OXFORD

Briefings in Bioinformatics, 2024, **25(1)**, 1–10
https://doi.org/10.1093/bib/bbad440
Problem Solving Protocol

xQTLbiolinks: a comprehensive and scalable tool for integrative analysis of molecular QTLs

Ruofan Ding†, Xudong Zou†, Yangmei Qin, Lihai Gong, Hui Chen, Xuelian Ma, Shouhong Guang, Chen Yu, Gao Wang and Lei Li

Corresponding authors: Lei Li, Shenzhen Bay Laboratory, Institute of Systems and Physical Biology, Shenzhen 518055, China. E-mail: Lei.Li@szbl.ac.cn; Gao Wang, Department of Neurology, The Gertrude H. Sergievsky Center, Columbia University, NY 10032, USA. Tel.: +86 0755 2684 9284; E-mail: wang.gao@columbia.edu

†Ruofan Ding and Xudong Zou contributed equally

Abstract

Genome-wide association studies (GWAS) have identified thousands of disease-associated non-coding variants, posing urgent needs for functional interpretation. Molecular Quantitative Trait Loci (xQTLs) such as eQTLs serve as an essential intermediate link between these non-coding variants and disease phenotypes and have been widely used to discover disease-risk genes from many population-scale studies. However, mining and analyzing the xQTLs data presents several significant bioinformatics challenges, particularly when it comes to integration with GWAS data. Here, we developed xQTLbiolinks as the first comprehensive and scalable tool for bulk and single-cell xQTLs data retrieval, quality control and pre-processing from public repositories and our integrated resource. In addition, xQTLbiolinks provided a robust colocalization module through integration with GWAS summary statistics. The result generated by xQTLbiolinks can be flexibly visualized or stored in standard R objects that can easily be integrated with other R packages and custom pipelines. We applied xQTLbiolinks to cancer GWAS summary statistics as case studies and demonstrated its robust utility and reproducibility. xQTLbiolinks will profoundly accelerate the interpretation of disease-associated variants, thus promoting a better understanding of disease etiologies. xQTLbiolinks is available at https://github.com/lilab-bioinfo/xQTLbiolinks.

Keywords: eQTL; GWAS; xQTL; bioinformatics; disease variants

INTRODUCTION

The explosion of GWAS discovery and applications across multiple disciplines yielded many risk loci mainly located in non-coding regions, prompting the great need for functional interpretation of these variants by revealing the underlying mechanisms and susceptibility genes [1]. The large-scale molecular QTLs data have been widely used as an essential intermediate link of the non-coding disease risk variants to disease phenotypes. For example, more than four million common genetic variants (minor allele frequency > 0.01) associated with the gene expression of more than 23 000 genes across 49 human tissues have been identified by Genotype-Tissue Expression (GTEx) Consortium [2], representing a valuable resource for the molecular interpretation of disease risk variants. eQTL catalog [3] is another useful resource that includes uniformly processed expression QTLs (eQTLs) and splicing QTLs (sQTLs), which currently includes 21 studies, and the data are still increasing. Recently, single-cell expression quantitative trait locus (sc-eQTL) studies have emerged, offering a significant opportunity to gain insights into the biological mechanisms of diseases at the cellular level. Two representative projects,

Ruofan Ding is a postdoctoral scholar of Institute of Systems and Physical Biology in Shenzhen Bay Laboratory. His overall career goal is to develop bioinformatics tools and apply computational modeling.
Xudong Zou is currently an assistant researcher scholar of Institute of Systems and Physical Biology in Shenzhen Bay Laboratory. His research interests include Bioinformatics, disease genetics and rare diseases.
Yangmei Qin is a postdoctoral scholar of Institute of Systems and Physical Biology in Shenzhen Bay Laboratory. Her research interests include bioinformatics, biomedical big data mining and common variations.
Lihai Gong is currently a research assistant at Institute of Systems and Physical Biology in Shenzhen Bay Laboratory. His research interests include developing bioinformatical tools and pipelines for annotating genetic variation.
Hui Chen is currently a PhD student at Institute of Systems and Physical Biology in Shenzhen Bay Laboratory. Her research interests include bioinformatics, and disease genomics.
Xuelian Ma is currently a postdoctoral scholar of Institute of Systems and Physical Biology in Shenzhen Bay Laboratory. Her scholarly interests encompass bioinformatics, developmental genetics and database construction.
Shouhong Guang is a professor of University of Science and Technology of China. His research interests include bioinformatics and the identification of nuclear RNAi. Nuclear RNAi mediates pre-mature transcription termination.
Chen Yu is a principal investigator of Institute of Cancer Research in Shenzhen Bay Laboratory. He is interested in developing tools for CRISPR genome engineering.
Gao Wang is an assistant professor of Gertrude H. Sergievsky Center and the Department of Neurology in Columbia University. His work on computational methods includes multi-variate regression analysis, statistical fine-mapping and colocalization in association studies.
Lei Li is a principal investigator of Institute of Systems and Physical Biology in Shenzhen Bay Laboratory. He mainly focuses on development and application of bioinformatics algorithms to elucidate genetic and epigenetic regulatory mechanisms in human complex traits and diseases.

Received: June 27, 2023. **Revised:** October 23, 2023. **Accepted:** November 11, 2023

the eQTLGen Consortium and the DICE project, are focused on investigating the genetic architecture of blood gene expression and regulatory genes in 13 human immune cell types, respectively [4, 5]. In addition to the traditional transcriptomic phenotype associated QTLs, there has been an expansion in the identification of other molecular QTLs associated with epigenetic phenotypes, such as DNA methylation QTLs (mQTLs) [6] and histone QTLs (hQTLs) [7]. These molecular QTLs hold significant potential in providing valuable insights into the functional implications of non-coding disease risk variants. However, exploring and mining such a massive volume of xQTLs data remains computationally challenging. Firstly, xQTLs summaries typically contain millions of associations between genetic variants and molecular phenotypes, making retrieving all or a subset of xQTLs data of interest computationally expensive. Secondly, xQTLs data generated by different analysis pipelines with varying tools often did not correctly harmonize with rigorous quality control to boost statistical power and reduce false discovery in downstream analysis. Thirdly, currently, no tools can seamlessly annotate the xQTLs data of their underlying function.

Another challenge for analyzing xQTL data is that xQTLs are often integrated with GWAS summary statistics data for colocalization analysis, widely used to link potentially causal genes to GWAS risk loci. Several available tools can perform probabilistic colocalization analysis between xQTLs and GWAS summary statistics, including Coloc [8], HyPrColoc [9], ColocQuiaL [10], ezQTL [11], etc. Despite these tools being frequently used to identify disease-risk genes in many studies, many limitations remain. For instance, relying solely on a single colocalization tool may fail to identify reliable disease-related genes due to insufficient detection power, especially in cases where the sample size is small, or the causal variants have a relatively small effect size [12]. In addition, all these tools only focus on statistical methods without considering the upstream data processing and downstream visualization of the results, making the colocalization analysis still computationally challenging. To our knowledge, no comprehensive analysis pipeline can seamlessly mine and analyze both xQTLs and GWAS data.

