in 8-well chamber slides (FisherScientific 08-774-26)
195 at 500 cells/well in full medium as described. After adhering to the wells for 24 hours,
196 the medium was changed to phenol red-free Improved Minimum Essential Medium
197 (Richter's Mod.) (Fisher Scientific MT10026CV) plus 10% CSS for 5 days, with the
198 medium changed 2 times per day. For +E2 conditions, 1nM E2 was added to the
199 medium for 24 hours. The cells were fixed in 4% PFA (FisherScientific 50-980-487),
200 permeabilized for 10 minutes in 0.2% Triton X-100 (Sigma-Aldrich X100-100ML) in PBS
201 (Fisher Scientific MT21031CV), then blocked for 30 minutes in 1% BSA (Sigma-Aldrich
202 A9647-500G), 10% goat serum (Sigma-Aldrich G9023), 0.1% Tween20 (Sigma P7949-
203 100ML) in PBS. Primary antibodies for anti-MDK (1:1,000 ThermoFisher MA5-32538)

204 and anti-E-cadherin (1:200 BD Biosciences 610182) were diluted in 1% BSA, 0.1%
205 Tween20 in PBS and incubated overnight at 4 C in a humidified chamber. The cells
206 were then incubated in secondary antibodies for anti-rabbit Alexa Fluor 647 (Invitrogen
207 A32733) and anti-mouse Alexa Fluor 555 (Invitrogen A32727) for 1 hour at room
208 temperature in the dark. Counterstaining was performed with Hoechst 33342
209 (ThermoScientific 62249) in PBS for 10 minutes, and cells were imaged on a Nikon AX
210 confocal microscope with 10 image z-stacks in a minimum of 3 locations per sample at
211 60x oil magnification. Quantification was performed with CellProfiler and ImageJ.

212 **Code Availability**

213 The CITEgeist package, including the tool itself, benchmarking scripts, and code for
214 simulating test datasets, is available on GitHub at
215 https://github.com/leeoesterreich/CITEgeist. The repository provides the complete
216 implementation for generating the spatially resolved CITE-seq data simulations and
217 evaluating deconvolution accuracy and gene expression (GEX) profile inference
218 performance. It includes workflows for benchmarking CITEgeist against state-of-the-art
219 methods and instructions for replicating the analysis and customizing simulations for
220 new datasets. Comprehensive documentation and example scripts are included to
221 ensure usability and reproducibility. A persistent version of CITEgeist will be made
222 available via FigShare at the time of publication (doi: 10.6084/m9.figshare.28385675).

223 **Clinical Trial Sample Collection**

224 As part of a prospective, pragmatic, hybrid de-centralized non-randomized clinical trial
225 (NCT05914792) [14] designed to evaluate circulating tumor DNA (ctDNA) levels as a

226 means to augment longitudinal monitoring of older patients with ER+ breast cancer
227 receiving primary endocrine therapy (pET), tumor tissue samples were collected at
228 baseline and at the time of surgical intervention [15] . In brief, patients aged 70 years
229 and older with ER+/HER2- non-metastatic breast cancer signed informed consent
230 (protocol approved by the University of Pittsburgh IRB under STUDY21100091 and
231 conducted through the UPMC Hillman Cancer Center under protocol 22-088) and chose
232 to forego upfront surgery in favor of pET. Because this trial focused on surgical de-
233 escalation, most patients did not undergo surgery during the study follow-up. However,
234 a subset of six patients, four responding and two progressing, based on imaging
235 assessment and ctDNA results, underwent surgery after 3-50 months of pET (12 total
236 specimens). We collected matched core biopsy and surgical specimens from these six
237 patients (12 total specimens) for this analysis, providing a use case for the CITEgeist
238 tool. To preserve patient anonymity, all samples are labelled HCC22-088-P#-S#, for
239 HillmanCancerCenter trial 22-088 -Patient#-Sample#. Sample number is chronological;
240 in all cases, the biopsy samples are one, and the surgical samples are 2.

241 **For specific statistical and sequencing methods, please refer to our attached**
242 **Supplementary Methods.**

243

244

245 **Results**

246 **CITEgeist is uniquely suited for accurate deconvolution in mixed dysplastic**
247 **cancer architectures**

248 We developed CITEgeist, a computational method designed specifically for accurate
249 cell-type deconvolution in heterogeneous cancer tissues, inspired by a unique
250 opportunity from a clinical trial at our institution (Figure 1A). As illustrated in Figure 1B,
251 our method integrates multiple data layers to achieve high-precision cell-type
252 identification. The core algorithm implements a neighborhood-based approach that
253 utilizes the same-slide proteomic and transcriptomic data to capture cellular context at
254 local, regional, and global scales, providing a comprehensive view of the tumor
255 microenvironment.

256 Unlike many existing tools, which are primarily validated on homogeneous tissues such
257 as the mouse brain [8,16] we rigorously evaluated CITEgeist on cancer tissues with
258 their characteristic heterogeneity. To this end, we generated a series of simulated
259 datasets with varying degrees of cellular admixture and spatial segmentation (Figure
260 1C). These simulated benchmarks represent the complex architecture typical of
261 dysplastic tissues, with intermingled malignant and non-malignant cells exhibiting
262 different levels of differentiation. We then validate novel findings in cancer cell line
263 models via ELISA and immunofluorescent imaging as outlined in Figure 1D.

264 **Benchmarking using simulated data demonstrates that CITEgeist outperforms**
265 **state-of-the-art algorithms in deconvoluting cellular proportions.**

266 Using these simulated datasets, our comparative analysis revealed increased error in
267 the mixed datasets compared to the segmented datasets with state-of-the-art methods
268 such as Cell2Location and Seurat, both seeing increases in error from 0.08 RMSE to
269 0.167 RMSE and 0.10 RMSE to 0.133 RMSE for deconvoluting cellular proportions,
270 respectively, when comparing highly segmented test sets vs mixed test sets. CITEgeist
271 notably does not suffer from this drop, performing with a 0.08 RMSE in both
272 circumstances.

273 To test the sensitivity of different methods to the size of the single cell reference, we
274 also performed tests with reference datasets corresponding to single sample paired
275 single cell sequencing (8,000 cells) or atlas-level objects (30,000 cells). Based on this
276 test, RCTD was noted to have a severe reference dependency based on the
277 comprehensiveness of the reference dataset used. All methods were first run with the
278 comprehensive 30,000 cells, which in cost terms is about 4x the size of a standard
279 single-cell sequencing run, and then run with 8,000 cells sampled from that distribution
280 to capture what would be approximately as if a spatial sample had been submitted
281 jointly with a singular single-cell sequencing kit. Most methods had minimal alterations
282 in performance (Supplemental Files 2); however, RCTD showed deterioration under this
283 test in both text contexts. With RMSEs of 0.21 on average with the 8,000 cell reference,
284 compared to 0.05 with the 30,000 cell reference. This confirmed our hypothesis
285 regarding the difficulty in evaluating performance on real cancer data when there are no
286 clear guidelines on the necessary size and quality of single-cell references.

287 Tangram also notably and visibly proved a different issue of over-calling T-cells due to
288 their known over-representation in single-cell datasets [17] (Figure 2B). It reported a

JSD on average of 0.56 in the highly segmented dataset, and 0.53 in the mixed dataset, compared to an average of 0.15 and 0.34 in all other methods.

CITEgeist deconvolutes gene expression counts and outperforms state-of-the-art algorithms.

We then extended our benchmarking to evaluate the performance to deconvolute the spatially-derived gene expression of each cell type, a key function of spatial transcriptomics. At the time, only two methods had documented methods for deconvoluting RNA signals per cell type from their model, Cell2Location and Tangram (Figure 3).

