

课程概览

1 数据库基本概念

2 NCBI & EMBL-EBI & 在线分析工具

3 TCGA GTEx & GEPIA 1/2/2021

4 单细胞/空间转录组数据库

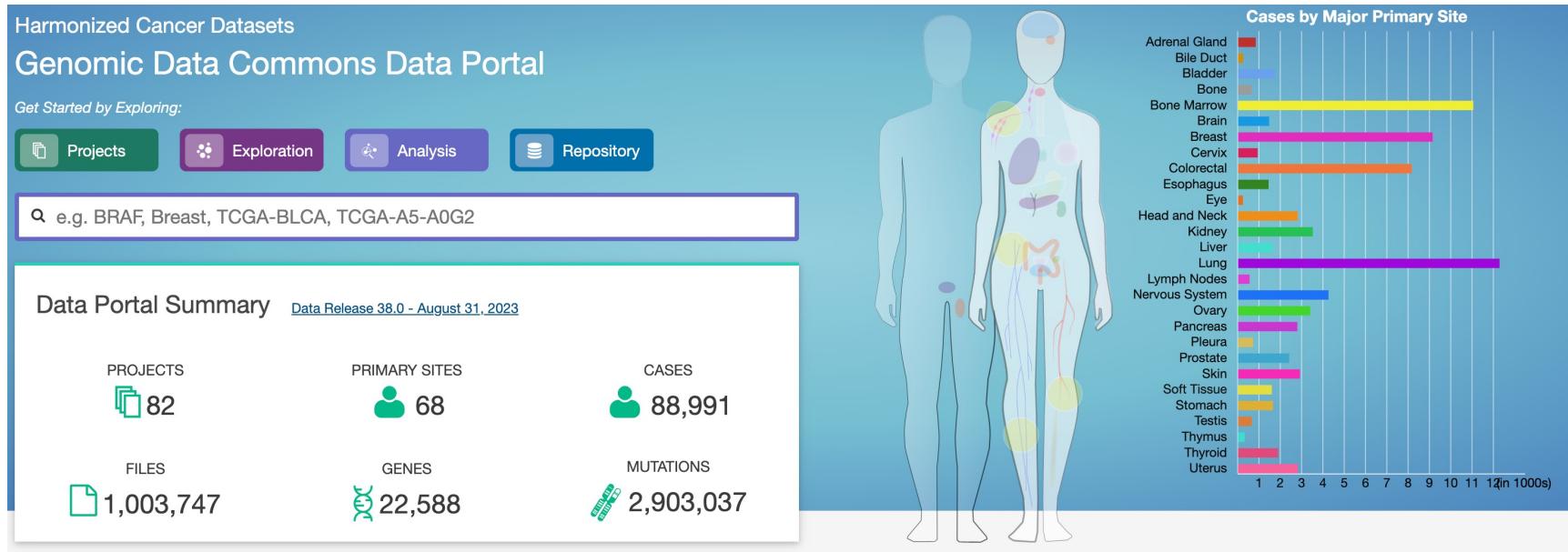


- 肿瘤基因组图谱 (The Cancer Genome Atlas, TCGA)计划由美国 National Cancer Institute(NCI)和National Human Genome Research Institute (NHGRI) 于2006年联合启动的项目。

<https://www.cancer.gov/ccg/research/genome-sequencing/tcga>

TCGA 成果与影响

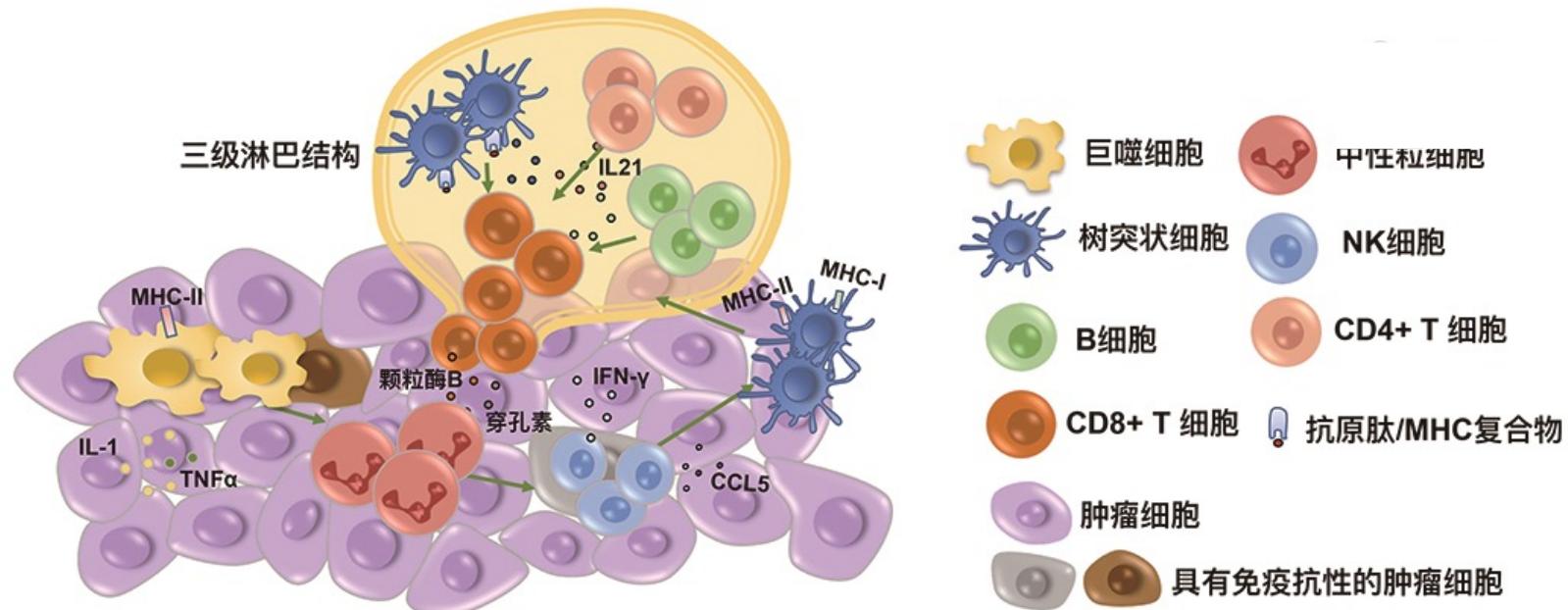
1. 建立了丰富的基因组学数据资源，支撑了计算生物学领域



- 8W+患者样本数据，33种癌症类型，5种测序类型数据(WXS, **RNA-Seq**, Genotyping Array, Targeted Sequencing, WGS)以及对应的临床信息(药物响应, 生存期等)
- TCGA 产生的海量数据刺激了计算生物学领域的巨大发展。研究人员针对开发很多生物信息学数据挖掘工具，例如识别体细胞和种系突变、预测具有预后意义的基因、构建基因调控网络、以及癌症图像的自动分析等。

TCGA 成果与影响

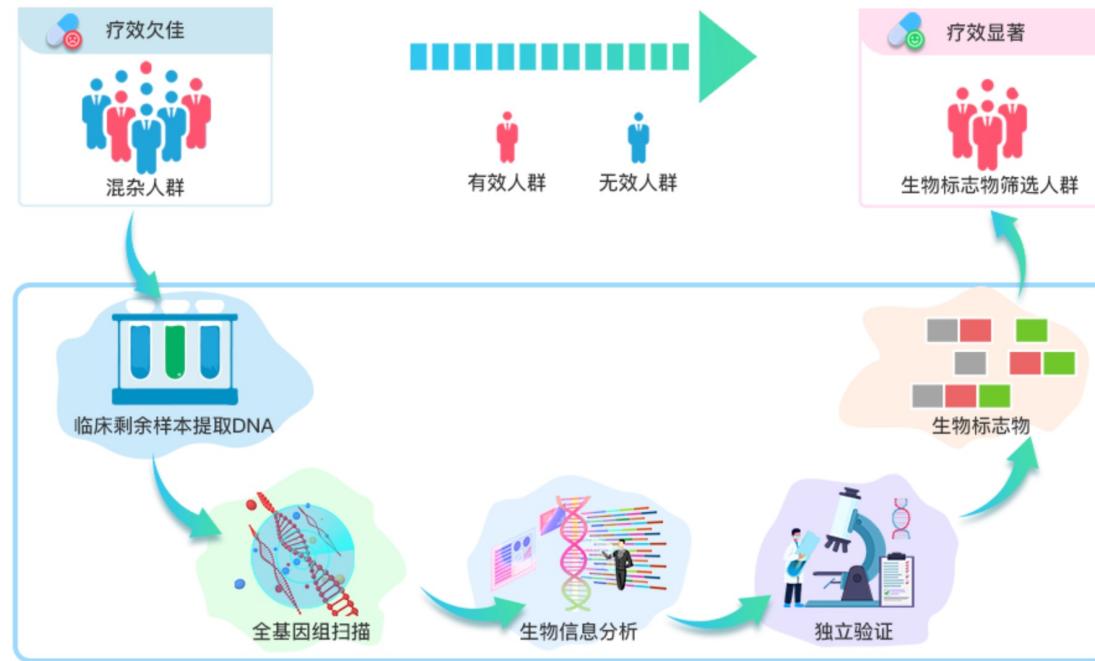
2. 加深了癌症的分子机制的探索



- 理解DNA各种类型的改变, 例如置换, 嵌合, 融合, 拷贝数改变以及其他复杂的结构变异
- 肿瘤异质性来源: 癌细胞和免疫细胞组成的微环境理解
- 不同癌种之间共有的分子演化规律

TCGA 成果与影响

3. 推进医疗相关科学进步



- 推动了新兴测序技术的巨大发展，提高数据质量并降低测序成本。比如反相蛋白阵列、福尔马林固定石蜡包埋的样品分析物提取等测序建库分子技术。
- 通过分子和临床数据对疾病进行更精确的诊断,分类和预后，推动精准医疗
- 通过海量数据挖掘癌症治疗潜在靶点分子以及对应分子机制。

TCGA RNA-seq数据介绍



- **RNA-seq** 数据中 TCGA 规范定义了33种常见癌种命名方式
- 每种癌种对应均有上百样本数

<https://portal.gdc.cancer.gov/repository>

TCGA RNA-seq数据介绍

The screenshot shows a search interface for TCGA RNA-seq data. At the top left, under 'Experimental Strategy', 'RNA-Seq' is selected with 22,548 files. Below it, under 'Workflow Type', 'STAR - Counts' is selected with 22,548 files, while other options like STAR-Fusion, Arriba, etc., have 0 files. To the right, under 'Data Format', 'tsv' is selected with 43,808 files, while bam and bedpe have 32,534 and 21,259 files respectively. At the bottom, a specific file entry is shown: '7d14ddef-7a1b-4515-9536-3fc4a9b85702.rna_seq.augmented_star_gene_counts.tsv' is listed as controlled, with 1 TCGA-OV transcriptome profiling TSV file of 4.24 MB.

| Workflow Type | # Files |
|---------------------------|---------|
| STAR - Counts | 22,548 |
| STAR-Fusion | 0 |
| Arriba | 0 |
| STAR 2-Pass Genome | 0 |
| STAR 2-Pass Chimeric | 0 |
| STAR 2-Pass Transcriptome | 0 |

| Data Format | # Files |
|-------------|---------|
| tsv | 43,808 |
| bam | 32,534 |
| bedpe | 21,259 |

| File Details | Download Options |
|--|---------------------|
| 7d14ddef-7a1b-4515-9536-3fc4a9b85702.rna_seq.augmented_star_gene_counts.tsv d0cb6e31-ce63-4c74-ad17-b2bcc5bdc053.rna_seq.star_splice_junctions.tsv.gz | open controlled |

