# Comprehensively Integrated Human Gastric Single-Cell Atlas reveals Lineage States, Tumor Microenvironment, and Subtype-Specific Expression Programs of Gastric Cancer

Peng Yu Zhong

## Thesis statement

Single-cell technologies have revolutionized our comprehension of gastric cancer. However, individual studies to date have typically captured only a restricted number of donors, leading to discrepancies in cell type definitions. Additionally, a comprehensively integrated analysis of these single-cell RNA sequencing datasets is not yet available, as there is no standardized process for harmonizing results across studies. This approach enables capturing the inherent variability within the population, discovering rare and previously unidentified cell types, investigating intratumoral heterogeneity (ITH) encompassing the tumor immune microenvironment and malignant cells, and establishing connections between these factors and their clinical relevance for patients. Bringing together disparate single-cell cancer datasets in a harmonized manner has the potential to provide novel insights into gastric cancer biology at an unprecedented level of detail. Besides, The use of patient-derived organoids (PDOs) or patient-derived xenografts (PDXs) could help validate insights generated from integrated single-cell analysis and advance our understanding of what drives heterogeneity in gastric cancer at a functional level.

## Background

Rapid technological advancements in the past decade have facilitated the exponential growth of single-cell datasets in terms of both size and quantity. Consequently, various resources have emerged to accommodate this vast generation of single-cell data. Prominent examples include the Human Cell Atlas [1], PanglaoDB [2], scRNAseqDB [3], Single Cell Expression Atlas [4], SC2diseases [5], and CELLxGENE [6], which serve as repositories for these extensive datasets. Additionally, specific databases such as SpatialOmics [7], Aquila [8], and STellaris [9] specialize in storing spatial transcriptomics data. For cancer-specific donors, databases like CancerSEA [10], TISCH [11], and CancerSCEM [12] are available. Moreover, DISCO [13] provides an integrated cell atlas encompassing harmonized metadata and employing uniform methods for quality control and expression quantification. Despite these advancements, a comprehensive scRNA atlas specifically for gastric cancer remains elusive. By integrating datasets generated by the broader research community, it becomes feasible to capture the diverse cellular landscape across individuals in gastric cancer research.

Integrated single-cell atlases provide novel insights that cannot be obtained from individual studies. Recent reference atlases have led to the discovery of previously unknown cell types [14–17]. Large-scale analyses can also offer new perspectives on the origins of intra-tumoral heterogeneity (ITH) in malignant cells [18] and enable exploration of immune microenvironment subtypes [19]. Leveraging the number and diversity of individuals sampled, an integrated atlas can help identify gene modules or cell subtypes associated with demographic covariates such as age, sex, body mass index [16], as well as variations in spatial location [20,21], and other clinical relevance, including implications for patient survival and treatment response [22,23]. Additionally, the discovery of shared cell states across multiple diseases may enhance our understanding of disease progression at a more detailed level [16].

Studies have proven that patient-derived models such as PDOs or PDXs could be used as an in vitro experimental model to investigate intratumoral heterogeneity in gastric cancer [24]. These models provide important opportunities to explore the molecular functions underlying our findings [22], and allow assessment of therapeutic effects by targeting specific tumor-associated cell subtypes [25,26]. Their ability to model intrinsic and extrinsic tumor complexity supports further investigation using this approach.

The accompanying metadata for harmonization across datasets are also retrieved and manually curated. Achieving metadata consistency facilitates efficient data retrieval and downstream data integration.

## Data collection, pre-processing and subsequent analysis

The data will be collected from GEO (<https://www.ncbi.nlm.nih.gov/geo/>) [27], ArrayExpress (<https://www.ebi.ac.uk/arrayexpress/>) [28], Single Cell Expression Atlas – EBI (<https://www.ebi.ac.uk/gxa/sc/home>) [4], GSA (<https://ngdc.cncb.ac.cn/gsa/>) [29], and other public resources.

The collected datasets will be divided into two parts: 1) with raw reads (fastq, SRA, or bam file), will perform re-alignment to a single reference genome, this data will be used to construct an integrated gastric cancer cell atlas; 2) with processed matrix with or without annotations, will be utilized to be the validation cohort or referenced in the annotation steps.

For the raw reads fastq from 10x Genomics platform, Cell Ranger v7.1 is used (<https://support.10xgenomics.com/single-cell-gene-expression/software/overview/welcome>) used, from BD Rhapsody, BD Genomics Rhapsody Analysis pipeline cwl v2.0 is used (<https://bitbucket.org/CRSwDev/cwl/src/master/v2.0/>), for other sequencing platforms, the process procedure follows the platform corresponded software or pipeline, for data acquired with other technologies without a consensus practice, Salmon Alevin v1.10.2 [30] is used, and all the raw sequencing data will be realigned to GRCh38 using Ensembl84.

