Supplementary Material

Chuhanwen Sun, Alper Eroglu, Jean Hausser

Supplementary discussion 1

We are still missing a quantitative understanding of the experimental factors that cause batch effects, and of how each of these factors biases the quantification of each gene. This limited understanding of the quantitative nature of batch effects limits the possibility to correct them.

Such an understanding could benefit from generating scRNAseq samples with well-controlled, identical biology, processed so as to span the range of causes of batch effects. The data should span multiple biological systems (tissues, cell lines, species) and batch effect sources (kit versions, technician, processing speed, etc..).

Armed with such data, one could empirically survey factors responsible for batch effects, all-the-while being aware that certain factors may not be explicitly observed but still implicitly present through their impact on gene expression quantification. Explicing and implicit factors responsible for batch effects could be inferred from the data and represented as eigen-cells, vectors or functions with dimensionality equal to the number of the genes in the dataset. The space of eigencells could then span a null space where gene expression is entirely explained by batch effects. Knowing the null space spanned by these eigencells, one could collapse gene expression variation on the complement of this null space.

Yet, such datasets are rare and expire as scRNAseq technology develops[1–3]. For example, in this study, we analyzed one dataset of samples produced using different versions of 10x Genomics' 3' end gene expression kit[3]. And even so, any correction can add noise, as illustrated by a recent benchmark which found that most batch effect correction methods introduce measurable artefacts to the data[4].

Thus, rather than seeking to correct batch effects, we can take advantage of the trend that scRNAseq data is becoming abundantly available and more affordable [5], with different samples typically showing significant overlap in transcriptional states[6,7]. Experimental design can seek to minimize batch effects by processing samples in the one batch. Following that, samples with strong batch effects can be flagged and excluded using the present approach.

Supplementary Discussion 2

bioLUCID could benefit multiple applications of scRNAseq and downstream analyses, by (i) computing cell embeddings that are less confounded by batch effects and that better preserve sample-specific biology, and (ii) flagging and excluding samples with strong batch effects.

Batch-correcting embeddings could benefit applications such as

- visualizing transcriptional heterogeneity with less confounding by batch effects and better preservation of sample-specific biology (Fig. 2a-b)
- identify new cell types and subtypes with less false positives,
- infer more accurate pseudotime and differentiation trajectories.

Excluding samples with strong batch effects could benefit applications that necessitate gene expression counts:

- imputing more accurate spatial transcriptomes by multi-omics integration (Fig. 2c, Fig. S6)
- more accurate differentially expression calls by excluding replicates with strong batch effects

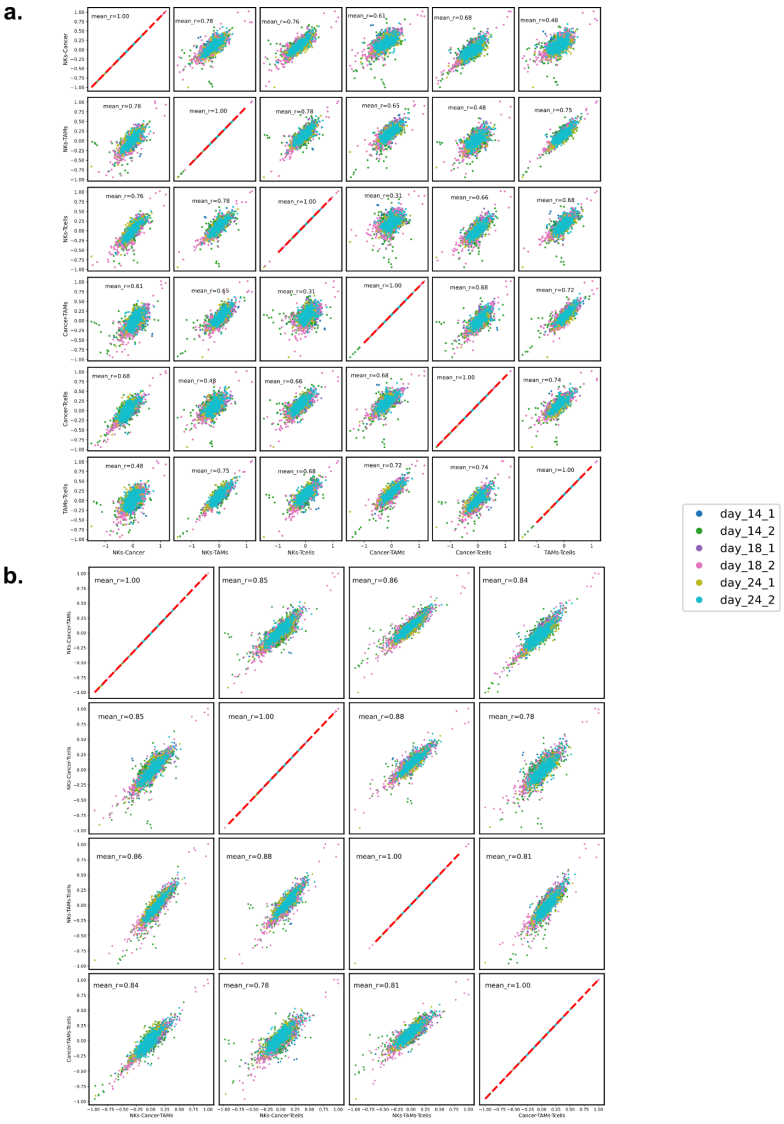
In the absence of clear gold standards for some of the above applications, we focused here on visualizing transcriptional heterogeneity (Fig. 2a-b) and on multi-omics integration (Fig. 2c) where gold standards are available.

