Nucleic Acids Research, 2023, 1–8
https://doi.org/10.1093/nar/gkad781
Database issue

scQTLbase: an integrated human single-cell eQTL database

Ruofan Ding [1,†], Qixuan Wang [1,†], Lihai Gong [1,†], Ting Zhang [1], Xudong Zou [1], Kewei Xiong [1], Qi Liao [2], Mireya Plass [3,4] and Lei Li [1,*]

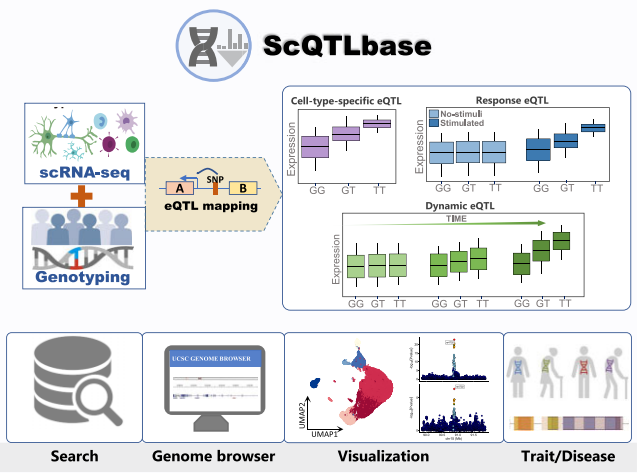
[1]Institute of Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen 518055, China
[2]School of Public Health, Health Science Center, Ningbo University, Ningbo 315211, China
[3]Gene Regulation of Cell Identity Group, Regenerative Medicine Program, Bellvitge Institute for Biomedical Research (IDIBELL), and Program for Advancing Clinical Translation of Regenerative Medicine of Catalonia, P-CMR[C], L'Hospitalet de Llobregat, Barcelona, Spain
[4]Center for Networked Biomedical Research on Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Madrid, Spain

*To whom correspondence should be addressed. Tel: +86 0755 2684 9284; Email: Lei.Li@szbl.ac.cn
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Abstract

Genome-wide association studies (GWAS) have identified numerous genetic variants associated with diseases and traits. However, the functional interpretation of these variants remains challenging. Expression quantitative trait loci (eQTLs) have been widely used to identify mutations linked to disease, yet they explain only 20–50% of disease-related variants. Single-cell eQTLs (sc-eQTLs) studies provide an immense opportunity to identify new disease risk genes with expanded eQTL scales and transcriptional regulation at a much finer resolution. However, there is no comprehensive database dedicated to single-cell eQTLs that users can use to search, analyse and visualize them. Therefore, we developed the scQTLbase (http://bioinfo.szbl.ac.cn/scQTLbase), the first integrated human sc-eQTLs portal, featuring 304 datasets spanning 57 cell types and 95 cell states. It contains ∼16 million SNPs significantly associated with cell-type/state gene expression and ∼0.69 million disease-associated sc-eQTLs from 3 333 traits/diseases. In addition, scQTLbase offers sc-eQTL search, gene expression visualization in UMAP plots, a genome browser, and colocalization visualization based on the GWAS dataset of interest. scQTLbase provides a one-stop portal for sc-eQTLs that will significantly advance the discovery of disease susceptibility genes.

Graphical abstract



Introduction

Genome-wide association studies (GWAS) have successfully identified numerous genetic variants linked to various diseases and traits. However, the functional characterization of these variants remains limited and challenging (1). To elucidate the mechanism linking genetic variants to diseases, expression quantitative trait loci (eQTLs) analysis has been widely employed. This analysis helps identifying causal variants that co-localize with GWAS variants, providing explanations for some of them (2,3). A particularly noteworthy effort is the Genotype-Tissue Expression Consortium, which has extensively mapped eQTLs across 49 tissues from 838 donors (4). Another significant work, the eQTL Catalogue, which compiles re-computed eQTLs from 21 studies encompassing 69 distinct cell types and tissues (5). However, traditional bulk studies typically assess average expression levels across

Received: July 30, 2023. **Revised:** August 24, 2023. **Editorial Decision:** September 9, 2023. **Accepted:** September 15, 2023

millions of cells from whole tissues or samples, potentially obscuring biologically regulatory relationships that are specific to certain cell types (6,7). These limitations have significantly restricted the applicability of bulk eQTLs in understanding the biology of disease-associated variants, resulting in only 20–50% of common disease alleles colocalize with eQTLs (6).