In the mixed cancer-similar test set, CITEgeist had at least a 4-fold improvement over the other methods with an average NRMSE of 0.04, while Cell2Location had an average NRMSE of 0.16, and Tangram had an average NRMSE of 0.20 (Figure 3). We show a 2- to 4-fold increase in performance in the highly segmented test set (CITEgeist: 0.04 average NRMSE, Cell2Location: 0.066 average NRMSE, Tangram: 0.17 average NRMSE).

The magnitude of this significant performance improvement is visualized in Figure 3, which quantitatively illustrates CITEgeist's accuracy metrics across different tissue composition scenarios compared to existing methods.

Interpretability in CITEgeist output

A key issue in spatial deconvolution is also an over-reliance on simulated benchmarks [18]. Hence, interpretability and verifiable results are key, and thus we developed several intrinsically-validating tests for CITEgeist deconvoluted samples, as

311 demonstrated in Figure 4, which allow for confidence in CITEgeist output on a sample-
312 by-sample basis, independent of simulated testing or ground truth. Figure 4A illustrates
313 our deconvoluted cell-type distributions, which were subsequently reviewed in
314 conjunction with the underlying H&E image, demonstrating consistency in the results
315 and expected proportions; with cancer cell proportion predictions matching prominent
316 clusters of epithelial cells, fibroblasts matching fibrotic regions, and CD4 and CD8 T
317 cells in expected proportions consistent with the known biology of these cell types.

318 Additionally, to demonstrate the accuracy of the gene expression deconvolution, we
319 took the gene layer for the three most abundant cell types: macrophages, cancer cells,
320 and fibroblasts, and used the GSEApy package to assess against the CellMarker_2024
321 gene set, and showed excellent concurrence with 60%, 70%, and 60% gene set overlap
322 respectively, as well as 3/3 aligned cell type calls for the top most pathway.

323 These methods demonstrate how even in the absence of ground truth, CITEgeist
324 interpretability allows for excellent traceable results for downstream analysis.

325 **CITEgeist preservation of spatial context and RNA informational landscape**
326 **allows for seamless integration with additional analysis tools**

327 A key strength of CITEgeist is the interpretability of its output, facilitating seamless
328 integration with various downstream analytical tools due to its use of linear assignment.
329 Because we utilize strict integer assignments of RNA transcripts, we maintain what are
330 effectively raw counts for programs such as DeSeq2 [19]. In Figure 4C, we illustrate the
331 functional effect of CITEgeist deconvolution within the interoperable AnnData format. By
332 stacking the deconvoluted compartments vertically and maintaining integer counts, we

333 connect CITEgeist objects to COMMOT [19], a signaling prediction package designed
334 for spatial data. We validated the biological relevance of CITEgeist results by examining
335 all detected secretion signals and comparing them with established biological findings
336 (Figure 4). To validate the accuracy of this output, we conducted a correlation analysis
337 using the output from our samples for secreted signals and their assigned cell types, as
338 well as known secretion amounts from the Human Protein Atlas (Figure 4E). This
339 analysis showed a consistent correlation (Spearman Rho = 0.46, p-value = 2.86e-37).

340 **Signaling analysis using CITEgeist deconvoluted spatial transcriptomics**
341 **identifies an upregulation in MDK signaling, confirmed with in-vitro experiments**

342 After validating CITEgeist performance on simulated ground truth datasets and also
343 real-world samples for interpretability, we applied CITEgeist to analyze clinical samples,
344 focusing on a breast cancer case with an *ESR1 D538G* mutation initially identified
345 through pilot bulk-RNA sequencing and subsequently confirmed by droplet digital PCR
346 (ddPCR) (Supplementary Figure 1). This mutation was of particular interest, given the
347 clinical relevance of *ESR1* mutations, and our laboratory's previous research in this area
348 [13].

349 We have previously shown that *ESR1* mutant breast cancers express basal
350 cytokeratins, and using CITEgeist on spatial transcriptomics from a case with an *ESR1*
351 mutation, we confirmed the presence of basal genes in the Cancer Cell layer (Figure
352 5A). However, we were intrigued to note that these signals were not clustered in one
353 location among the cancer cells as one would expect from a single subclone (Figure
354 5B), despite exhibiting all of the same gene mutation signatures from Estrogene [20]

355 (Figure 5C) and known pathway alterations from our previous work (Figures 5D, 5E)
356 [13]. As one of the first spatial samples of *ESR1* mutant breast cancer in the native
357 breast microenvironment, we examined the signaling effect of these mutant breast
358 cancer cells. We thus employed the validated COMMOT signaling analysis on
359 CITEgeist-deconvoluted data, which revealed increased MDK, PTN, and MIF signaling
360 in the mutant regions (Figure 5F). MDK has recently been shown to have a role in
361 breast cancer , age, and tumorigenesis [21,22], and to validate our in silico findings, we
362 performed ELISA on media from *ESR1* D538G MCF7 cell lines, and conducted analysis
363 of the MDK immunofluorescence of MCF7 *ESR1* mutant cell lines. ELISA confirmed that
364 the mutant cell line secreted double the MDK in environments both with and without
365 estradiol ($p < 0.0001$) (Figure 5H). Additionally, the literature review identified that
366 midkine has a strong pericellular effect, with a significant proportion being retained near
367 the cell in non-cancerous settings [23]. Hence, we conducted image analysis of
368 immunofluorescence on the same cell lines, which showed that in the vehicle, estrogen-
369 deprived environment (akin to a treated patient), the mutant cell lines have
370 approximately double the midkine at the cell membrane ($p < 0.001$) as well as in the cell
371 as a whole ($p < 0.01$) (Figure 5I).

372 **Discussion**

373 Our study demonstrates that CITEgeist offers a robust computational approach for
374 deconvoluting complex cellular architectures in cancer tissues. Here, we discuss the
375 broader implications of our findings and contextualize them within the current landscape
376 of spatial transcriptomics analysis.

377 **CITEgeist avoids many of the difficulties in using single-cell references in**
378 **heterogeneous cancer environments.**

379 The integration of same-slide proteomics with our deconvolution method significantly
380 enhances both interpretability and analytical robustness. By simultaneously capturing
381 transcriptomic and proteomic profiles from the same tissue section, CITEgeist mitigates
382 the technical variability inherent in comparing sequential slides, which is particularly
383 valuable in heterogeneous cancer tissues where cellular composition can vary
384 substantially across minimal spatial distances. This multi-modal approach provides
385 internal validation and enables more confident cell-type assignment, especially in
386 ambiguous cases where transcriptional profiles alone may be insufficient for accurate
387 classification.

388 For example, we generated the reference size dependence test because Cell2Location
389 has previously been demonstrated to be reasonably robust even in cases where the
390 reference may not come from the exact same informational distribution [8], and hence
391 wanted to replicate this test to test for the performance of each method in
392 circumstances mimicking the effects of cancer heterogeneity. RCTD was noticeably
393 challenged by this test, with a 4-fold increase in error – emphasizing the unique

394 challenge of applying tools benchmarked on common public normal tissue datasets to a
395 heterogeneous cancer dataset.

396 **CITEgeist is designed with interoperability and clinical considerations for**
397 **maximum utility in analysis that allow for interpretable laboratory validation**

398 A key contribution of our work lies in the methodology's capacity to not only confirm
399 existing biological knowledge but also to uncover novel insights during tool
400 development. The identification of upregulated MDK signaling in *ESR1*-mutated breast
401 cancer exemplifies how computational tools can drive biological discovery when
402 designed with interpretability as a core principle. This finding emerged from our
403 analytical pipeline rather than being a predetermined hypothesis, highlighting the value
404 of data-driven approaches in uncovering unexpected biological mechanisms.