- 支持多种常见分析软件的结果(STAR等) 同时提供多种文件格式(bam,bed等)
- 基本的基因表达矩阵支持公开下载，其他上游数据需要额外申请

TCGA RNA-seq数据下载方式

1. `gdc-client`(官网可执行程序下载方式)

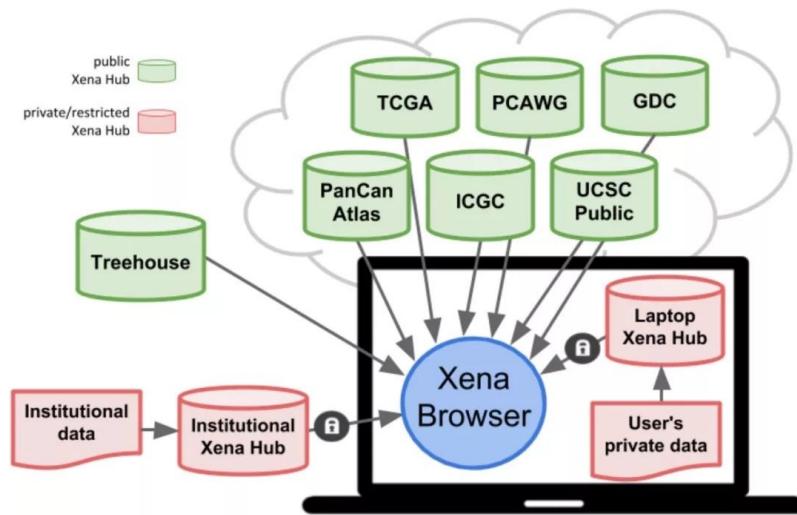
The screenshot shows a dark-themed web page for the GDC Data Transfer Tool (gdc-client). At the top, the title "GDC Data Transfer Tool (gdc-client)" is displayed with a small "🔗" icon. Below the title, there is a section titled "Overview of the GDC Data Transfer Tool". The main content area contains text about the tool's purpose and a bulleted list of links:

- [GDC Data Transfer Tool \(gdc-client\)](#)
 - [Building the gdc-client](#)
 - [Instructions](#)
 - [Executing unit tests](#)
 - [Install pre-commit](#)
 - [Update secrets baseline for detect-secrets](#)
 - [Contributing](#)

- 官方原生数据，不存在二次修改
- 以上两种下载方式需要 从官网对应筛选数据下载 (有多个文件需要对应)
- 下载后的数据需要一一对照整理 (费时费力)

TCGA RNA-seq数据下载方式

2. Xena (网页工具) 新手入门推荐

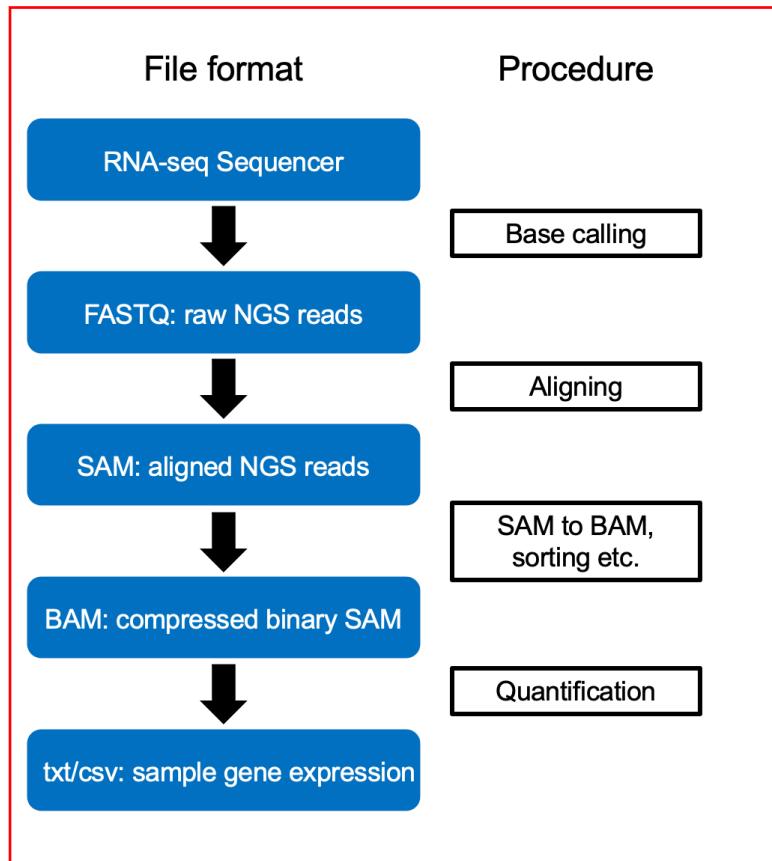


GDC TCGA Acute Myeloid Leukemia (LAML) (15 datasets)
GDC TCGA Adrenocortical Cancer (ACC) (14 datasets)
GDC TCGA Bile Duct Cancer (CHOL) (14 datasets)
GDC TCGA Bladder Cancer (BLCA) (14 datasets)
GDC TCGA Breast Cancer (BRCA) (20 datasets)
GDC TCGA Cervical Cancer (CESC) (14 datasets)
GDC TCGA Colon Cancer (COAD) (15 datasets)
GDC TCGA Endometrioid Cancer (UCEC) (15 datasets)
GDC TCGA Esophageal Cancer (ESCA) (14 datasets)
GDC TCGA Glioblastoma (GBM) (15 datasets)
GDC TCGA Head and Neck Cancer (HNSC) (14 datasets)
GDC TCGA Kidney Chromophobe (KICH) (14 datasets)
GDC TCGA Kidney Clear Cell Carcinoma (KIRC) (15 datasets)
GDC TCGA Kidney Papillary Cell Carcinoma (KIRP) (15 datasets)
GDC TCGA Large B-cell Lymphoma (DLBC) (14 datasets)
GDC TCGA Liver Cancer (LIHC) (14 datasets)

- UCSC Xena 整理了很多常用数据库的数据，其中包括TCGA
- 只需要在线网页选择数据点击即可下载 处理好的 TCGA数据
- 需要注意单位换算，该网页存储方式一般取对数化 $\log_2(X + 1)$ 存储

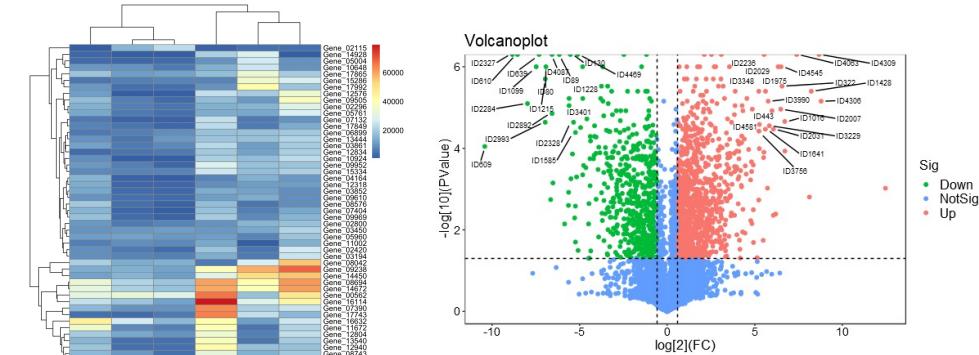
TCGA RNA-seq 数据 常见处理流程和下游分析

• 自测数据处理流程



• TCGA RNA-seq数据处理流程

- 下载获得上游**处理好的**count 矩阵, 一般
不需要重新上游处理
 - 根据需求类别进行整理和**标准化**
 - 下游差异分析, 生存期分析等



- TCGA 数据库中正常人对照的 RNA-seq 数据较少？

联合GTEx数据库

GTEX 数据库

Genotype-Tissue Expression (GTEX)

The screenshot shows the GTEx Portal homepage. At the top is a logo with a DNA helix and the text "GTEx Portal". Below the logo is a navigation bar with links: Home, Downloads ▾, Expression ▾, Single Cell ▾, and QTL ▾. The main content area is divided into three sections: "Single Cell", "Expression", and "QTL".

- Single Cell:** Includes "Data Overview" (Learn more about available single cell data) and "Multi-Gene Single Cell Query" (Browse and search single cell expression by gene and tissue).
- Expression:** Includes "Multi-Gene Query" (Browse and search expression by gene and tissue) and "Transcript Browser" (Visualize transcript expression and isoform structures).
- QTL:** Includes "Locus Browser (Gene-centric)" (Visualize QTLs by gene in the Locus Browser), "Locus Browser (Variant-centric)" (Visualize QTLs by variant in the Locus Browser VC (Variant Centric)), "IGV Browser" (Visualize tissue-specific eQTLs and coverage data in the IGV Browser), "eQTL Dashboard" (Batch query eQTLs by gene and tissue), and "eQTL Calculator" (Test your own eQTLs).

<https://gtexportal.org/home/>

- 主要由美国NIH（国立卫生研究院）的公共基金计划资助
- 其主要针对正常人（非疾病患者）死亡后的尸体组织进行相关基因测序（**RNA-seq, scRNA-seq, WGS等**）

GTEX 数据库 第一阶段成果

RESEARCH ARTICLE

The Genotype-Tissue Expression (GTEX) pilot analysis:
Multitissue gene regulation in humans

THE GTEX CONSORTIUM, KRISTIN G. ARDLIE, DAVID S. DELUCA, AYELLET V. SEGRÈ, TIMOTHY J. SULLIVAN, TAYLOR R. YOUNG, ELLEN T. GELFAND, CASANDRA A. TROWBRIDGE, JULIAN B. MALLER, [...], AND EMMANOUIL T. DERMITZAKIS +130 authors Authors Info & Affiliations

SCIENCE · 8 May 2015 · Vol 348, Issue 6235 · pp. 648-660 · DOI: 10.1126/science.1262110

7,075 2,755 CHECK ACCESS

REPORT

The human transcriptome across tissues and individuals

MARTA MELÉ, PEDRO G. FERREIRA, FERRAN REVERTER, DAVID S. DELUCA, JEAN MONLONG, MICHAEL SAMMETH, TAYLOR R. YOUNG, JAKOB M. GOLDMANN, DMITRI D. PERVOUCHINE, [...], AND RODERIC GUIGÓ +12 authors Authors Info & Affiliations

SCIENCE · 8 May 2015 · Vol 348, Issue 6235 · pp. 660-665 · DOI: 10.1126/science.aaa0355

3,383 791 CHECK ACCESS

REPORT

Effect of predicted protein-truncating genetic variants on the human transcriptome

MANUEL A. RIVAS, MATTI PIRINEN, DONALD F. CONRAD, MONKOL LEK, EMILY K. TSANG, KONRAD J. KARCZEWSKI, JULIAN B. MALLER, KIMBERLY R. KUKURBA, DAVID S. DELUCA, [...], AND XAVIER ESTIVILL +190 authors Authors Info & Affiliations

SCIENCE · 8 May 2015 · Vol 348, Issue 6235 · pp. 666-669 · DOI: 10.1126/science.1261877

1,416 168 CHECK ACCESS

- **2015年，GTEX发布了第一阶段成果，一次性在Science杂志上发表3篇研究成果，该成果还被选为当年封面文章。**
- **从175名死者身上采集到了1641个尸检样本，这些样本来自54个不同的身体部位**
- **这三篇文章分别从基因表达模式和对基因组中影响基因表达的特定区域(增强子,启动子等)以及组织特异性基因的表达调控模式和截短的蛋白变异体对不同组织中的基因表达影响。**