Beyond these applications, the present findings could also make machine-learning more effective. While machine-learning has a degree of robustness to batch effects, observations suggest that there is still room for improvement. For example, spatial transcriptome imputation accuracy is increased by excluding samples with strong batch effects (Fig. 2c). And machine learning approaches to modeling scRNAseq data have used non-parametric transformations to provide "[robustness] against technical artefacts that may systematically bias the absolute transcript counts values" [8], thus trading off the more precise quantitative gene expression information provided by scRNAseq for robustness to technical artefacts.

In the context of machine-learning, bioLUCID could thus serve to exclude samples with strong batch effects, potentially reducing the need for such non-parametric transformations. The finding that batch effects tend to be shared across cell types could also inspire batch effect-aware deep learning architectures that deprioritize gene expression shared across all cell types of a given sample. Both approaches could enable simpler, more effective deep learning models, an avenue we leave for future research.

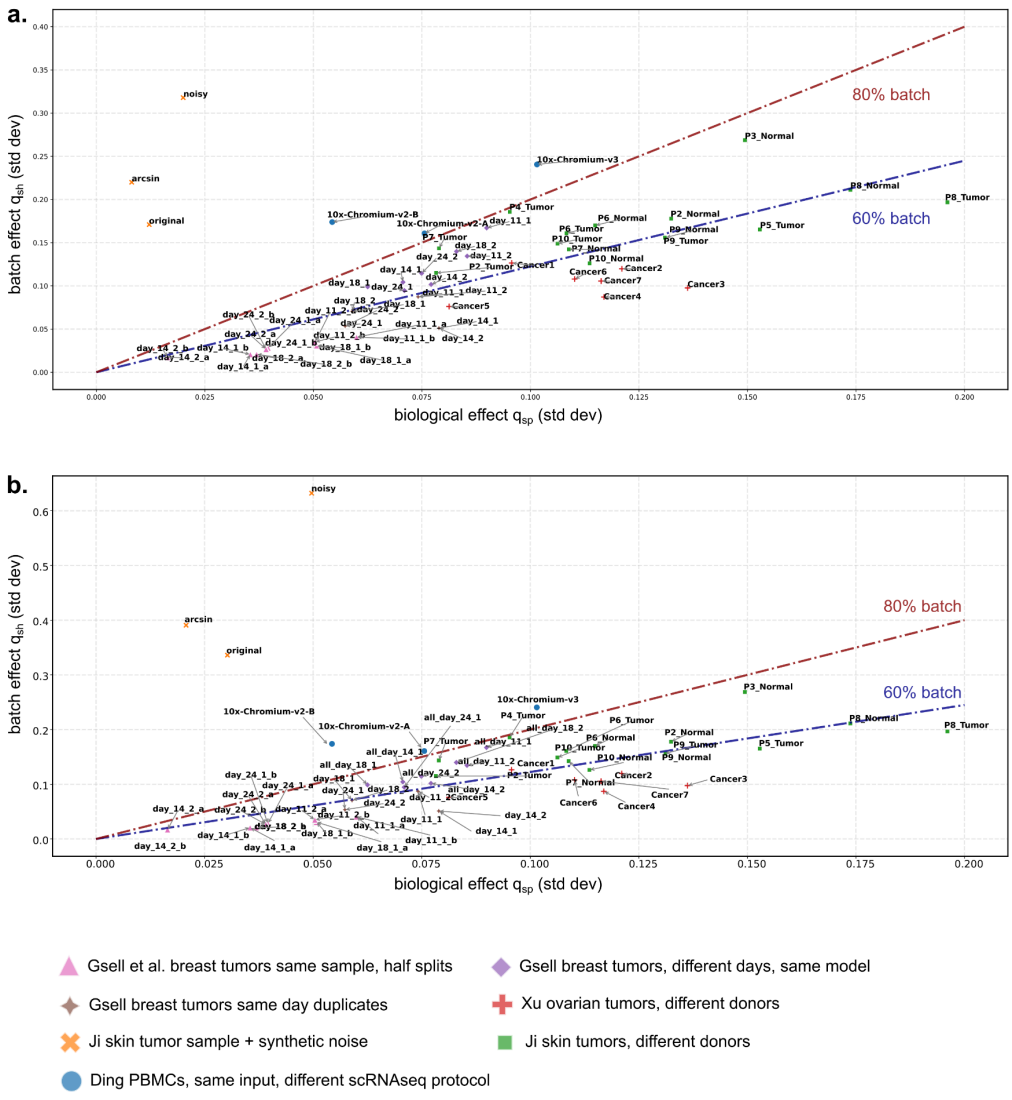
Supplementary figures

Supplementary Figure 1



Supplementary Figure 1: bioLUCID identifies consistent batch effects regardless of which cell types are used for the inference. **a-b.** Batch effects inferred from different abundant cell type combinations are highly correlated, supporting the hypothesis that batch effects are shared across cell types within the same sample. Batch effects were inferred from pairs (**a.**) or triplets (**b.**) of cell types in the 6 breast tumor samples of Gsell et al.[6] Dots: genes. Colors: samples. In panel a, we don't infer batch effects from the T/NK cells pair because both cell types have low abundance, which leads to noisier batch effect estimates.