405 The *ESR1* D538G mutation identified in our study has significant implications for
406 understanding tumor microenvironment dynamics. This mutation, occurring in the
407 ligand-binding domain of the estrogen receptor, is known to confer constitutive activity
408 and resistance to endocrine therapies. Our findings suggest that beyond its cell-
409 autonomous effects, this mutation influences the broader tumor ecosystem through
410 altered paracrine signaling, particularly via MDK. The increased MDK secretion
411 observed in mutant cells potentially modulates immune cell recruitment, angiogenesis,
412 and stromal remodeling, creating a microenvironment more conducive to tumor
413 progression and therapy resistance. This expanded understanding of mutation effects
414 beyond cancer cell-intrinsic properties represents an important shift in how we
415 conceptualize the consequences of genomic alterations.

416 **Limitations**

417 There are several limitations associated with CITEgeist. It is currently dependent on the
418 variety and accuracy of the protein panel sequenced with spatial transcriptomics. It must
419 be ordered together, which limits its utility in already sequenced datasets that lack
420 protein information. However, as a first-in-class tool, it fills a necessary gap in the field
421 where such samples already exist and have been sequenced, but researchers lack the
422 tool to integrate proteomics and transcriptomics.

423 Secondly, our clinical trial has a small number of samples and is limited to breast
424 cancer; hence, there may be other concerns or issues that emerge with other cancer
425 types or tissue samples processed differently from other centers, and thus, careful
426 usage and incorporation of feedback are critical to the long-term sustainability of the
427 method.

428 Lastly, we also repeated our ELISA and imaging analysis in T47D, a different breast
429 cancer cell line, which did not recapitulate the results (Supplemental Figure S3).
430 However, previous work has established that the effects of these mutants are often
431 context-specific [24,25]. It is important to note that breast cancer is a notably
432 heterogeneous disease [26], hence, the purpose of our wet lab validation is not to find
433 something true for all breast cancer, but to demonstrate that CITEgeist output can be
434 directly input into external tools developed by others, and still provide analysis that can
435 be confirmed in an isogenic model, proving that CITEgeist deconvolution provides a
436 robust statistical foundation for further abstractions via interoperable tool usage. Further
437 work is needed to establish the whole pathway that supports increased midkine in
438 certain ESR1 mutant breast tumors.

Future Directions

Looking forward, several promising directions emerge from this work. First, the integration of temporal dynamics into spatial analysis would provide valuable insights into how signaling patterns evolve during disease progression and treatment response. Second, expanding our approach to incorporate additional data modalities, such as chromatin accessibility or metabolomic profiles, could further refine cell-type characterization and functional state assessment. Finally, the application of CITEgeist to larger patient cohorts across multiple cancer types would enable the identification of conserved and cancer-specific microenvironmental signatures, potentially revealing new therapeutic vulnerabilities.

Additionally, the computational framework established here could be adapted for other complex tissue architectures beyond cancer, such as in inflammatory conditions or developmental contexts, where cellular heterogeneity and spatial organization profoundly influence tissue function. The principles of neighborhood-based analysis and multi-modal integration are broadly applicable across diverse biological systems.

Conclusions

In this study, we present CITEgeist, a computational method designed explicitly for accurate cell-type deconvolution in heterogeneous cancer tissues. Our approach demonstrates superior performance in mixed cellular architectures compared to existing tools, particularly in the challenging context of dysplastic cancer tissues. The integration of neighborhood context at multiple scales enables robust cell-type identification even in regions of high cellular admixture.

The interpretability of CITEgeist outputs facilitates seamless integration with complementary analytical tools, enhancing flexibility in downstream analyses. This interoperability allowed us to leverage CITEgeist's deconvolution capabilities to identify and validate a novel biological finding: increased MDK signaling associated with the ESR1 D538G mutation, with potential implications for understanding endocrine resistance mechanisms.

By combining computational innovation with biological validation, our work contributes both a valuable methodological advancement and a specific mechanistic insight relevant to cancer biology. CITEgeist represents a step toward more comprehensive characterization of the complex cellular ecosystems in cancer, ultimately supporting the development of more effective therapeutic strategies that account for the full complexity of the tumor microenvironment.

List of abbreviations

ESR1 = Estrogen Receptor 1

CITEgeist = Cellular Indexing of Transcriptomes and Epitopes for Guided Exploration of Intrinsic Spatial Trends

MDK = Midkine

Declarations

Ethics approval and consent to participate

The protocol for the related clinical trial was approved by the University of Pittsburgh IRB under STUDY21100091 and conducted through the UPMC Hillman Cancer Center under protocol 22-088

Consent for publication

Not applicable.

Availability of data and materials

The Wu et al. breast cancer atlas can be accessed according to the instructions in their original publication; the processed scRNA-seq data utilized in this study can be downloaded through GEO (accession number: GSE176078, accessed 10/25/24). The downsampled reference and simulated test datasets will be available via FigShare at publication (doi: 10.6084/m9.figshare.28385675).

Visium datasets used in this paper will be published at the time of publication and are currently embargoed. Reviewers can access the data via the link and private access token provided to the journal in the cover letter at the time of submission.

All correspondence should be addressed to Adrian V. Lee at leeav@upmc.edu.

Competing interests

All authors have no competing interests to declare.

Funding

500 This research was partly supported by the University of Pittsburgh Center for Research
501 Computing (RRID: SCR 022735) through the resources provided. We utilized the HTC
502 cluster, supported by NIH award S10OD028483, and the H2P cluster, supported by
503 NSF award OAC-2117681. Additionally, this work was funded by US NIH Grant
504 R01HG010589, NIH Grant 5F30CA264963-03, and a Developmental Pilot Award from
505 the UPMC Hillman Cancer Center, supported through NIH grant P30CA047904.

506 Histology sectioning was performed in the Pitt Biospecimen Core (RRID: SCR 025229),
507 and the services and instruments used in this project were supported, in part, by the
508 University of Pittsburgh and the Office of the Senior Vice Chancellor for Health
509 Sciences. Work conducted at the UPMC Hillman Cancer Center Tissue and Research
510 Pathology Services (TARPS) Shared Resource Facility was also supported, in part, by
511 the University of Pittsburgh and the National Cancer Institute of the NIH under Award
512 Number P30CA047904.

513 Lastly, 10x Visium library generation, Takara Smart Stranded Total RNA library
514 generation, and Illumina sequencing were performed in the Health Sciences
515 Sequencing Core (RRID: SCR 023116) at UPMC Children's Hospital, Rangos Research
516 Center. These services and instruments were generously supported, in part, by the
517 University of Pittsburgh, the Office of the Senior Vice Chancellor for Health Sciences,
518 the Department of Pediatrics, the Institute for Precision Medicine, and the Richard K
519 Mellon Foundation for Pediatric Research.

520 **Authors' contributions**

521 A.C.C. conceived the project, designed the computational framework, wrote the
522 CITEgeist code, conducted the bioinformatic analyses, and prepared the initial
523 manuscript draft. B.T.S. performed the CITEgeist benchmarking, simulation studies, and
524 statistical analysis and contributed to discussing benchmarking and simulation
525 strategies. N.C. facilitated the collection of clinical samples and coordinated the clinical
526 trial. P.F.A. served as the clinical trial's principal investigator and performed breast
527 surgery, collecting the samples. H.W completed the wet lab validation of MDK signaling
528 in ESR1 mutant cell lines. S.O. oversaw the overall project management and shaped
529 the project's goals. R.S. contributed to manuscript revisions and provided expertise in
530 mathematical modeling and computational methodologies. A.V.L. guided the project
531 design and objectives. All authors reviewed and edited the manuscript.