To address these challenges, we developed xQTLbiolinks, a user-friendly R package, as the first end-to-end bioinformatics tool for efficient mining and analyzing public and user-customized xQTLs data for the discovery of disease susceptibility genes. xQTLbiolinks offers the following unique and practical advantages: (i) enables flexible access to bulk xQTLs data from 23 839 samples across 75 human tissues or cell types and provides quality control and annotation modules for summary statistics data; (ii) provides fast querying of sc-eQTLs from manually curated 16 studies and 304 datasets; (iii) offers a robust colocalization pipeline that utilizes two popular colocalization methods to streamline the identification of colocalized disease-associated genes; (iv) compatible with external tools for downstream analysis and can flexibly generate detailed outputs and ready-to-publish figures. Our package is freely available at https://github.com/lilab-bioinfo/xQTLbiolinks.

METHODS

Implementation of xQTLbiolinks

xQTLbiolinks, a user-friendly R package under the General Public License (GPLv3) license, can be installed in any operating system supporting R through the general function install.packages ('xQTLbiolinks') from the Comprehensive R Archive Network (CRAN: https://cran.r-project.org/). All functions of xQTLbiolinks are available by standard R commands to manipulate xQTLs and GWAS data after installation and loading of the package. A comprehensive user manual introducing all functions and corresponding usages can be found in the Github repository (https://github.com/lilab-bioinfo/xQTLbiolinks). Briefly, xQTLbiolinks implements four modules: data retrieval, pre-processing, colocalization analysis and data visualization (Figure 1). Users can query and download gene expression, sample and variant details, xQTLs and cell type-specific eQTL (sc-eQTLs) data through the APIs of GTEx [2], eGTEx [6], eQTL catalog [3], BLUEPRINT epigenome project [7] and our curated sc-eQTL resource (Table S1 available online at http://bib.oxfordjournals.org/). Available bulk xQTLs include eQTL, sQTL, 3'aQTL, mQTL and haQTL, while sc-eQTLs consist of cell type-specific eQTL (genetic variants that are associated with gene expression in a specific cell type), response eQTL (genetic variants that are associated with changes in gene expression in response to stimuli) and dynamic eQTL (genetic variants that are associated with changes in gene expression over time). xQTLbiolinks offers a data pre-processing panel that allows users to perform quality control on GWAS/xQTL summary statistics datasets using the QQ plot, PZ plot and inflation factor. Users can perform genomic annotation of GWAS/xQTLs signals using enrichment analysis or conduct functional annotation by incorporating customized ChIP-seq data, such as enhancer and transcription binding site (TFBS). The colocalization analysis is facilitated through a three-step pipeline, enabling the easy detection of trait/disease-relevant genes. Moreover, xQTLbiolinks visualizes the results with publication-ready plots, such as heatmap, boxplot, scatter plot and locusZoom plot.

Compatibility of xQTLbiolinks

The functions and outputs in xQTLbiolinks are compatible with other functions and packages. We have provided expression data objects as the Bioconductor specified 'SummarizedExperiment' class, which is a matrix-like container with rows representing genes of interest (as a GRanges or GRangesList object), columns representing samples (with sample data summarized as a DataFrame) and the matrix is filled with normalized expression profile. 'SummarizedExperiment' is critical for allowing the full integration and use of other popular Bioconductor packages. We also provided the function 'xQTL_export' to export the data to the specified format, which can be used as direct input for other tools, including 'to_clusterP', 'to_deseq' and 'to_wgcna'. Users can perform customized analysis with external R packages, including functional enrichment analysis using clusterProfiler [11], differential expression analysis using DESeq2 and edgeR [13, 14], weighted Correlation Network Analysis using WGCNA [15], etc.

Data retrieval

In this study, we select the available datasets of xQTLs that are most directly relevant to understanding the regulation of gene expression, splicing and 3'UTR alternative polyadenylation. These datasets were generated from the extensive GTEx cohorts, ensuring a strong representation of the genetic regulation of these molecular traits. Additionally, we have included single-cell eQTL datasets to provide insights into the regulation of gene expression at the single-cell level. Furthermore, we have expanded our analysis to include important epigenetic modifications, such as DNA methylation QTLs and histone modification QTLs, which play a crucial role in gene regulation. The data retrieval module enables users to query and download publicly available xQTLs summary data from GTEx and eQTL catalog, and also sc-eQTL

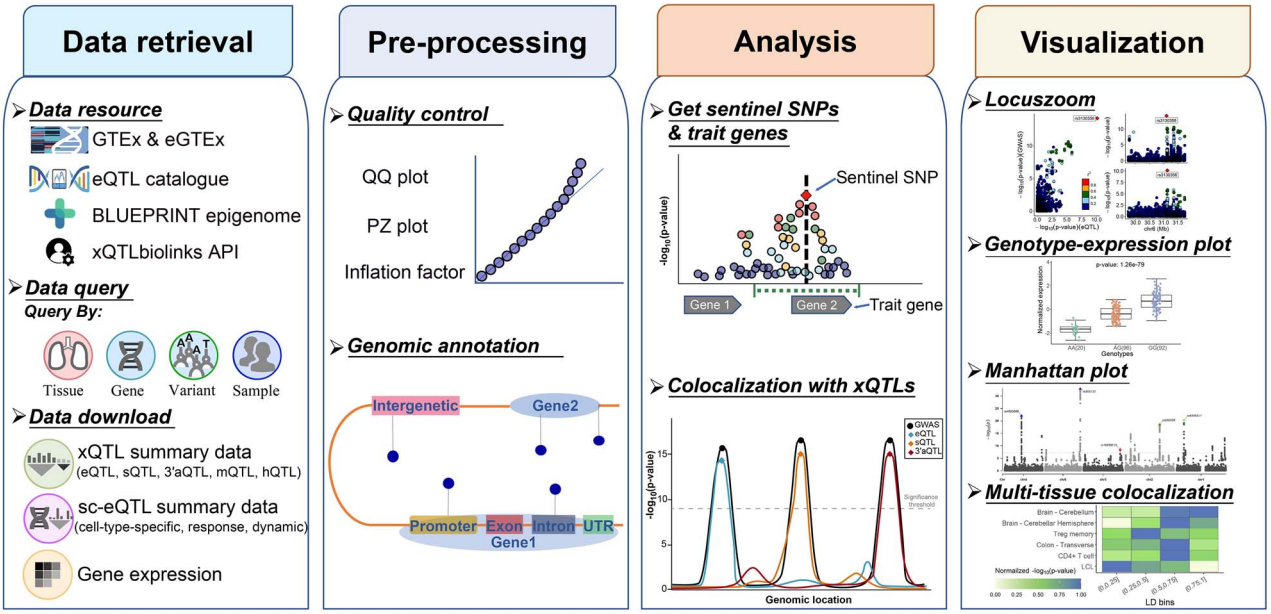

Figure 1. Overview of xQTLbiolinks data and functions, including four main function categories: data retrieval, pre-processing, analysis and visualization.

data from our server. The current version supports xQTLs data from 23 839 samples across 75 human tissues/cell types, and sc-eQTL data from 16 studies across 57 cell types. Two commands, *xQTLquery* and *xQTLdownload*, provide flexible user interfaces to query and download xQTLs data by tissue, gene, SNP or combination. For example, specifying a specific tissue name to the *xQTLquery* function will display all xQTLs data associated with the tissue; users can also focus only on xQTLs related to one single gene by specifying gene ID to the function. *xQTLquery* is the primary function that executes the query of entities in xQTLs data, including genes, variants and samples. At the same time, *xQTLdownload* allows users to retrieve xQTLs data with tailored demands, including eGenes/sGenes, associations between variants and expression (eQTLs), splicing (sQTLs), 3′UTR alternative polyadenylation (3′aQTLs), DNA methylation QTLs (mQTLs) and histone modification (hQTLs), normalized gene expressions and linkage disequilibrium of xQTLs in specified genes. The *xQTLquery_sc* and *xQTLdownload_sc* functions have been developed for the query and download of sc-eQTLs. These functions allow users to retrieve sc-eQTLs by gene name, sc-eQTL type, cell type, cell state and study name. All retrieved xQTLs and sc-eQTL data can be handled by *xQTLanalyze*, *xQTLvisual* and their sub-functions.