Supplementary Figure 2

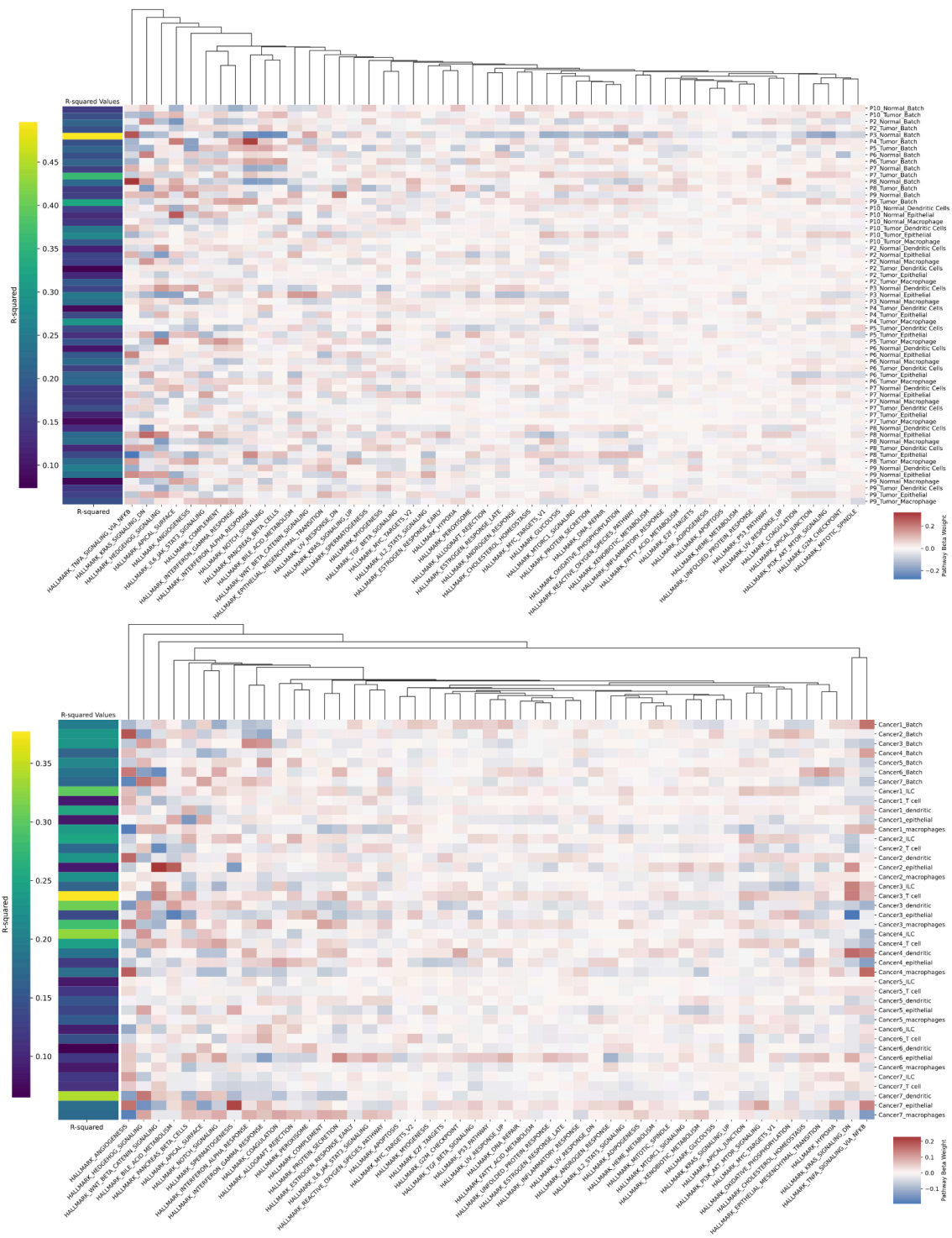


Supplementary Figure 2: The magnitude of batch and biological effects inferred by bioLUCID matches their expected contribution for each dataset. **a.** Same as Fig. 1c, with individual samples labeled by sample name for completeness. **b.