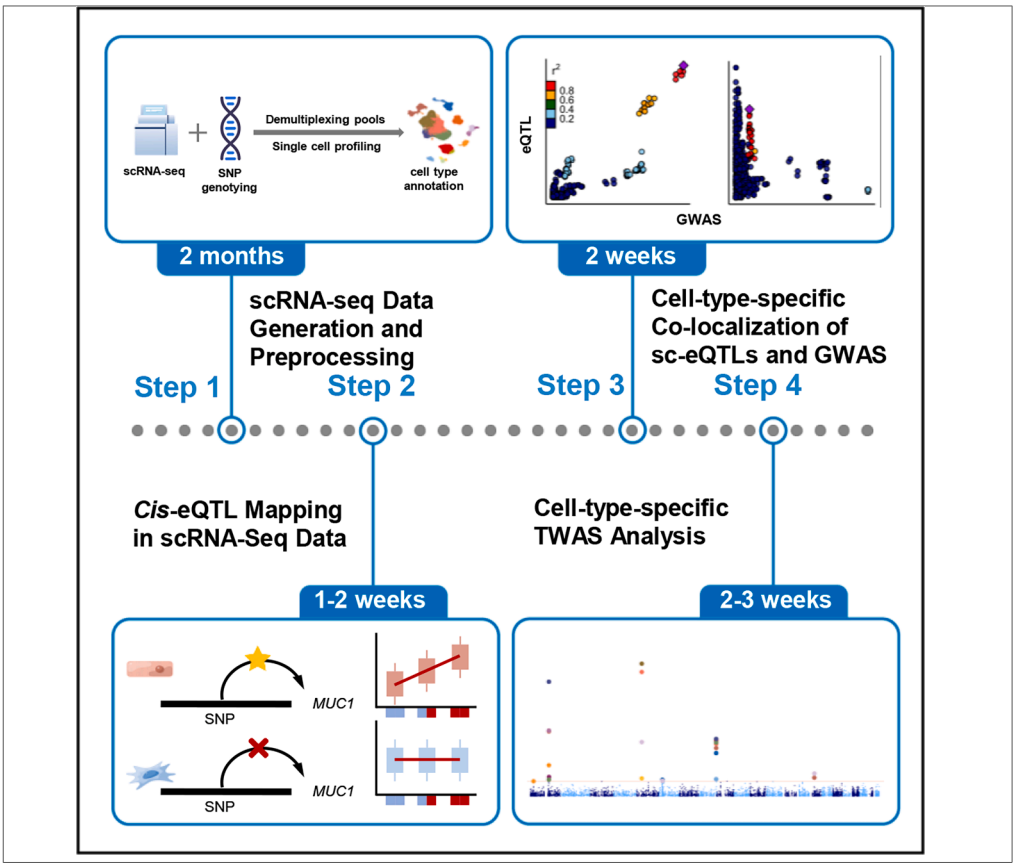
STAR Protocols

CellPress
OPEN ACCESS

Protocol

Protocol for identifying cell-type-specific genes associated with disease risk using single-cell eQTL data



Here, we present a protocol for mapping cell-type-specific *cis*-acting expression quantitative trait loci (*cis*-eQTLs) in a target tissue and integrating them with genome-wide association study (GWAS) data. We describe pooled-sample multiplexed sequencing, genotype-based cell demultiplexing, *cis*-eQTL detection in specific cell types, and integration of single-cell eQTLs with GWAS. This protocol facilitates the identification of candidate risk genes and mechanisms underlying disease susceptibility.

Publisher's note: Undertaking any experimental protocol requires adherence to local institutional guidelines for laboratory safety and ethics.

Xin Fang, Lijun Bian, Caiwang Yan, Guangfu Jin

guangfujin@njmu.edu.cn

Highlights

Multiplexed sequencing of pooled samples and genotype-based cell demultiplexing

Processing scRNA-seq data, identifying cell types, and mapping *cis*-eQTLs

Integrating cell-type-specific eQTL results with GWAS data

Linking regulatory variants to cell-type-specific regions using scATAC-seq data

Fang et al., STAR Protocols 7, 104545
June 19, 2026
https://doi.org/10.1016/j.xpro.2026.104545

STAR Protocols

CellPress
OPEN ACCESS

Protocol

Protocol for identifying cell-type-specific genes associated with disease risk using single-cell eQTL data

Xin Fang,[1,2,3] Lijun Bian,[1,2,3] Caiwang Yan,[1] and Guangfu Jin[1,2,4,5,*]

[1]Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing 211166, China

[2]State Key Laboratory Cultivation Base of Biomarkers for Cancer Precision Prevention and Treatment, Collaborative Innovation Center for Cancer Personalized Medicine and China International Cooperation Center for Environment and Human Health, Nanjing Medical University, Nanjing 211166, China

[3]These authors contributed equally

[4]Technical contact

[5]Lead contact

*Correspondence: guangfujin@njmu.edu.cn
https://doi.org/10.1016/j.xpro.2026.104545

SUMMARY

Here, we present a protocol for mapping cell-type-specific *cis*-acting expression quantitative trait loci (*cis*-eQTLs) in a target tissue and integrating them with genome-wide association study (GWAS) data. We describe pooled-sample multiplexed sequencing, genotype-based cell demultiplexing, *cis*-eQTL detection in specific cell types, and integration of single-cell eQTLs with GWAS. This protocol facilitates the identification of candidate risk genes and mechanisms underlying disease susceptibility.