532 **Acknowledgements**

533 We want to thank Jingyang Qian for their gracious guidance in utilizing the scCube
534 method to simulate and benchmark spatially resolved Visium-like datasets from a
535 single-cell reference. We also thank Jian Chen from the Lee-Oesterreich Lab for her
536 help with RNA sequencing prep and ddPCR.

537 We acknowledge the assistance of the physicians and clinical support staff at the
538 University of Pittsburgh Medical Center health system for their contributions to the
539 clinical trial—special thanks to Natera Inc. for their support with the clinical trial
540 NCT05914792.

541

542

Supplementary Files

Supplementary File 1: Full mathematical derivation of the CITEgeist method

Supplementary Files 2: Tables reporting the metrics of all compared and benchmarked methods.

Supplementary Files 3: Raw data from all wet lab experiments (ELISA and IF).

Supplementary Files 4: Supplementary Statistical and Sequencing Methods

References

1. Huang H, Wei T, Zhang A, Zhang H, Kong L, Li Y, et al. Trends in the incidence and survival of women with hormone receptor-positive breast cancer from 1990 to 2019: a large population-based analysis. Sci Rep. 2024;14:23690.

2. Patel R, Klein P, Tiersten A, Sparano JA. An emerging generation of endocrine therapies in breast cancer: a clinical perspective. Npj Breast Cancer. 2023;9:1–12.

3. Regan C, Preall J. Practical Considerations for Single-Cell Genomics. Curr Protoc. 2022;2:e498.

4. Chang AC-C, Balic M, Bartholow T, Bhargava R, Brown DD, Brown L, et al. Hope for OTHERS (Our Tissue Helping Enhance Research & Science): research results from the University of Pittsburgh rapid autopsy program for breast cancer. Breast Cancer Res [Internet]. 2025 [cited 2025 July 9];27. Available from: https://link.springer.com/epdf/10.1186/s13058-025-02014-9

5. Wess M, Andersen MK, Midtbust E, Guillem JCC, Viset T, Størkersen Ø, et al. Spatial integration of multi-omics data from serial sections using the novel Multi-Omics Imaging Integration Toolset. GigaScience. 2025;14:giaf035.

6. Zaidi M, Fu F, Cojocari D, McKee TD, Wouters BG. Quantitative Visualization of Hypoxia and Proliferation Gradients Within Histological Tissue Sections. Front Bioeng Biotechnol [Internet]. 2019 [cited 2025 Aug 23];7. Available from: https://www.frontiersin.org/journals/bioengineering-and-biotechnology/articles/10.3389/fbioe.2019.00397/full

7. Chen J, Liu W, Luo T, Yu Z, Jiang M, Wen J, et al. A comprehensive comparison on cell-type composition inference for spatial transcriptomics data. Brief Bioinform. 2022;23:bbac245.

8. Kleshchevnikov V, Shmatko A, Dann E, Aivazidis A, King HW, Li T, et al. Cell2location maps fine-grained cell types in spatial transcriptomics. Nat Biotechnol. 2022;40:661–71.

9. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

10. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36:411–20.

11. Wu SZ, Al-Eryani G, Roden DL, Junankar S, Harvey K, Andersson A, et al. A single-cell and spatially resolved atlas of human breast cancers. Nat Genet. 2021;53:1334–47.

583 12. Qian J, Bao H, Shao X, Fang Y, Liao J, Chen Z, et al. Simulating multiple variability in
584 spatially resolved transcriptomics with scCube. Nat Commun. 2024;15:5021.

585 13. Li Z, McGinn O, Wu Y, Bahreini A, Priedigkeit NM, Ding K, et al. ESR1 mutant breast
586 cancers show elevated basal cytokeratins and immune activation. Nat Commun.
587 2022;13:2011.

588 14. McAuliffe P. Longitudinal ctDNA Monitoring in Older Women With ER+ Breast Cancer
589 Who Forego Upfront Surgery in Favor of Primary Endocrine Therapy [Internet].
590 clinicaltrials.gov; 2025 Apr. Report No.: NCT05914792. Available from:
591 https://clinicaltrials.gov/study/NCT05914792

592 15. Carleton N, Chang AC, Chen F, Puhalla SL, Foldi J, Waltermire H, et al. Longitudinal
593 ctDNA Surveillance in Older Women with ER+ Breast Cancer to Facilitate Surgical De-
594 Escalation: A Prospective, Hybrid-Decentralized Trial with Correlative Studies [Internet].
595 medRxiv; 2025 [cited 2025 Sept 11]. p. 2025.08.23.25332468. Available from:
596 https://www.medrxiv.org/content/10.1101/2025.08.23.25332468v1

597 16. Biancalani T, Scalia G, Buffoni L, Avasthi R, Lu Z, Sanger A, et al. Deep learning and
598 alignment of spatially resolved single-cell transcriptomes with Tangram. Nat Methods.
599 2021;18:1352–62.

600 17. Li X, Wang C-Y. From bulk, single-cell to spatial RNA sequencing. Int J Oral Sci.
601 2021;13:36.

602 18. Li H, Zhou J, Li Z, Chen S, Liao X, Zhang B, et al. A comprehensive benchmarking with
603 practical guidelines for cellular deconvolution of spatial transcriptomics. Nat Commun.
604 2023;14:1548.

605 19. Anders S, Huber W. Differential expression analysis for sequence count data. Genome
606 Biol. 2010;11:R106.

607 20. Li Z, Chen F, Chen L, Liu J, Tseng D, Hadi F, et al. EstroGene2.0: A multi-omic database
608 of response to estrogens, ER-modulators, and resistance to endocrine therapies in breast
609 cancer. bioRxiv. 2024;2024.06.28.601163.

610 21. Visvader JE. Midkine links aging with breast cancer—A new predictor of cancer risk.
611 Cancer Cell. 2024;42:1815–7.

612 22. Yan P, Jimenez ER, Li Z, Bui T, Seehawer M, Nishida J, et al. Midkine as a driver of age-
613 related changes and increase in mammary tumorigenesis. Cancer Cell. 2024;42:1936-
614 1954.e9.

615 23. Muramatsu T. Midkine, a heparin-binding cytokine with multiple roles in development,
616 repair and diseases. Proc Jpn Acad Ser B Phys Biol Sci. 2010;86:410–25.

24. Arnesen S, Blanchard Z, Williams MM, Berrett KC, Li Z, Oesterreich S, et al. Estrogen receptor alpha mutations in breast cancer cells cause gene expression changes through constant activity and secondary effects. Cancer Res. 2021;81:539–51.

25. Li Z, Wu Y, Yates ME, Tasdemir N, Bahreini A, Chen J, et al. Hotspot ESR1 Mutations Are Multimodal and Contextual Modulators of Breast Cancer Metastasis. Cancer Res. 2022;82:1321–39.

26. Guo L, Kong D, Liu J, Zhan L, Luo L, Zheng W, et al. Breast cancer heterogeneity and its implication in personalized precision therapy. Exp Hematol Oncol. 2023;12:3.