GTEX 数据库 第二阶段成果

[Open access](#) | Published: 12 October 2017

Landscape of X chromosome inactivation across human tissues

Taru Tukiainen , Alexandra-Chloé Villani, Angela Yen, Manuel A. Rivas, Jamie L. Marshall, Rahul Satija, Matt Aguirre, Laura Gauthier, Mark Fleharty, Andrew Kirby, Beryl B. Cummings, Stephane E. Castel, Konrad J. Karczewski, François Aguet, Andrea Byrnes, **GTEX Consortium**, Tuuli Lappalainen, Aviv Regev, Kristin G. Ardlie, Nir Hacohen & Daniel G. MacArthur 

[Nature](#) 550, 244–248 (2017) | [Cite this article](#)

63k Accesses | 549 Citations | 494 Altmetric | [Metrics](#)

- 2017年，GTEX发布了第二阶段成果，在Nature杂志上发表4篇研究成果。
- 收集来自449名生前健康的人类捐献者的7000多份尸检样本，涵盖44个组织（42种不同的组织类型），包括31个实体器官组织
- 结合GTEX数据研究了与基因表达相关联的基因变异调节RNA编辑和X染色体失活现象的机制

基于TCGA分析的数据库：GEPIA 1



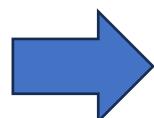
Accessing Single Cell Gene
Expression Data on the GTEx Portal

癌症样本RNA-seq
(比如脑肿瘤)

+

正常人对应组织样本RNA-seq
(比如大脑)

1. 联合分析可**高通量**地筛选相关肿瘤特异性相关表达基因
2. 可以在多个肿瘤疾病中寻找异同表达基因
- ...
- 对于**不熟悉生信的生物研究者**如何获取数据推动相关研究呢?



GEPIA1/2

基于TCGA分析的数据库：GEPIA 1



张泽民

邮箱: zemin@pku.edu.cn

职称: 教授

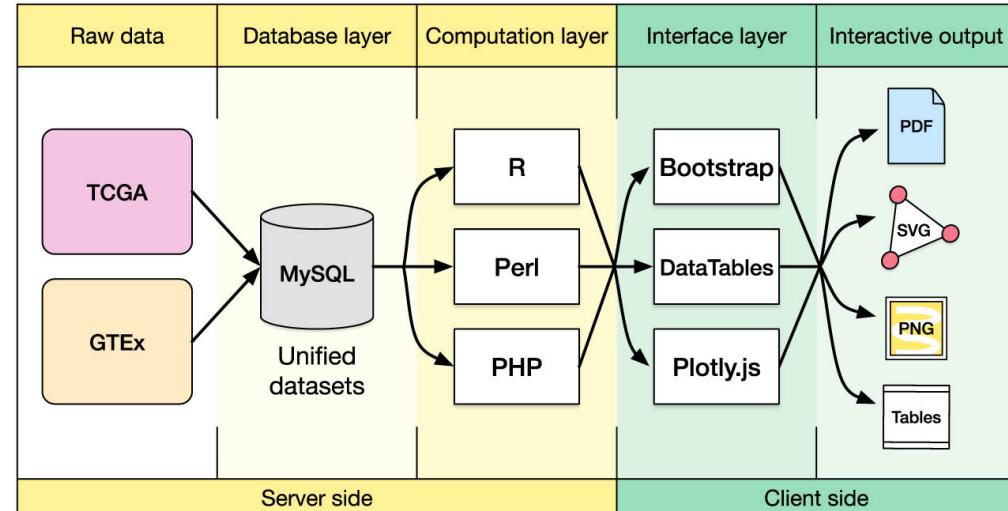
办公室地址: 北京市海淀区颐和园路5号, 综合科研2号楼, 100871

所属实验室: 张泽民实验室, 生物动态光学成像中心 (BIOPIC), 生命科学联合中心

个人主页: <http://cancer-pku.cn/>

个人介绍:

张泽民, 北京大学生物医学前沿创新中心研究员、副主任, 北京大学未来基因诊断高精尖创新中心研究员, 中国“海外高层次人才”计划, 长江特聘教授, 北大-清华生命科学联合中心高级研究员。1988年本科毕业于南开大学生物系, 1989年由CUSBEA项目赴美留学。1995年研究生毕业于滨州州立大学生物化学和分子生物学专业, 获得博士学位; 1995-1998年在美国旧金山加州大学进行博士后研究工作。1998年起任职于美国 GENENTECH 公司开始生物信息学研究, 曾任博士后导师, 生物信息学部主任, 生物信息首席科学家。前期主要工作包括: 在世界上首次报道实体癌的全基因组测序; 首次在全基因组水平研究病毒插入事件在肝癌发生发展中的作用; 首次通过计算方法鉴别癌症的驱动突变。通过对癌症高通量数据的分析, 成功地发现了多种特异性抗癌靶点, 获得60多项美国专利授权。2014 年加入北京大学。2017年获 The Boehringer Ingelheim Investigator Award, 2017年获 Bayer Investigator Award, 相关成果入选2017年度中国十大医学科技新闻、2017年度中国生命科学十大进展、细胞出版社2017 和2019中国年度论文、2018和2019年度中国生物信息十大进展、2020年细胞杂志最佳论文。



GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses

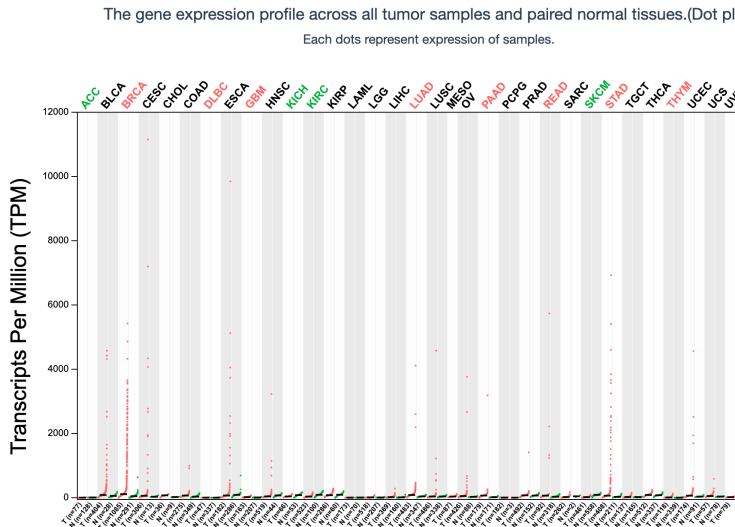
Z Tang, C Li, B Kang, G Gao, C Li... - Nucleic acids ..., 2017 - academic.oup.com

... GEPIA provides key interactive and customizable ... through GEPIA greatly facilitate data mining in wide research areas, scientific discussion and the therapeutic discovery process. GEPIA ...

☆ 保存 ⚡ 引用 被引用次数: 7235 相关文章 所有 10 个版本

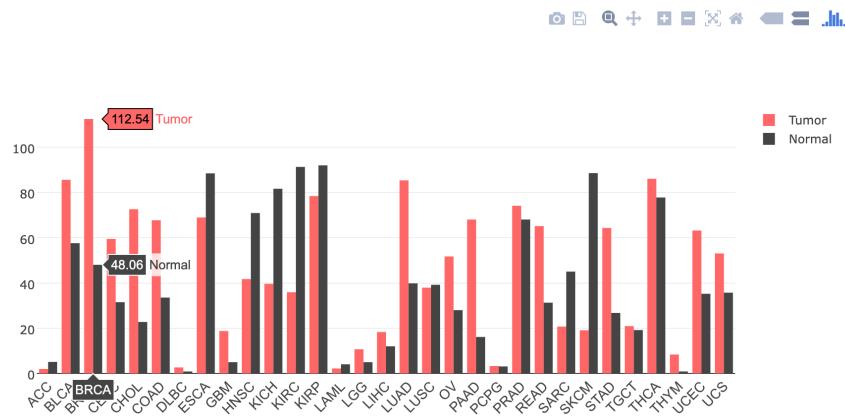
- 北大的张泽民教授课题组2017年基于上述思想 开发了一键化操作 数据库
用于方便生物领域内研究者便捷地获得相关数据分析结果

基于TCGA分析的数据库：GEPIA 1



The gene expression profile across all tumor samples and paired normal tissues.(Bar plot)

The height of bar represents the median expression of certain tumor type or normal tissue.



<http://gepia.cancer-pku.cn/>

- 通过输入基因可直接查看对应癌症和正常样本中该基因的TPM的情况

基于TCGA分析的数据库：GEPIA 1

Differential Expression Analysis

---- Help ----

Dataset (Cancer name)

ACC

|Log₂FC| Cutoff: q-value Cutoff:

1 0.01

Differential Methods

ANOVA
 LIMMA
These two methods are used for Tumor vs Paired Normal samples.
 Top 10
This method are used for Tumor vs All Normal Samples.

Chromosomal Distribution

Over-expressed

Default color: Over-Red; Under-Green

List Plot

Show 10 entries

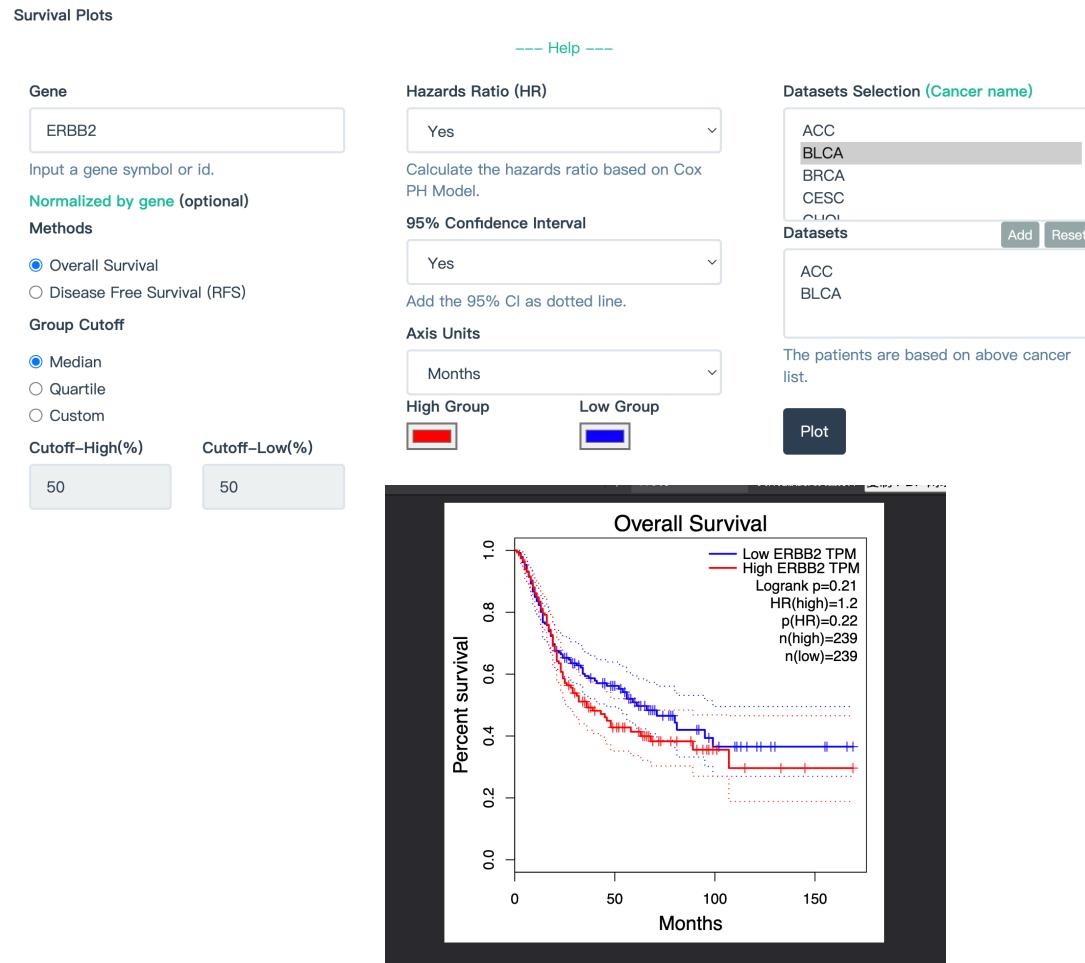
| Gene Symbol | Gene ID | Median (Tumor) | Median (Normal) | Log2(Fold Change) | adjp |
|---------------|--------------------|----------------|-----------------|-------------------|----------|
| RP11-40C6.2 | ENSG00000219928.2 | 495.347 | 0.000 | 8.955 | 1.57e-71 |
| RP5-940J5.9 | ENSG00000269968.1 | 251.933 | 0.000 | 7.983 | 1.69e-24 |
| IGF2 | ENSG00000167244.17 | 2348.949 | 19.734 | 6.824 | 6.90e-25 |
| AC132217.4 | ENSG00000240801.1 | 5394.224 | 46.691 | 6.822 | 2.22e-24 |
| RP11-316M1.12 | ENSG00000259357.2 | 79.981 | 0.990 | 5.347 | 6.25e-27 |
| CTB-63M22.1 | ENSG00000229119.3 | 547.531 | 17.294 | 4.906 | 3.81e-44 |
| NPTX2 | ENSG00000106236.3 | 450.879 | 24.130 | 4.168 | 1.38e-15 |
| CTC-425F1.4 | ENSG00000267458.1 | 44.261 | 1.590 | 4.127 | 1.34e-21 |
| LLNLR-284B4.1 | ENSG00000274177.1 | 15.850 | 0.000 | 4.075 | 3.73e-14 |
| HSPB1P1 | ENSG00000236060.2 | 14.160 | 0.000 | 3.922 | 2.57e-46 |