** Batch and biological effects are inferred robustly for different batch effect magnitudes. Same as **a.**, using 3x stronger batch effects in the Ji synthetic noise samples.

Supplementary Figure 3

a.

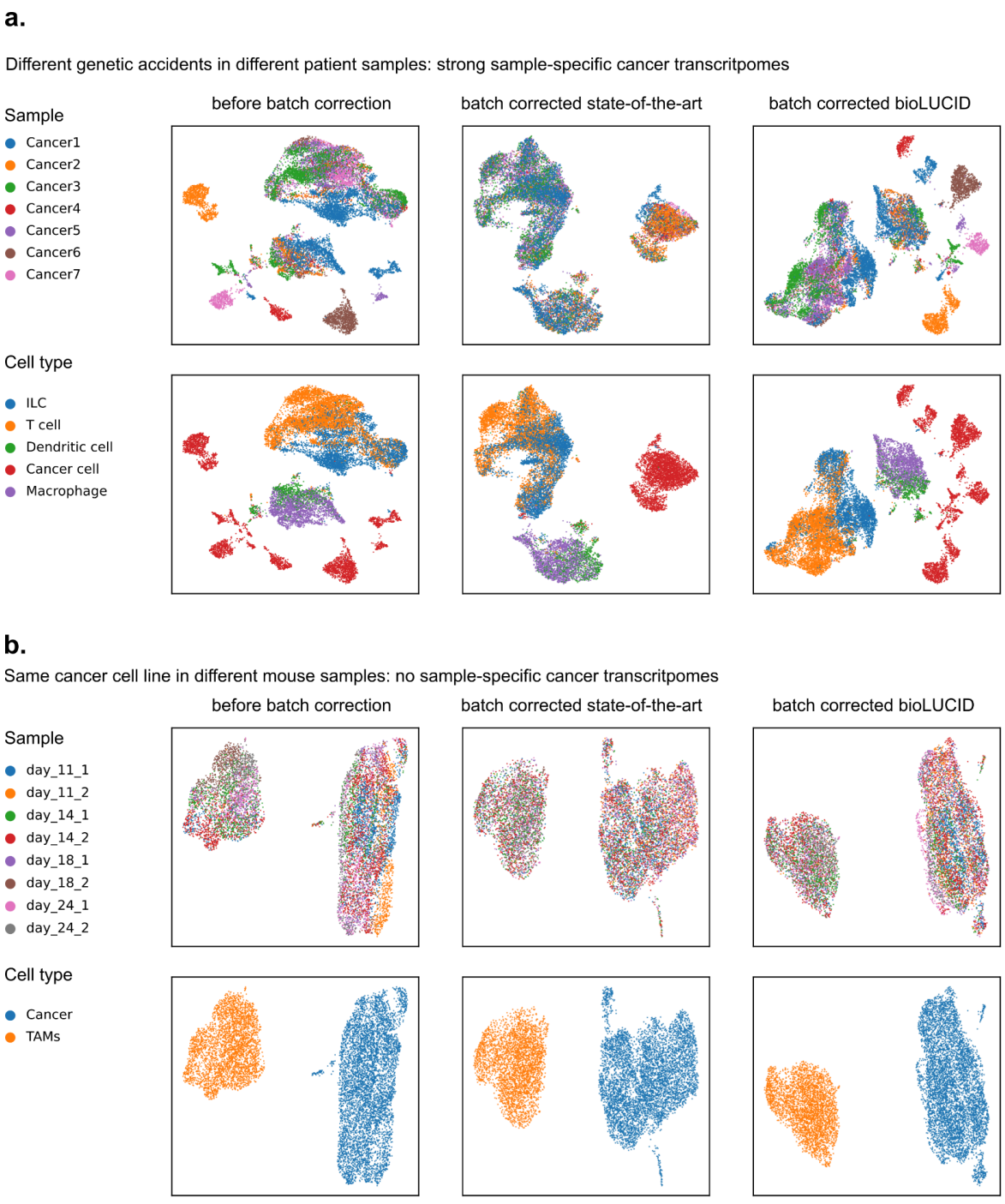
b.