For complete details on the use and execution of this protocol, please refer to Bian et al.[1]

BEFORE YOU BEGIN

Innovation

Traditional eQTL analyses have primarily been performed on various bulk tissue samples, which limits the resolution of regulatory effects that are unique to specific cell types. This protocol introduces a novel pooled multiplexing strategy to construct a comprehensive single-cell atlas, enabling detailed eQTL analyses at the cell-type level. To relate cell-type-specific eQTLs to disease-associated risk loci, we integrate GWAS summary statistics and perform co-localization analysis. Additionally, a cell-type-specific transcriptome-wide association study (TWAS) is conducted to prioritize candidate effector genes at susceptibility loci.

The key innovation of this protocol lies in its ability to link genetic variations to cell-type-specific regulatory networks, which are often masked in bulk tissue analyses. By operating at the single-cell level, this method delineates cell-type-specific patterns of genetic regulation of gene expression. Moreover, the integration of eQTLs with TWAS offers a powerful approach to discover novel susceptibility genes, which may be missed in non-cell-type-specific eQTL studies. In addition, this protocol provides an easy downloadable example dataset, enabling users to quickly test and run the complete protocol.

In summary, this protocol can provide insights into cell-type-specific eQTLs and clarify regulatory mechanisms underlying cell-type-specific risk genes. Here, the protocol describes the specific steps for mapping single cell eQTLs of gastric tissue. It could also be applied to other tissues, such as liver and kidney.

Table 1. Source of GWAS summary data

Source	URL
GWAS catalog	https://www.ebi.ac.uk/gwas/
KoGES	https://koges.leelabsg.org/pheno/KoGES_GCA

Integration of GWAS datasets

Timing: 1 week

1. Search and download several published GWAS public databases.
2. Integrate and compile datasets with a focus on gastric cancer-related data (Table 1).[2]

Set up the computational environment

Timing: 3 days

This section describes how to set up a computational environment with the necessary software tools for data analysis.

3. Install all required software and key tools on a suitable computational platform.
4. Ensure R and Python environments are set up with needed packages.

Note: High-memory servers (RAM > 64G) or a computing cluster are recommended for large single-cell datasets. It is recommended to run the relevant programs on a Linux system.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Gastric mucosal tissues	Yangzhou screening program[3]	This study
Adjacent normal gastric mucosal tissue from GC patients	The First Affiliated Hospital of Nanjing Medical University	This study
Blood samples	Yangzhou screening program[3]	This study
Critical commercial assays		
sCelLive Tissue Preservation Solution	Singleron	Cat# 1010012
sCelLive Tissue Dissociation Solution	Singleron	Cat# 1020012
GEXSCOPE red blood cell lysis buffer	Singleron	Cat# 1200050002
GEXSCOPE Single Cell RNA Library Kit Tissue for HD V2	Singleron	Cat# 1110032
Qiagen miRNeasy Kit	Qiagen	Cat# 217084
RNA Nano 6000 Assay Kit	Agilent	Cat# 5067-1511
Nuclei Isolation for Single Cell ATAC + Gene Expression Sequencing	10× Genomics	CG000365
Diluted Nuclei Buffer	10× Genomics	PN-2000153
Chromium Next GEM Chip J Single Cell Kit	10× Genomics	PN-1000234
Chromium Next GEM Single Cell Multiome ATAC + Gene Expression Reagent Bundle	10× Genomics	PN-1000283
Deposited data		
The results of eQTLs in different gastric cell types	This paper	http://ccra.njmu.edu.cn/scgate/
The summary data reported in this study	This paper	OMIX: OMIX008014 at https://ngdc.cncb.ac.cn/omix
The raw sequencing data in this study	This paper	GSA: HRA009277 at https://ngdc.cncb.ac.cn/gsa-human

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Human genome reference (GRCh37)	Ensembl	ftp://ftp.ensembl.org/pub/release-75/fasta/homo_sapiens/dna/Homo_sapiens.GRCh37.75.dna.primary_assembly.fa.gz
1000 Genomes Project (Phase 3)	The 1000 Genomes Project	https://www.internationalgenome.org/
GENCODE human gene reference (v19)	GENCODE project	https://www.gencodegenes.org/human/release_19.html
ScATAC-seq data sets of gastric tissues	Dong et al., 2023[4]	ftp://download.big.ac.cn/gsa-human/HRA002917
Software and algorithms		
Codes and scripts	This paper	https://github.com/NMUGCEPI/sc_eQTL
PLINK (v1.9)[a]	Purcell et al., 2007[5]	https://www.cog-genomics.org/plink/
SHAPEIT (v2.12)[a]	Delaneau et al., 2011[6]	https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html
IMPUTE2 (v2.3.1)[a]	Howie et al., 2009[7]	https://mathgen.stats.ox.ac.uk/impute/impute_v2.html
CeleScope (v1.1.8)[a]	Singleron	https://github.com/singleron-RD/CeleScope
Cutadapt (v1.17)[a]	Martin, 2011[8]	https://github.com/marcelm/cutadapt
STAR (v2.6.1a)[a]	Dobin et al., 2013[9]	https://github.com/alexdobin/STAR
featureCounts (v2.0.1)[a]	Liao et al., 2014[10]	https://rnnh.github.io/bioinfo-notebook/docs/featureCounts.html
Demuxlet[a]	Kang et al., 2018[11]	https://github.com/statgen/demuxlet
Seurat (v4.2.0)[b]	Satija et al., 2015[12]	https://satijalab.org/seurat/
FastQTL[a]	Ongen et al., 2016[13]	https://github.com/francois-a/fastqtl
ClusterProfiler (v4.2.2)[b]	Yu et al., 2012[14]	https://github.com/YuLab-SMU/clusterProfiler
ComplexHeatmap (v2.10.0)[b]	Gu et al., 2016[15]	https://github.com/jokergoo/ComplexHeatmap
COLOC (v5.2.2)[b]	Giambartolomei et al., 2014[16]	https://github.com/chr1swallace/coloc
FUSION[a]	Gusev et al., 2016[17]	http://gusevlab.org/projects/fusion/
GCTA-GREML[a]	Yang et al., 2010[18]	https://yanglab.westlake.edu.cn/software/gcta/#GREML
Cell Ranger ARC (v2.0.2)[a]	10× Genomics	https://www.10xgenomics.com/support/single-cell-multiome-atac-plus-gene-expression
Signac (v1.9.0)[b]	Stuart et al., 2021[19]	https://stuartlab.org/signac/
MACS2[a]	Zhang et al., 2017[20]	https://github.com/macs3-project/MACS

[a]command line in Linux.
[b]R packages.