Showing 1 to 10 of 3,093 entries

Previous 1 2 3 4 5 ... 310 Next

- 通过选择相关癌种和差异分析方法可直接获得对应癌症和正常样本 RNA-seq 的差异分析结果 (无需手动运行deseq2/edgeR等分析软件)

基于TCGA分析的数据库：GEPIA 1



- 通过选择相关癌种,对应基因和生存分析方法可直接获得对应癌症和正常样本生存分析结果(无需手动导入样本生存期信息并用相关软件分析)

基于TCGA分析的数据库：GEPIA 2

JOURNAL ARTICLE

GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis

Zefang Tang, Boxi Kang, Chenwei Li, Tianxiang Chen, Zemin Zhang  Author Notes

Nucleic Acids Research, Volume 47, Issue W1, 02 July 2019, Pages W556–W560,

<https://doi.org/10.1093/nar/gkz430>

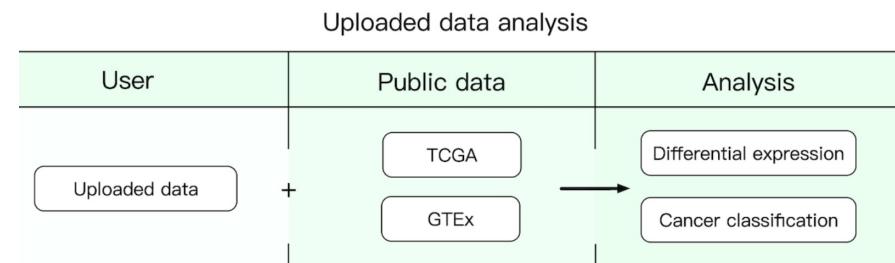
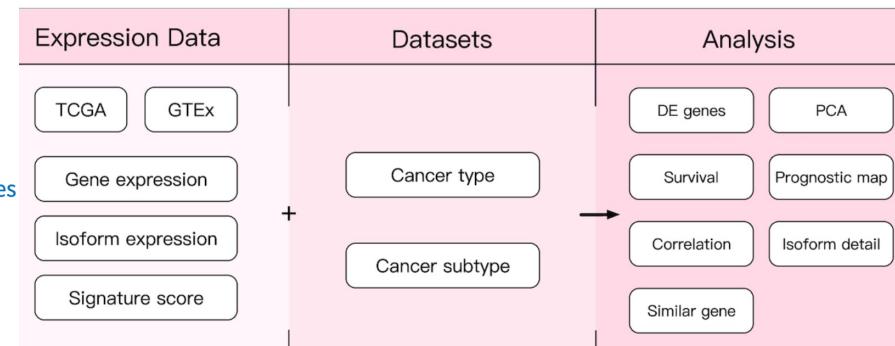
Published: 22 May 2019 Article history ▾

GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis

Z Tang, B Kang, C Li, T Chen, Z Zhang - Nucleic acids research, 2019 - academic.oup.com

... of different cancer subtypes, GEPIA2 allows users to fine-... GEPIA2 as we provide signature-based functionalities with a curated list of signatures for efficient investigation. Finally, ...

☆ 保存 引用 被引用次数: 2381 相关文章 所有 10 个版本



2019年，GEPIA2更新发布第二版

- 数据方面主要更新了不同剪切体水平(isoform)的定量结果以及癌症的亚型(84种)
- 提供用户上传自己的RNA-seq数据并和TCGA/GTEx进行比较以及对应批处理的API。

基于TCGA分析的数据库：GEPIA 2

The screenshot displays the GEPIA 2 web application interface. On the left, a dark sidebar lists various functions: Home, Expression Analysis, Custom Data Analysis, Docs, Examples, Dataset Sources, and Deconvolution Analysis. The main content area features the GEPIA 2 logo and navigation tabs: Single Gene Analysis (selected), Cancer Type Analysis, Custom Data Analysis, and Multiple Gene Analysis. A search bar prompts users to "Enter gene/isoform name" with examples like ERBB2/ENSG00000141736 or ERBB2-001/ENST00000584601. Below the search bar are buttons for Profile, Boxplots, Stage Plots, Survival Analysis, and Similar. The interface includes several data visualizations: a heatmap titled "Survival map of Hazardous Ratio" showing log10(HR) values across cancer types (BLCA, BRCA, ACC, LUAD, MESO, USC, TGCT, THCA, UCEC, LIHC) and genes; an "Overall Survival" Kaplan-Meier plot comparing Low Signatures Group (blue) and High Signatures Group (orange); and a "Signature-based statistics" box containing a box plot for ACC with sample sizes num(T)=77 and num(N)=128. At the bottom, there are links to compare uploaded data with TCGA and GTEx, and a "Cancer-type classifier for uploaded data" section.

<http://gepia2.cancer-pku.cn>

- 更新了前端框架并集成了GEPIA1的全部功能
- 侧边栏中包含所有支持的功能以及数据下载和上传功能

基于TCGA分析的数据库：GEPIA 2

Isoform Usage

Gene

ERBB2

Input a gene set using symbol or id.

Axis Option

Default (X: Cancer; Y: Isoform)

Which labels should be on X axis.

Datasets Selection (Cancer name)

Add

Tips: Ctrl/Command + A: select all cancer types.

ACC

BLCA

BRCA

CESC

CHOL

COAD

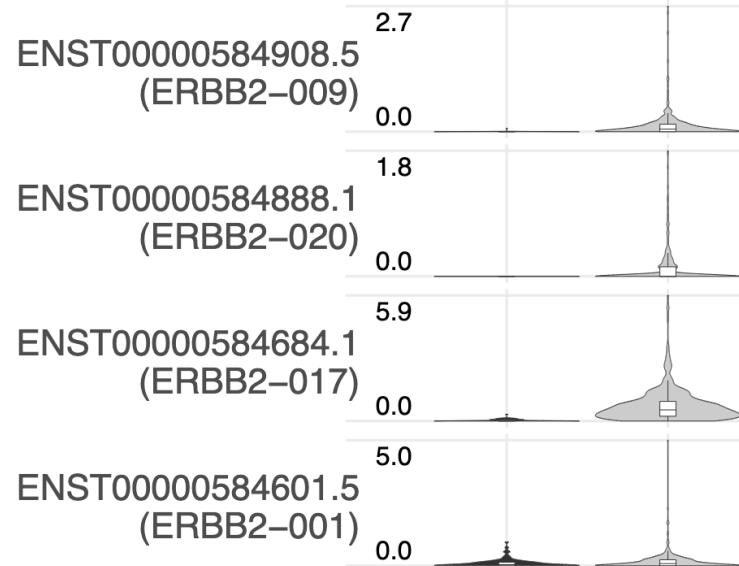
Datasets

Reset

ACC
BLCA

The patients are based on above cancer list.

Plot



- 通过输入基因可直接查看该基因在对应癌症的样本中该基因已知的isoform的表达情况

基于TCGA分析的数据库：GEPIA 2

The cancer subtype classifier takes an RNA-seq profile and makes a prediction.

Choose your model to be tested on and you will get a probability matrix by your sample for each subtype.

The uploaded gene expression profile should be TPM values with Hugo gene names.

This classifier is based on the naive bayes algorithm implemented by `sklearn` package in python with some modification (e.g., feature selection)

WARNING: please select the correct cancer type of your file uploaded, or you will get meaningless results.

If you want to determine the origin of the cancer metastasis, please select 'TCGA_Subtype', which has the overall subtypes of all cancer types provided by TCGA officially.

Step 1: Choose the cancer type for comparation :

▼

[Click here](#) to show details for cancer subtype information.

Step 2: Choose one file to upload

未选择任何文件

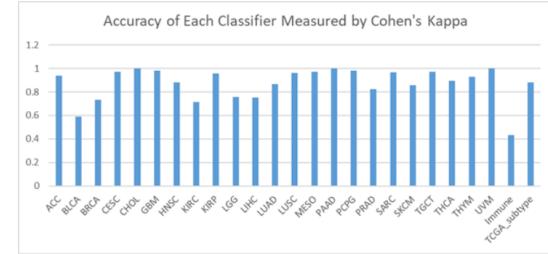
Note: You need the csv format like this.

| | A1BG | ADA | AKT3 | ZBTB11-AS1 | MED6 |
|---------|----------|----------|----------|------------|----------|
| sample1 | 5.443278 | 1.008734 | 1.008734 | 0 | 0 |
| sample2 | 0 | 8.684895 | 0 | 0 | 0 |
| sample3 | 0 | 0 | 5.439709 | 0 | 7.926596 |
| sample4 | 0 | 0 | 0 | 0 | 0 |
| sample5 | 0 | 0 | 1.362182 | 0 | 0 |

An example can be downloaded [here](#).

Step 3: Click to submmmit your file.

Here is the result of perfomance measured by a robust static [cohen's kappa](#) for each classifier.



- 通过上传自己的RNA-seq矩阵可获得预测的癌症分型

基于TCGA分析的数据库：GEPIA 2

You can upload your file and choose a cancer type for comparison.

Quantile normalization will be performed by default, based on the median value of the cancer type chosen.

Then the session will keep until you refresh the window, so you can enter one gene for multiple times without re-upload the file.

The uploaded gene expression profile should be TPM values with Hugo gene names.

We recommend you upload your expression profile processed by XENA pipeline, which is used by GEPIA.