To elucidate the connections between gene regulation and disease, single-cell eQTLs (sc-eQTLs) have emerged as a valuable resource. By profiling transcriptomes at the individual cell level, sc-eQTL analysis minimizes cell dilution and enables studying regulatory relationships at a higher resolution (8–10). Various types of sc-eQTLs have been discovered and analysed in different contexts, including cell-type-specific eQTLs (genetic variants that are associated with the expression of specific cell types), dynamic eQTLs (genetic variants that are associated with changes in gene expression over continuous or time), response eQTLs (genetic variants that are associated with changes in gene expression in response to stimuli) (6,11). For instance, in 2018, van der Wijst *et al.* identified single-cell eQTLs in ∼25 000 peripheral blood mononuclear cells (PMBCs) from 45 individuals (10). Cuomo *et al.* explored dynamic eQTLs during iPSC differentiation toward definitive endoderm using 125 individuals across four time points (12). Oelen *et al.* found that response eQTLs were typically more cell-type specific than pathogen-specific when exposing 1.3 million peripheral blood mononuclear cells (PBMCs) *in vitro* to *Mycobacterium tuberculosis*, *Candida albicans* and *Pseudomonas aeruginosa* (13). In addition, Sarkar *et al.* investigated variance eQTLs by analysing cell-to-cell gene expression variability in 5 447 induced pluripotent stem cells (iPSCs) from 53 Yoruba individuals (14). Over a dozen sc-eQTL studies have been published so far (11), and large collaborative efforts, such as the eQTLGen Consortium and the DICE project, have provided valuable resources and greatly facilitated our understanding of how genetic variants influence gene regulation at the cellular level (15,16). Despite these advancements, there is still a lack of a comprehensive database that integrates available sc-eQTLs from various studies, enabling seamless querying, browsing, downloading, and visualization at the cellular level. Establishing such a comprehensive large-scale sc-eQTL database would create an atlas of disease-related variants that were overlooked by traditional bulk studies, allowing for a more powerful and in-depth analysis of GWAS loci. Hence, integrating large-scale sc-eQTL datasets is crucial for identifying causal variants and advancing our understanding of gene regulation in the context of human disease.

In this work, we conducted a systematic collection of summary-level results from 16 published studies and manually curated 304 independent sc-eQTL datasets spanning 57 cell types and 95 cell states/simulations following a standardized approach. Through the implementation of a highly interactive web interface, we developed a versatile database named scQTLbase (http://bioinfo.szbl.ac.cn/scQTLbase), representing the first sc-eQTL database offering convenient cataloging, searching, browsing, downloading, and visualization of single-cell regulatory data. scQTLbase also serves as an open resource specifically designed for meta-analyses, revisiting GWAS findings, and deciphering disease-associated risk SNPs at single-cell resolution. By providing an accelerated platform for interpreting the mechanisms underlying the relationship between genetic variants and diseases, scQTLbase significantly contributes to our understanding of complex disease biology.

Materials and methods

Data collection and processing

We conducted a manual collection of single-cell eQTL datasets from published literature, adhering to a predefined set of rules: (i) the selected studies utilized actual samples from a diverse range of biological contexts, including normal, treated or disease conditions. We excluded datasets from meta-analysis or secondary analyses. To ensure sufficient power for sc-eQTLs, we included only studies with a minimum of 40 samples or 5 000 cells. (ii) We included only single-cell transcriptome data generated using well-established and reliable sequencing technologies such as 10×, Smart-seq/Smart-seq2 and CITE-seq. (iii) We considered both eQTL datasets derived from genome-wide primary eQTL mapping and local-region eQTL mapping. After applying these stringent inclusion criteria and thoroughly reviewing the literature, we identified and retained 16 studies in our database (Table 1). There are two types of data obtained from the collected literature, including sc-eQTL summary statistics data and gene expression. The summary statistics data were preprocessed by collecting the following information: genetic variant (identified by dbSNP rsID or genomic coordinates), the genes that are affected by the variant, the strength and direction (effect size) of the association between the variant and gene expression, and statistical measures including *P*-values and standard error. Gene expression data are preprocessed into a 2D matrix where each row corresponds to a gene, and each column corresponds to a cell barcode. The values in the matrix represent the expression level of each gene in each cell, typically measured as normalized expression values.

Variant normalization

Variants with a dbSNP identifier were harmonized based on dbSNP build 151 (17) and aligned to the hg38 human genome version. During this process, positional information, as well as the alternative and reference alleles, was extracted and recorded for each variant. By aligning variants to a standardized reference, scQTLbase ensures consistency and simplifies cross-study and cross-dataset comparisons of variants. Moreover, the effective allele of each sc-eQTL is preserved from its original publication.

Statistical value normalization

The sc-eQTL summary statistics dataset contains essential statistical values, including *P*-value, effect size (beta), standard error and false discovery rate. In instances where specific statistical values are not available in the original dataset, we calculate the missing value based on other additional information. For instance, if *P*-values are not provided, we derive them from *z*-scores, and in cases where standard error values are missing, we calculate them based on available *P*-values and effect sizes.