STEP-BY-STEP METHOD DETAILS

Gastric tissue sample collection

Timing: 2 days

This section describes how to collect the gastric tissue sample used for single-cell RNA sequencing.

1. Prepare phosphate-buffered saline (PBS) and sCelLive Tissue Preservation Solution (Singleron).
2. Collect approximately 100–200 mg of gastric mucosal tissues from each individual for the scRNA-seq analysis.
3. Collect matched tissues from a region adjacent to the scRNA-seq sampling site for subsequent histopathological assessment.

⚠ CRITICAL: Immediately after rinsing fresh tissues in phosphate-buffered saline (PBS), transfer them into pre-chilled sCelLive Tissue Preservation Solution (Singleron) and keep them on ice.

4. Also collect peripheral blood from each individual for genotyping.

Note: The study should be approved by the Ethical Review Committee and written informed consent should be obtained from all participants prior to their inclusion in the study.

Genotyping and imputation of genotyping data

Timing: 1 month

This section describes the process of genotyping and imputing genotyping data to prepare for subsequent data analysis

5. Genotype DNA using a suitable SNP genotyping array, such as the Infinium Asian Screening Array (ASA) or Global Screening Array (GSA), in accordance with the manufacturer's instructions.

Note: This array captures common variants across the genome specific to Asian populations.

6. Perform standard genotype quality control before downstream analyses.

⚠ CRITICAL: Exclude genotyped variants with a call rate of < 95%, Hardy-Weinberg Equilibrium (HWE) $p < 0.001$ or a minor allele frequency (MAF) < 5%. Remove samples with overall call rates of < 95%, extreme values in population stratification analysis and heterozygosity analysis (>6 s.d. from the mean), or evidence of duplication or cryptic relatedness (PI_HAT > 0.25).[2]

7. Phase genotypes that pass quality control for each chromosome with SHAPEIT (v2.12).
8. Impute variants with IMPUTE2 (v2.3.1) for each 5-Mb interval by using the 1000 Genomes Project Phase III haplotype panel as the reference.
9. Perform a dedicated imputation of the HLA region (Chr6: 28Mb-34Mb) with SNP2HLA and a Chinese reference panel to ensure accurate imputation for the highly polymorphic HLA region.

⚠ CRITICAL: Exclude poorly imputed variants (INFO < 0.5) and low-frequency variants (MAF < 0.05) from further downstream analysis.[2]

Single-cell sample preparation and pooling

Timing: 1–2 days

This section describes how to prepare and pool single-cell samples from gastric mucosal tissues. The preparation includes tissue dissociation, removal of red blood cells, cell counting, and pooling from multiple donors to ensure even representation across samples.

10. Rinse the specimens three times with Hanks Balanced Salt Solution (HBSS). Transfer the specimens into 3 mL sCelLive Tissue Dissociation Solution (Singleron) and digest at 37 °C for 15 min using Singleron PythoN Tissue Dissociation System.
11. Add the GEXSCOPE red blood cell lysis buffer (RCLB, Singleron) to the cell suspension at a 1:2 (cell:RCLB) volume ratio. Mix gently by inversion and incubate at room temperature for 5–8 min until the suspension visibly clears.
12. Centrifuge the mixture at 300 × *g* at 4°C for 5 min and then suspend them softly with PBS.
13. Determine cell concentration and viability using TC20 automated cell counter (Bio-Rad) according to the manufacturer's instructions.
14. Profile gastric mucosal tissues in multiple pools with each pool containing equal numbers of cells from 8–9 donors (Figure 1).

Single-cell RNA sequencing

Timing: 1–2 weeks

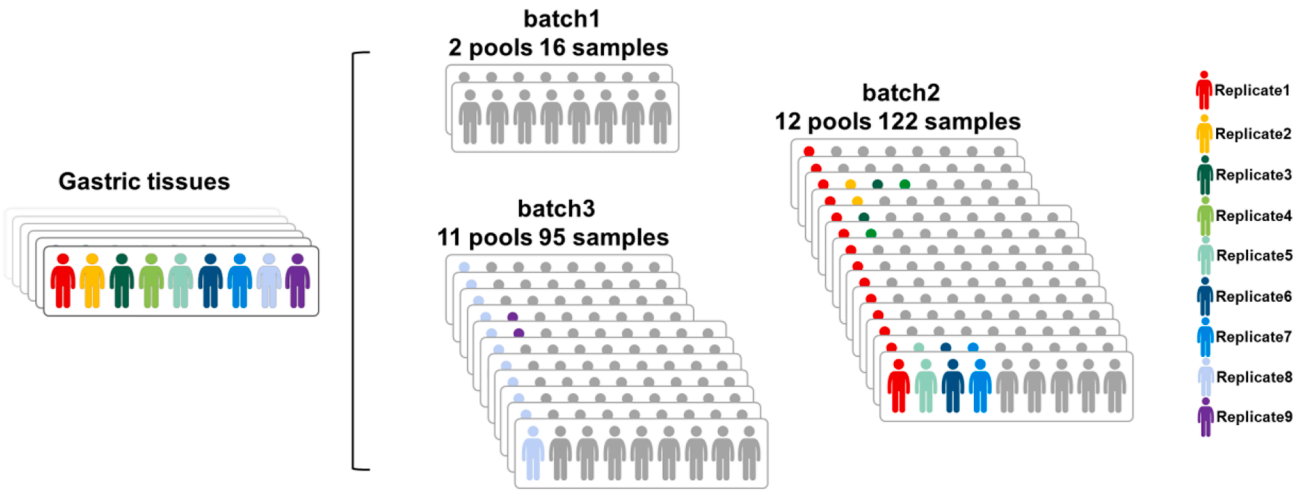


Figure 1. Pooled multiplexing strategy applied to 27 pools of 8–9 donors per pool over three processing batches

This section describes the process of preparing and sequencing single-cell RNA from pooled samples.

