A Novel Imputation-Based Computational Strategy for the Identification of Tumor-Macrophage Hybrid Cells in Single-Cell RNA-sequencing Data Reveals a Transcriptional Phenotype Consistent with Metastatic Function

Alissa Cait[1], Haider Hassan[1], Erica Scott[1] , Jessica Corrado [1], Dan Ryder [1, 2], Seng-Lai Tan[3]
[1] Bridge Informatics, Inc., 160 Alewife Brook Pkwy, Cambridge, MA 02138; [2]Corresponding author and Chief Executive Officer (CEO) of Bridge Informatics, Inc.; [3] TFC Therapeutics, 430 East 29th Street, Floor 14 New York, NY 10016

Tumor Macrophage Hybrid Cells

- Metastasis remains the primary cause of cancer mortality
- Tumor-macrophage hybrid (TMH) cells co-express markers of both epithelial tumor cells and myeloid macrophages
- TMH cells arise from rare somatic cell fusion or phenotypic plasticity
- TMH cells have been implicated **in cancer initiation, metastatic dissemination, and disease recurrence** across multiple cancer types
- TMH presence correlates with poor survival outcomes, but **no therapeutic strategy** currently targets them

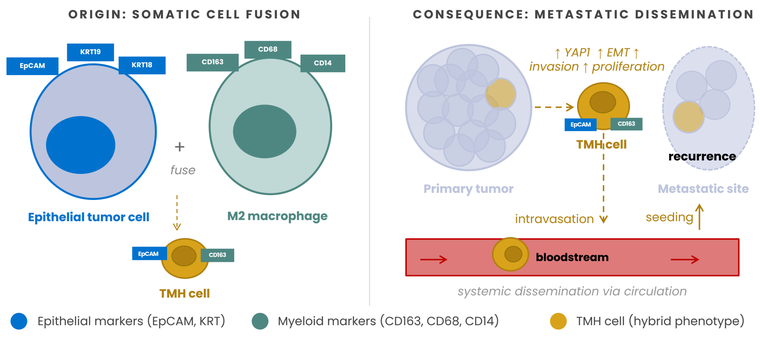


Figure 1: Tumor Macrophage Hybrid (TMH) Cells. TMH cells arise via somatic fusion between epithelial tumor cells and M2 macrophages, acquiring co-expression of epithelial and myeloid surface markers. The resulting hybrid phenotype is associated with intravasation and systemic dissemination to metastatic sites.

The Core Problem

- A major challenge in identifying TMH cells is distinguishing true hybrid biology from technical doublets
- **Single-Cell RNA-seq** is sparse (dropout heavy)
- Identification requires identifying cells with co-expression of markers from two transcriptionally distinct lineages
  - Epithelial EpCAM
  - Myeloid CD163
- Co-expression is routinely obscured by dropout events
- TMH cells are **missed or misclassified** by standard pipelines

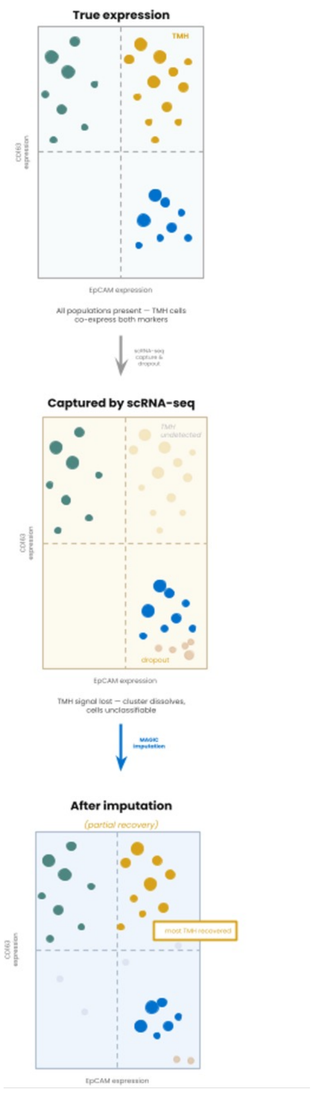


Figure 2: Conceptual illustration of scRNA-seq dropout and signal recovery. TMH cells expressing both EpCAM and CD163 are present in true biology (panel 1) but lost to dropout in raw sequencing data (panel 2). Imputation is predicted to recover the majority of the co-expression signal, enabling TMH cell classification (panel 3).

The Strategy

5 independent publicly available colorectal cancer (CRC) scRNA-seq cohorts were processed using **Seurat.** A novel imputation-based pipeline was developed to recover sparse co-expression signals. Parameters were selected empirically and reproduced across the independent cohorts

1 **MAGIC Imputation**
Diffusion-based imputation (Rmagic V2; npca = 40, batched X10) recovers co-expression signal across the cell-to-cell similarity graph

2 **Multi-criterion Classification**
Applied to imputed expression matrix; three simultaneous criteria required:

TMH Classification Criteria:
- Myeloid CD68 ≥ median + ≥ 3 of {CD163, CD14, MRC1, APOC1, SPP1, CD209, TREM2}
- Epithelial ≥ 2 of [EpCAM, KRT19, KRT18] above median
- Doublet score ≤ 0.3 (artifact exclusion