Supplementary Figure 3: Known biological pathways — TNF and hedgehog signaling, interferon response, angiogenesis, complement — associate with gene regulation shared across cell types σ_s. Gene regulation shared across cell types thus likely captures a mix

of batch and biological effects. a-b. To explore whether σ_s associates with known pathways *P*, we regressed σ_s (dimension: *G* genes) on a binary matrix *H* (dimensions: *G*x*P)* indicating which genes belong to the 50 Hallmark pathways of MSigDB: entries of this matrix were set to 1 if gene *g* belongs to pathway *p* (10-300 genes per pathway) and 0 otherwise. Pathway coefficients *β* (dimension: *P* pathways) were computed using the ordinary least squares estimator, to explore which pathways most contributed to shared gene regulation (see heatmap). We further determined how well known pathways captured shared gene regulation σ_s by computing the coefficient of determination R^2 (see left-most column of heatmap). Potential collinearity between pathways (columns of H) were addressed by computing an orthonormal basis of the columns of H (Gram-Schmidt), projecting σ_s on that basis, and decomposing σ_s into projected and residual vectors. To serve as positive control, the same analysis was performed on cell-specific gene regulation ε_{st} (dimension: *G* genes). **a.** In the skin tumor samples of Ji et al.[9] , σ_s batch effect vectors (top rows) mainly associate with pathways such as interferon response and TNF signaling. This suggests that σ_s captures some sample-specific biology alongside batch effects. In sample P3 Normal, pathways *H* capture 49.6% of batch effect σ_s. This associates with low expression of proliferation-associated genes specific to this sample (MTOR, G2M, mitotic spindle, fatty acid metabolism, oxidative phosphorylation, DNA repair). Yet, in most samples, pathways *H* typically poorly explain σ_s (R^2 < 0.2 +/- 0.063), consistent with the view that σ_s represents batch effects. **b.** In the ovarian tumor samples of Xu et al.[10], σ_s batch effect vectors (top rows) mainly associate with pathways such as angiogenesis, TNF signaling and interferon response. This suggests that σ_s captures some sample-specific biology alongside batch effects. However, when regression batch and biological effects on pathways *H*, highest R^2 values are not found for batch effects σ_s (R^2 < 0.2 +/- 0.027) but for biological effects ε_{st} , consistent with the view that σ_s represent batch effects.