15. Load the prepared single-cell suspensions onto a microwell chip using the Singleron Matrix Single Cell Processing System, following the manufacturer's protocol.
16. Recover the barcoded capture beads from the microwell chip and perform reverse transcription to synthesize cDNA from the captured mRNA.
17. Amplify the cDNA by PCR to increase its quantity.
18. Fragment the amplified cDNA to an appropriate size for sequencing and ligate sequencing adapters to these fragments.
19. Construct single-cell RNA-seq libraries using the GEXSCOPE Single Cell RNA Library Kits (Singleron).
    a. Quantify and normalize individual libraries to 4 nM.
    b. Pool them equimolarly.
    c. Sequence on an Illumina NovaSeq 6000 platform with a 150 bp paired-end read configuration.

Single-cell RNA-seq data preprocessing, quality control, and cell type annotation

Timing: 3–4 weeks

This section describes the processing of single-cell RNA-seq data, including preprocessing, quality control steps, clustering of cells and cell type annotation.

Note: This procedure ensures that high-quality data is retained for further analysis and that potential biases due to technical artifacts are minimized.

20. Preprocess the Single-cell RNA-seq data with the celescope pipeline (https://github.com/singleron-RD/CeleScope).

```
#Sample information extraction and quality control

>celescope rna sample --outdir .//example/00.sample --sample example --thread 8 --chemistry
auto --wells 384 --fq1 ./example_L2_1.fq.gz

#Align sequencing reads to the reference genome, perform cell identification and gene expres-
sion quantification, and generate the core expression matrix
```

```
>celescope rna starsolo --outdir .//example/01.starsolo --sample example --thread 8 --chemistry auto --adapter_3p AAAAAAAAAAAA --genomeDir ./ensembl_75 --outFilterMatchNmin 50 --soloCellFilter "EmptyDrops_CR 30000 0.99 10 45000 90000 500 0.01 20000 0.001 10000'' --starMem 32 --soloFeatures "Gene GeneFull_Ex50pAS'' --fq1 ./example_L2_1.fq.gz --fq2 /example_L2_2.fq.gz

#Downstream Analysis and Visualization

>celescope rna analysis --outdir .//example/02.analysis --sample example --thread 8 --genomeDir ./ensembl_75 --matrix_file .//example/outs/filtered
```

21. Running demuxlet method to assign cells to individuals.

```
>demuxlet \
  --sam sorted.bam \
  --vcf sorted.vcf \
  --field GT \
  --geno-error 0.05 \
  --doublet-prior 0.1 \
  --alpha 0 --alpha 0.5 \
  --out demuxlet_info
```

Note: The method demuxlet classifies droplets based on genotyping results: droplets most compatible with a single donor (singlets), droplets best explained by a mixture of genotypes from two donors (doublets), and droplets with insufficient informative SNPs (ambiguous). For singlets, the most likely donor is assigned using exonic variants with imputation quality INFO > 0.9 and MAF > 0.05), assuming a prior doublet rate of 0.1. Only droplets with confidently assigned donor identities are retained for downstream analysis.

22. Conduct single cell quality control, followed by dimensionality reduction and unsupervised clustering under the R environment.

```
>library(Seurat)

>data <- Read10X(data.dir = input) #The variable "input'' represents the directory path where the matrix files are stored.

>data <- CreateSeuratObject(counts = data ,project = 'seurat', min.cells = 3, min.features = 200,names.delim = '_')

>data[["percent.mt'']]<-PercentageFeatureSet(object = data, pattern = "^MT-'')

>data<-subset(x=data,subset=nFeature_RNA>200&nFeature_RNA<4000&percent.mt<20&nCount_RNA>1000&nCount_RNA<20000)

>data <- NormalizeData(object = data, normalization.method = 'LogNormalize',scale.factor = 10000)

>data <- FindVariableFeatures(object = data,selection.method='vst',mean.function = ExpMean,dispersion.function=LogVMR,mean.cutoff=c(0.125,3),dispersion.cutoff=c(0.5,Inf))

>data <- ScaleData(data,vars.to.regress = c("percent.mt'',"nCount_RNA'',"orig.ident''))

>data <- RunPCA(object=data,npcs=100,pc.genes=VariableFeatures(object=data))

>data <- FindNeighbors(object=data,dims=1:50)

>data <- FindClusters(object=data,resolution=0.3)

>data <- RunTSNE(object=data,dims=1:50)
```