After expression matrix construction, a series of analyses was utilized to conduct the cell quality control: 1) mRNA counts > 500, gene counts > 200, mitochondrial mRNA counts percentage < 15%; 2) gene counts and mRNA counts both in their 5 MAD (median absolute deviations) ranges; 3) remove doublets by Scrublet [31] with default parameters. 4) We further filtered out samples with less than 500 cells remaining. Though excluded from the data integration, these samples are marked as low quality in clinical annotation table.

State-of-the-art techniques will be employed to integrate and annotate the data, as well as to develop analytical methods for characterizing lineage relationships, cellular heterogeneity, and rare cell populations. Subpopulations will be carefully annotated, and molecular profiles associated with clinical variables will be constructed. To validate the key findings, functional experiments will be conducted using patient-derived organoid models. Further details regarding the utilization of research approaches will be elaborated by referencing high-level articles.

## Expected Outcomes

The Integrated Human Gastric Single-Cell Atlas projects include several expected outcomes:

1) Providing an integrated gastric cancer atlas for reference work, which involves identifying rare and previously unidentified cell types;

2) Investigating intratumoral heterogeneity (ITH) that encompasses the tumor immune microenvironment and malignant cells, which aims to establish connections between these components and their clinical relevance for patients. For example, identifying cell subtypes associated with patient survival or treatment response;

3) Proposing novel gastric cancer subtypes at the single-cell resolution to guide clinical treatment decisions;

4) Exploring shared cell states across multiple diseases to enhance our understanding of disease progression.

Furthermore, patient-derived organoids (PDOs) or patient-derived xenografts (PDXs) models can be utilized to validate insights generated from integrated single-cell analysis and enhance our understanding of the factors driving heterogeneity in gastric cancer at a functional level. For instance, specific cell types associated with tumor progression can be targeted in PDOs to assess the clinical therapeutic potential.

## Reference

[1] Lindeboom RGH, Regev A, Teichmann SA. Towards a Human Cell Atlas: Taking Notes from the Past. Trends Genet 2021;37:625–30. https://doi.org/10.1016/j.tig.2021.03.007.

[2] Franzén O, Gan L-M, Björkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database (Oxford) 2019;2019:baz046. https://doi.org/10.1093/database/baz046.

[3] Cao Y, Zhu J, Jia P, Zhao Z. scRNASeqDB: A Database for RNA-Seq Based Gene Expression Profiles in Human Single Cells. Genes (Basel) 2017;8:368. https://doi.org/10.3390/genes8120368.

[4] Papatheodorou I, Moreno P, Manning J, Fuentes AM-P, George N, Fexova S, et al. Expression Atlas update: from tissues to single cells. Nucleic Acids Res 2020;48:D77–83. https://doi.org/10.1093/nar/gkz947.

[5] Zhao T, Lyu S, Lu G, Juan L, Zeng X, Wei Z, et al. SC2disease: a manually curated database of single-cell transcriptome for human diseases. Nucleic Acids Res 2021;49:D1413–9. https://doi.org/10.1093/nar/gkaa838.

[6] Chan Zuckerberg Initiative. CZ CELLxGENE Discover. CZ CELLxGENE Discover n.d. https://cellxgene.cziscience.com/ (accessed October 31, 2023).

[7] Yuan Z, Pan W, Zhao X, Zhao F, Xu Z, Li X, et al. SODB facilitates comprehensive exploration of spatial omics data. Nat Methods 2023;20:387–99. https://doi.org/10.1038/s41592-023-01773-7.

[8] Zheng Y, Chen Y, Ding X, Wong KH, Cheung E. Aquila: a spatial omics database and analysis platform. Nucleic Acids Research 2023;51:D827–34. https://doi.org/10.1093/nar/gkac874.

[9] Li X, Xiao C, Qi J, Xue W, Xu X, Mu Z, et al. STellaris: a web server for accurate spatial mapping of single cells based on spatial transcriptomics data. Nucleic Acids Res 2023;51:W560–8. https://doi.org/10.1093/nar/gkad419.

[10] Yuan H, Yan M, Zhang G, Liu W, Deng C, Liao G, et al. CancerSEA: a cancer single-cell state atlas. Nucleic Acids Res 2019;47:D900–8. https://doi.org/10.1093/nar/gky939.

[11] Sun D, Wang J, Han Y, Dong X, Ge J, Zheng R, et al. TISCH: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment. Nucleic Acids Res 2021;49:D1420–30. https://doi.org/10.1093/nar/gkaa1020.

[12] Zeng J, Zhang Y, Shang Y, Mai J, Shi S, Lu M, et al. CancerSCEM: a database of single-cell expression map across various human cancers. Nucleic Acids Res 2021;50:D1147–55. https://doi.org/10.1093/nar/gkab905.

[13] Li M, Zhang X, Ang KS, Ling J, Sethi R, Lee NYS, et al. DISCO: a database of Deeply Integrated human Single-Cell Omics data. Nucleic Acids Research 2022;50:D596–602. https://doi.org/10.1093/nar/gkab1020.