Cell type mapping

To address the challenge of inconsistent cell type names across different studies and datasets, we obtained expert-annotated cell type reference datasets from Azimuth (18). Subsequently, we conducted a manual review of each cell type name, accurately mapping them to the appropriate reference label. This process enabled us to align cell types with multiple names in different studies to their corresponding references. For instance, 't4', 'CD4T' and 'CD4' are

Table 1. 16 sc-eQTL studies collected in scQTLbase

Study	Number of cells	Number of donors	Type
Bryois-2022-Nat. Neurosci.	36 000	192 individuals	Genome-wide
Jerber-2021-Nat. Genet.	>1 million	215 iPSC lines	Genome-wide
Nathan-2022-Nature	>500 000	259 individuals	Genome-wide
Natri-2023-bioRxiv	437 618	116 individuals	Genome-wide
Neavin-2021-Genome Biol.	83 985	79 fibroblast and 31 iPSC	Genome-wide
Oelen-2022-Nat. Commun.	1.3 million	120 individuals	Genome-wide
Perez-2022-Science	>1.2 million	162 SLE and 99 healthy	Genome-wide
Randolph-2021-Science	0.2 million	89 individuals	Genome-wide
Resztak-2022-Genome Res.	292 394	96 Individuals	Genome-wide
Schmiedel-2022-Sci. Immunol.	>1 million	89 individuals	Genome-wide
Soskic-2022-Nat. Genet.	655 349	119 individuals	Genome-wide
YAZAR-2022-Science	1.27 million	982 individuals	Genome-wide
Cuomo-2020-Nat. Commun.	36 044	125 individuals	Genome-wide
Van der Wijst-2018-Nat. Genet.	25 000	45 individuals	Genome-wide
Kang-2017-Nat. Biotechnol.	22 000	23 individuals	Genome-wide
Wills-2013-Nat. Biotechnol.	1440	15 individuals	Targeted gene sets

normalized to 'CD4 T'. We also named cell types using both their full names and common abbreviations, if available. For example, 'monocyte' and 'Mono' are normalized to 'Monocyte (Mono)'. Furthermore, for cell types not present in the reference dataset, such as 'iPSC', we have standardized them to 'Induced pluripotent stem cells (iPSC)'. Altogether, we standardized 154 previously non-unified cell types to 57 canonical cell types (Supplementary Table S1). To provide more details about cell types, we have linked cell type names with the cell IDs from Cell Ontology (https://www.ebi.ac.uk/ols/ontologies/cl), a widely used database for cell type annotation (19). For instance, 'CD4+ T cell' cell type is now linked with 'CL_0000624'. To facilitate easy access to additional information, we have added hyperlinks to each cell type name in the search result table. These hyperlinks direct users to dedicated pages with more comprehensive descriptions of the cell types, including details on cell types, origins, biological processes, and function roles.

Database design

The scQTLbase platform was developed using a flask-based web framework, ensuring efficient data storage and retrieval. To ensure retrieval speed, we stored gene expression profile and genotype data as flat files, while eQTL data was stored in a MySQL database. For data analyses and visualization, we leveraged our recently developed xQTLbiolinks (20) and in-house R scripts. Interactive web pages were developed using a combination of HTML, CSS, JavaScript, and Python languages along with various JavaScript libraries, including react-ploty.js and IGV.js, and the widely-used react-TDesign component library for building interactive websites. To provide users with a seamless experience, we implemented a highly interactive UMAP plot for directly viewing specific cell types of interest. Additionally, we incorporated an interface for GWAS-sc-eQTL colocalization to visualize colocalized signals. The scQTLbase platform is freely accessible online, and there is no requirement for registration or login. For optimal performance, we recommend using Chrome as the web browser.

Visualization of UMAP panel

In order to visualize single-cell dataset, we followed two approaches: first, we utilized the principal components (PCs) provided in the literature, and if not available, we performed principal component analysis (PCA) to derive the top 20 PCs. Subsequently, we employed Uniform Manifold Approximation and Projection (UMAP) (21) to project the PCs into a lower-dimensional space, positioning the cells based on UMAP dimensions and color-coding them to represent specific cell types or clusters. Additionally, we identified highly variable expression genes using Seurat, and users now have the option to select genes of interest from a drop-down box, enabling the visualization of their expression patterns across different cell types.

Marker gene and eGene

To determine marker genes for UMAP clusters in each study, we first filtered genes that are expressed in at least 25% of cells in both the target cell cluster and all other cells. Then, we performed differential expression analysis between each cell type and all other cells. Marker genes are identified with $\log_2$(fold change) > 1 and adjusted P-value < 0.05. We identified eGenes whose expression level has been associated with at least one genetic variation at a specific genetic locus.