Table 2. Classical markers used for cell type identification

Cell type	Abbreviation	Markers
Epithelium		*EPCAM, KRT18*
Mucous neck cells	Neck	*MUC6*
Mucous pit cells	Pit	*MUC5AC, TFF1*
Chief cells	Chief	*PGA3, PGA4*
Parietal cells	Parietal	*ATP4A, ATP4B*
Endocrine cells	Endocrine	*CHGA, CHGB*
Intestinal-type cells	Enterocytes	*FABP1, TFF3*
T cells	–	*CD3D, CD2*
$CD4^+$ T cells	CD4	*CD4, IL17A*
$CD8^+$ T cells	CD8	*CD8A, GZMA, GZMB*
B cells	–	*CD79A, MZB1*
Plasma cells	Plasma	*MZB1*
Naive B cells	Naive	*MS4A1*
Myelocyte	–	*CSF1R, CD14*
M1-like macrophages	M1	*S100A8, IL1B*
M2-like macrophages	M2	*CD163, C1QA, APOE*
Dendritic cells	DCs	*CD1C, FCER1A*
Mastocyte	Mast	*TPSB2, CPA3*
Fibroblasts	–	*DCN, COL1A2*
Myofibroblasts	Myo	*ACTA2*
Inflammatory fibroblasts	Inflammatory	*PDGFRA*
Endothelial cells	ECs	*VWF, ENG*
Tip-like endothelial cells	Tip	*RGCC*
Stalk-like endothelial cells	Stalk	*SELP*
Artery-derived endothelial cells	Arterial	*SEMA3G*

23. Re-cluster the filtered cells using unsupervised clustering methods and annotate major cell types based on marker gene expression (Table 2).

Note: For other tissues, researchers can refer to databases like CellMarker (http://117.50.127.228/CellMarker/index.html). And cell-type labels for clusters are assigned based on the known marker genes expression levels. Cells co-expressing markers from two distinct major cell types are defined as ambiguous and are removed by manual curation prior to downstream analysis.

Single-cell *cis*-eQTL mapping

Timing: 3–4 days

This section describes how to perform *cis*-eQTL mapping for single-cell RNA-seq data, resolve conditionally independent regulatory signals and assess the extent to which *cis*-regulatory effects are shared across cell types.

24. Retain genes expressed in at least 1% of the corresponding cell type, and donors with a minimum of five cells per cell type.

Note: Using normalized counts from Seurat, derive expression profiles by averaging the expression of each gene across all cells from the same donor within each cell type.

25. Perform *cis*-eQTL mapping with FastQTL, restricting analysis to variants located ±1 Mb of the transcription start site (TSS) of each gene.

```
>run_FastQTL_threaded.py \
sample.vcf.gz\
celltype_exp.bed.gz\
celltype_eqtl\
--  covariates covariates.txt \
--  window 1e6 --chunks 100 --threads 16 \
--  output_dir output_path
```

Note: For each gene, apply FastQTL in nominal mode, fitting a linear model with gene expression as the dependent variable and genotype as the main predictor while adjusting for covariates. Then record the most strongly associated variant within the *cis* window. Covariates such as sex, age, and *H. pylori* infection, along with PCs and PEER factors, are included to correct for confounding factors and batch effects in gene expression data.

⚠ CRITICAL: To maximize the detection of *cis*-eQTLs, the number of PEER factors are increased incrementally and observing the number of eQTLs began to decrease after adding two factors (Figure 2). Therefore, two PEER factors are included for the sc-eQTL analysis.

26. For each eGene (i.e., a gene with at least one significant cis-eQTL) and cell type, perform stepwise conditional cis-eQTL analysis to delineate independent loci influencing gene expression.

Note: In this process, indentify the top-ranked eSNP(eSNP1) based on the smallest nominal *p* value initially. Then, regress out the effect of eSNP1 and repeat the analysis to detect secondary eQTLs with the same criteria. The eSNP with the strongest association is designated as eSNP2.

27. For eGenes marked by distinct eSNPs across various cell types, test for shared signals by reciprocal conditioning.
    a. Analyze the lead eSNP from the first cell type.
    b. Include the lead eSNP from the second cell type as a covariate in the model.
    c. Evaluate the change in the test statistic or effect size of the lead eSNP from the first cell type after conditioning on the second cell type's eSNP.

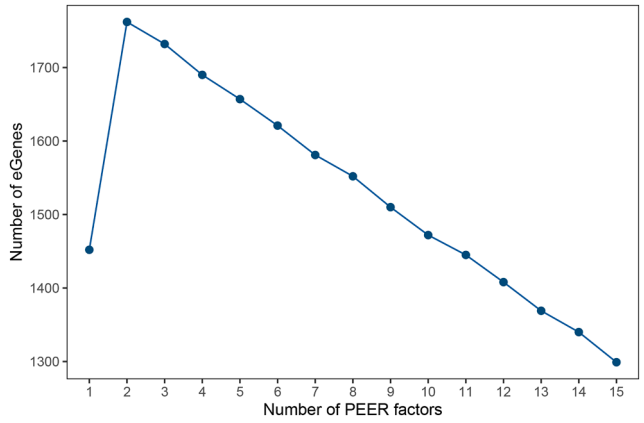


Figure 2. Changes in the number of detected *cis*-eQTLs with increasing PEER factors

Table 3. Example header of the GWAS summary data for co-localization

SNP_ID	BP	Beta_GWAS	Varbeta_GWAS
4:125418536	125418536	0.1236	0.00061009
4:125414757	125414757	−0.1196	0.00061504
4:125419479	125419479	0.1237	0.00061009
4:125415534	125415534	−0.1217	0.00061504

⚠ CRITICAL: Shared eQTLs are identified if the allelic effect size of the original eSNP significantly decreases after conditioning, indicating the eSNPs are tagging the same genetic variant across cell types. If the allelic effect size remains unchanged after conditioning, the eQTLs are considered independent between the cell types.

Integration of single-cell eQTLs with genome-wide association studies for gastric cancer

Timing: 1–2 weeks

This section describes how to assess the functional relevance of cell-type-specific eQTLs for gastric cancer (GC) by performing co-localization analysis to identify shared causal variants.