Collection and processing of sc-eQTLs

We conducted a manual collection of sc-eQTL datasets from published literature: (i) The selected studies utilized samples from a diverse range of biological contexts, including normal samples, treated samples or samples of disease conditions. Datasets from meta-analysis or secondary analyses are excluded, and we only included studies with a minimum of 40 samples or 5 000 cells to ensure sufficient power for sc-eQTLs. (ii) We only included single-cell transcriptome data generated from well-established sequencing technologies such as 10x Genomics, Smart-seq/Smart-seq2 or CITE-seq. In total, 304 sc-eQTL datasets from 16 studies were collected, which contain ~2 750 individuals and approximately 8.04 million cells (Table S2 available online at http://bib.oxfordjournals.org/). Variants with a dbSNP identifier are harmonized with the dbSNP build 151 based on the hg38 genome version. The names of cell types across different studies are standardized based on expert-annotated cell type reference [16]. The statistical values of sc-eQTL summary statistics data are contained, including *P*-value, effect size (beta), standard error and false discovery rate. We stored the sc-eQTL summary statistics data in MySQL database and developed open-source API using a flask-based web framework.

Data pre-processing

xQTLbiolinks allows users to detect possible inflation or deflation of test statistics due to population stratification, genotyping errors or other sources of bias for GWAS/QTL summary statistics datasets. *xQTLvisual_qqPlot* and *xQTLanno_calLambda* can plot quantile-quantile plot (QQ-plot) and calculate genomic control inflation factor, respectively. *xQTLvisual_PZPlot* plots the concordance correlation of observed *P*-values and *P*-values calculated from the Z-score derived from beta (representing effect size) and se (representing standard error of effect size). A P-Z plot is used to examine the discrepancy between the *P*-value reported by association results and the *P*-value calculated manually from the Z statistic. In the absence of discrepancies, the points should fall along the diagonal line, indicating the consistency between observed and expected *P*-values. Besides, *xQTLanno_genomic* enables users to functionally annotate variants in GWAS/xQTLs datasets by calculating fold enrichment scores. The functional categories are referred to ANNOVAR [17] including intron, exon, 3′UTR, 5′UTR, splicing site, intergenic region promoter (1-kb region upstream of transcription start site) and downstream (1-kb region downstream of transcription end site). The fold enrichment score for each functional category is calculated as the proportion of the significant SNPs with a certain annotation divided by the proportion of SNPs with the same annotation in the background, and the corresponding *P*-value is calculated by performing Fisher's exact test (one-tailed test) against the entire genome [18].

Colocalization analysis

xQTLbiolinks implements a pipeline that contains three sub-functions to perform colocalization analysis following the steps

(Figure S1 available online at http://bib.oxfordjournals.org/): (i) *xQTLanalyze_getSentinelSnp* extracts sentinel SNP, which represents the most prominent signal in a specific genome region, from GWAS summary statistics data. By default, it selects the variants with a *P*-value less than 5×10^{-8} within 1 million base pairs. The *P*-value threshold of 5×10^{-8} is decided according to many previous studies [19–21] where this threshold was widely used to identify significant disease/trait-associated variants such as GWAS catalog (https://www.ebi.ac.uk/gwas/). (ii) *xQTLanalyze_getTraits* identify trait genes nearby sentinel SNPs within a 1 Mb region by default. The 1 Mb region is a commonly used threshold for identifying potential SNP-gene associations, such as in GTEx and eQTLgen [2, 5]. (iii) *xQTLanalyze_coloc* and *xQTLanalyze_coloc_diy* perform two commonly used colocalization methods, Coloc [8] and HyPrColoc [9], for each trait gene.

Visualization

The visualization module contains a main function *xQTLvisual* and several sub-functions that allow users to visualize the results with publication-ready plots, such as heatmap, boxplot, scatter plot and locusZoom plot. *xQTLvisual_genesExp* and *xQTLvisual_geneExpTissues* are used to plot the distribution of gene expression for the queried gene(s). Sub-functions *xQTLvisual_eqtlExp* and *xQTLvisual_sqtlExp* can plot the association between genotypes and molecular phenotypes with a grouped boxplot by genotypes. xQTLbiolinks also contains three sub-functions *xQTLvisual_locusZoom, xQTLvisual_locusCompare* and *xQTLvisual_coloc* visualizing the colocalization results. The first two sub-functions can plot publication-ready locusZoom on specific GWAS and xQTLs signals. Finally, *xQTLvisual_coloc* visualizes the regulatory sharing effects of colocalized variants across multiple tissues or cell types.

RESULTS

xQTLbiolinks as a comprehensive tool for exploring and mining xQTLs data

xQTLbiolinks presents the first end-to-end solution for molecular QTL data mining and analyzing. Compared to previous tools, xQTLbiolinks provides comprehensive and versatile approaches to accessing and manipulating xQTLs summary data (Table 1). It streamlines the querying and retrieval of bulk and single-cell xQTL data to meet user-customized demand from public repositories (i.e. GTEx and eQTL catalog) and our xQTLbiolinks server. Notably, it can also analyze user-customized xQTLs summary data. xQTLbiolinks is characterized by its exceptional adaptability and user-friendliness for the analysis of xQTLs data. Its key strengths can be attributed to the following factors: (i) xQTLbiolinks was developed as a standardized R package which makes it easily accessible to users who are familiar with other R packages; (ii) xQTLbiolinks provides the first querying and mining single-cell genetic effect from 16 independent studies across 57 cell types (Table S2 available online at http://bib.oxfordjournals.org/). (iii) the utilization of xQTLs datasets in public repositories, i.e. largest atlas of human gene expression, eQTL and sQTL data from GTEx [2], uniformly processed xQTLs summary statistics across 75 distinct cell types and tissues from the eQTL Catalog [3]. xQTLbiolinks allows users to conveniently retrieve xQTLs data and metainformation for further analysis through gene names/IDs, tissue/cell type names or genomic regions of interest. It also facilitates easy query and retrieval of either bulk xQTL or sc-eQTL for a specific gene in a particular tissue or cell type. Such flexibility saves running time and decreases the requirement of computational resources, thus taking full advantage of comprehensive xQTLs data for GWAS integration of varying scales from candidate genes to genome-wide.

xQTLbiolinks provides a robust colocalization module through integration with GWAS data