3 **DBSCAN Spatial Clustering**
Candidates projected onto UMAP only spatially coherent clusters retained (eps = 0.03, minPts =5)

4 **Pathway Analysis**
Non-imputed normalized counts used; hybrid vs non-hybrid tumor cells compared across curated gene sets (YAP1, EMT, migration/invasion, proliferation)

Key Papers

Gast et al. (2018) Cell fusion potentiates tumor heterogeneity and reveals circulating hybrid cells that correlate with stage and survival. *Science Advances.*
van Dijk et al. (2018) Recovering gene interactions from single-cell data using data diffusion. *Cell.*
Hao et al. (2021) Integrated analysis of multimodal single-cell data. *Cell*
Ester et al. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*

Let's Connect

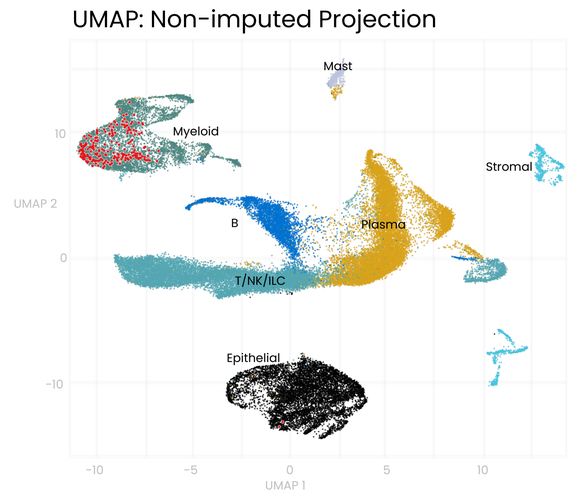
www.bridgeinformatics.com
/company/bridge-informatics
@Bridge_Info_USA
dan.ryder@bridgeinformatics.com

Acknowledgments

We want to thank TFC Therapeutics for collaborating with us on this project and poster. In particular, we thank Seng-Lai (Thomas) Tan, TFC's Chief Scientific Officer, for his collaboration.

The Signal

Without imputation, the hybrid population is largely invisible to standard threshold-based approaches. Imputation is a necessary preprocessing step for reliable TMH identification.



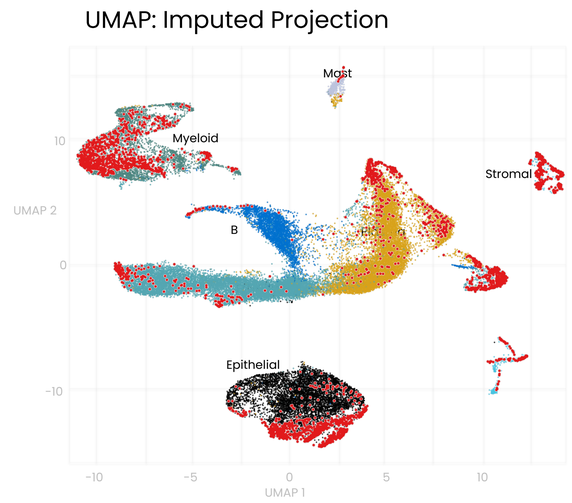


Figure 3: Imputation-based identification of TMH cells in colorectal cancer scRNA-seq data. Single-cell RNA-seq data from 5 independent colorectal cancer (CRC) cohorts were processed using the described method to recover co-expression signal lost to sequencing drop-out. UMAP projections show the cell populations and highlight cells meeting the TMH classification criteria. A representative dataset is shown.

Pathway analyses were performed using non-imputed normalized counts to reduce imputation-driven artifacts, and revealed a striking, coherent function profile in confirmed TMH cells.

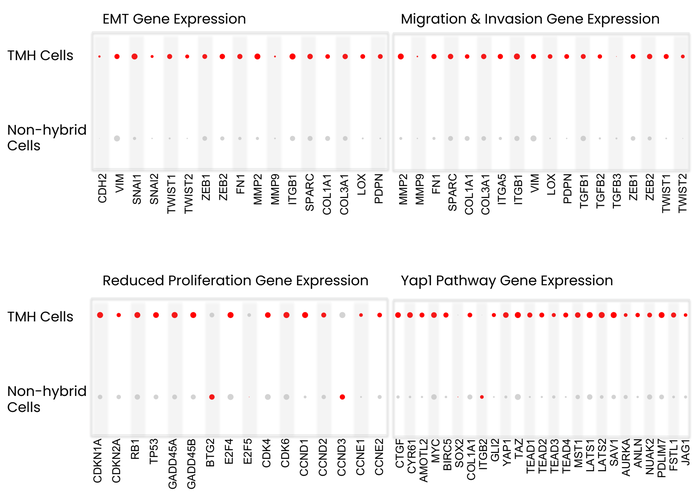


Figure 4. TMH cells display a transcriptional phenotype consistent with pro-metastatic function. Dot plots comparing normalized gene expression (non-imputed values) between TMH cells and non-hybrid tumor cells across four curated pathway gene sets. Dot size represents the percentage of cells expressing each gene; dot color represents scaled average expression. A representative dataset is shown.

Conclusions

- Gene expression imputation is **essential** for detecting rare hybrid populations in scRNA-seq data
- MAGIC-based imputation + multi-criterion classification + DBSCAN enables reproducible TMH identification across independent patient cohorts