Figure 1: Graphical abstract of the CITEgeist development process

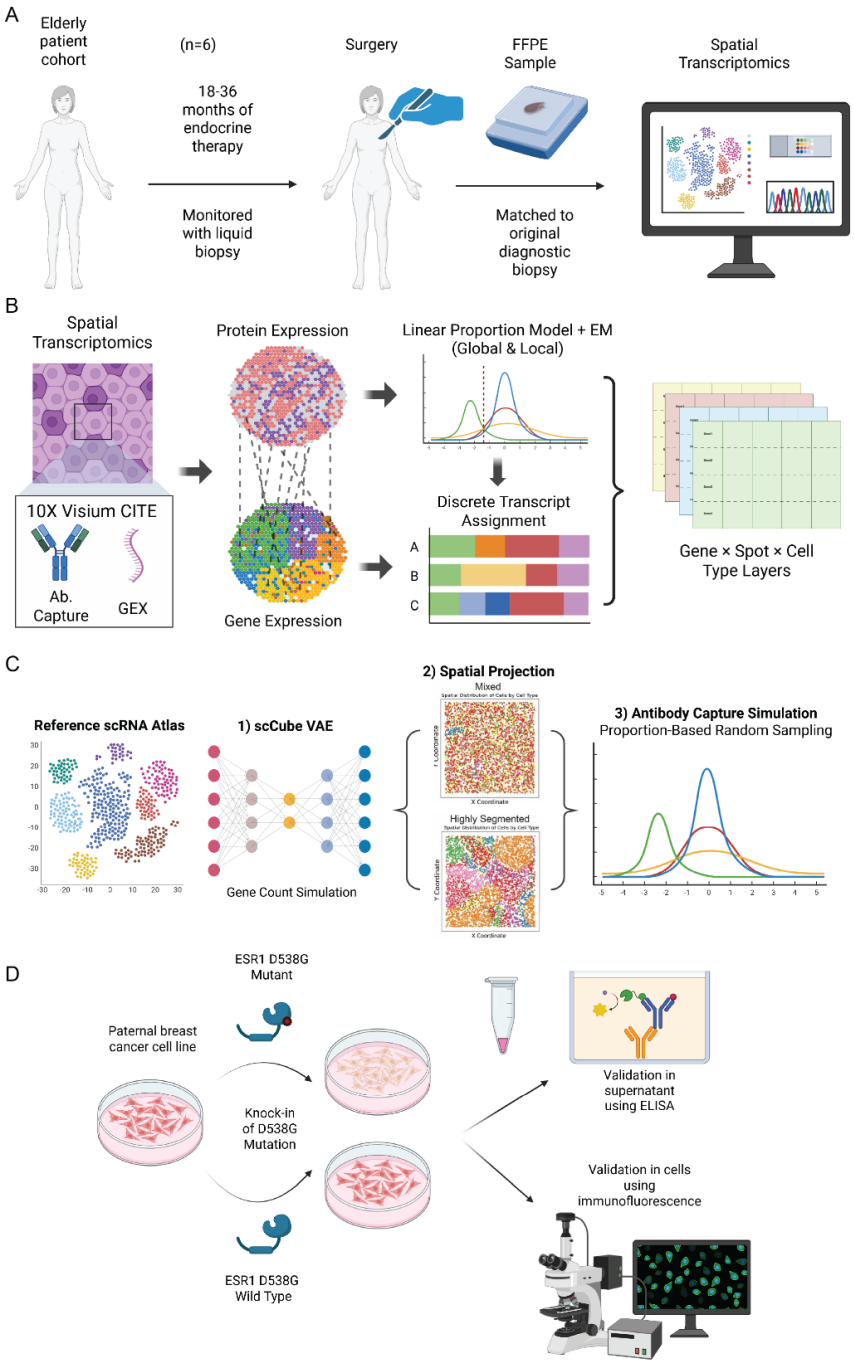

Figure 1. (A) Graphical abstract describing the clinical trial tissue collection process. (B) Graphical abstract describing the CITEgeist approach. (C) Graphical abstract describing the complete simulation pipeline for benchmarking CITEgeist against other state of the art single cell reference based spatial deconvolution methods. (D) Graphical abstract describing the wet lab validation process.

Figure 2: Performance evaluation of CITEgeist and reference-based cell type proportion deconvolution methods on simulated Visium datasets.

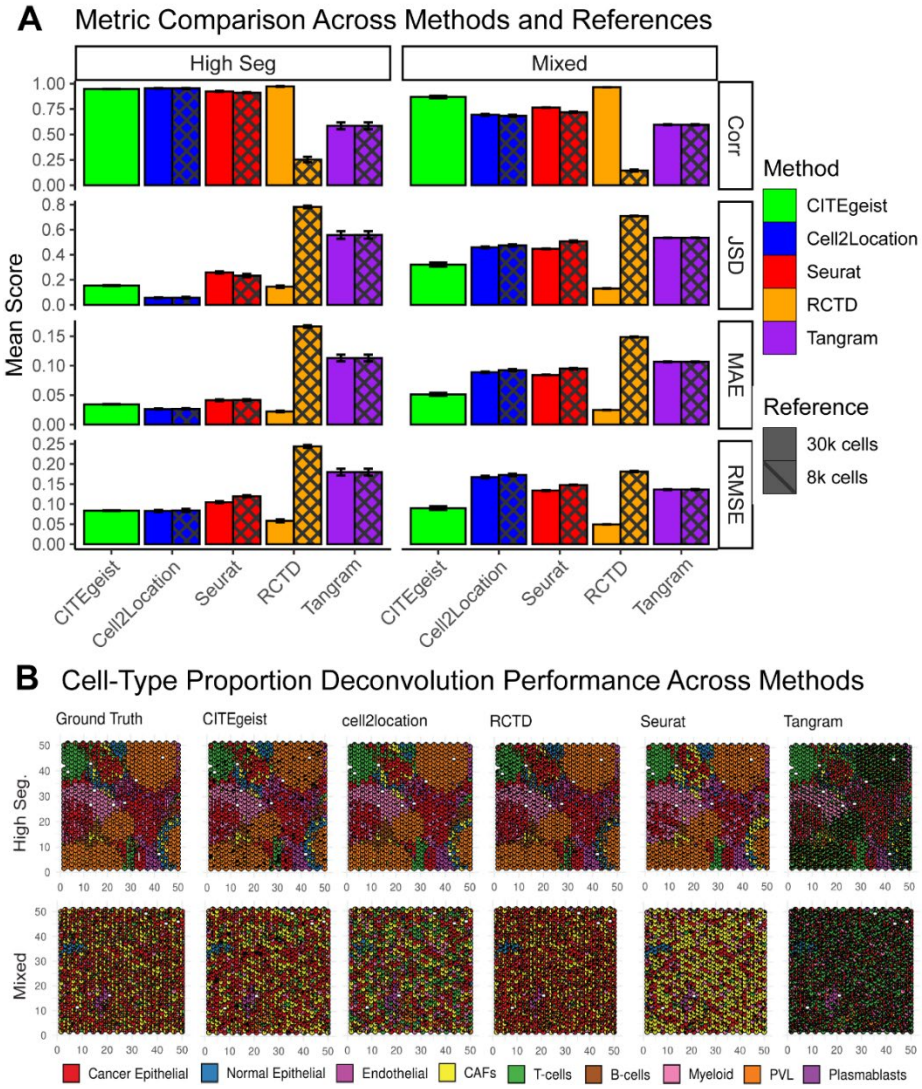

Figure 2. (A) Quantitative comparison of deconvolution accuracy across different spatial patterns ("Highly Segmented" and "Mixed") and reference atlas sizes (30,000 or 8,000 cells). Metrics include Pearson's Correlation (Corr, higher values indicate better performance), Jensen-Shannon Divergence (JSD), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE), with lower values indicating better performance for the latter three error metrics. Error bars represent standard error (n = 5 independent replicates per condition). (B) Representative deconvolution results from all benchmarked methods showing cell-type proportion estimates across spatial spots for both Highly Segmented (top row) and Mixed (bottom row) tissue patterns. Each spot is colored according to the estimated proportions of nine cell types (legend at bottom). Ground truth (leftmost column) shows the actual cell-type distributions, while subsequent columns display predictions from each benchmarked method. Samples were selected from n = 5 independent simulations per condition to represent typical performance.

Figure 3: Metric comparison of gene expression (GEX) deconvolution per spot, evaluated across all benchmarked methods and reference cell types.