Step 1: Choose the cancer type for comparation :

[Click here](#) to show details for cancer type information.

dataset for comparation :

normalization methods :

Step 2: Choose one file to upload

Upload

未选择任何文件

Note: You need the csv format like this.

| | A1BG | ADA | AKT3 | ZBTB11-AS1 | MED6 |
|---------|----------|----------|----------|------------|----------|
| sample1 | 5.443278 | 1.008734 | 1.008734 | 0 | 0 |
| sample2 | 0 | 8.684895 | 0 | 0 | 0 |
| sample3 | 0 | 0 | 5.439709 | 0 | 7.926596 |
| sample4 | 0 | 0 | 0 | 0 | 0 |
| sample5 | 0 | 0 | 1.362182 | 0 | 0 |

An example can be downloaded [here](#).

- 通过上传自己的RNA-seq矩阵可获得自己的数据和数据库的数据
基因表达比较情况 (相当于不需要用户自己做标准化)

基于TCGA分析的数据库：GEPIA 2

JOURNAL ARTICLE

GEPIA2021: integrating multiple deconvolution-based analysis into GEPIA ⑧

Chenwei Li, Zefang Tang, Wenjie Zhang, Zhaochen Ye, Fenglin Liu ✉ Author Notes

Nucleic Acids Research, Volume 49, Issue W1, 2 July 2021, Pages W242–W246,
<https://doi.org/10.1093/nar/gkab418>

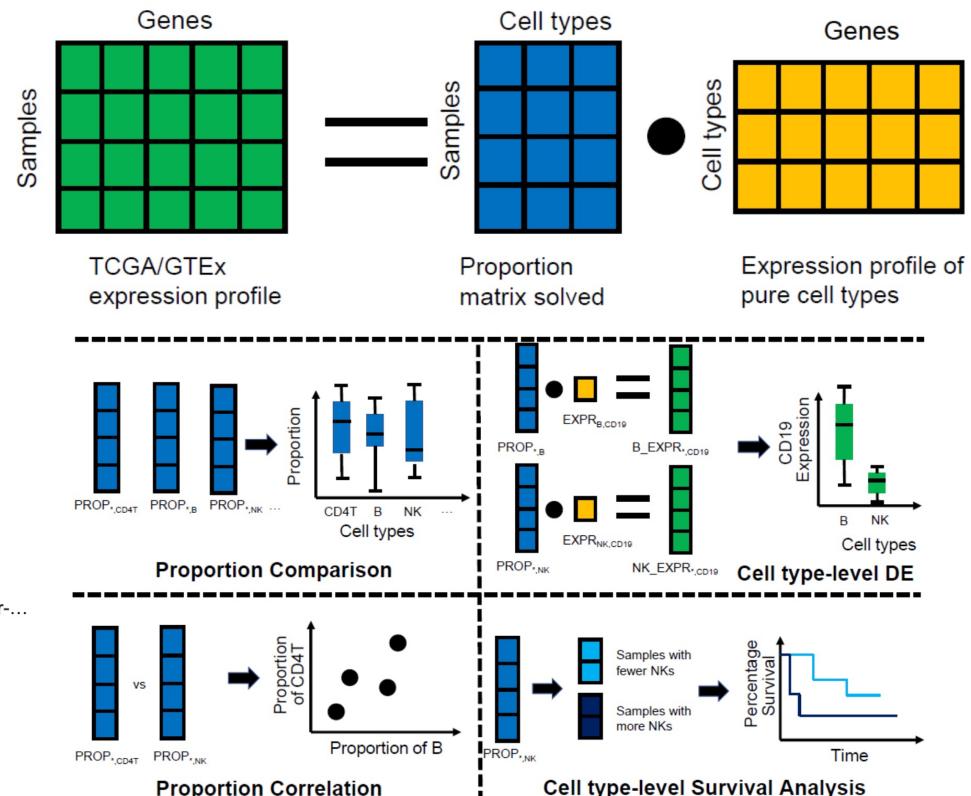
Published: 29 May 2021 Article history ▾

GEPIA2021: integrating multiple deconvolution-based analysis into GEPIA

C Li, Z Tang, W Zhang, Z Ye, F Liu - Nucleic Acids Research, 2021 - academic.oup.com

... Thus, we present GEPIA2021, a standalone extension of GEPIA, ... With GEPIA2021, experimental biologists could easily ... GEPIA2021 is publicly accessible at <http://gepia2021.cancer-pku.cn/>...

☆ 保存 引用 被引用次数: 214 相关文章 所有 9 个版本



详细信息参照 <http://gepia2021.cancer-pku.cn/>

2021年，GEPIA-2021更新发布第三版

- 主要更新基于scRNA-seq数据拆分大规模RNA-seq数据获得单细胞维度后的一系列分析(包括细胞亚型的差异分析，细胞类型的生存分析等)

课程概览

1 数据库基本概念

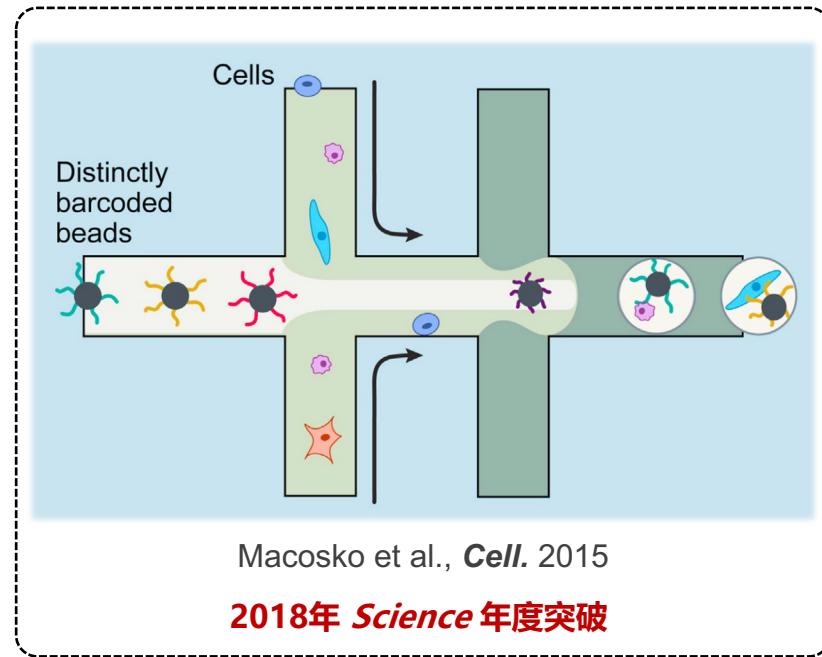
2 NCBI & EMBL-EBI & 在线分析工具

3 TCGA GTEx & GEPIA 1/2/2021

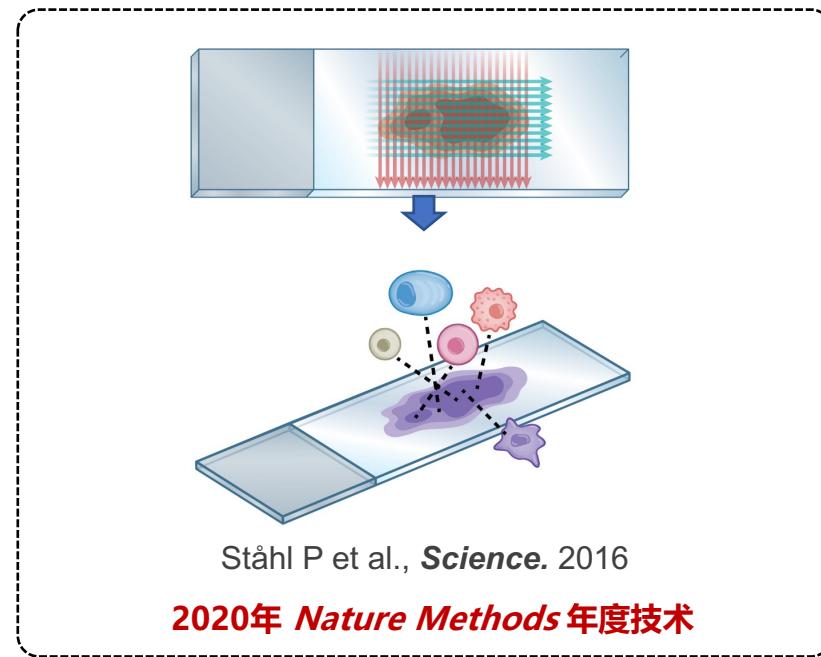
4 单细胞/空间转录组数据库

单细胞转录组(scRNA-seq) 和空间转录组(ST) 背景介绍

单细胞转录组测序技术 Single Cell RNA-seq



空间转录组测序技术 Spatial Transcriptome



- 检测**单个细胞的转录组**
- 发现稀有特殊细胞亚群
- 获得复杂组织细胞构成
- 检测组织原位的转录组
- 发现细胞**空间位置关系**
- 识别重要空间组织区域

单细胞转录组(scRNA-seq) 数据库

TISCH: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment

D Sun, J Wang, Y Han, X Dong, J Ge, R Zheng, X Shi, B Wang, Z Li, P Ren, L Sun, Y Yan...

Nucleic acids research, 2021 · academic.oup.com

Abstract

Cancer immunotherapy targeting co-inhibitory pathways by checkpoint blockade shows remarkable efficacy in a variety of cancer types. However, only a minority of patients respond to treatment due to the stochastic heterogeneity of tumor microenvironment (TME). Recent advances in single-cell RNA-seq technologies enabled comprehensive characterization of the immune system heterogeneity in tumors but posed computational challenges on integrating and utilizing the massive published datasets to inform

展开 ▾

☆ 保存 ⚙ 引用 被引用次数: 404 相关文章 所有 12 个版本



Chenfei Wang

Professor of Bioinformatics, [Tongji University](#)
在 tongji.edu.cn 的电子邮件经过验证 - 首页

Bioinformatics Artificial Intelligence Single Cell Genomics

同济大学 王晨飞教授

- 2021年对标TCGA开发了scRNA-seq技术的肿瘤疾病组学数据库 **Tumor Immune Single Cell Hub(TISCH)**
- 数据包含**27**种癌症类型, **76**个高质量肿瘤数据集, 总计**200**万个细胞的单细胞转录组图谱。

单细胞转录组(scRNA-seq) 数据库

JOURNAL ARTICLE

TISCH2: expanded datasets and new tools for single-cell transcriptome analyses of the tumor microenvironment

Ya Han, Yuting Wang, Xin Dong, Dongqing Sun, Zhaoyang Liu, Jiali Yue, Haiyun Wang,
Taiwen Li ✉, Chenfei Wang ✉ Author Notes

Nucleic Acids Research, Volume 51, Issue D1, 6 January 2023, Pages D1425–D1431,
<https://doi.org/10.1093/nar/gkac959>

Published: 02 November 2022 Article history ▾

- 2022年12月 **Tumor Immune Single Cell Hub2(TISCH2)** 第二版发布
- 数据增加至 **50种癌症类型, 190个高质量肿瘤数据集, 总计 600 万个细胞** 的单细胞转录组图谱(细胞数为之前三倍)。
- 增加了**细胞通讯分析**以及**转录因子预测的可视化分析**