Colocalization is a powerful approach for integrating xQTLs and GWAS signals and has been widely used to identify novel disease susceptibility genes [22]. This approach evaluates whether xQTLs and GWAS signals statistically share putative causal variants and can provide valuable insights into the genetic mechanisms underlying complex diseases. xQTLbiolinks employs different colocalization methods to achieve more solid results since single method has limitations in certain scenarios. For example, the widely used colocalization software, Coloc, is a valuable tool for estimating posterior probabilities of colocalization between two traits. However, its effectiveness may be limited when studying smaller sample sizes [8]. HyPrColoc is a Bayesian method that utilizes clustering to group traits and identify shared genetic associations for each cluster. However, the assumption of a single causal SNP hypothesis may lead to potential omissions of causal signals [9]. Moreover, visualization tools such as eQTpLot and LocusCompare are solely used for visualization [23, 24]. xQTLbiolinks provides a comprehensive pipeline that can perform colocalization analysis across multiple tissues or cell types and handle upstream data processing and downstream visualization (Figure S1 available online at http://bib.oxfordjournals.org/). This framework offers a one-step solution of functions that can be used for quality control of GWAS significant variants, extraction of sentinel SNP, identification of trait genes, preparation of curated or user-customized xQTLs datasets, colocalization analysis and visualizing the GWAS/xQTLs signals using the locusZoom plot. We benchmarked the running time of the colocalization module on 10 GWAS datasets from UK Biobank [25]. The mean running time for each colocalization analysis is approximately 30.5 minutes when using both methods simultaneously (Figure S2 available online at http://bib.oxfordjournals.org/). Furthermore, to investigate the regulatory effect of colocalized xQTLs across multiple tissues/cell types, we have made a new plot by correlating xQTLs *P*-values with linkage disequilibrium (LD) bins across multiple tissues.

Case study 1: quality control and functional characterization of breast cancer risk SNPs using xQTLbiolinks

We first downloaded the Breast cancer GWAS summary statistics from the literature, representing the largest GWAS study on breast cancer conducted on more than 80 000 individuals [26]. We then performed a quality control analysis of the data. We first use *xQTLanno_calLambda* to estimate the *P*-values' inflations, and it returns a lambda value of 1.147, indicating no strong population stratification exists. Then we evaluate the quality of GWAS data by examining whether the observed distribution of *P*-values follows the expected distribution under the null hypothesis of no association between the genetic variants and the disease using *xQTLvisual_qqplot*; we found a significant deviation from the diagonal line, which indicates potential variations from the null hypothesis that may result from true associations or LD (Figure 2A). *xQTLvisual_PZplot* is further used to investigate the normality of the distribution of Z-scores derived from beta and se, which outputs the strong concordance between the observed *P*-values and those calculated from Z-scores (Figure S3 available online at http://bib.oxfordjournals.org/). We further annotated all significant GWAS variants by genomic locations

Table 1: A comparison of different tools for retrieving and analysis of xQTLs data

Features	Sub-features	Coloc	ColocQuiaL	ezQTL	eQTpLot	locuscomparer	xQTLbiolinks
Availability	Platform	R/ C++	R	web	R	R	R
Query xQTLs	sc-eQTL						✓
	Multiple tissues/cell types						✓
	Gene/variant/sample						✓
Download	xQTLs summary data						✓
	Gene expression						✓
Quality control	QQ-plot/PZ-plot/Inflation factor						✓
Annotation	Genomic annotation						✓
Analysis	Colocalization analysis	✓	✓	✓			✓
Visualization	LocusZoom			✓	✓	✓	✓
	Genotype-expression/splicing boxplot						✓
	Manhattan plot						✓

Note: The first two columns of the table represent features and detailed features for each tool, respectively. The cell checked with '✓' indicates features that exist in the tool.

using *xQTLanno_genomic* (Figure 2B). We also retrieved eQTLs and sQTLs summary data from GTEx Breast - Mammary tissue using *xQTLdownload_eqtlAllAsso* and *xQTLdownload_sqtlAllAsso functions*, respectively. In addition, we included our recently developed 3′UTR alternative polyadenylation quantitative trait loci (3′aQTLs) as a user-customized xQTLs dataset using *xQTLdownload_xqtlAllAsso*. We later performed similar quality control analyses for these xQTLs data and found no inflation or quality issues on these datasets. The genomic control inflation values of the xQTLs summary data are 1.149 (eQTLs), 1.049 (sQTLs) and 1.022 (3′aQTLs), and the corresponding QQ-plots are shown in Figure 2C–E. We used *xQTLvisual_manhattan* to generate a Manhattan plot, which exhibits strong signals of associations across all chromosomes at a genome-wide level (Figure 2F). By integrating with eQTL signals, we detected some significant loci that showed significant associations with disease susceptibility but also exerted regulatory influence on gene expression. For instance, rs8018155 was a breast cancer risk variant and an eQTL, as illustrated in Figure 2G. This suggests that risk variant rs8018155 plays a crucial role in modulating the expression of gene *CCDC88C* and may have potentially important implications for breast cancer. Furthermore, the distribution of expression of 12 eQTL-colocalized genes in TCGA breast cancer samples is shown in Figure S4A available online at http://bib.oxfordjournals.org/.

Case study 2: identification of prostate cancer susceptibility genes using xQTLbiolinks

Prostate cancer (PCa) is one of the most common cancers, the pathogenesis of which involves both heritable and environmental factors [27]. The molecular events involved in the development or progression of PCa are still unclear. Here, we applied xQTLbiolinks to integrate the PCa GWAS dataset [28] with xQTLs data from GTEx prostate tissue and aimed to identify putative causal variants and susceptibility genes associated with PCa. We first extracted 94 sentinel SNPs with $P < 5 \times 10^{-8}$ using *xQTLanalyze_getSentinelSnp*. We then identified 835 genes for eQTLs, 1 676 genes for sQTLs and 209 genes for 3′aQTLs using *xQTLanalyze_getTraits*. Later, for each trait gene, we analyzed the colocalization pattern between PCa risk variants and xQTLs using *xQTLanalyze_coloc* for eQTLs and *xQTLanalyze_coloc_diy* for sQTLs and 3′aQTLs. By default, two colocalization methods (Coloc and HyPrColoc) are used. The colocalization analysis returns four posterior probabilities corresponding to four different null hypotheses; notably, the posterior probability under hypothesis 4 (PPH4), representing the potential same causal variants shared by GWAS variants with xQTLs data, was used to define significant colocalization events. Using a PPH4 threshold of 0.75, we identified 47 genes, including 27 eQTLs genes, 17 sQTLs genes and seven 3′aQTLs genes that are colocalized with 38 PCa risk loci. Among these colocalized genes, many have been previously reported to be associated with PCa susceptibility (Table S3 available online at http://bib.oxfordjournals.org/). For example, the gene *MMP7*, which is strongly colocalized by eQTLs, encodes a member of the peptidase M10 family of matrix metalloproteinases and is involved in the breakdown of extracellular matrix in normal physiological processes. It has been evidenced that PCa can be promoted via *MMP7*-induced epithelial-to-mesenchymal transition by Interleukin-17 [29], and serum *MMP7* levels could guide metastatic therapy for PCa [30]. Besides, we also observed that 28 of the 47 colocalized genes have been reported as known susceptibility genes in prostate cancer (Table S3 available online at http://bib.oxfordjournals.org/), and the remaining 19 genes without evidence could be considered novel candidates of prostate cancer susceptibility genes. For instance, we identified the eQTL-colocalized gene *GGCX* (PP4 = 0.9905), which encodes an enzyme called gamma-glutamyl carboxylase that is responsible for post-translational modifications of vitamin K-dependent (VKD) proteins [31]. Carboxylation is essential for the biological function of VKD proteins that control blood coagulation, vascular calcification, bone metabolism, signal transduction and cancer cell proliferation. However, the specific mechanisms and functions of the *GGCX* gene in tumor growth