Identification of trait-associated sc-eQTLs

To identify potential trait-associated sc-eQTLs, we first employed a Bayesian co-localization approach using the coloc package (22). We collected 26 GWAS datasets corresponding to the populations and tissues of sc-eQTL studies included in our database (Supplementary Table S2). For each GWAS, we determined independent loci by extracting variants with at least genome-wide significance (P-value $< 5 \times 10^{-8}$) and located at least 1 Mb away from all other variants with higher statistical significance. Subsequently, we extracted a list of all eGenes (genes that have at least one significant sc-eQTL) within 1 Mb of each genome-wide significant variant for further colocalization analyses. To ensure the robustness of our results, for each eGene we excluded any variants lacking both eQTL and GWAS association statistics, including effect size estimate, standard error, and P-value. Additionally, colocalization analyses were conducted for matched tissues of GWAS and sc-eQTLs using coloc with default parameters (19). A region or eGene was considered to show evidence of co-localization when the region- or gene-based posterior probability of co-localization (PP4) was >0.75.

Next, we gathered significant SNPs from 3 333 GWAS datasets (Supplementary Table S3) collected in NHGRI GWAS catalog (1) and filtered out those without dbSNP identifiers.

Subsequently, we retrieved significant sc-eQTLs from scQTLbase and assessed linkage disequilibrium (LD) in the corresponding population between GWAS lead SNPs and sc-eQTL variants. We define the lead SNP as the SNP within a locus having the lowest *P*-value. sc-eQTLs with an R2 (a measure of linkage disequilibrium between alleles at two loci) value ≥0.8, which were found to be in strong LD with a GWAS SNP, were considered as overlapping with disease-associated loci.

Results

Data summary of scQTLbase

The scQTLbase database provides a comprehensive summary of sc-eQTL studies conducted up to June 2023. These studies have expanded to include hundreds or even thousands of individuals across various tissues and cell types (13). By manual curation and processing, scQTLbase has successfully incorporated a total of 16 single-cell eQTL studies, including an impressive number of ∼2 750 individuals and approximately 8.04 million cells (Table 1). Notably, blood samples have been predominantly used due to their ease of obtainment and well-established protocols for isolating high-quality single cells. As a result, many investigations have focused on PBMCs to explore cellular genetic effects, while other studies have delved into samples like iPSCs, brain and lung, encompassing a diverse range of cell types. For example, PBMCs comprise various immune cells, red blood cells and platelets, whereas iPSCs possess the potential to differentiate into multiple cell types, including neurons, cardiomyocytes and hepatocytes (23). In scQTLbase, a total of 57 cell types have been included through manual curation. Additionally, scQTLbase encompasses 95 distinct cell states (Figure 1), representing a diverse range of cellular contexts, such as gene exposure (e.g. 'SIX5+' in fibroblasts, 'CD45RA+' in CD4+ Effector Memory T cells), cell proliferation at different time points (e.g. 'day 11', 'day 30' in iPSCs), and cell differentiation at various stages (e.g. 'Pulmonary alveolar type I' and 'Pulmonary alveolar type II' in Epithelial cells). Furthermore, the scQTLbase database includes three main types of eQTLs: cell-type-specific eQTLs (63.82%), response eQTLs (24.34%), and dynamic eQTLs (5.26%). The database also covers 6.58% variance eQTLs.

Web design and interface

scQTLbase incorporates ∼16 million SNPs significantly associated with gene expression at the cell-type or cell-state level. The database's homepage offers a concise summary and detailed description of its contents. Users can effortlessly navigate the database through an intuitive and user-friendly top navigation menu, ensuring quick access to various functions and features. scQTLbase offers five main functional interfaces for users to interact effectively with the database: (i) 'Search' enables users to easily look for specific genes and SNPs of interest within the database (Figure 2A). (ii) 'UMAP' allows users to explore the gene expression patterns across different cell types clustered by marker genes in each study (Figure 2B). (iii) 'Genome Browser' provides a detailed view that enables users to examine the genomic regions of interest and explore the associations between SNPs and gene expression in a genomic context (Figure 2C). (iv) 'Colocalization' facilitates the visualization of potential colocalization events to gain insights into the shared genetic signals between sc-eQTLs and user-defined GWAS dataset (Figure 2D). (v) 'Trait/Disease' provides an atlas of GWAS-associated sc-eQTLs identified through colocalization analysis and linkage disequilibrium measurement between GWAS lead SNPs and sc-eQTL variants (Figure 2E). Overall, scQTLbase provides a comprehensive and user-friendly platform, empowering users to explore, visualize and analyze sc-eQTLs, which enhances our understanding of gene regulation and its implications for cellular function and disease mechanisms.