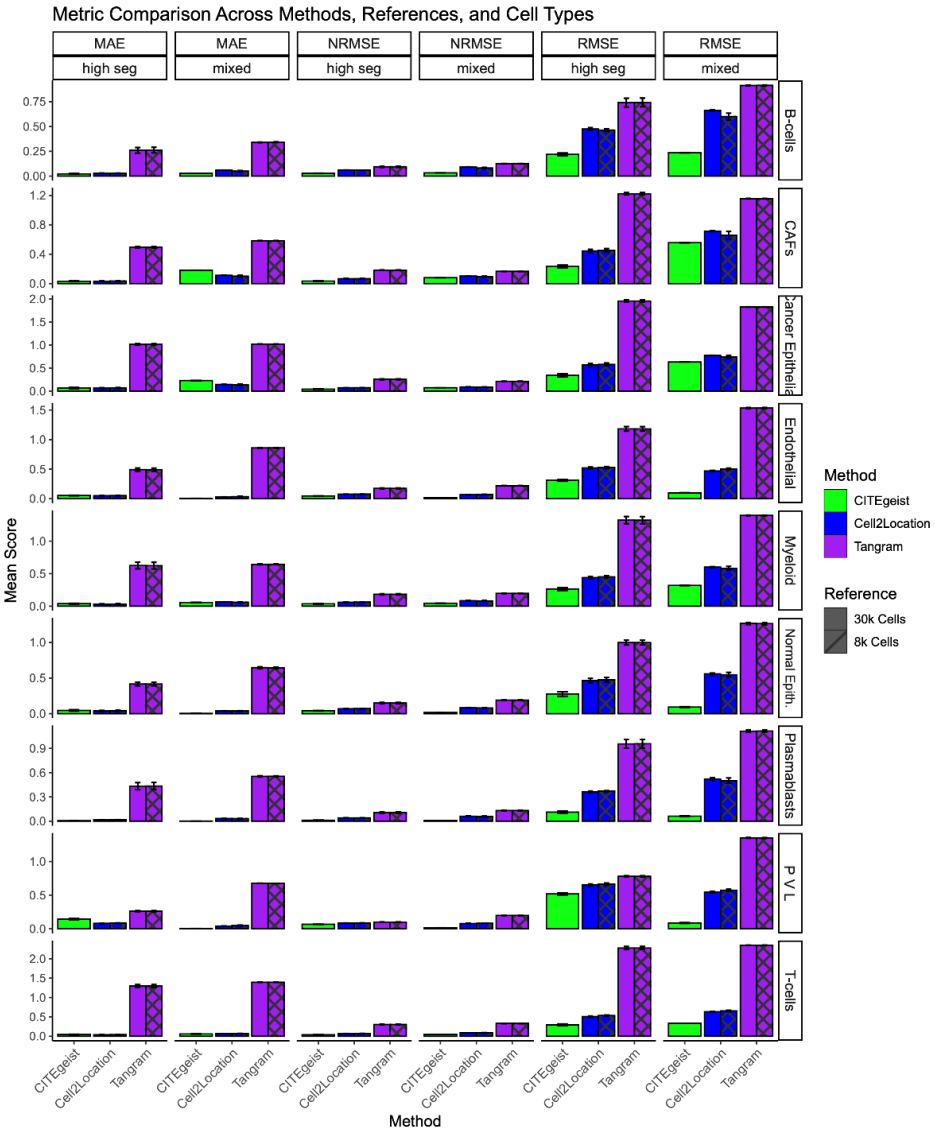

Figure 3. Metric comparison of gene expression (GEX) deconvolution per spot, evaluated across all benchmarked methods and reference cell types. Metrics include Pearson's Correlation (Corr, higher values indicate better performance), Jensen-Shannon Divergence (JSD), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE), with lower values indicating better performance for the latter three error metrics. Error bars represent standard error (n = 5 independent replicates per condition).

Figure 4: Demonstration of CITEgeist validity when applied to real cancer data.

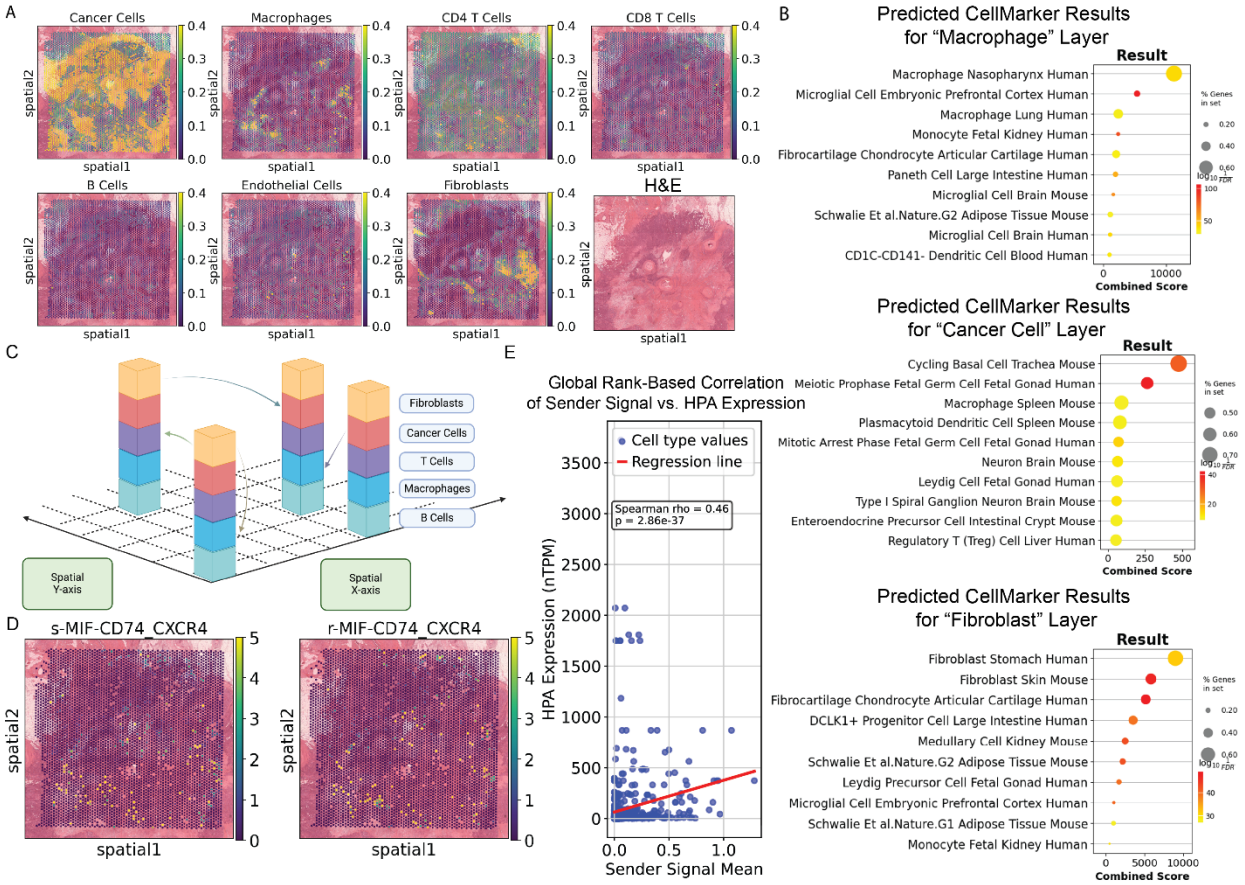

Figure 4. (A) Spatial plots showing the predicted proportions from CITEgeist output of the celltypes in sample HCC22-088-P1-S2 (B) CellMarker 2024 GSEApy calls, showing the top predicted celltypes for their respective 'assigned' gene expression layer. (C) Graphical abstract demonstrating how AnnData objects permit stacking of cell 'compartments' in the same adata spatial spot, and how COMMOT calculates using this information (D) Spatial plot showing an example of the detected sender and receiver signal for pathway MIF-CD74 CD44. (E). Global correlation regression plot showing the correlation between cell type assigned sender signal and said ligands expression in the Human Protein Atlas.