[14] Schupp JC, Adams TS, Cosme C, Raredon MSB, Yuan Y, Omote N, et al. Integrated Single-Cell Atlas of Endothelial Cells of the Human Lung. Circulation 2021;144:286–302. https://doi.org/10.1161/CIRCULATIONAHA.120.052318.

[15] Steuernagel L, Lam BYH, Klemm P, Dowsett GKC, Bauder CA, Tadross JA, et al. HypoMap—a unified single-cell gene expression atlas of the murine hypothalamus. Nat Metab 2022;4:1402–19. https://doi.org/10.1038/s42255-022-00657-y.

[16] Sikkema L, Ramírez-Suástegui C, Strobl DC, Gillett TE, Zappia L, Madissoon E, et al. An integrated cell atlas of the lung in health and disease. Nat Med 2023;29:1563–77. https://doi.org/10.1038/s41591-023-02327-2.

[17] Nie H, Lin P, Zhang Y, Wan Y, Li J, Yin C, et al. Single-cell meta-analysis of inflammatory bowel disease with scIBD. Nat Comput Sci 2023;3:522–31. https://doi.org/10.1038/s43588-023-00464-9.

[18] Wang R, Dang M, Harada K, Han G, Wang F, Pool Pizzi M, et al. Single-cell dissection of intratumoral heterogeneity and lineage diversity in metastatic gastric adenocarcinoma. Nat Med 2021;27:141–51. https://doi.org/10.1038/s41591-020-1125-8.

[19] Xue R, Zhang Q, Cao Q, Kong R, Xiang X, Liu H, et al. Liver tumour immune microenvironment subtypes and neutrophil heterogeneity. Nature 2022:1–7. https://doi.org/10.1038/s41586-022-05400-x.

[20] Sun C, Wang A, Zhou Y, Chen P, Wang X, Huang J, et al. Spatially resolved multi-omics highlights cell-specific metabolic remodeling and interactions in gastric cancer. Nat Commun 2023;14:2692. https://doi.org/10.1038/s41467-023-38360-5.

[21] Wu L, Yan J, Bai Y, Chen F, Zou X, Xu J, et al. An invasive zone in human liver cancer identified by Stereo-seq promotes hepatocyte–tumor cell crosstalk, local immunosuppression and tumor progression. Cell Res 2023:1–19. https://doi.org/10.1038/s41422-023-00831-1.

[22] Wang R, Song S, Qin J, Yoshimura K, Peng F, Chu Y, et al. Evolution of immune and stromal cell states and ecotypes during gastric adenocarcinoma progression. Cancer Cell 2023;41:1407-1426.e9. https://doi.org/10.1016/j.ccell.2023.06.005.

[23] Kang B, Camps J, Fan B, Jiang H, Ibrahim MM, Hu X, et al. Parallel single-cell and bulk transcriptome analyses reveal key features of the gastric tumor microenvironment. Genome Biology 2022;23:265. https://doi.org/10.1186/s13059-022-02828-2.

[24] Kumar V, Ramnarayanan K, Sundar R, Padmanabhan N, Srivastava S, Koiwa M, et al. Single-Cell Atlas of Lineage States, Tumor Microenvironment, and Subtype-Specific Expression Programs in Gastric Cancer. Cancer Discov 2022;12:670–91. https://doi.org/10.1158/2159-8290.CD-21-0683.

[25] Binnewies M, Pollack JL, Rudolph J, Dash S, Abushawish M, Lee T, et al. Targeting TREM2 on tumor-associated macrophages enhances immunotherapy. Cell Reports 2021;37:109844. https://doi.org/10.1016/j.celrep.2021.109844.

[26] Luo Q, Dong Z, Xie W, Fu X, Lin L, Zeng Q, et al. Apatinib remodels the immunosuppressive tumor ecosystem of gastric cancer enhancing anti-PD-1 immunotherapy. Cell Reports 2023;42:112437. https://doi.org/10.1016/j.celrep.2023.112437.

[27] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res 2013;41:D991-995. https://doi.org/10.1093/nar/gks1193.

[28] Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, et al. ArrayExpress update - from bulk to single-cell expression data. Nucleic Acids Res 2019;47:D711–5. https://doi.org/10.1093/nar/gky964.

[29] Chen T, Chen X, Zhang S, Zhu J, Tang B, Wang A, et al. The Genome Sequence Archive Family: Toward Explosive Data Growth and Diverse Data Types. Genomics Proteomics Bioinformatics 2021;19:578–83. https://doi.org/10.1016/j.gpb.2021.08.001.

[30] Srivastava A, Malik L, Smith T, Sudbery I, Patro R. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. Genome Biology 2019;20:65. https://doi.org/10.1186/s13059-019-1670-y.

[31] Wolock SL, Lopez R, Klein AM. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Systems 2019;8:281-291.e9. https://doi.org/10.1016/j.cels.2018.11.005.