Single-cell eQTL searching and querying

In the search interface of scQTLbase, users can easily query sc-eQTLs using either a gene symbol or a dbSNP ID. Once a query is performed, the total number of sc-eQTLs associated with queried genes or SNPs are displayed, and four summary plots are shown to provide an overview of the sc-eQTL landscape. The results are presented in an interactive table with informative columns such as gene name, cell type, cell state, sc-eQTL type, SNP, *P*-value, beta, standard error (SE), study, GTEx eQTL, browse gene, and browser SNP. The Gene name and SNP columns provide details of the specific gene and SNP associated with the sc-eQTL, while the cell type column specifies the cell type in which the sc-eQTL was observed. The cell state column offers information about the particular state or condition of the cell during the sc-eQTL analysis. To categorize the sc-eQTL, the sc-eQTL type column classifies them into various types, including cell-type-specific eQTL, response eQTL, dynamic eQTL, variance eQTL. The Study column displays relevant information about the research study from which the sc-eQTL data originated, formatted in an author-year-journal abbreviation style (e.g. Bryois-2022-Nat. Neurosci.). The GTEx eQTL column shows whether the sc-eQTL is identified in GTEx eQTL from bulk RNA-seq. Two values are filled in this column: 'Yes' indicates that the sc-eQTL is identified in GTEx eQTL data, while 'No' indicates that the sc-eQTL is not identified in GTEx eQTL data. For each 'Yes' entry, a hyperlink is provided to access detailed information on GTEx eQTL associated with the same SNP-gene pair as the sc-eQTL. Additionally, each sc-eQTL record in the table is accompanied by two fixed buttons labelled 'Browse gene' and 'Browse SNP' on the right-hand side, enabling users to explore the specific sc-eQTL in the genome browser for further investigation.

To enhance exploration capabilities, users can apply filters to refine the sc-eQTLs based on specific criteria, such as sc-eQTL type, cell type, cell state or study. By simply clicking the search icon or filter icon in the header column (e.g. cell type), users can input custom filter keywords (e.g. 'CD8+ T cell') to narrow down the results based on their specific interests. Moreover, the table can be sorted based on any provided fields, allowing users to arrange the data in either ascending or descending order. Table can be exported, enabling users to obtain the data in a format suitable for further customized analysis. For instance, a gene query for 'BIN3' yields 1008 significant sc-eQTL records, comprising 864 cell-type-specific eQTLs and 144 response eQTLs across 23 distinct cell types. To delve deeper into the data, users can further filter the sc-eQTLs by clicking the filter icon in the header column (e.g. cell type) and inputting custom keywords (e.g. CD8+ T cell) to focus specifically on the sc-eQTLs associated with CD8+ T cells or any other relevant cell type of interest.