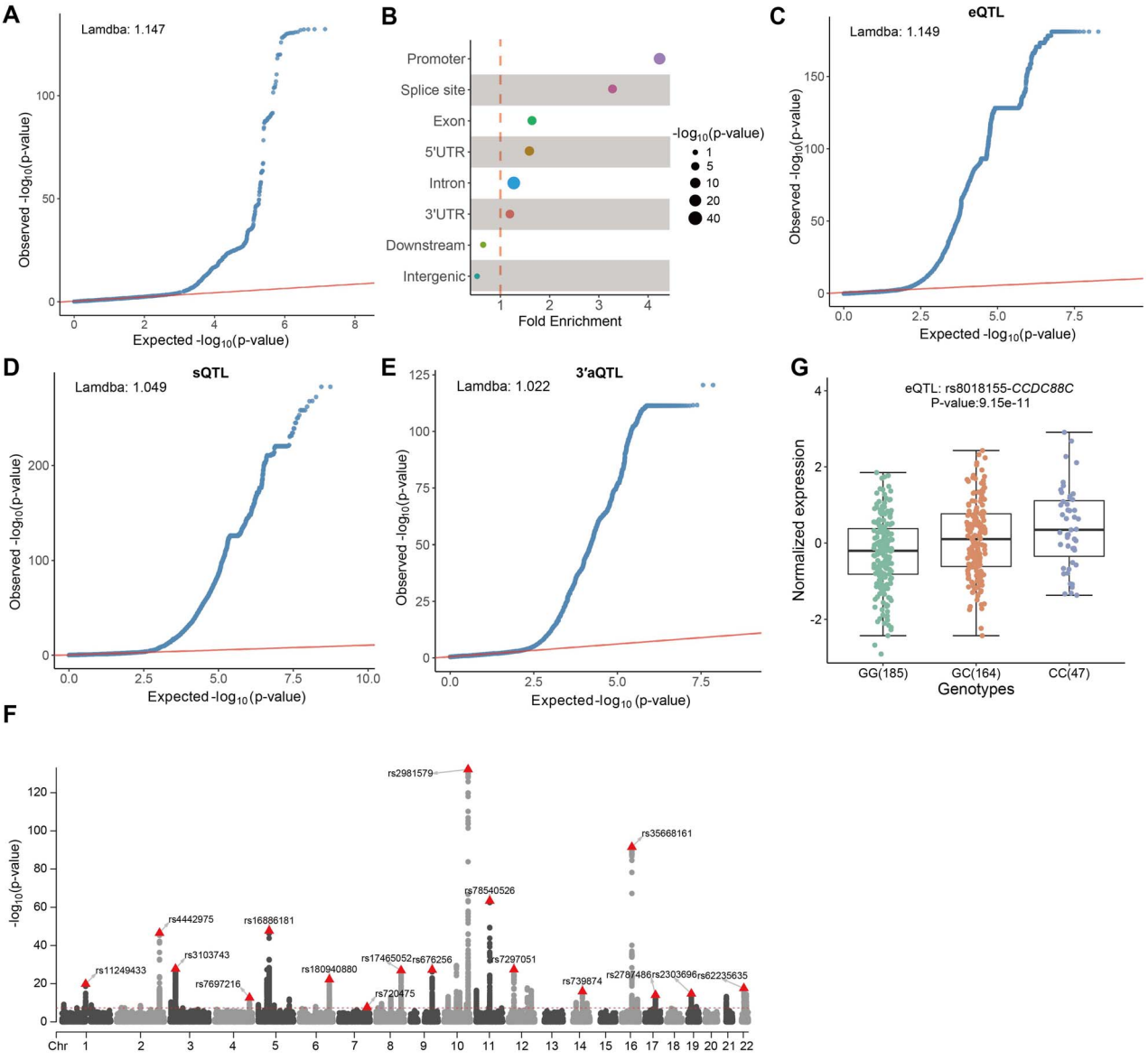

Figure 2. Quality control and annotation of breast cancer and xQTLs summary statistics data. (**A**) QQ-plot labeled with inflation factor. (**B**) PZ-plot, the x-axis stands for the normalized *P*-values estimated by the z-score derived from beta and standard error and the y-axis stands for the raw normalized *P*-values. (**C**) Genomic annotation of significant breast cancer risk SNPs. (**D**) QQ-plot for eQTLs. (**E**) QQ-plot for sQTLs. (**F**) QQ-plot for 3′aQTLs. (**G**) Manhattan plot of the GWAS study of breast cancer. The strongest signals on each chromosome are labeled. (**H**) Boxplot of normalized expression of eQTL rs8018155–*CCDC88C* in the Breast-Mammary Tissue.

and development require further study [32]. We further examined the dependency score which represents the essentiality of these genes in cancer cell lines, obtained from CRISPR screening assay [33]. As shown in the added Figure S5A available online at http://bib.oxfordjournals.org/, four of the 19 candidate genes are ranked in the top 10 essential genes. Notably, xQTL-colocalized genes are largely non-overlapped (Figure 3A), such as *MMP7* was only significantly colocalized by eQTLs (Figure S5B available online at http://bib.oxfordjournals.org/), *AGAP10P* was only colocalized by sQTLs rather than eQTLs and 3′aQTLs. Additionally, *HOXB2* is colocalized by 3′aQTLs instead of eQTLs and sQTLs (Figure S5C available online at http://bib.oxfordjournals.org/). Moreover, five genes in Table S3 available online at http://bib.oxfordjournals.org/ are identified by colocalization with 3′aQTL that have never been detected by traditional eQTLs or sQTLs (Figure 3A). Among them, two known cancer-relevant genes (*ARNT* and *CFLAR*) are founded, and the remaining three genes (*MRPL52*, *HOXB2* and *CCDC97*) are potentially novel susceptibility genes.

To understand the colocalized results, we first visualize the expression distribution of all eQTL-colocalized genes in the TCGA prostate samples (Figure S4B available online at http://bib.oxfordjournals.org/). Besides, we checked *MMP7* gene expression across 49 GTEx tissues by function *xQTLvisual_geneExpTissues*. As shown in Figure 3B, *MMP7* has a relatively high expression level in the prostate and relevant tissues, indicating a potential essential role in these tissues. The distribution of *MMP7* expression level in prostate tissue stratified by the genotype of the lead SNP was also presented by function *xQTLvisual_eqtlExp* (Figure 3C). Then, we visualized the *MMP7*-colocalized signals by function *xQTLvisual_locusZoom*, which reveals a high correlation between GWAS variants and *MMP7* eQTLs (Figure 3D).