28. Prepare a GWAS summary data (gwas_coloc) with the following columns (Table 3).
    a. SNP_ID: The identifier for each SNP.
    b. BP: The genomic position of each SNP.
    c. Beta_GWAS: The beta value of GWAS.
    d. Varbeta_GWAS: The variance of beta value.
29. Prepare sc-eQTL result data (eqtl_coloc) with the following columns (Table 4).
    a. SNP_ID: The identifier for each SNP.
    b. BP: The genomic position of each SNP.
    c. Beta_eQTL: The beta value of eQTL.
    d. Varbeta_eQTL: The varience of beta value.
30. Conduct a co-localization analysis between *cis*-eQTLs from each cell type and gastric cancer (GC) GWAS signals using the coloc (v5.2.2) package under the R environment.[16] Additionally, the analysis utilizes the approximate Bayes factor method to integrate the summary statistics.

```
>library(coloc)

>result <- coloc.abf(dataset1 = eqtl_coloc, dataset2 = gwas_coloc)
```

Note: Calculate the posterior probabilities (PP0-PP4) corresponding to five hypotheses through the coloc package: PP0 represents the probability of the null model, indicating no association; PP1 reflects the probability that causal SNPs are associated only with GWAS signals; PP2 represents the probability that causal SNPs are linked solely to eQTLs; PP3 evaluates the probability that causal SNPs for GWAS signals and eQTLs are independent; and PP4 quantifies the probability that GWAS and eQTL signals share common causal SNPs. To evaluate pairwise co-localization, we examined GWAS and eQTL summary statistics for SNPs located

Table 4. Example header of the sc-eQTL result data for co-localization

SNP_ID	BP	Beta_eQTL	Varbeta_eQTL
4:125418536	125418536	0.0419397	9.23E−05
4:125414757	125414757	−0.0422436	9.60E−05
4:125419479	125419479	0.0414326	9.29E−05
4:125415534	125415534	−0.0392966	9.16E−05

within a ±1 Mb window around each gene. If the posterior probability of co-localization (PP4) > 0.7, we conclude that the GWAS and eQTL signals share a causal variant.

Cell-type-specific TWAS analysis

Timing: 2–3 weeks

This section describes how to use transcriptome-wide association studies (TWAS) to explore gene expression associations with GC risk factors.

31. For each gene, define a *cis* window as the ±1 Mb interval flanking its transcription start site (TSS). Use GCTA-GREML to estimate the proportion of expression variance attributable to SNPs within this *cis* window.

```
>gcta64 --bfile $gene --autosome --make-grm --out gene_grm/$gene

>gcta64 --reml --grm gene_grm/$gene --pheno pheno/$gene --covar ./ccovar --qcovar ./qcovar --reml-no-constrain --out heritability/$gene
```

Note: "ccovar" contains discrete covariates, such as sex and *H. pylori* infection; "qcovar" contains quantitative covariates, such as age and PCs.

⚠ CRITICAL: This approach calculates the heritability (h^2) of each gene expression, and we select genes with significant *cis*-heritable SNPs ($p < 0.05$ and $h^2 > 0.04$) for further analysis.

32. Select from a variety of regularized linear models to construct cell-type-specific reference panels, including top1, blup, lasso, and enet. Subsequently, compute SNP-expression weights using FUSION.[17]

```
>Rscript FUSION.compute_weights.R \
--  bfile $gene_plink_file \
--  tmp $TMP \
--  covar $covar_file \
--  out $gene_out \
--  PATH_gcta fusion_twas-master/gcta_nr_robust \
--  PATH_gemma gemma --PATH_plink plink \
--  models top1,blup,lasso,enet
```

Note: Use 5-fold cross-validation to evaluate the accuracy of each model by randomly sampling 1,000 highly heritable genes for testing.

33. Prepare a GWAS summary data with the following columns (Table 5).
    a. SNP_ID: The identifier for each SNP.
    b. A1: The effect allele.
    c. A2: The other allele.
    d. Z: The Z scores, sign with respect to effect allele.

Table 5. Example header of the GWAS summary data for TWAS

SNP_ID	A1	A2	Z
10:60071446	C	G	0.21
13:43570490	T	C	0.01
9:127577087	A	T	−1.30
8:31574074	T	C	0.07

34. Combine cell-type-specific gene expression prediction weights with gastric cancer GWAS summary statistics to obtain the association statistics between genetically predicted expression and gastric cancer risk.

```
>Rscript FUSION.assoc_test.R \
-- sumstats gwas.sumstats \
-- weights Exp_weights_file.list \
-- weights_dir ./weight\
-- ref_ld_chr ./LDREF_EAS/1000G.EAS. \
-- chr 1 \
-- GWASN gwas_sample_size \
-- min_r2pred 0.4 \
-- out twas.chr1.dat
```

Note: Genes are considered significant if the FDR < 0.05, indicating a meaningful association between gene expression in specific cell types and GC risk.

Mapping of sc-eQTLs with cell-type-specific chromatin accessibility

Timing: 2–3 weeks

This section outlines the approach to use scATAC-seq data to characterize cell-type-specific regulatory elements that underlie cis-eQTL effects. Specific example data of scATAC data and the complete code will be provided on GitHub.

