Bioinformatics, 38(18), 2022, 4442–4445
https://doi.org/10.1093/bioinformatics/btac524
Advance Access Publication Date: 21 July 2022
Applications Note

Data and text mining

REALGAR: a web app of integrated respiratory omics data

Mengyuan Kan †, Avantika R. Diwadkar†, Supriya Saxena, Haoyue Shuai, Jaehyun Joo and Blanca E. Himes*

Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104, USA

*To whom correspondence should be addressed.

†The authors wish it to be known that these authors contributed equally.

Associate Editor: Karsten Borgwardt

Received on February 9, 2022; revised on June 22, 2022; editorial decision on July 19, 2022; accepted on July 20, 2022

Abstract

Motivation: In the post genome-wide association study (GWAS) era, omics techniques have characterized information beyond genomic variants to include cell and tissue type-specific gene transcription, transcription factor binding sites, expression quantitative trait loci (eQTL) and many other biological layers. Analysis of omics data and its integration has in turn improved the functional interpretation of disease-associated genetic variants. Over 170 000 transcriptomic and epigenomic datasets corresponding to studies of various cell and tissue types under specific disease, treatment and exposure conditions are available in the Gene Expression Omnibus resource. Although these datasets are valuable to guide the design of experimental validation studies to understand the function of disease-associated genetic loci, in their raw form, they are not helpful to experimental researchers who lack adequate computational resources or experience analyzing omics data. We sought to create an integrated re-source of tissue-specific results from omics studies that is guided by disease-specific knowledge to facilitate the design of experiments that can provide biologically meaningful insights into genetic associations.

Results: We designed the Reducing Associations by Linking Genes and omics Results web app to provide multi-layered omics information based on results from GWAS, transcriptomic, epigenomic and eQTL studies for gene-centric analysis and visualization. With a focus on asthma datasets, the integrated omics results it contains facilitate the formulation of hypotheses related to airways disease-associated genes and can be addressed with experimental validation studies.

Availability and implementation: The REALGAR web app is available at: http://realgar.org/. The source code is available at: https://github.com/HimesGroup/realgar.

Contact: bhimes@pennmedicine.upenn.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Following the successful completion of thousands of genome-wide association studies (GWASs), it remains clear that linking statistical signals to biological mechanisms underlying complex traits is difficult. The reasons for this include that (i) most associated loci are not in protein-coding genes or in/near genes with functions linked to the trait studied, (ii) functional studies are time-consuming as each association follow-up experiment has to be tailored to a particular phenotype that represents a complex disease and type of polymorphisms in a local association region and (iii) in order to test genes and variants for their function, complex diseases have to be simplified into assays that may not capture the cell-specific, developmental and/or environmental context necessary for functional elucidation of their role. Thus, novel *in silico* approaches that screen genes and variants for potential function based on biological information are helpful to guide the efficient validation of top GWAS findings.

Functional validation studies of gene associations often begin with searches for what is known about a specific locus, including whether associated variants are expression quantitative trait loci (eQTL) and/or are located within transcription regulatory regions of nearby genes. This valuable single-nucleotide polymorphism (SNP)-level information has become increasingly available thanks to large-scale efforts such as the genotype-tissue expression (GTEx) project and ENCODE. However, these resources are limited in their ability to provide results for the design of disease-specific experiments because their results mostly reflect baseline conditions and were obtained with immortalized cell lines or whole tissues rather than prominent disease-specific cell types. To obtain mechanistic clues

about how loci in/near genes may modify biological pathways relevant to the trait under consideration, researchers often search public databases to find out the tissue(s) where identified genes are expressed and under what disease and treatment conditions they are differentially expressed. Public gene expression and transcription factor binding site data from resources such as the Gene Expression Omnibus (GEO)—which as of April 2022, contains over 170 000 individual studies—are a primary resource for answering these questions, but many experimental researchers do not have the expertise or dedicated computational resources necessary to obtain and integrate gene expression microarray, RNA-Seq and epigenomic results. Even researchers who do have such resources may repeat similar analytical tasks every time a new association study is performed. Having integrated resources of tissue-specific *results* from omics studies that are guided by disease-specific knowledge facilitates the prioritization and design of experiments that can provide biologically meaningful insights. In addition, integration of omics data is helpful to generate hypotheses based on the ranking of genes across experiments and biological layers. Motivated in part by frequent queries made to us by experimental investigators who sought to validate asthma-related genetic association findings, we previously designed the Reducing Associations by Linking Genes and omics Results (REALGAR) web app (Shumyatcher *et al.*, 2017) to display integrated asthma-related transcriptomic and GWAS results for 25 transcriptomic datasets and 3 asthma GWAS. Here, we describe how we have expanded REALGAR to include more respiratory omics datasets and to provide visualization and retrieval of results corresponding to specific genes or SNPs in multiple cell and tissue types, as well as disease and exposure conditions.

2 REALGAR design

REALGAR (http://realgar.org/) was developed with the RStudio R Shiny package. A detailed description of its design and features can be found in Supplementary Note S1 and its full code is available at https://github.com/HimesGroup/realgar.