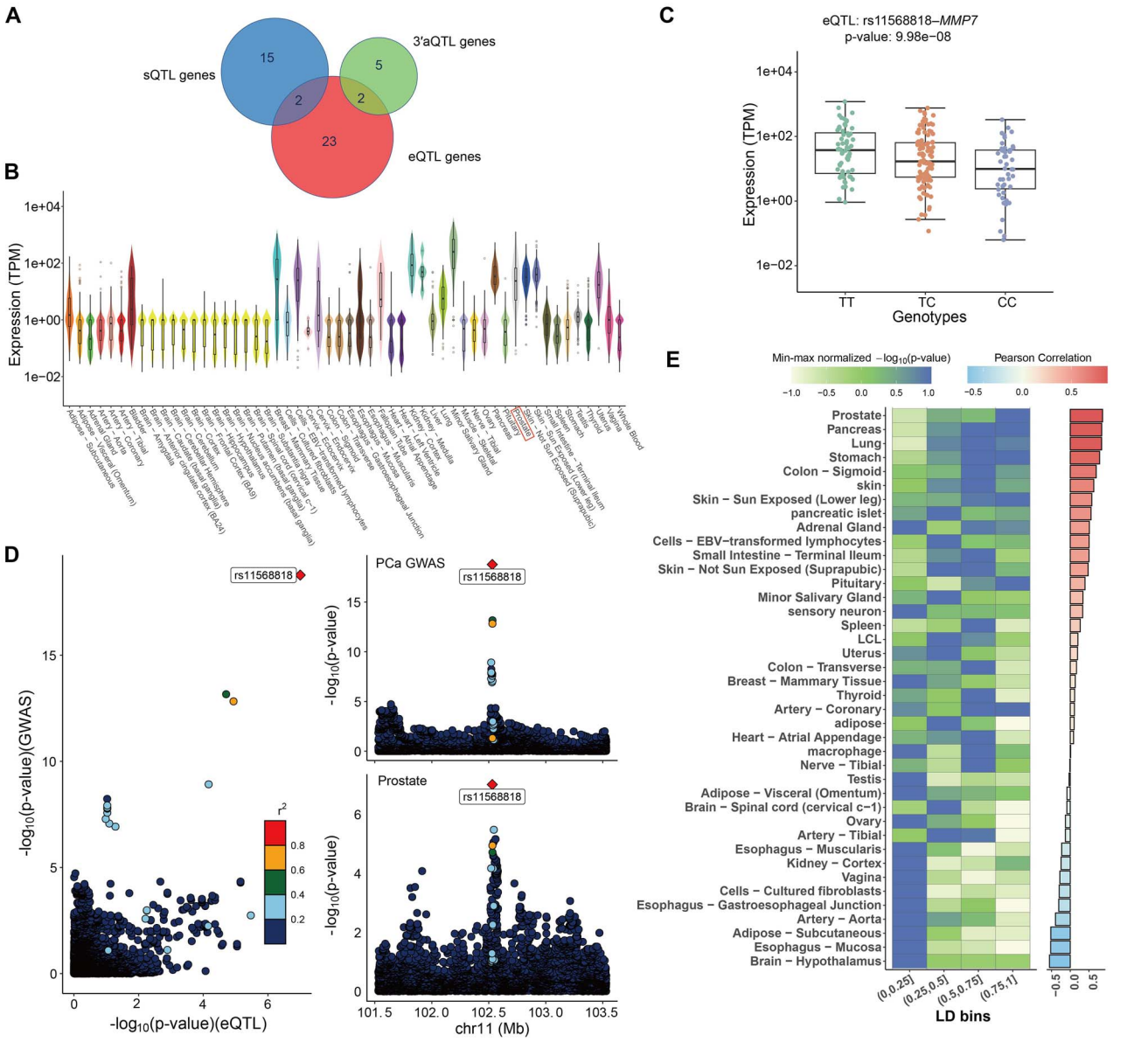

Figure 3. Integrative analysis of GWAS study of prostate cancer. (**A**) Venn plot of 47 xQTL-colocalized genes. (**B**) Gene expression levels (TPM) of *MMP7* among 54 GTEx tissues. (**C**) Boxplot of expression (TPM) of eQTL rs11568818–*MMP7* in the prostate. (**D**) Distribution of GWAS and eQTL signals within a genome region of *MMP7*. PCa, Prostate Cancer. (**E**) Heatmap of eQTL (rs11568818–*MMP7*) *P*-values in different LD bins across 40 tissues/cell types. The y-axis represents the tissues/cell types, and the x-axis represents LD bins. The left panel indicates the median normalized eQTL *P*-values in different LD bins. The right panel represents the Pearson correlation between normalized eQTL *P*-values and values of *r*-squared of LD.

xQTLbiolinks is highly compatible, allowing for seamless integration of its outputs with external packages. For example, we performed gene ontology (GO) enrichment analyses on the eQTL-colocalized genes with external package clusterProfiler [11]. Cancer-related GO terms are significantly enriched, including 'positive regulation of T cell receptor signaling pathway', 'Execution phase of apoptosis' and 'DNA replication checkpoint signalling' (Figure S5D available online at http://bib.oxfordjournals.org/). We conducted gene set enrichment analysis (GSEA) to identify enriched gene sets and pathways from the Molecular Signatures Database (MSigDB), and identified three cancer-related pathways relevant to colocalized genes: 'Regulation of cell death', 'Prostate cancer' and 'Tumor invasiveness up' (Figure S5E available online at http://bib.oxfordjournals.org/). Moreover, we also perform co-expression analysis using the corrplot package [34] (Figure S5F available online at http://bib.oxfordjournals.org/). To investigate the regulatory sharing of colocalized variants across multiple tissues, we implemented *xQTLvisual_coloc* to visualize the correlation between *P*-values of xQTLs LD across numerous tissues/cell types (Figure 3E). We observed that prostate tissues showed the strongest correlation indicating that the heatmap can reveal the potential disease-relevant tissues.

Case study 3: xQTLbiolinks reveals novel cell type-specific susceptibility genes

Systemic lupus erythematosus (SLE) is a complex autoimmune disease, with occurrence heavily influenced by genetics [35]. Joint analysis of immune cell sc-eQTL and SLE GWAS results enable the identification of susceptibility genes and cell types relevant to this immune-mediated disease. Here, we applied xQTLbiolinks to integrate SLE GWAS dataset [36] with sc-eQTL data across