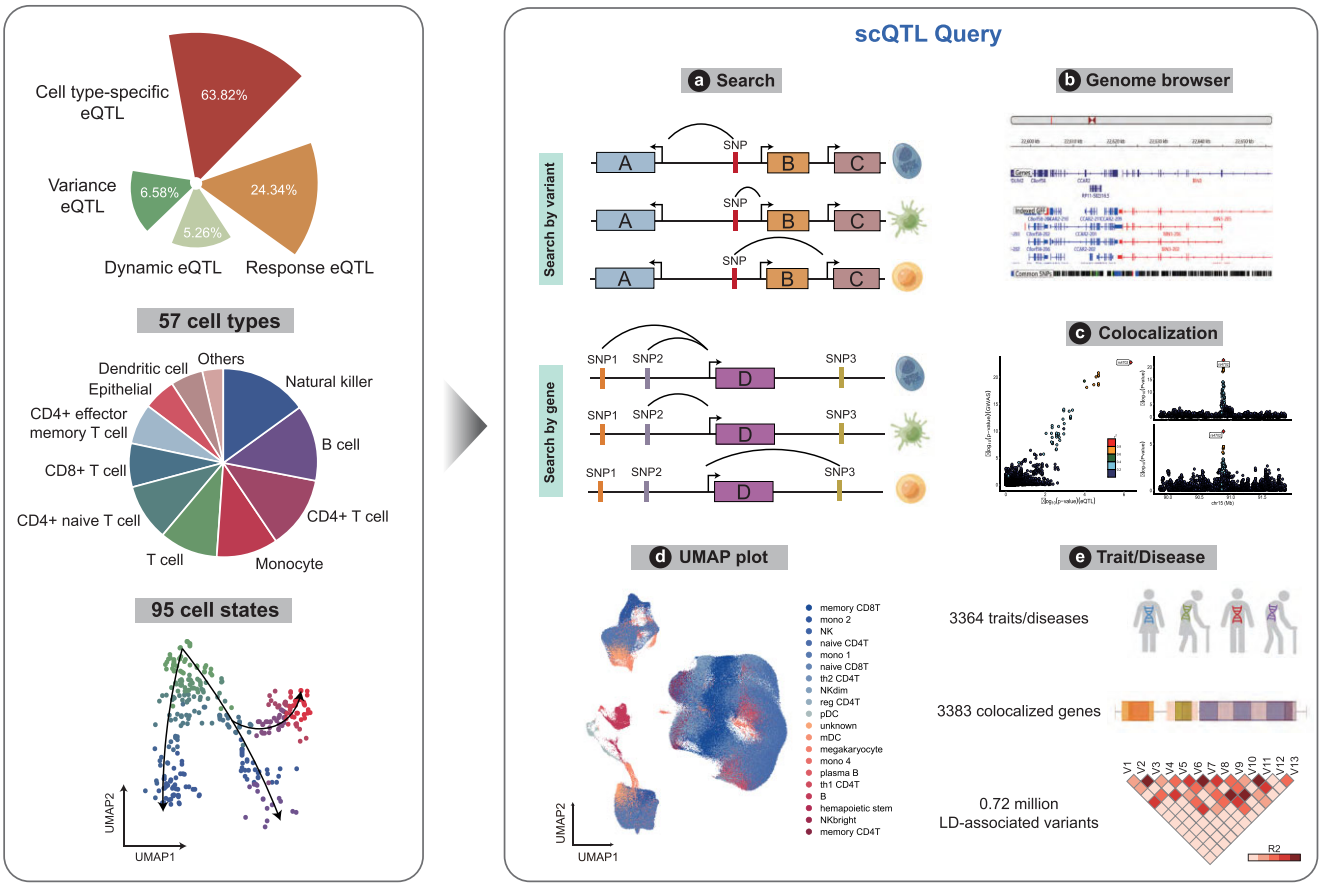


Figure 1. The data structure and general function of scQTLbase. The left panel is the sc-eQTL data summary and the right panel is the functionality of the database.

Visualization of cells in UMAP

In the UMAP section, users can explore the cellular landscape by visualizing individual cells colored either by their cell type or gene expression values. The left panel of the interface presents an interactive UMAP embedding, where each point represents a specific cell. By selecting the study or gene of interest from the pull-down list at the top of the panel, users can observe cellular heterogeneity and gene expression patterns. If no specific gene is chosen, the UMAP will display distinct cell types using different colors. Conversely, selecting a gene will color the UMAP based on normalized gene expression levels. On the right panel of the interface, an interactive table lists the marker genes of cell types and sc-eGenes (genes significantly associated with sc-eQTLs) related to the chosen study. By clicking on the genes or variants within each record, users can access detailed information about significant QTLs across studies. This feature provides valuable insights into the regulatory mechanisms underlying cellular processes and disease associations.

Sc-eQTL browsing in the genome browser

In the 'Genome Browser' section, users can actively explore sc-eQTLs across various cell types using an interactive genome browser by inputting the gene symbol (e.g. HLA-DRB5), SNP ID (e.g. rs113683581), or genome position (e.g. chr6:32492342–32555316). Users can select the browse mode by clicking radio button 'Browse by study' or 'Browse by cell type'. Within the genome browser, three tracks, namely Genes, Common SNPs, and GWAS are shared across all studies or cell types. However, upon selecting a specific study, sc-eQTLs across cell types will be displayed in separate tracks. Conversely, choosing a particular cell type will showcase sc-eQTLs across studies in distinct tracks. This allows users to focus on and compare sc-eQTLs efficiently based on their specific area of interest. For instance, if users wish to explore the genetic regulation to the gene 'HLA-DRB5' across cell types in the study 'Bryois-2022-Nat. Neurosci.', they can follow these steps: select the 'Browse by study' radio button, choose 'Bryois-2022-Nat. Neurosci' from the pull-down list, enter 'HLA-DRB5' in the search box and initiate the search. As a result, all significant sc-eQTLs related to HLA-DRB5 across eight cell types will be presented in eight individual tracks. By clicking on a specific data point, users can access detailed information about the corresponding SNP, including its ID and *P*-value. Notably, sc-eQTLs specific to the queried gene are highlighted in red in the genome browser, while those associated with other genes are marked in grey. The genome browser also includes gene structure annotation, GWAS Catalog (1) risk SNPs and dbSNP (17) variants. Furthermore, users have the option to download the browser tracks' figures in SVG format by clicking on the 'Save SVG' button situated at the top-right corner of the genome browser.

Traits/diseases-relevant sc-eQTLs

To bridge the gap between human genetic effects and disease observed, we focused on identifying sc-eQTLs associated

Figure 2. The web interface of scQTLbase. (**A**) Query interface and result visualization for 'Gene/SNP Search' in sc-eQTLs. (**B**) Example of UMAP view (top panel) and corresponding marker genes/egenes (bottom panel) from the study 'Van der Wijst-2018-Nat. Genet.' (**C**) Genome browser view showing sc-eQTLs across cell types in the study 'Jerber-2021-Nat. Genet.' (**D**) Interface for 'Colocalization' and an example of the LocusCompare plot at the gene 'BIN1,' displaying Alzheimer's disease GWAS *P*-values and sc-eQTLs *P*-values in Microglia (Micro) from the study 'Bryois-2022-Nat. Neurosci.' (**E**) Interface for 'Traits/Diseases' and an example of sc-eQTLs related to asthma.