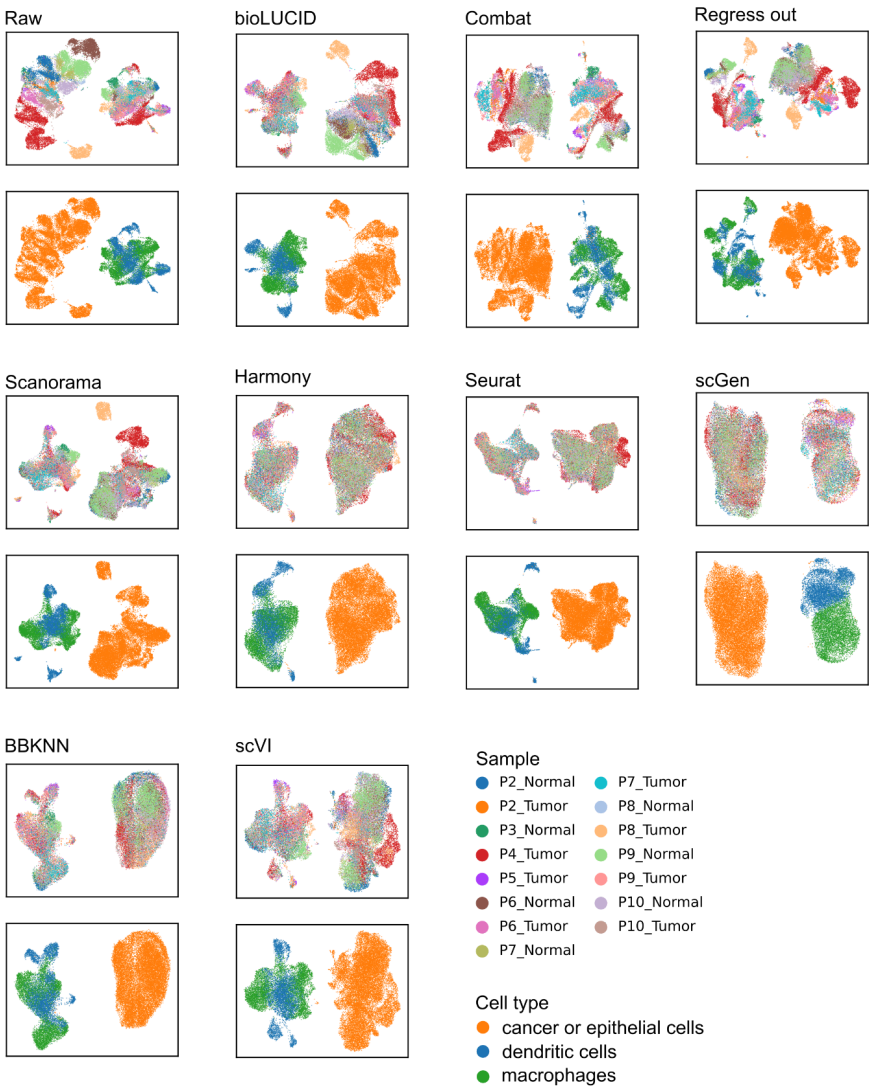
Supplementary Figure 4



Supplementary Figure 4: Using the bioLUCID hypotheses to correct for batch effects recovers known sample- and cell-type-specific biological effects. a-b. To validate the bioLUCID model, we tested if bioLUCID could accurately recover the magnitude of biological effects using samples where biological effects have known, cell-type-specific magnitudes. **a.** To