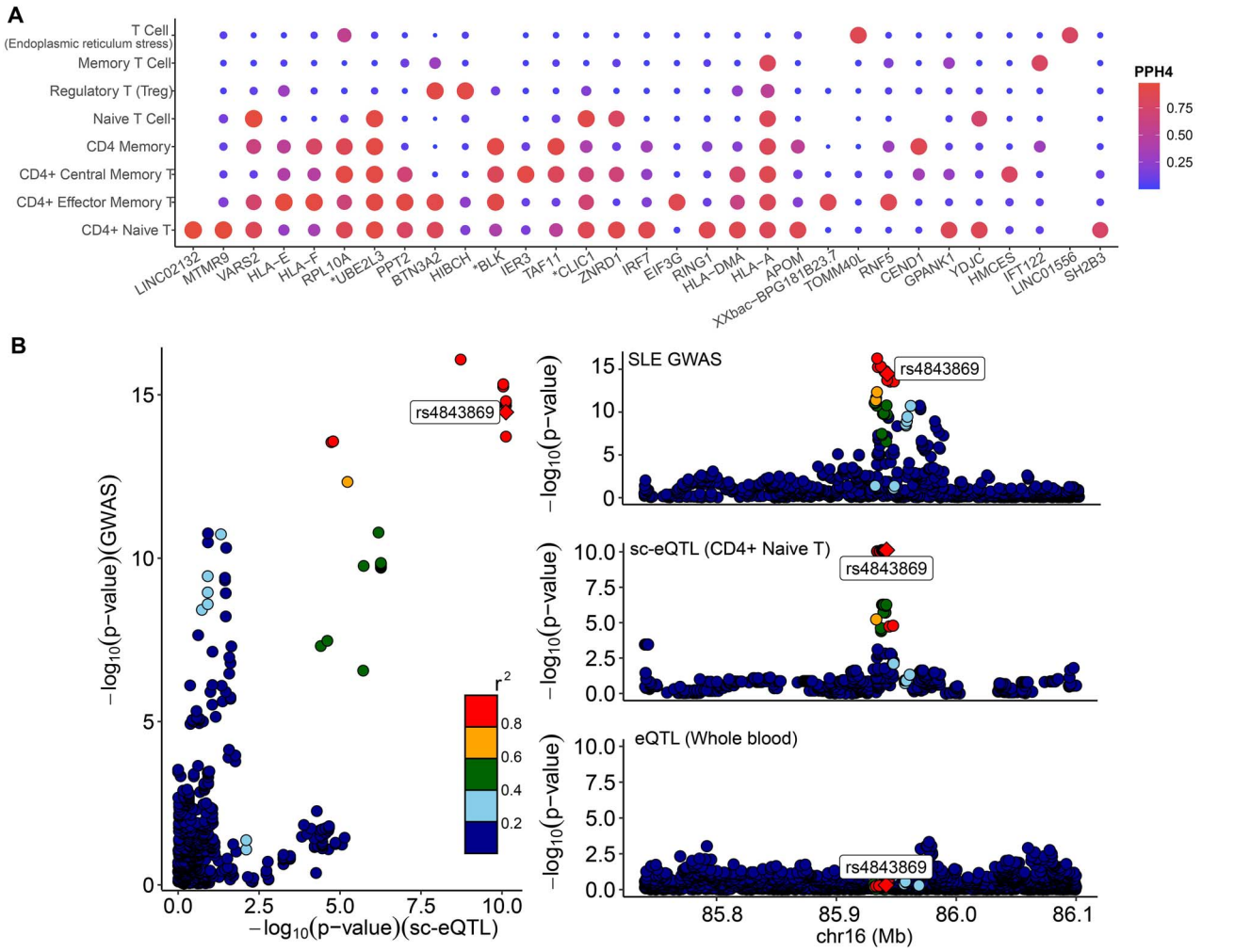

Figure 4. Integrative analysis of GWAS study of systemic lupus erythematosus (SLE) using both sc-eQTL and bulk eQTL. (**A**) Posterior probability (PPH4) of the colocalized genes identified from sc-eQTLs across eight cell types. The x-axis represents the gene names. The y-axis represents the cell types. (**B**) Colocalization of signals from GWAS, sc-eQTL and eQTL for gene *LNIC02132* with SLE-associated loci.

eight immune cells [37] and aimed to identify potential susceptibility genes associated with SLE. We first extracted 46 sentinel SNPs with $P < 5 \times 10^{-8}$ using *xQTLanalyze_getSentinelSnp* and identified 2 690 trait genes using *xQTLanalyze_getTraits*. For each trait gene, we performed colocalization analysis using *xQTLanalyze_coloc_diy* with sc-eQTL from eight immune cells and bulk eQTL data from whole blood tissue. There were 31 sc-eQTL genes, and 17 eQTL genes were colocalized with 18 SLE risk signals using a PPH4 threshold of 0.75 (Tables S4 and S5 available online at http://bib.oxfordjournals.org/). Three genes including *UBE2L3*, *BLK* and *CLIC1* were shared between sc-eQTL and bulk eQTL (Figure 4A). Interestingly, we found 28 sc-eQTL-colocalized genes rather than bulk eQTL. For instance, *LNIC02132*, a long intergenic non-protein coding RNA 2132, is strongly colocalized by sc-eQTL (PPH4 = 0.98) in CD4+ Naive T cells but not colocalized by bulk eQTL (PPH4 = 0.01) (Figure 4B). Additionally, these sc-eQTL-specific genes enable the potential identification of novel cell type-specific susceptibility genes and provide insights into disease genetics and biology.

DISCUSSION

Molecular Quantitative Trait Loci is a crucial step toward better understanding the effects of non-coding genetic variants on genes, pathways and their function mechanism and serves as an essential link between genotype and disease phenotype. Although many xQTLs summary statistics are available, mining and analyzing these xQTLs data remains several major bioinformatics challenges, such as data retrieval, quality control and preprocessing, which are essentially required steps to promote the reproducible use of xQTLs resources and accelerate disease susceptibility genes' identification remains challenging. Here, we developed xQTLbiolinks, which is motivated by TCGAbiolinks that provides several useful functions to search, download and prepare TCGA samples for data analysis [38]. Here, xQTLbiolinks aims to 'link' xQTLs data to disease genomics research by providing flexible interfaces to allow users to access xQTLs data from GTEx and eQTL catalog without having to navigate through different data portal sites or download whole tables of millions of xQTLs associations. xQTLbiolinks also provides the functions for comprehensive querying and mining sc-eQTLs. The current version of xQTLbiolinks provides access to over 4 million bulk eQTLs and sQTLs, and 16 million sc-eQTLs. It enables manipulating user-customized xQTLs data, such as our recently developed 3′UTR alternative polyadenylation QTLs (3′aQTLs) [39]. To date, >20 molecular traits have been profiled [40]. We acknowledge our selection does not fully represent the breadth of xQTL data, such as chromatin accessibility QTLs (caQTLs), RNA

editing QTLs (edQTLs) and ribosome occupancy QTLs (riboQTL). However, our plan includes collecting more data resources, such as the eQTLGen and MetaBrain project [5, 41], and continuously updating and integrating additional QTL data for other molecular traits. This will allow us to provide a comprehensive understanding of the molecular regulatory mechanisms underlying complex diseases and traits by incorporating multiple layers of information.

In addition, colocalization analysis is a powerful approach widely used to identify new susceptibility genes in disease analysis by integrating xQTLs and GWAS signals. However, current colocalization tools have limitations in that they only focus on colocalization methods without considering the whole analysis pipeline and the effects of multiple tissue or cell types. Employing several colocalization tools in the context of numerous tissues or cell types is a superior option to enhance the robustness and reliability of the findings [12]. To facilitate the identification of robust susceptibility genes in colocalization analysis, xQTLbiolinks provides a comprehensive pipeline that employs two popular colocalization methods and handles upstream data processing and downstream visualization of results in a one-step solution. Notably, for diseases with no obvious relevant tissue, we recommend using previously published method TCSC [42] to identify the most relevant tissue or cell type.

xQTLbiolinks is a scalable tool that facilitates integrating and utilizing external tools or packages. We will also actively maintain xQTLbiolinks and respond to user inquiries. As the first end-to-end bioinformatics framework for mining and analyzing xQTL data for discovering disease susceptibility genes, it will make significant contributions toward our understanding of human non-coding variants, thus promoting a better understanding of disease etiologies.

Key Points

- xQTLbiolinks is the first end-to-end bioinformatics tool for mining and analyzing bulk and single-cell xQTL data for discovering disease susceptibility genes.
- xQTLbiolinks provides flexible interfaces for xQTLs data retrieval, pre-processing and quality controls from 75 human tissues and cell types and our own integrated resources.
- xQTLbiolinks provided a robust colocalization module through integration with GWAS data. The result generated by xQTLbiolinks can be flexibly visualized or stored in standard R objects that can easily be integrated with other R packages and custom pipelines.