with a wide range of traits and diseases. We manually curated 26 GWAS summary statistics and all associations, and gathered significant SNPs from 3 333 GWAS datasets collected in NHGRI GWAS catalogue (1). Through this process, we identified 3 133 sc-eQTLs co-localized with GWAS signals and approximately 0.69 million sc-eQTLs were in linkage disequilibrium (LD) with trait signals. To ensure easy access and exploration of these associations, we have presented the data through two interactive tables within the 'Trait/Disease' section, namely the 'co-localized genes' and 'LD-associated variants' tables. By default, these tables display all available data, empowering users to rapidly identify relevant associations between GWAS loci and sc-eQTLs. In the 'co-localized genes' table, a fixed column on the right-hand side labeled 'Locus compare' is incorporated. This column features a set of buttons, each corresponding to a specific colocalized gene. Upon selecting any button, the LocusCompare plot related to the chosen gene is displayed. To further refine the results, users have the flexibility to apply various filters based on specific sc-eQTL types, cell types, cell states and sources of interest. Additionally, the table can be sorted based on any provided fields, allowing users to organize and examine the data in a manner that aligns with their specific research needs. This comprehensive approach facilitates an enhanced understanding of the intricate connections between genetic variants and diseases at the single-cell level.

Summary and future directions

We have developed the user-friendly database scQTLbase by systematically curating and harmonizing sc-eQTL summary statistic datasets from various studies. This database offers ∼16 million SNPs significantly associated with gene expression at the cell-type or cell-state level, making it a one-stop portal for sc-eQTL search, UMAP, and genome browsing. Moreover, users can visualize the colocalization results based on GWAS datasets of their interest. By integrating sc-eQTL and GWAS data, we identify and display ∼0.69 million GWAS-associated sc-eQTLs, providing insights into the molecular mechanisms underlying complex traits and diseases at the cellular resolution. Although the field of sc-eQTLs is still in its early stages (6), with a current focus on immune cells or specific tissues such as the brain and lung, we recognize the need to expand these investigations to encompass all major cell types in the human body. Such expansion will yield valuable insights into the broader landscape of sc-eQTLs and their significance across diverse cellular populations. To keep pace with the increasing number of sc-eQTL datasets from large consortium projects, continuous updates to scQTLbase are essential. We are dedicated to maintaining and upgrading our database in line with advances in data, technologies, and methods in this field. In conclusion, scQTLbase is the first integrated sc-eQTL database, providing a valuable resource for understanding the genetic basis of complex human traits and diseases at cellular resolution. Through ongoing updates, this database promises to be an indispensable tool, significantly advancing our understanding of the fundamental principles of gene regulation and their implications in complex traits and diseases.

Data availability

scQTLbase is freely available at: http://bioinfo.szbl.ac.cn/scQTLbase.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We acknowledge all members of the Li lab for constructive discussions and help. We also thank Qin Wang at Shenzhen Bay Laboratory supercomputing center for high-computing support.

Author contributions: Ruofan Ding: Writing—original draft, formal analysis, methodology. Qixuan Wang & Lihai Gong: Formal analysis, methodology, database construction. Ting Zhang & Xudong Zou: Validation, writing-review & editing. Qi Liao & Mireya Plass & Kewei Xiong: Writing-review & editing. Lei Li: Conceptualization, methodology, writing—review & editing.

Funding

National Natural Science Foundation of China [32100533, 32370721 to L.L.]; Open grant funds from Shenzhen Bay Laboratory [SZBL2021080601001 to L.L.]; A Ramón y Cajal contract of the Spanish Ministry of Science and Innovation [RYC2018-024564-I to M.P.]. Funding for open access charge: Natural Science Foundation of China and Shenzhen Bay Laboratory.

Conflict of interest statement

None declared.