Figure 5: Wet lab validation of predictions based on CITEgeist output using external analytical packages.

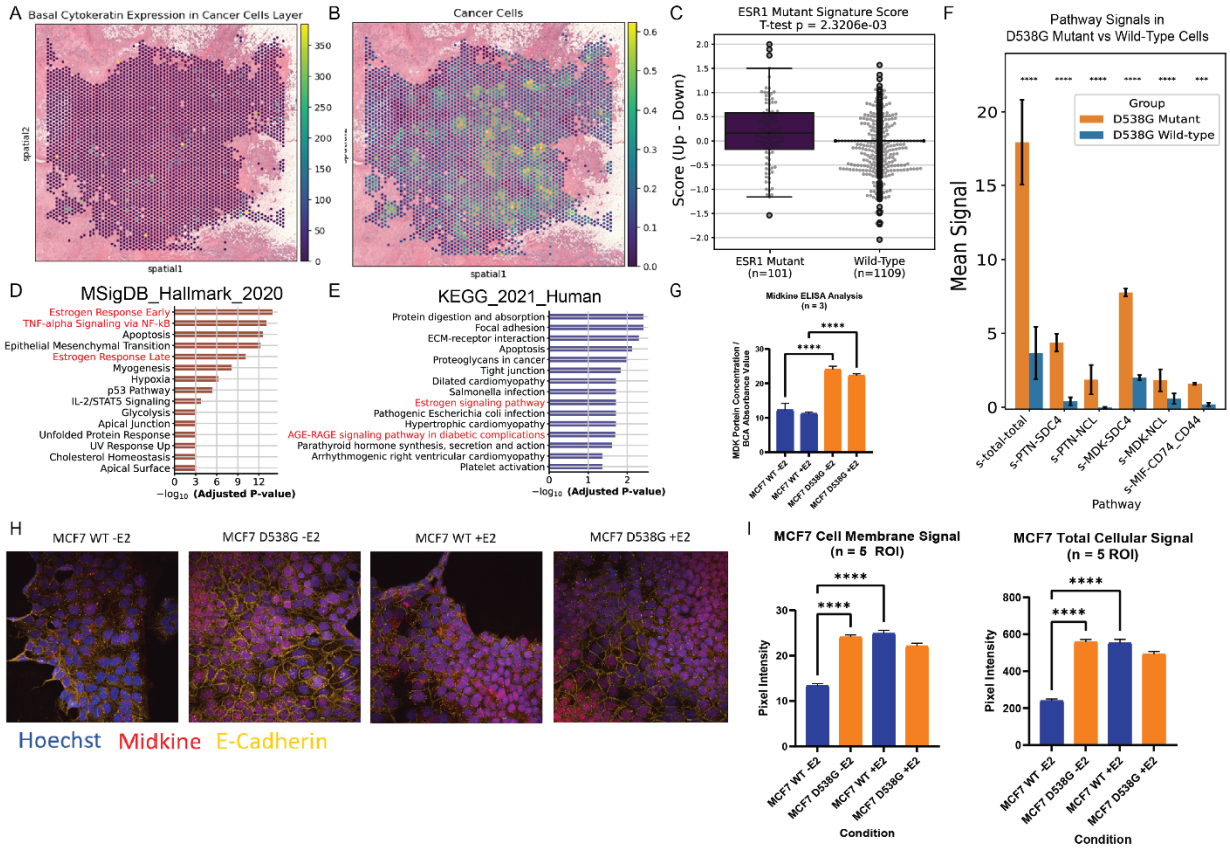

Figure 5. (A) Spatial plots showing the heterogenous deconvoluted result of elevated basal cytokeratin expression in the (B) Cancer Cells layer in the surgical sample from the same patient in sample HCC22-088-P4-S2, which has an ESR1 D538G mutation. (C) Combined box and scatter plot showing the average ESR1 Mutant Gene Signature Score between ESR1 D538G mutant cancer cells and WT cells. (T-test value 3.1229, p-value 2.3206e−03 (D) Gene pathways upregulated in the D538G cancer spots from the MSigDB Hallmark 2020 gene set. Red text pathways are previously confirmed upregulated activities in ESR1 mutant cells from Li et al. [20] (E) Gene pathways upregulated in the D538G cancer spots from the KEGG 2021 Human gene set. Red text pathways are previously confirmed upregulated activities from Li et al. (F) Bar plot showing the statistically significant COMMOT sender signal differences between D538G cancer spots and WT cancer spots. A Mann-Whitney U test was performed to compare pathway-specific sender signals between cells with and without the D538G mutation. Error bars are SEM. The two-way test statistic and p-value were computed, and pathways were ranked by significance. The mean difference in signal between mutation-positive and mutation-negative cells was calculated. Benjamini-Hochberg FDR correction was applied to control for multiple comparisons, and pathways with FDR < 0.05 were identified as significantly different. (G) Quantification of midkine ELISA results from MCF7 wild type and with ESR1 D538G mutation. Statistical testing was done with one-way ANOVA with Tukey's post-hoc multiple

comparison (H) Image examples of MCF7 WT and ESR1 D538G with and without 24 hours of E2 treatment after 5 days of E2 starvation. (I) Quantification of image examples assessing the midkine signal at the cell membrane and for the total cell signal, with values compared and tested via unpaired T-test. CellProfiler was used to assess values for each cell after Z Max Projection across 5 ROIs per slide.