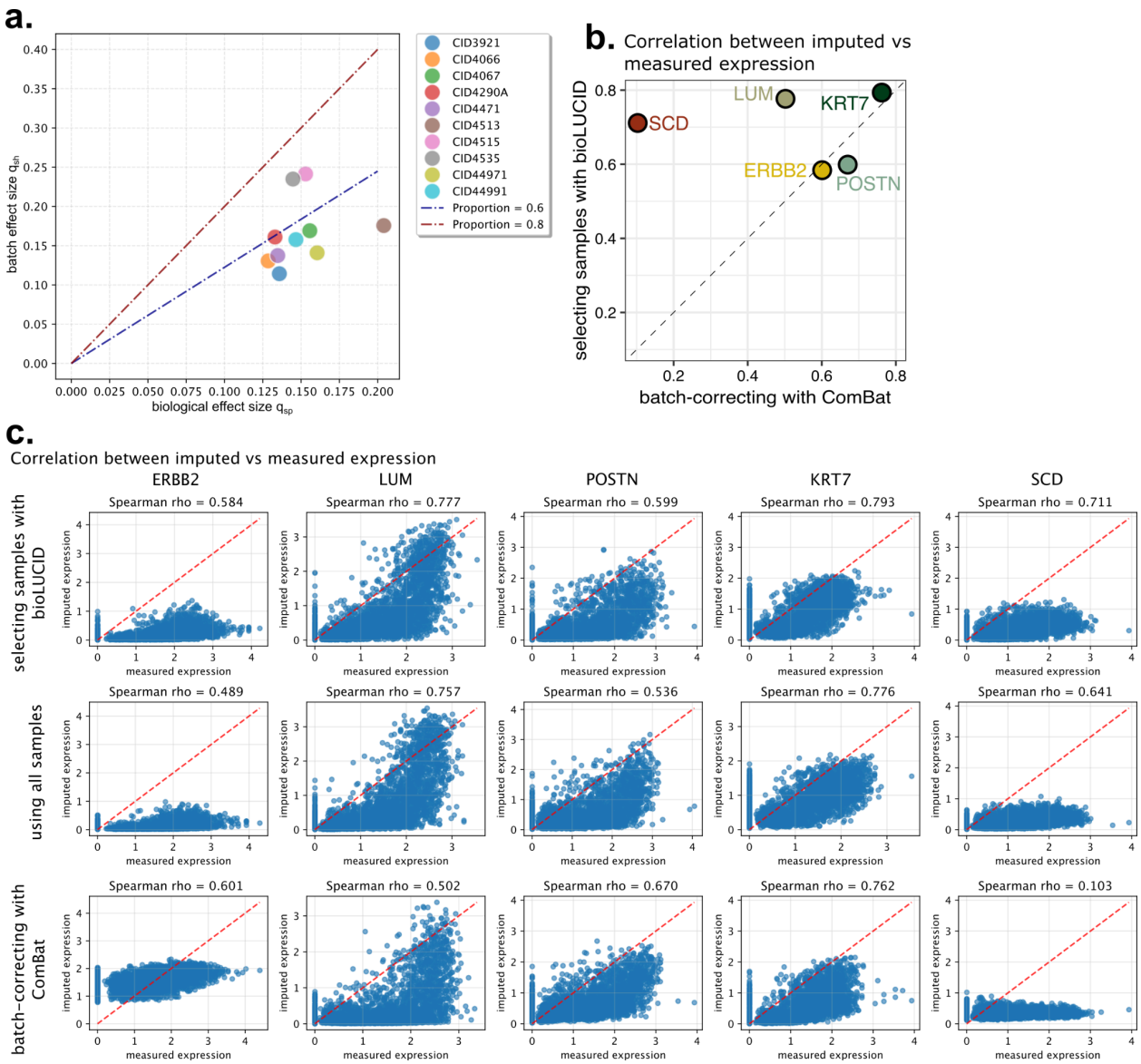
serve as a positive control, we used ovarian tumor samples from Xu et al.[10]: cancer cells are known to have strong sample-specific biology due to sample-specific genetic alterations, whereas the transcriptomes of host cells (immune, endothelial, stromal) are expected to overlap more across samples. Prior to batch correction (i.e. using uncorrected gene expression X), both cancer and host cells cluster by sample, consistent with the presence of batch effects. Performing batch effect correction using a state-of-the-art method (here Harmony) removes all sample specificity, including in cancer cells. This contradicts the known sample-specific biology of cancer cells. The sample-specific biology of cancer is preserved by bioLUCID: after subtracting the sample-specific, cell type-unspecific effect ($X - \sigma$), cancer cells cluster by sample. The transcriptome of host cells shows residual sample-specificity and integrates better than prior to batch correction. **b.** To serve as a negative control, we used tumor samples from an allograft mouse breast cancer model where all samples stem from the same cell line (Gsell et al.)[6]. We thus expect little sample-specific biology across cell types in contrast to patient-derived samples. Prior to batch correction (X), both cancer cells and macrophages show sample-specificity. After batch correction guided by the bioLUCID hypotheses ($X - \sigma$), neither cell types show strong sample-specificity, as expected in this model. Thus, bioLUCID preserves sample-specific biology while removing batch effects across datasets. Dots: cells. Axes: UMAP components.

Supplementary Figure 5



Supplementary Figure 5: A comparison with 9 batch correction methods shows that explicitly modeling biological and batch effects is necessary to integrate immune cells from multiple tumor samples while preserving the sample-specificity of cancer cells. Data: squamous cell carcinoma scRNAseq from Ji et al.. Dots: cells. Axes: UMAP components. Regression-based methods regress_out and ComBat tend to struggle to integrate immune cells from different samples. Embedding alignment based methods Harmony and Seurat appear to over-integrate cancer cells from different samples. Scanorama, an embedding alignment method which aligns datasets but without needing one to be a reference and optimizes for speed and memory fairs better at this test. Machine learning-based methods scGen and BBKNN also tend to over-integrate cancer cells from different samples. Of this family of methods, scVI performs best here. BioLUCID performs well at both challenges, achieving high sample mixing for immune cells while preserving the low sample mixing of cancer cells (see Fig. 2b).