References

1. Sollis,E., Mosaku,A., Abid,A., Buniello,A., Cerezo,M., Gil,L., Groza,T., Gunes,O., Hall,P., Hayhurst,J., *et al.* (2023) The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.*, **51**, D977–D985.
2. Zeng,B., Bendl,J., Kosoy,R., Fullard,J.F., Hoffman,G.E. and Roussos,P. (2022) Multi-ancestry eQTL meta-analysis of human brain identifies candidate causal variants for brain-related traits. *Nat. Genet.*, **54**, 161–169.
3. Zhu,Z., Zhang,F., Hu,H., Bakshi,A., Robinson,M.R., Powell,J.E., Montgomery,G.W., Goddard,M.E., Wray,N.R., Visscher,P.M., *et al.* (2016) Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.*, **48**, 481–487.
4. Consortium,G.T. (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, **369**, 1318–1330.
5. Kerimov,N., Hayhurst,J.D., Peikova,K., Manning,J.R., Walter,P., Kolberg,L., Samovica,M., Sakthivel,M.P., Kuzmin,I., Trevanion,S.J., *et al.* (2021) A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.*, **53**, 1290–1299.
6. Cuomo,A.S.E., Nathan,A., Raychaudhuri,S., MacArthur,D.G. and Powell,J.E. (2023) Single-cell genomics meets human genetics. *Nat. Rev. Genet.*, **24**, 535–549.
7. Wills,Q.F., Livak,K.J., Tipping,A.J., Enver,T., Goldson,A.J., Sexton,D.W. and Holmes,C. (2013) Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat. Biotechnol.*, **31**, 748–752.
8. Nathan,A., Asgari,S., Ishigaki,K., Valencia,C., Amariuta,T., Luo,Y., Beynor,J.I., Baglaenko,Y., Suliman,S., Price,A.L., *et al.* (2022) Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. *Nature*, **606**, 120–128.
9. Yazar,S., Alquicira-Hernandez,J., Wing,K., Senabouth,A., Gordon,M.G., Andersen,S., Lu,Q., Rowson,A., Taylor,T.R.P.,

Clarke,L., *et al.* (2022) Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science*, **376**, eabf3041.
10. van der Wijst,M.G.P., Brugge,H., de Vries,D.H., Deelen,P., Swertz,M.A., LifeLines Cohort,S., Consortium,B. and Franke,L. (2018) Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.*, **50**, 493–497.
11. Kang,J.B., Raveane,A., Nathan,A., Soranzo,N. and Raychaudhuri,S. (2023) Methods and insights from single-cell expression quantitative trait loci. *Annu. Rev. Genomics Hum. Genet.*, **24**, 277–303.
12. Cuomo,A.S.E., Seaton,D.D., McCarthy,D.J., Martinez,I., Bonder,M.J., Garcia-Bernardo,J., Amatya,S., Madrigal,P., Isaacson,A., Buettner,F., *et al.* (2020) Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat. Commun.*, **11**, 810.
13. Oelen,R., de Vries,D.H., Brugge,H., Gordon,M.G., Vochteloo,M., single-cell e,Q.c., Consortium,B., Ye,C.J., Westra,H.J., Franke,L., *et al.* (2022) Single-cell RNA-sequencing of peripheral blood mononuclear cells reveals widespread, context-specific gene expression regulation upon pathogenic exposure. *Nat. Commun.*, **13**, 3267.
14. Sarkar,A.K., Tung,P.Y., Blischak,J.D., Burnett,J.E., Li,Y.I., Stephens,M. and Gilad,Y. (2019) Discovery and characterization of variance QTLs in human induced pluripotent stem cells. *PLos Genet.*, **15**, e1008045.
15. Vosa,U., Claringbould,A., Westra,H.J., Bonder,M.J., Deelen,P., Zeng,B., Kirsten,H., Saha,A., Kreuzhuber,R., Yazar,S., *et al.* (2021) Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.*, **53**, 1300–1310.
16. Schmiedel,B.J., Singh,D., Madrigal,A., Valdovino-Gonzalez,A.G., White,B.M., Zapardiel-Gonzalo,J., Ha,B., Altay,G., Greenbaum,J.A., McVicker,G., *et al.* (2018) Impact of genetic polymorphisms on human immune cell gene expression. *Cell*, **175**, 1701–1715.
17. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
18. Hao,Y., Hao,S., Andersen-Nissen,E., Mauck,W.M. 3rd, Zheng,S., Butler,A., Lee,M.J., Wilk,A.J., Darby,C., Zager,M., *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573–3587.
19. Diehl,A.D., Meehan,T.F., Bradford,Y.M., Brush,M.H., Dahdul,W.M., Dougall,D.S., He,Y., Osumi-Sutherland,D., Ruttenberg,A., Sarntivijai,S., *et al.* (2016) The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J. Biomed. Semantics*, **7**, 44.
20. Ding,R., Zou,X., Qin,Y., Chen,H., Ma,X., Yu,C., Wang,G. and Li,L. (2023) xQTLbiolinks: a comprehensive and scalable tool for integrative analysis of molecular QTLs. bioRxiv doi: https://doi.org/10.1101/2023.04.28.538654, 29 April 2023, preprint: not peer reviewed.
21. McInnes,L., Healy,J. and Melville,J.J.a.p.a. (2018) Umap: uniform manifold approximation and projection for dimension reduction. *J. Open Source Softw.*, **3**, 861.
22. 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.
23. Shi,Y., Inoue,H., Wu,J.C. and Yamanaka,S. (2017) Induced pluripotent stem cell technology: a decade of progress. *Nat. Rev. Drug Discov.*, **16**, 115–130.

Received: July 30, 2023. **Revised:** August 24, 2023. **Editorial Decision:** September 9, 2023. **Accepted:** September 15, 2023