<http://tisch.comp-genomics.org/>

单细胞转录组(scRNA-seq) 数据库

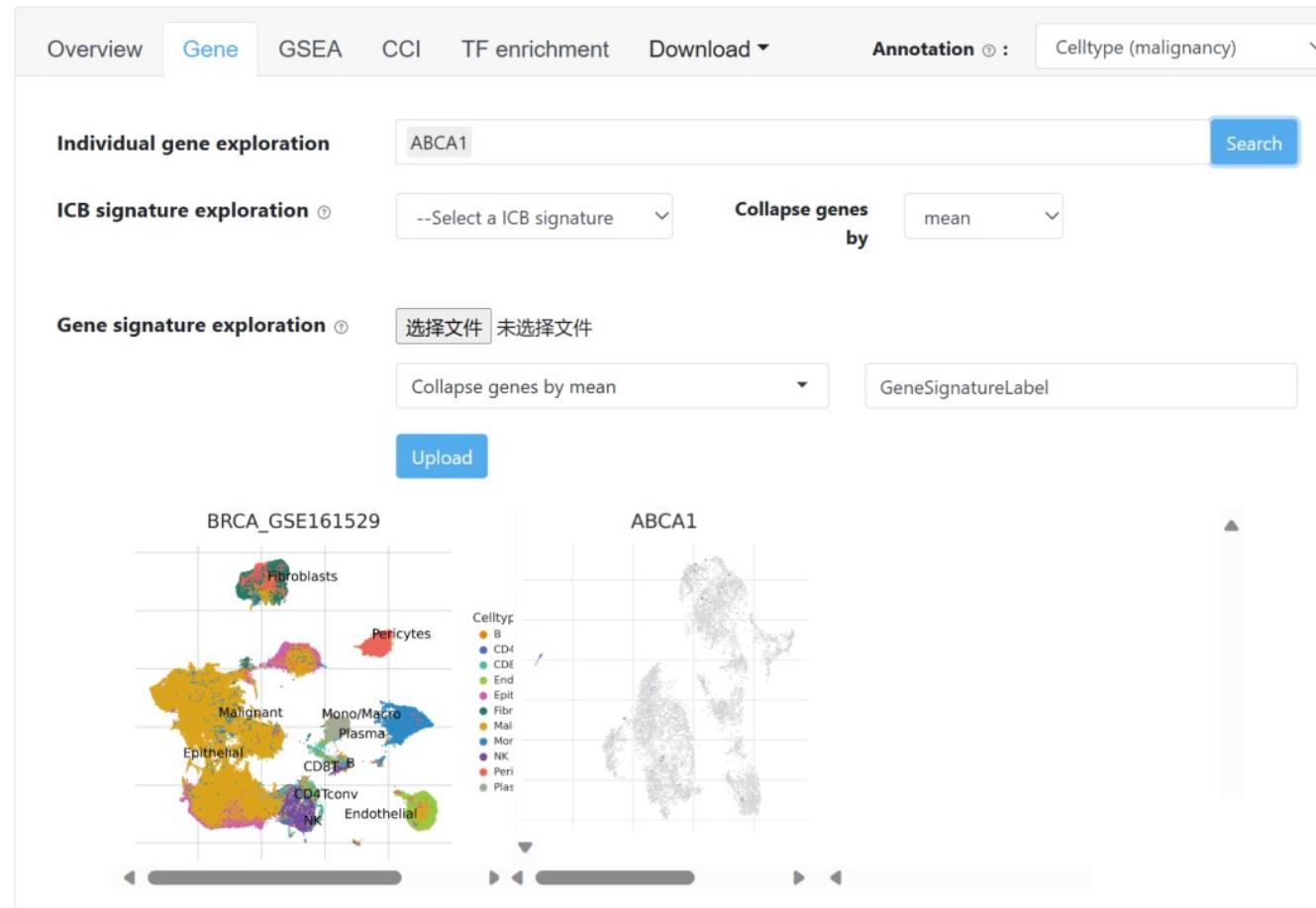
The screenshot shows the TISCH2 database homepage. At the top is a dark blue header with the logo "TISCH2" and navigation links for Home, Dataset, Gene, Documentation, Statistics, and TISCH1. Below the header is a light gray main area. In the center, it says "Welcome to TISCH2" with "BETA" and "190 datasets 6,297,320 cells" nearby. Below this is a search bar with the placeholder "Search a gene of interest" and a blue "Explore" button. To the left of the search bar is a stylized human body diagram highlighting internal organs. To the right is a grid of circular icons representing various tumor types, each with its name and dataset count:

| Tumor Type | Dataset Count |
|----------------|--|
| Bladder | BLCA (3) |
| Blood | AEL (1) AML (5) PBMC (3) ALL (3) CLL (6) |
| Bone | MM (2) GCTB (1) OS (2) |
| Brain | Glioma (17) MB (2) |
| Breast | BRCA (12) |
| Colorectum | CRC (11) |
| Esophagus | ESCA (3) |
| Eye | RB (1) UVM (4) |
| Head & Neck | HNSC (10) |
| Kidney | KIRC (6) KIRP (1) KIPAN (2) |
| Liver | CHOL (3) LIHC (8) HB (1) |
| Lung | NSCLC (17) SCLC (1) |
| Lymph node | NHL (2) DLBC (1) |
| Nervous system | MPNST (1) NET (1) NB (1) Neurofibroma (1) |
| Pancreas | PAAD (9) |
| Pelvic cavity | ESCC (1) OV (9) UCEC (2) |

- 数据库主页显示支持在线分析的肿瘤类型种类(命名与TCGA一致)

<http://tisch.comp-genomics.org/>

单细胞转录组(scRNA-seq) 数据库

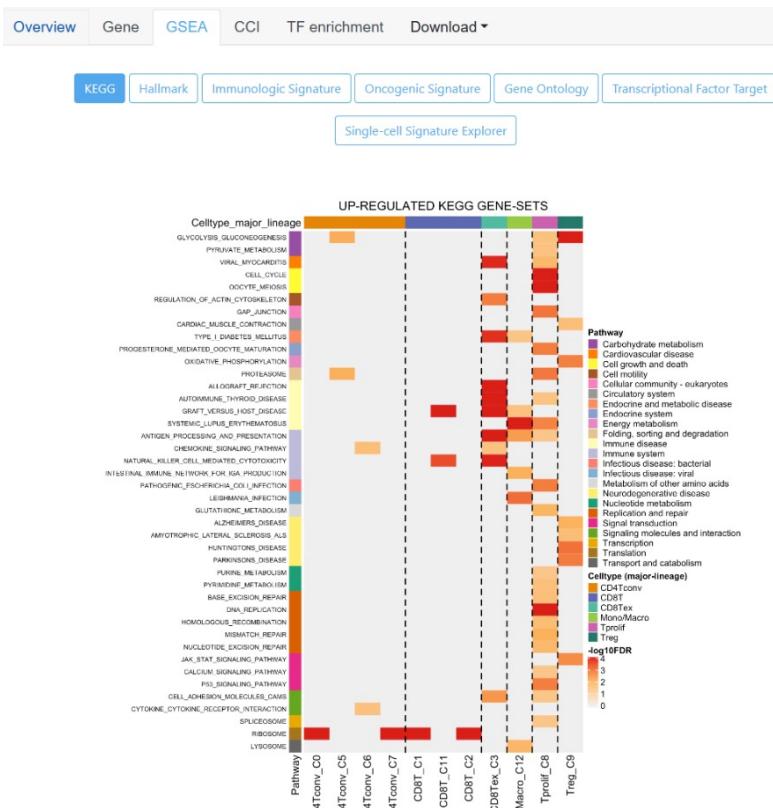


- 支持查看**自定义基因**在单细胞数据中**分群结果和基因表达情况**

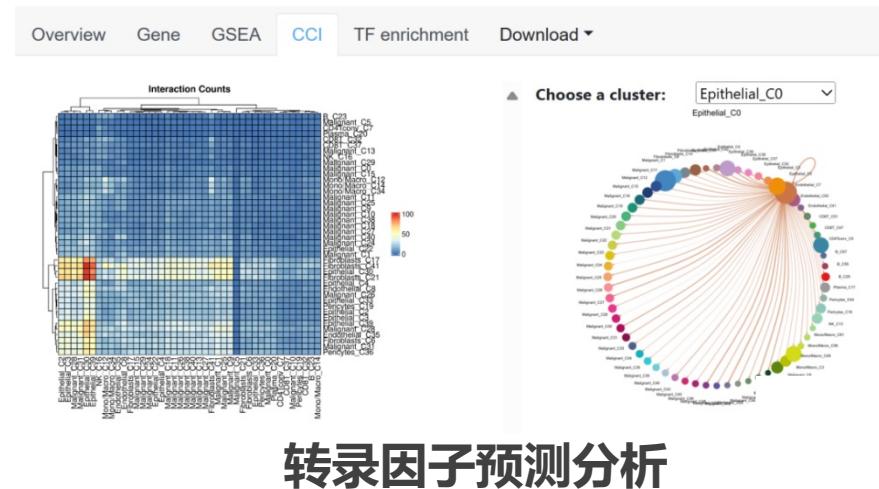
<http://tisch.comp-genomics.org/>

单细胞转录组(scRNA-seq) 数据库

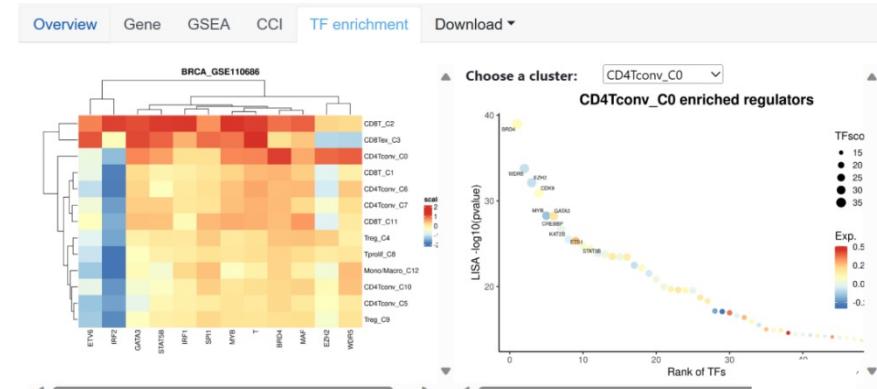
GSEA 通路分析



细胞互作分析



转录因子预测分析



- 支持查看在单细胞数据中通路富集分析、细胞互作分析以及TF预测分析

单细胞转录组(scRNA-seq) 数据库



- 来自美国博德研究所(隶属于MIT和哈佛大学)建立了更加全面scRNA-seq检索收录数据库
- 数据搜集包括**618**项研究中的**3700W+**细胞
- 支持**单个**研究的数据对应查看基因表达情况，缺少大规模整合

https://singlecell.broadinstitute.org/single_cell

空间转录组(ST) 数据库

nature methods

Resource

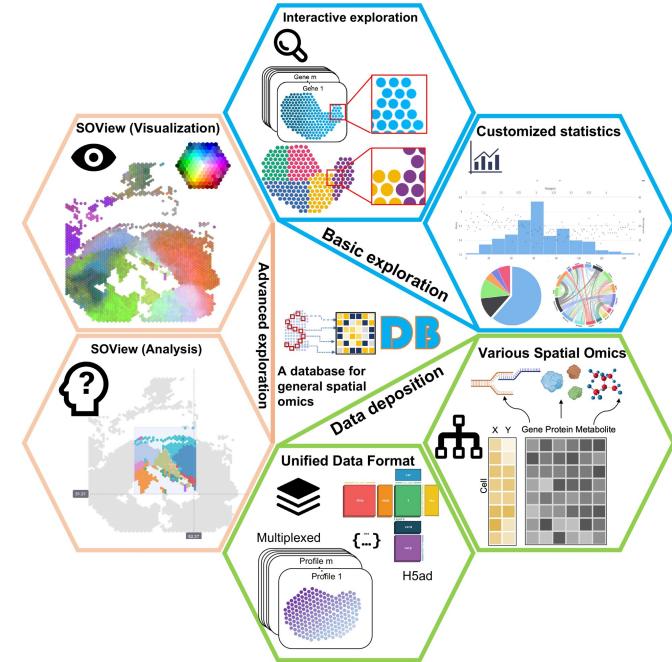
<https://doi.org/10.1038/s41592-023-01773-7>