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.

SUPPLEMENTARY DATA

Supplementary data are available online at https://academic.oup.com/bib.

FUNDING

This work was supported by the National Natural Science Foundation of China (no. 32100533, 32370721 to L.L.) and Open grant funds from Shenzhen Bay Laboratory (no. SZBL2021080601001 to L.L.).

AUTHOR CONTRIBUTIONS

Ruofan Ding: Conceptualization, Formal analysis, Methodology, Validation, Writing—original draft. Xudong Zou: Conceptualization, Formal analysis, Validation, Writing—review & editing. Xuelian Ma & Yangmei Qin: Formal analysis, Validation. Hui Chen & Lihai Gong: Writing-review. Shouhong Guang: Review & editing. Chen Yu: Writing—review & editing. Gao Wang: Methodology, Writing—review & editing. Lei Li: Conceptualization, Methodology, Validation, Writing—review & editing.

DATA AVAILABILITY

xQTLbiolinks and the standard manual (version 1.6.2) are publicly available at CRAN https://cran.r-project.org/web/packages/xQTLbiolinks. The source codes of the latest version of xQTLbiolinks can be found in GitHub repository https://github.com/lilab-bioinfo/xQTLbiolinks.

REFERENCES

1. Hormozdiari F, Gazal S, van de Geijn B, *et al.* Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat Genet* 2018;**50**:1041–7.
2. Consortium GT. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science* 2020;**369**:1318–30.
3. Kerimov N, Hayhurst JD, Peikova K, *et al.* A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat Genet* 2021;**53**:1290–9.
4. Schmiedel BJ, Singh D, Madrigal A, *et al.* Impact of genetic polymorphisms on human immune cell gene expression. *Cell* 2018;**175**:1701, e1716–15.
5. Vosa U, Claringbould A, Westra HJ, *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet* 2021;**53**:1300–10.
6. Oliva M, Demanelis K, Lu Y, *et al.* DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nat Genet* 2023;**55**:112–22.
7. Chen L, Ge B, Casale FP, *et al.* Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell* 2016;**167**:1398, e1324–414.
8. Giambartolomei C, Vukcevic D, Schadt EE, *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* 2014;**10**:e1004383.
9. Foley CN, Staley JR, Breen PG, *et al.* A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nat Commun* 2021;**12**:764.
10. Chen BY, Bone WP, Lorenz K, *et al.* ColocQuiaL: a QTL-GWAS colocalization pipeline. *Bioinformatics* 2022;**38**:4409–11.
11. Zhang T, Klein A, Sang J, *et al.* ezQTL: a web platform for interactive visualization and colocalization of QTLs and GWAS loci. *Genomics Proteomics Bioinformatics* 2022;**20**:541–8.
12. Hukku A, Pividori M, Luca F, *et al.* Probabilistic colocalization of genetic variants from complex and molecular traits: promise and limitations. *Am J Hum Genet* 2021;**108**:25–35.

13. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550.
14. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**:139–40.
15. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;**9**:559.
16. Hao Y, Hao S, Andersen-Nissen E, *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**:3573, e3529–87.
17. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;**38**:e164.
18. Watanabe K, Stringer S, Frei O, *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat Genet* 2019;**51**:1339–48.
19. Pe'er I, Yelensky R, Altshuler D, *et al.* Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol* 2008;**32**:381–5.
20. Benjamin DJ, Berger JO, Johannesson M, *et al.* Redefine statistical significance. *Nat Hum Behav* 2018;**2**:6–10.
21. International HapMap C. A haplotype map of the human genome. *Nature* 2005;**437**:1299–320.
22. Hormozdiari F, van de Bunt M, Segre AV, *et al.* Colocalization of GWAS and eQTL signals detects target genes. *Am J Hum Genet* 2016;**99**:1245–60.
23. Drivas TG, Lucas A, Ritchie MD. eQTpLot: a user-friendly R package for the visualization of colocalization between eQTL and GWAS signals. *BioData Min* 2021;**14**:32.
24. Liu B, Gloudemans MJ, Rao AS, *et al.* Abundant associations with gene expression complicate GWAS follow-up. *Nat Genet* 2019;**51**:768–9.
25. Astle WJ, Elding H, Jiang T, *et al.* The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* 2016;**167**:1415, e1419–29.
26. Michailidou K, Beesley J, Lindstrom S, *et al.* Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat Genet* 2015;**47**:373–80.
27. Braga-Basaria M, Dobs AS, Muller DC, *et al.* Metabolic syndrome in men with prostate cancer undergoing long-term androgen-deprivation therapy. *J Clin Oncol* 2006;**24**:3979–83.
28. Schumacher FR, Al Olama AA, Berndt SI, *et al.* Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat Genet* 2018;**50**:928–36.
29. Zhang Q, Liu S, Parajuli KR, *et al.* Interleukin-17 promotes prostate cancer via MMP7-induced epithelial-to-mesenchymal transition. *Oncogene* 2017;**36**:687–99.
30. Tregunna R. Serum MMP7 levels could guide metastatic therapy for prostate cancer. *Nat Rev Urol* 2020;**17**:658.
31. Berkner KL, Runge KW. Vitamin K-dependent protein activation: normal gamma-glutamyl carboxylation and disruption in disease. *Int J Mol Sci* 2022;**23**:5759.
32. Hao Z, Jin DY, Chen X, *et al.* Gamma-Glutamyl carboxylase mutations differentially affect the biological function of vitamin K-dependent proteins. *Blood* 2021;**137**:533–43.
33. Meyers RM, Bryan JG, McFarland JM, *et al.* Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat Genet* 2017;**49**:1779–84.
34. Taiyun Wei VS. *R Package 'Corrplot': Visualization of a Correlation Matrix*, 2021. https://github.com/taiyun/corrplot.
35. Lawrence JS, Martins CL, Drake GL. A family survey of lupus erythematosus. 1. Heritability. *J Rheumatol* 1987;**14**:913–21.
36. Perez RK, Gordon MG, Subramaniam M, *et al.* Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. *Science* 2022;**376**:eabf1970.
37. Soskic B, Cano-Gamez E, Smyth DJ, *et al.* Immune disease risk variants regulate gene expression dynamics during CD4(+) T cell activation. *Nat Genet* 2022;**54**:817–26.
38. Colaprico A, Silva TC, Olsen C, *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* 2016;**44**:e71.
39. Li L, Huang KL, Gao Y, *et al.* An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. *Nat Genet* 2021;**53**:994–1005.
40. Zheng Z, Huang D, Wang J, *et al.* QTLbase: an integrative resource for quantitative trait loci across multiple human molecular phenotypes. *Nucleic Acids Res* 2020;**48**:D983–91.
41. de Klein N, Tsai EA, Vochteloo M, *et al.* Brain expression quantitative trait locus and network analyses reveal downstream effects and putative drivers for brain-related diseases. *Nat Genet* 2023;**55**:377–88.
42. Amariuta T, Siewert-Rocks K, Price AL. Modeling tissue co-regulation estimates tissue-specific contributions to disease. *Nat Genet* 2023;**55**:1503–11.