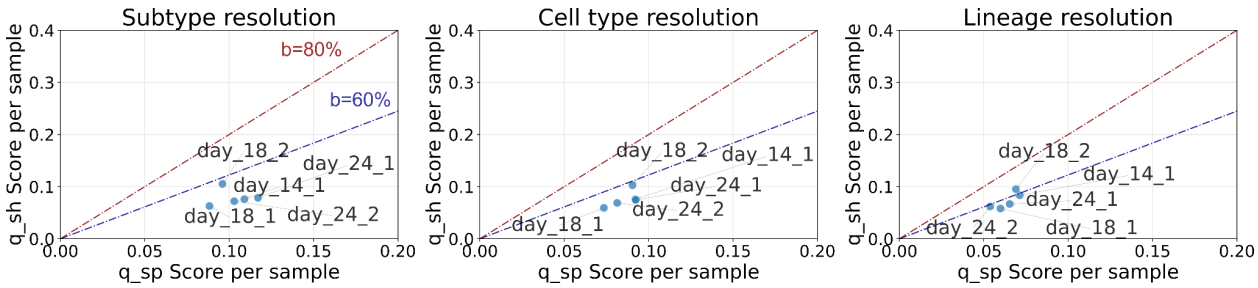
Supplementary Figure 6



Supplementary Figure 6: Excluding samples with strong batch effects imputes more accurate spatial transcriptomes. **a.** Of the 12 scRNAseq samples of Wu et al. that passed QC, bioLUCID infers that CID4515 and CID4535 have b=60-80% batch effects while all other samples have b<60% batch effects. This supports excluding these two samples to perform spatial transcriptome imputation. **b.** Batch-correcting scRNAseq samples prior to imputing spatial transcriptomes decreases the accuracy of gene expression imputation. Axes: Spearman correlation between gene expression imputed by gimVI and measured by Xenium. Dots: 5 genes held out from training. **c.** Comparing imputed expression of 5 genes held out from training to their measured expression by the Xenium assay. Imputation was performed using (i) all 12 scRNAseq samples of Wu et al. passing QC, (ii) 10 scRNAseq selected by bioLUCID to have limited batch effects, (iii) all 12 samples post batch effect correction by ComBat. ComBat

was used because it is among the few batch-correction methods that returns expression matrices suitable for downstream differential expression analysis instead of correcting cell embeddings.

Supplementary Figure 7



Supplementary Figure 7: The inferred q_{sh} and q_{sp} are robust to the resolution of cell type calling. Calling cell subtypes (TAMs, T CD8, T CD4, Tregs, cancer, …) produces similar estimates of q_{sh} and q_{sp} to coarser-grained cell types (TAMs, T cells, cancer, …). Calling cell types with even coarser resolution, at the level of lineage (myeloid, lymphoid, cancer) yields comparable estimates still, yet which begin to depart from estimates obtained from finer-grained cell types. Data: breast tumor scRNAseq from Gsell et al.[6].

1. Grün, D., Kester, L. & Van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11**, 637–640 (2014).
2. Svensson, V. *et al.* Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14**, 381–387 (2017).
3. Ding, J. *et al.* Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* **38**, 737–746 (2020).
4. Antonsson, S. E. & Melsted, P. Batch correction methods used in single-cell RNA sequencing analyses are often poorly calibrated. *Genome Res.* (2025) doi:10.1101/gr.279886.124.
5. Zhang, J. *et al.* Tahoe-100M: A Giga-Scale Single-Cell Perturbation Atlas for Context-Dependent Gene Function and Cellular Modeling. 2025.02.20.639398 Preprint at https://doi.org/10.1101/2025.02.20.639398 (2025).
6. Gsell, Louise *et al.* Multi-cellular phenotypic dynamics during the progression of breast

tumors. Preprint at http://hausserlab.org/assets/GsellEtAl2022.pdf (2022).

7. Grobecker, P., Sakoparnig, T. & Nimwegen, E. van. Identifying cell states in single-cell RNA-seq data at statistically maximal resolution. *PLOS Comput. Biol.* **20**, e1012224 (2024).
8. Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).
9. Ji, A. L. *et al.* Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. *Cell* **182**, 1661–1662 (2020).
10. Xu, J. *et al.* Single-Cell RNA Sequencing Reveals the Tissue Architecture in Human High-Grade Serous Ovarian Cancer. *Clin. Cancer Res.* **28**, 3590–3602 (2022).