SODB facilitates comprehensive exploration of spatial omics data

Received: 10 August 2022

Accepted: 6 January 2023

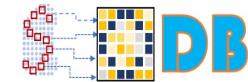
Zhiyuan Yuan^{①,2,7}✉, Wentao Pan^{2,3,7}, Xuan Zhao², Fangyuan Zhao^{4,5},
Zhimeng Xu², Xiu Li^③, Yi Zhao^{④,5}, Michael Q. Zhang^⑥✉ & Jianhua Yao²✉



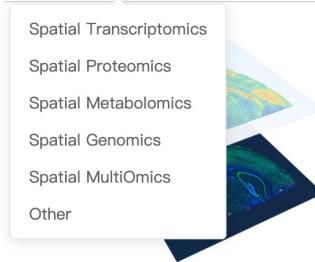
- 2023年1月 复旦大学类脑研究院联合德州大学达拉斯分校以及腾讯 AI lab 联合开发了空间转录组第一个公开发表的大规模数据库发表于Nature Methods
- 收录了26种空间组学技术, 细胞数5000W+, 同时开发优化了超大数据的内存读取效率工具包-pysodb, 时间效率和内存提升均在百倍以上。
- 支持多种在线数据分析和可视化模块, 包括基因空间表达、细胞类型注释、基因表达比较等

空间转录组(ST) 数据库

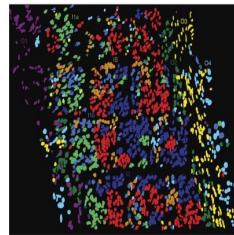
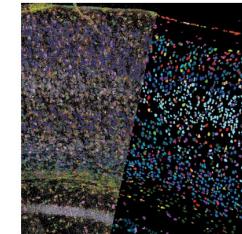
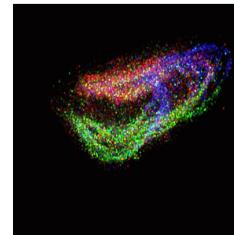
INTRODUCTION **BROWSE** SEARCH STATISTICS TUTORIAL NEWS



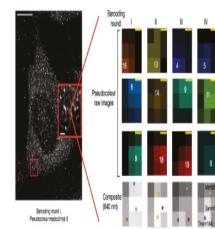
Biotech Categories ▾ Spatial Transcriptomics



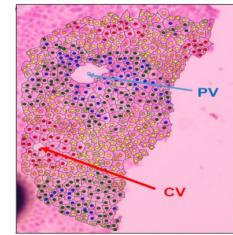
10X Visium



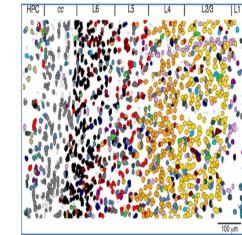
seqFISH



seqFISH+



seqScope



STARmap

<https://gene.ai.tencent.com/SpatialOmics/>

- 可以通过选择技术类型以及单个研究数据进行查看(类似于singlecell portal)
- 除了空间转录组数据，数据库同时支持其他空间代谢组，空间蛋白组等数据

空间转录组(ST) 数据库

Biotech Categories ▾ Spatial Transcriptomics / MERFISH

Date To Search

Country Select

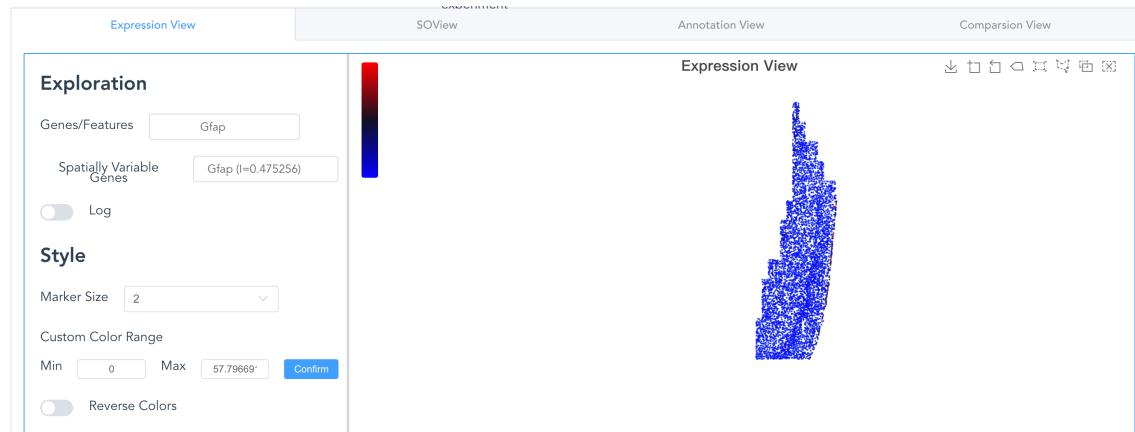
Journal Select

Tissue Select

Species Select

| Dataset Name | Date | doi | Country | Operations |
|--|------|------------------------------|---------|-------------------------------------|
| Shang2021spatially | 2021 | 10.1038/s41586-021-03705-x | USA | <input type="button" value="View"/> |
| Decoding molecular and cellular heterogeneity of mouse nucleus accumbens | 2021 | 10.1038/s41593-021-00938-x | USA | <input type="button" value="View"/> |
| Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression | 2019 | 10.1073/pnas.1912459116 | USA | <input type="button" value="View"/> |
| Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region | 2018 | 10.1126/science.aau5324 | USA | <input type="button" value="View"/> |
| Conservation and divergence of cortical cell organization in human and mouse revealed by MERFISH | 2022 | DOI: 10.1126/science.abn1741 | USA | <input type="button" value="View"/> |
| Merfish_Visp | 0 | None | None | <input type="button" value="View"/> |
| Molecular and spatial signatures of mouse brain aging at single-cell resolution | 2023 | 10.1016/j.cell.2022.12.010 | USA | <input type="button" value="View"/> |

- 支持选择近年来不同研究的数据进行单独查看



<https://gene.ai.tencent.com/SpatialOmics/> Gfap基因 Merfish技术

- 可以通过左边选择对应的可视化参数查看不同基因的空间表达以及细胞注释

空间转录组(ST) 数据库

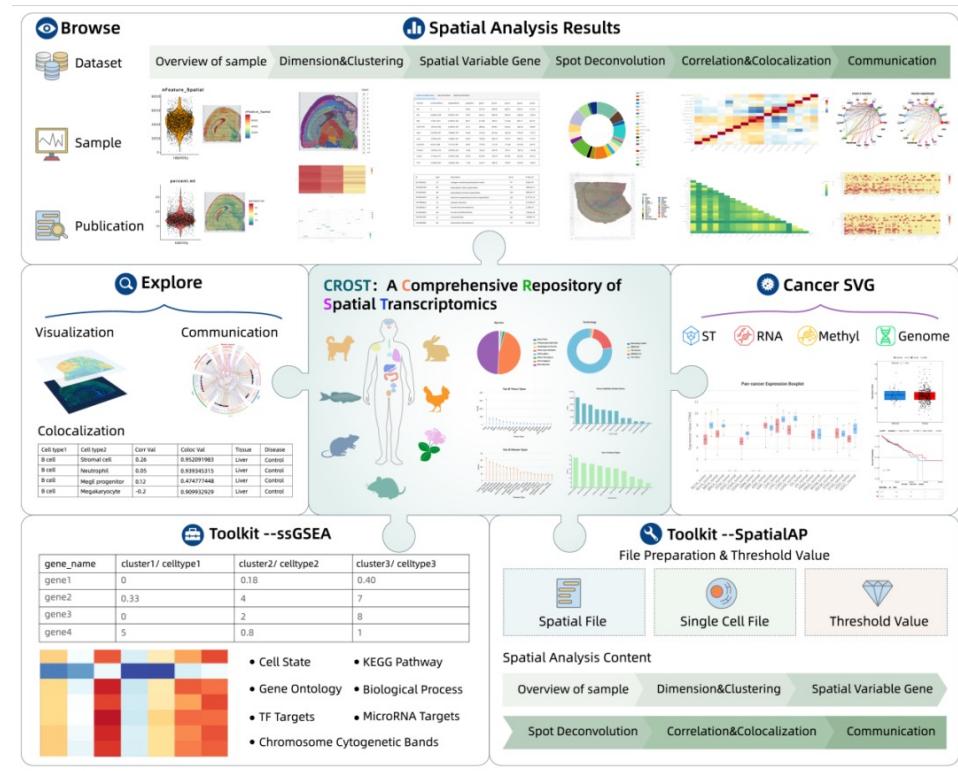
JOURNAL ARTICLE

CROST: a comprehensive repository of spatial transcriptomics

Guoliang Wang, Song Wu, Zhuang Xiong, Hongzhu Qu ✉, Xiangdong Fang ✉,
Yiming Bao ✉ Author Notes

Nucleic Acids Research, gkad782, <https://doi.org/10.1093/nar/gkad782>

Published: 04 October 2023 Article history ▾



- 2023年10月 北京基因组所国家基因组科学数据中心鲍一明和方向东研究员开发了更大规模的空间转录组数据库，发表于NAR
- 收录了8类物种, 35个组织类型, 56种疾病, 182个高质量空间转录组数据集

空间转录组(ST) 数据库

The screenshot shows the homepage of the CROST database. At the top, there is a navigation bar with links for Data Resources, Computing Analysis, Data Network, and Standards. Below the navigation bar is a main menu with Home, Browse, Cancer SVG, Explore, Tools, Download, Statistics, and Help. The central header reads "CROST: A Comprehensive Repository of Spatial Transcriptomics". Below the header is a search bar with a "Global Search" dropdown and a placeholder "Please input 1 query field. (e.g. VISDP000XXX, VISDS000XXX, Brain, Cancer, S100A9, Naive CD4+ T;)" followed by a search button. A note below the search bar provides examples of valid queries. The page is divided into several sections: "Highlight" (describing single-sample analysis, interactive visualization, multi-omics integration), "Statistics" (showing 182 Datasets, 1,033 Samples, 5 Technologies, 83 Publications, 8 Species, 56 Diseases, 33 Cancer Types, and 48,043 Cancer SVG), "Explore" (listing Spatial Visualization and Spatial Colocalization), "Tools" (listing ssGSEA, Single Sample GSEA, SpatialAP, and Spatial Analysis Pipeline), "Release Note" (listing two recent updates), and "Related Links" (listing STomicsDB, SGBDB, and AQUILA). The "Related Links" section is highlighted with a red border.

CROST offers single-sample analysis, interactive visualization, multi-omics integration for exploring cancer svg, and other powerful tools for spatial transcriptomics researches.