Supplementary Figures

Figure S1: ddPCR confirming ESR1 mutation in HCC22-088-P4-S2

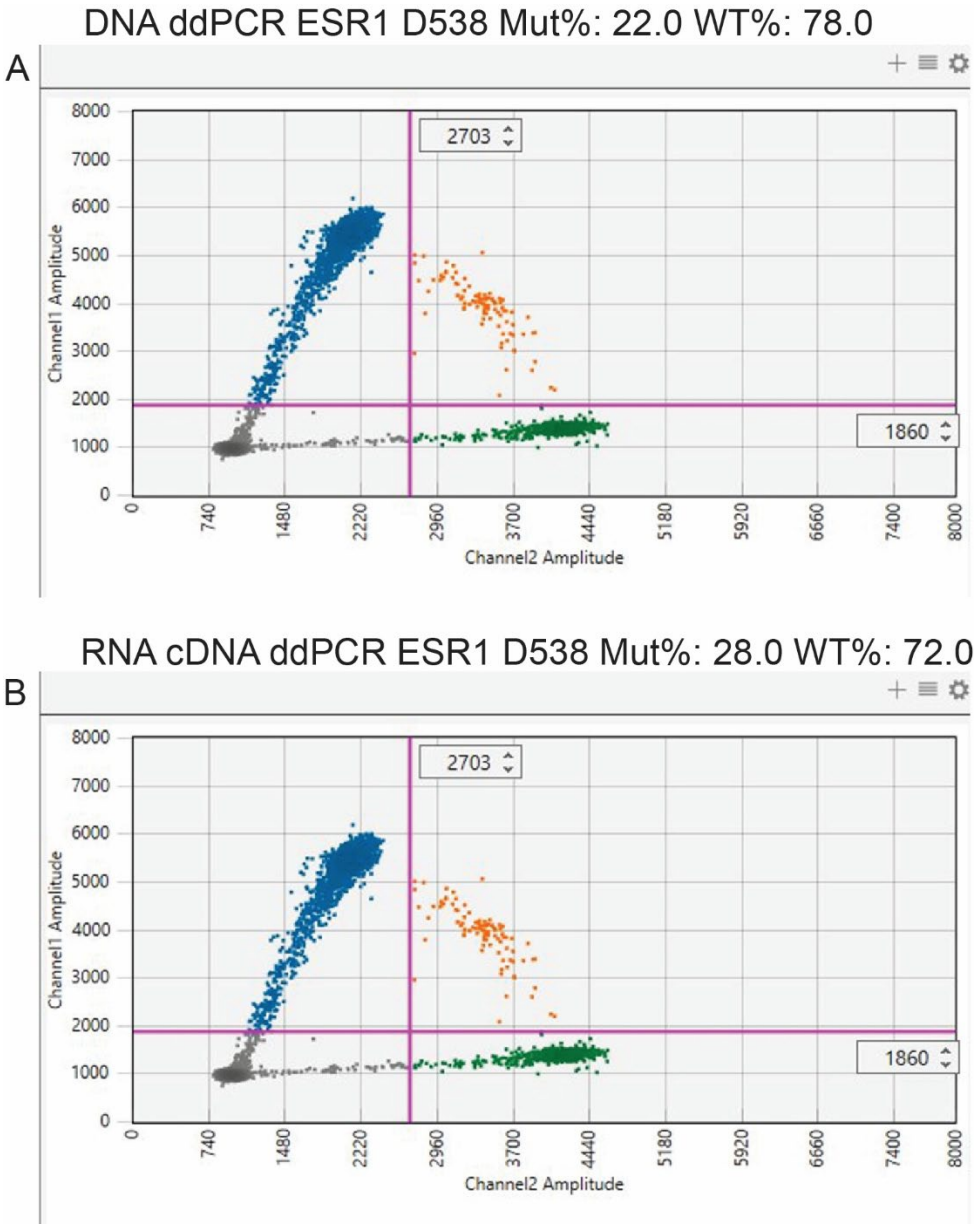

Figure S1. ddPCR confirmation of D538G mutation in Sample HCC22-088-P4-S2 (A) DNA, (B) RNA

Figure S2: Runtime comparison of CITEgeist against other SOTA methods.

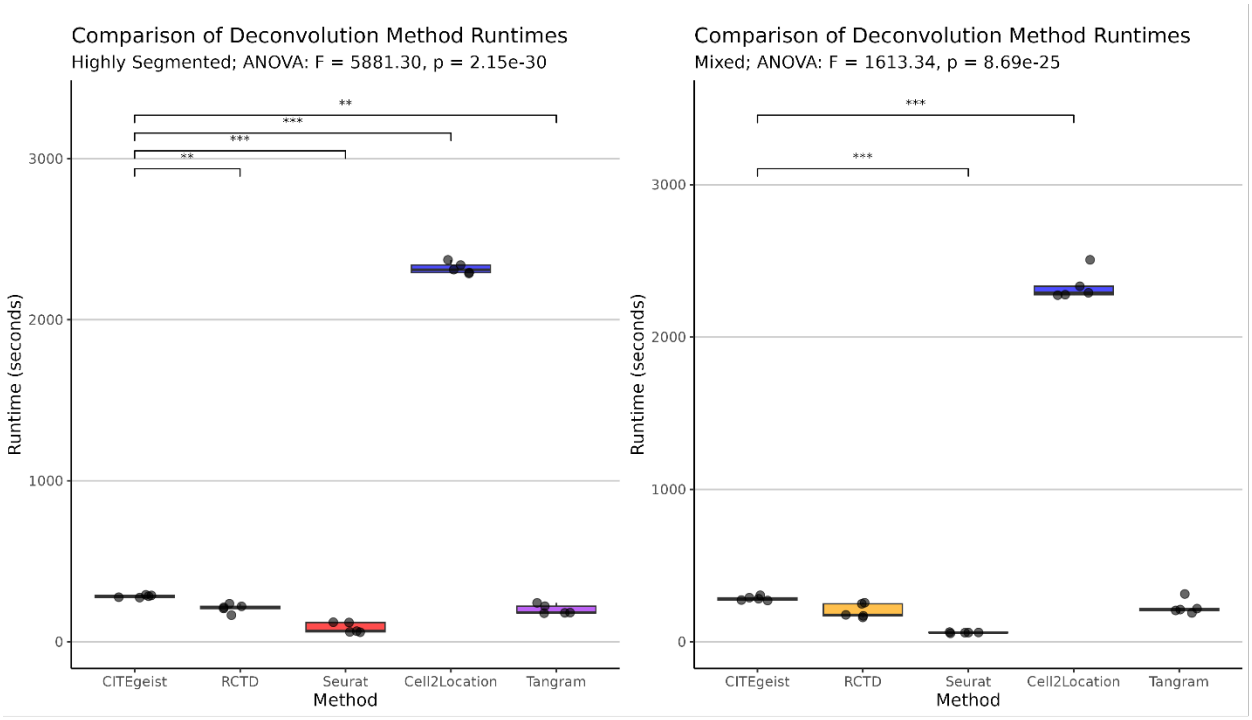

Figure S2. Runtime comparison across all benchmarked methods, in both the "Highly Segmented" **(left)** and "Mixed" **(right)** simulated conditions**.** Statistical significance was calculated via ANOVA followed by Tukey's Honest Significant Difference test.

Figure S3: Quantification of midkine ELISA and immunofluorescence signals in T47D

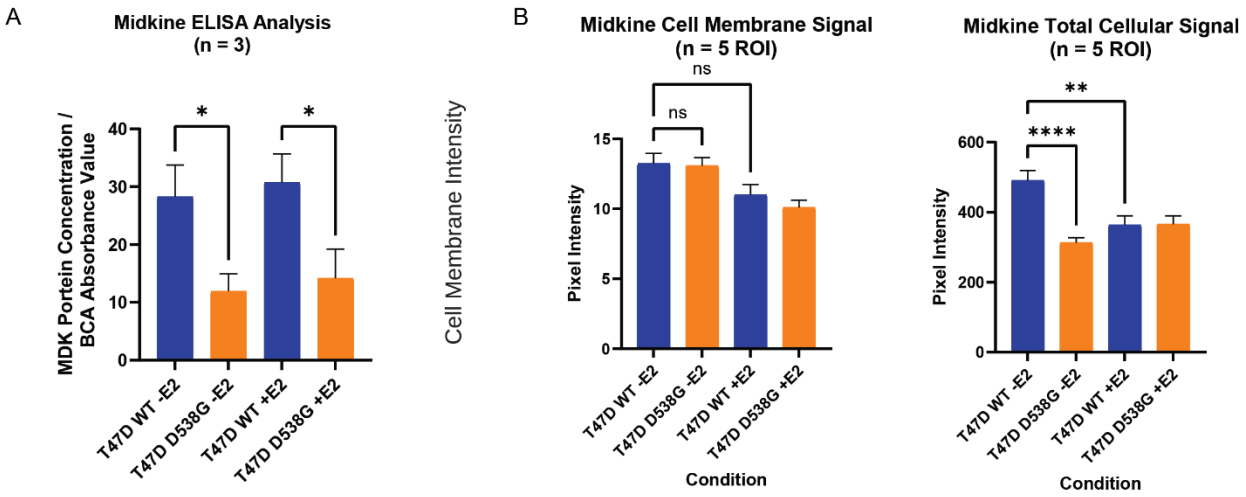

Figure S3. (A) Quantification of midkine ELISA results from T47D wild type and with ESR1 D538G mutation. Statistical testing was done with one-way ANOVA with Tukey's post-hoc multiple comparison. (B) Quantification of T47D midkine immunofluorescence assessing the midkine signal at the cell membrane and for the total cell signal, with values compared and tested via an unpaired T-test. CellProfiler was used to assess values for each cell after Z Max Projection.