35. Isolate nuclei from normal gastric mucosal tissue and generate scATAC-seq libraries on the 10x Genomics platform, following the manufacturer's protocol.
36. Conduct the Cell Ranger ARC pipeline to perform comprehensive analyses of gene expression and chromatin accessibility.

```
>cellranger-arc mkref --config=hg19.config
>cellranger-arc count --id=sample \
-- reference=hg19 \
-- libraries=library.csv \
-- localcores=16 \
-- localmem=64
```

37. Conduct quality control for scATAC-seq data under the R environment.

```
>scATAC <- NucleosomeSignal(scATAC)

>scATAC <- TSSEnrichment(scATAC, fast = FALSE)

>scATAC <- subset(x = scATAC, subset = nCount_data>1000 & nCount_data<50000 & nucleosome_signal < 4 & TSS.enrichment > 2)
```

38. Integrate the scATAC-seq and scRNA-seq datasets using canonical correlation analysis (CCA), and transfer cell type annotations from the scRNA-seq reference to the scATAC-seq object.

```
>transfer.anchors <- FindTransferAnchors(reference = scRNA_seq_data, query = scATAC, reduction = 'cca')

>predicted.labels <- TransferData(anchorset = transfer.anchors, refdata = scRNA_seq_data$celltype, weight.reduction = scATAC[['integrated_lsi']],dims = 2:30)
```

39. Run MACS2 to call chromatin accessibility peaks for each gastric mucosal cell type, generating cell-type-specific peak sets.

```
>peaks <- CallPeaks(object = scATAC, group.by = "celltype'')
```

40. Visualize gene loci by plotting aggregated scATAC-seq signal together with gene expression across gastric mucosal cell types using the CoveragePlot function in Signac.

```
>CoveragePlot(object = scATAC, region = 'gene_name', group.by="celltype'',ranges = peaks,

ranges.title = "MACS2'', region.highlight = ranges.show)
```

EXPECTED OUTCOMES

By following this protocol, researchers are equipped to construct a high-resolution atlas of gene expression and genetic regulation at the single-cell level. This approach significantly advances our understanding of tissue biology by enabling the detailed identification and characterization of cell-type-specific eQTLs within complex tissues. Such precise mapping is crucial for deciphering the genetic architecture underlying normal physiological functions and their dysregulation in disease states. Furthermore, the protocol facilitates the identification of cell-type-specific genomic regions associated with disease susceptibility. This analysis helps pinpoint genetic variants that may influence gene expression within specific cell contexts, contributing to a nuanced understanding of disease development at the cellular level. Additionally, the protocol supports a cell-type-specific transcriptome-wide association study (TWAS) framework. This framework links gene expression profiles to the disease risk, allowing for the identification of genes whose expression-modulated by genetic variants-correlates with increased or decreased disease risk. By connecting genetic variations to functional outcomes in specific cell types, this analysis offers a powerful tool for uncovering the molecular mechanisms through which genetic differences influence disease pathogenesis.

LIMITATIONS

While this protocol offers significant insights into cell-type-specific genetic regulation, several limitations must be acknowledged. Firstly, the effectiveness of this protocol largely depends on the availability of a substantial number of human donors. Adequate statistical power is crucial for the discovery of eQTLs. A limited sample size may result in missing subtle regulatory effects that could be significant. The protocol's reliance on large datasets underscores the need for extensive and diverse sample collections to capture a broad range of genetic interactions. Here, we recommend $\geq$ 100 donors, with each donor contributing

≥ 5 cells per cell type for analysis. Secondly, the scATAC data can initially be used to explore the regulatory mechanism of cell-type-specific eQTL, but determining the exact causal elements—such as a specific enhancer region disrupted by the variant—requires additional investigative techniques. Techniques such as reporter assays, chromatin conformation studies, or CRISPR-based perturbations might be necessary to elucidate the functional impact of identified variants. Lastly, the protocol enables the linkage of genes to disease through methods like colocalization and TWAS. While these analyses suggest potential causal relationships, they do not confirm them. Functional validation using cellular models or animal studies is essential to confirm that alterations in gene expression within a particular cell type directly influence disease phenotypes.

TROUBLESHOOTING

Problem 1

Low yield of viable single cells after dissociation (gastric tissue sample collection steps 1–2 and single-cell sample preparation and pooling steps 10–13).

Potential solution

To improve the yield of viable cells, ensure the tissue is kept cold during processing to preserve cell integrity. It is essential to mince the tissue into very small pieces to enhance enzyme penetration during the dissociation process. Consider reducing the enzyme concentration or shortening the digestion time if the cells appear damaged. Processing the sample promptly after collection, ideally within one hour of surgery or biopsy, helps prevent degradation. Adding gentle DNase to the dissociation mixture can decrease viscosity and facilitate cell recovery. If the yield remains low, consider extending the digestion time slightly or pooling multiple small samples from the same donor to increase the volume of the starting material.

Problem 2

Low cDNA library yield or quality (single-cell RNA sequencing steps 15–19).

Potential solution

Issues with reverse transcription or PCR can significantly impact library quality. Ensure that all reagents, especially enzymes and primers, are fresh and handled on ice to maintain their efficacy. If the initial cDNA yield is low, increasing the number of PCR cycles by 2–4 may help, but be cautious of amplification bias from excessive cycles. Check the success of droplet or partition generation. Failed emulsions or microfluidic chip issues (such as clogs or operational errors) can compromise the reverse transcription efficiency. If problems persist, consider repeating the experiment with a new chip and ensuring careful sample loading. Analyze the Bioanalyzer trace of the amplified cDNA. A lack of a broad smear on the trace may indicate a failure in the reverse transcription step. Optimizing RT conditions, such as increasing the RT enzyme concentration or adjusting the temperature profile, might be required.

Problem 3

Few or no eQTLs detected in a given cell type (single-cell cis-eQTL mapping step 24–27).

Potential solution

Low detection rates of eQTLs can be attributed to several factors. Assess if the sample size or expression variability for the cell type in question is too low. Consider if the significance thresholds are too stringent or if the cell type inherently has high measurement noise. Review the number of donors contributing cells to the analysis. A lack of sufficient cells from a sufficient number of donors can reduce statistical power. If necessary, slightly relax the multiple testing correction threshold for exploratory analysis to see if any suggestive signals emerge. To increase statistical power, consider obtaining more donor samples or aggregating similar cell types, though this may reduce the specificity of the findings. Ensure that the expression data aggregation was performed correctly. For example, using median values instead of means can diminish the detectable signal. Finally,

reevaluate the inclusion of covariates in the analysis. Overfitting covariates can inadvertently absorb genuine biological signals, masking true genetic influences.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Guangfu Jin (guangfujin@njmu.edu.cn).