- ❖ CROST stored 1033 samples, which include expression and spatial coordinate information;
- ❖ CROST implemented a standardized analysis pipeline that leverages raw

182 Datasets 1,033 Samples 5 Technologies 83 Publications

8 Species 56 Diseases 33 Cancer Types 48,043 Cancer SVG

Explore Tools Release Note Related Links

- Add 123 new samples (2023-07-12)
- Bugs Fix and code clean (2023-06-24)

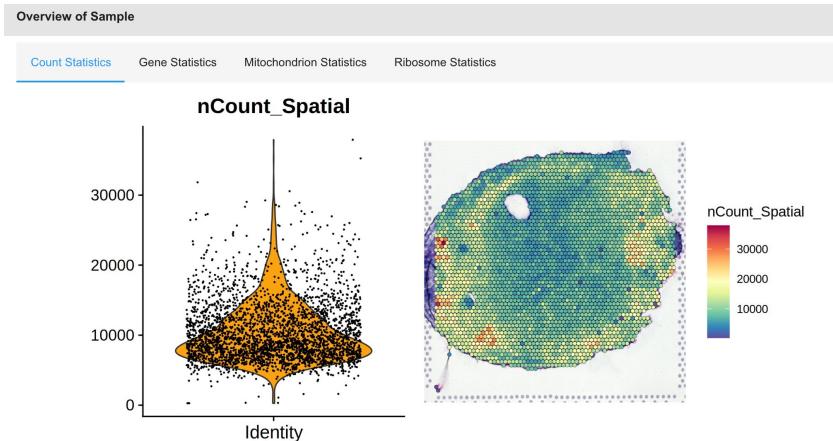
STomicsDB
SGBDB
AQUILA

- 该数据库质量和可操作性很高，用户友好度高，同时兼有相关数据库链接

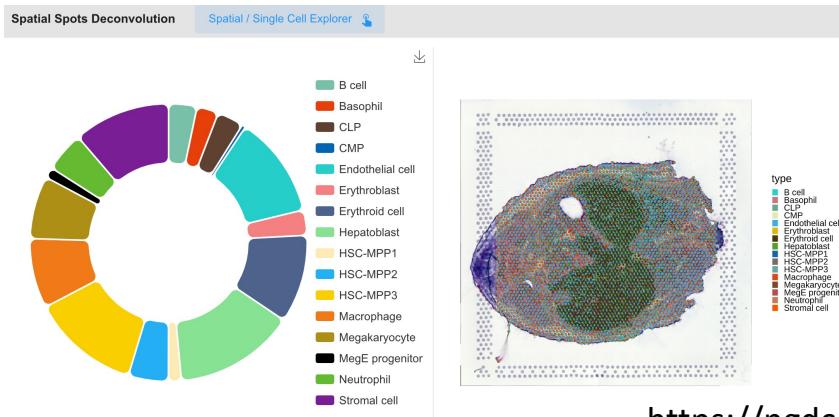
<https://ngdc.cncb.ac.cn/crost/>

空间转录组(ST) 数据库

基因查看空间分布

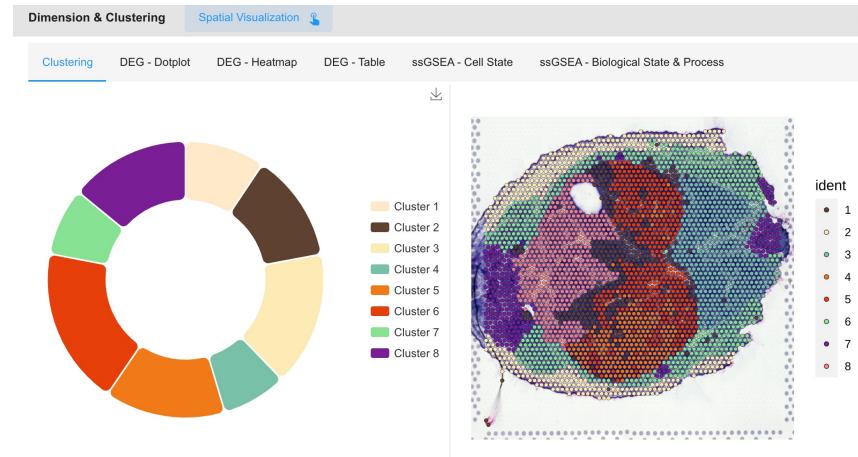


空间去卷积分析

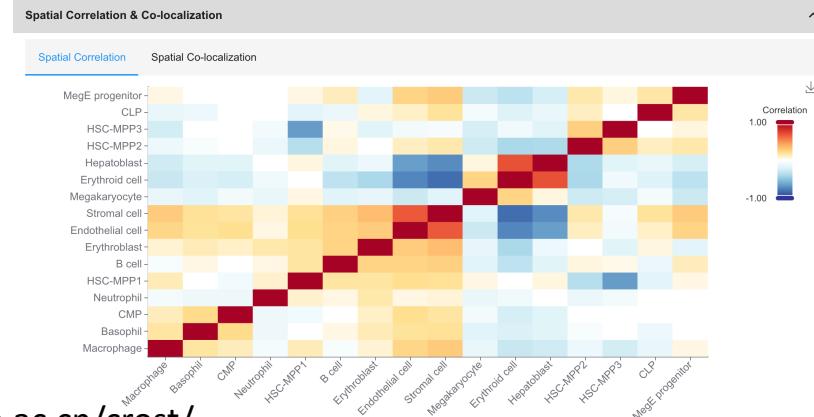


<https://ngdc.cncb.ac.cn/crost/>

分群与差异基因分析



空间共定位分析



- 支持多种在线分析方法，通过选择某一个研究的数据继而可以进行基因查看，分群和差异基因分析，空间去卷积，共定位，细胞互作等分析

空间转录组(ST) 数据库

The screenshot shows the homepage of the STOmics DB website. The header features the logo 'STOmicsDB' with a blue square icon. The main navigation menu includes 'Home', 'Datasets', 'Resources', and a dropdown menu. Below the header, the title 'STOmics DB' is prominently displayed. A brief description states: 'Spatial Transcript Omics DataBase (STOmics DB) is a comprehensive repository of literature and Datasets related to spatial transcriptomics topics, and provides convenient tools for Data analysis, search and visualization.' At the bottom of the page is a search bar with the placeholder 'Search' and a blue 'Search' button.

STOmicsDB: a database of spatial transcriptomic data

Z Xu, W Wang, T Yang, J Chen, Y Huang, J Gould, W Du, F Yang, L Li, T Lai, C Hua, S Hu...
bioRxiv, 2022 • biorxiv.org

Abstract

Recent technological development in spatial transcriptomics allows researchers to measure gene expression of cells and their spatial locations at the almost single-cell level, which generates detailed biological insight into biological processes. However, specialized spatial transcriptomics databases are rare. Here, we present the Spatial TranscriptOomics DataBase (STOmicsDB), a user-friendly database with multifunctions including search of relevant publications and tools, public dataset visualization, customized specialized

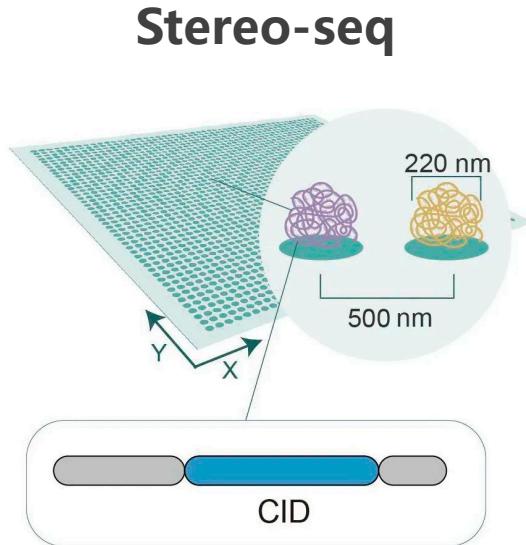
展开 ▾

☆ 保存 引用 被引用次数: 12 相关文章 所有 2 个版本 ⟲

<https://db.cngb.org/stomics/>

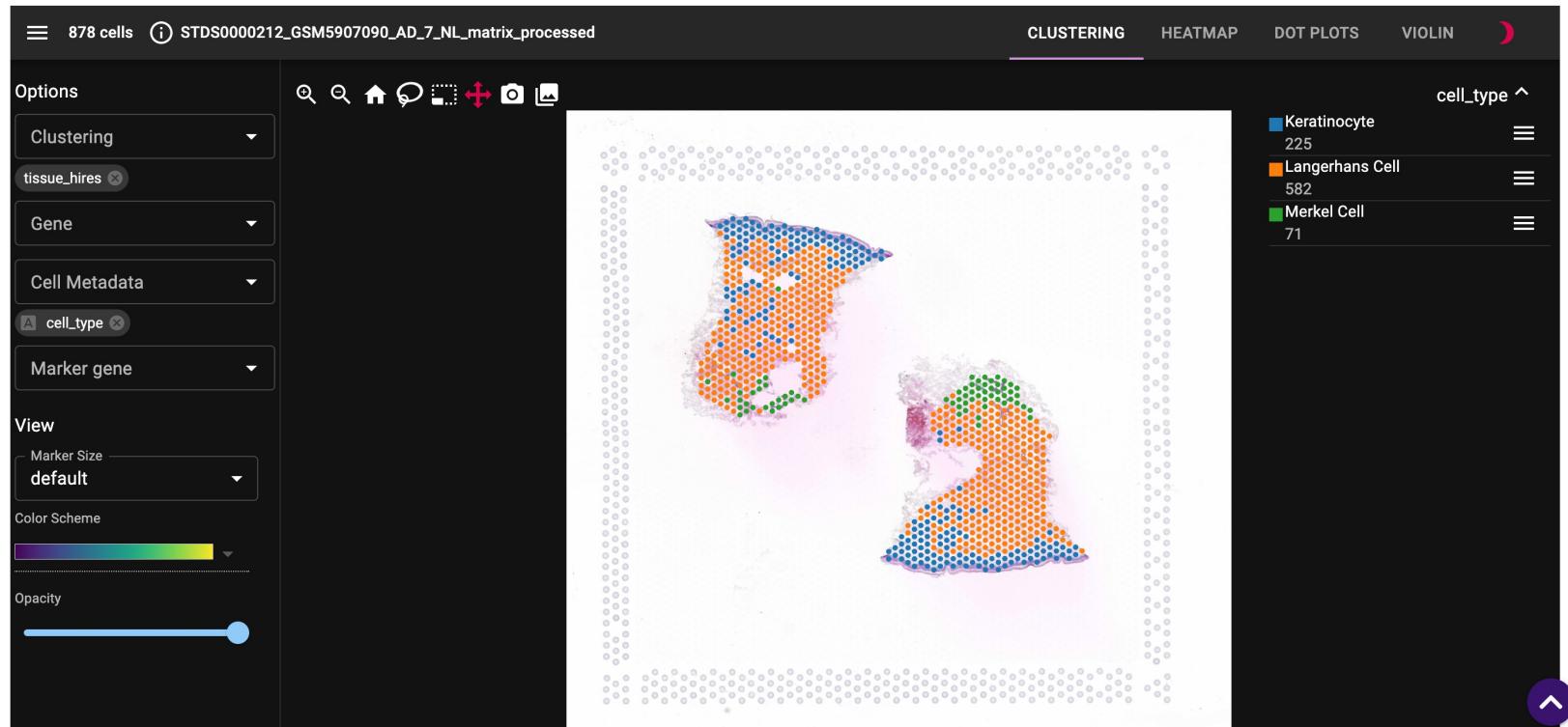
- 2022年 华大基因已经着手开发空间转录组大规模数据库，目前发表于预印本上，尚未正式发表
- 截止目前收录了17类物种，128个组织类型，218个高质量空间转录组数据集

空间转录组(ST) 数据库



- 近年来华大基因自主研发的空间转录组技术Stereo-seq 领域内影响很大，其多个图谱类工作均发表在CNS上(小鼠胚胎发育Cell, 蝾螈脑Science)，并且还有很多子刊
- STOmicsDB相比于其他数据库更多的亮点在于其自研技术衍生数据的展示

空间转录组(ST) 数据库



<https://db.cngb.org/stomics/>

- 支持多种在线分析方法，通过选择某一个研究的数据可以进行基因可视化交互
- 其他进阶诸如差异基因分析，空间互作等分析为结果展示，暂不支持交互展示

课程总结

1 数据库基本概念

2 NCBI & EMBL-EBI & GO

3 TCGA GTEx & GEPIA 1/2/2021

4 单细胞/空间转录组数据库