Technical contact

For technical specifics on executing this protocol, Guangfu Jin (guangfujin@njmu.edu.cn) will provide support to ensure its correct implementation.

Materials availability

This study did not generate new unique reagents.

Data and code availability

- A publicly accessible website, scGaTE (http://ccra.njmu.edu.cn/scgate/), has been developed to enable browsing of eQTLs and eGenes across various gastric cell types.
- The summary data reported in this study have been deposited in the OMIX, the China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences (accession no. OMIX: OMIX008014), and are available at https://ngdc.cncb.ac.cn/omix.
- The raw sequencing data have been deposited in the Genome Sequence Archive (GSA-Human) at the National Genomics Data Center, China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences (accession no. GSA: HRA009277), and are publicly accessible at https://ngdc.cncb.ac.cn/gsa-human.
- Analysis code used for the analyses is available in Zenodo: https://doi.org/10.5281/zenodo.19454281. Supplementary key data files are available in the GitHub release: https://github.com/NMUGCEPI/sc_eQTL/releases/tag/data. All data sharing complies with the regulations set forth by the Ministry of Science and Technology of the People's Republic of China.

ACKNOWLEDGMENTS

We would like to express our gratitude to all study participants and research staff for their contributions and commitment to the present study. This work was supported by the National Natural Science Foundation of China (82125033, 82388102, 82230110, 82273084, and 82273714); the Chinese Academy of Medical Sciences (2019RU038); the Excellent Youth Foundation of Jiangsu Province (BK20220100); the Noncommunicable Chronic Diseases-National Science and Technology Major Project (2023ZD0501500); the Major Program of Gusu School (GSKY20210103); and the Major Program of Wuxi Medical Center, Nanjing Medical University.

AUTHOR CONTRIBUTIONS

G.J. initiated, conceived, and supervised the study. X.F., L.B., and C.Y. performed the single-cell RNA sequencing and data analyses. X.F. wrote the protocol. L.B. scripted the single-cell RNA sequencing analysis code. G.J. and C.Y. revised the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Bian, L., Hu, B., Li, F., Gu, Y., Hu, C., Chen, Y., Deng, B., Fang, H., Zhu, X., Chen, Y., et al. (2025). Single-cell eQTL mapping reveals cell-type-specific genes associated with the risk of gastric cancer. Cell Genom. *5*, 100812. https://doi.org/10.1016/j.xgen.2025.100812.

2. Jin, G., Lv, J., Yang, M., Wang, M., Zhu, M., Wang, T., Yan, C., Yu, C., Ding, Y., Li, G., et al. (2020). Genetic risk, incident gastric cancer, and healthy lifestyle: a meta-analysis of genome-wide association studies and prospective cohort study. Lancet Oncol. *21*, 1378–1386. https://doi.org/10.1016/s1470-2045(20)30460-5.

3. Zhu, X., Lv, J., Zhu, M., Yan, C., Deng, B., Yu, C., Guo, Y., Ni, J., She, Q., Wang, T., et al. (2023). Development, validation, and evaluation of a risk assessment tool for personalized screening of gastric cancer in Chinese populations. BMC Med. *21*, 159. https://doi.org/10.1186/s12916-023-02864-0.

4. Dong, J., Wu, X., Zhou, X., Gao, Y., Wang, C., Wang, W., He, W., Li, J., Deng, W., Liao, J., et al. (2023). Spatially resolved expression landscape and gene-regulatory network of human gastric corpus epithelium. Protein Cell *14*, 433–447. https://doi.org/10.1093/procel/pwac059.

5. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575. https://doi.org/10.1086/519795.

6. Delaneau, O., Marchini, J., and Zagury, J.F. (2011). A linear complexity phasing method for thousands of genomes. Nat. Methods *9*, 179–181. https://doi.org/10.1038/nmeth.1785.

7. Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. *5*, e1000529. https://doi.org/10.1371/journal.pgen.1000529.

8. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. j. *17*, 10–12.

9. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

10. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. https://doi.org/10.1093/bioinformatics/btt656.

11. Kang, H.M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C.M., et al. (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat. Biotechnol. *36*, 89–94. https://doi.org/10.1038/nbt.4042.

12. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502. https://doi.org/10.1038/nbt.3192.

13. Ongen, H., Buil, A., Brown, A.A., Dermitzakis, E.T., and Delaneau, O. (2016). Fast and efficient QTL mapper for thousands of molecular phenotypes. Bioinformatics *32*, 1479–1485. https://doi.org/10.1093/bioinformatics/btv722.

14. Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS A J. Integr. Biol. *16*, 284–287. https://doi.org/10.1089/omi.2011.0118.

15. Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics *32*, 2847–2849. https://doi.org/10.1093/bioinformatics/btw313.

16. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. *10*, e1004383. https://doi.org/10.1371/journal.pgen.1004383.

17. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J.H., Jansen, R., de Geus, E.J.C., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252. https://doi.org/10.1038/ng.3506.

18. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569. https://doi.org/10.1038/ng.608.

19. Stuart, T., Srivastava, A., Madad, S., Lareau, C.A., and Satija, R. (2021). Single-cell chromatin state analysis with Signac. Nat. Methods *18*, 1333–1341. https://doi.org/10.1038/s41592-021-01282-5.

20. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137. https://doi.org/10.1186/gb-2008-9-9-r137.