2.1 Data pre-processing

We analyzed 72 transcriptomic datasets representing 25 cell/tissue types, 8 asthma endotypes and 9 asthma-related drug and environmental exposures with our transcriptomic pipeline RAVED (Kan *et al.*, 2018). In addition, we analyzed three glucocorticoid receptor (GR) ChIP-Seq datasets with our ChIP-Seq pipeline brocade (Diwadkar *et al.*, 2019) given the prominence of this transcription factor in the mechanism of action of a major asthma drug (i.e. glucocorticoids) (Kan and Himes, 2020). Glucocorticoid response element (GRE) motifs, where GRs may bind DNA, were identified with the FIMO tool (Grant *et al.*, 2011). GWAS results from five published asthma studies were downloaded from author-provided repositories, and in-house GWAS results from our UK Biobank study of asthma and COPD were included. Genes, SNPs, transcription factor binding sites, GRE motifs and eQTLs were mapped to the hg38 genome build to ensure use of consistent genomic coordinates. More information on datasets included can be found in Supplementary Note S2. Results of individual datasets analyzed were stored as a SQLite database.

2.2 User input interface

Users make queries by providing an official gene symbol or SNP identifier and selecting from among the tissue types, conditions, treatments and GWAS datasets available (Supplementary Fig. S1). Interactive results and graphs are generated with every query.

2.3 User output interface

2.3.1 Omics results tab

This section displays results based on user selections of (i) tissue- and disease-specific differential expression results for asthma endotypes, asthma-related drugs and environmental exposures as forest plots (Supplementary Fig. S2); (ii) boxplots of normalized gene expression levels across disease and exposure conditions of individual RNA-Seq datasets (Supplementary Fig. S3) and (iii) genomic tracks of GWAS, ChIP-Seq, GRE motif and eQTL results (Supplementary Fig. S4). Users can download the image files of forest plots, gene-tracks and boxplots displayed, as well as the results displayed for further analyses.

2.3.2 Datasets loaded tab

This section describes the data sources for which results are displayed, along with HTML reports of quality control metrics for transcriptomic and ChIP-Seq datasets (Supplementary Fig. S5).

3 Example use cases

REALGAR has been helpful to generate data for published studies of asthma drug response (Kan *et al.*, 2021, 2019). Next, we provide two novel examples that illustrate how REALGAR is helpful to generate hypotheses (i) related to disease-associated genetic loci and (ii) based on integrated transcriptomic data.

3.1 Potential function of *HDAC7* variants in asthma

A genetic locus near the histone deacetylase 7 (*HDAC7*) gene has been associated with allergic diseases and asthma (Ferreira *et al.*, 2019; Pividori *et al.*, 2019). Although histone deacetylases (HDACs) are potential targets for asthma therapy (Zwinderman *et al.*, 2019), the mechanisms by which *HDAC7* and its variants influence asthma are not fully understood. A query of *HDAC7* in REALGAR shows that several asthma-associated SNPs (P-value $<1\times10^{-5}$) are located within 20 kb of its transcription start site. Among these SNPs, three are eQTLs of an antisense transcript of *HDAC7* in lung tissue and are within glucocorticoid-responsive GR-binding sites that contain a putative GRE motif. In addition, *HDAC7* gene expression levels were significantly increased with glucocorticoid exposure in two airway smooth muscle RNA-Seq studies (q-value <0.05) (Fig. 1a). Together, these findings suggest that three SNPs near *HDAC7* contribute to asthma via their influence on GR-modulated glucocorticoid responses in airway smooth muscle cells. Functional validation studies could test whether alleles corresponding to these *HDAC7* variants differ in their GR-binding activity with glucocorticoid exposure, thereby promoting differential *HDAC7* transcription in a specific airway cell type.

3.2 Influence of cigarette smoke exposure on viral entry gene expression

Angiotensin-converting enzyme 2, which is encoded by the *ACE2* gene, is a known SARS-CoV-2 virus host receptor (Hoffmann *et al.*, 2020). Cigarette smoking has been associated with increased risk of COVID-19 and worse outcomes (Killerby *et al.*, 2020), which leads to the hypothesis that smoking increases *ACE2* gene expression in airway epithelium thereby promoting infectivity via increased SARS-CoV-2 entry into host cells. By querying REALGAR for *ACE2* and cigarette exposure in airway cell types, results are provided for 18 transcriptomic studies with comparisons involving (i) *in vitro* smoke exposure versus vehicle control and (ii) smokers versus non-smokers (Fig. 1b). These results show that *ACE2* gene expression levels were significantly increased in small airway epithelium of smokers in three studies (q-value<0.05), but they were not consistently changed in other airway cell types derived from smokers or in *in vitro* models where cells were exposed to cigarette smoke. Thus, the mechanism linking cigarette smoking to COVID-19 is more likely to involve long-term changes in *ACE2* expression in small airway epithelium than result from short-term induction of this gene's expression. Consistent with these findings, published studies have reported increased *ACE2* gene expression in small airways of smokers versus non-smokers (Cai *et al.*, 2020; Zhang *et al.*, 2020).

4 Discussion

REALGAR is focused on respiratory-related cell types, conditions and exposures, with an emphasis on asthma. We purposefully built REALGAR to focus on these areas to address a gap identified via

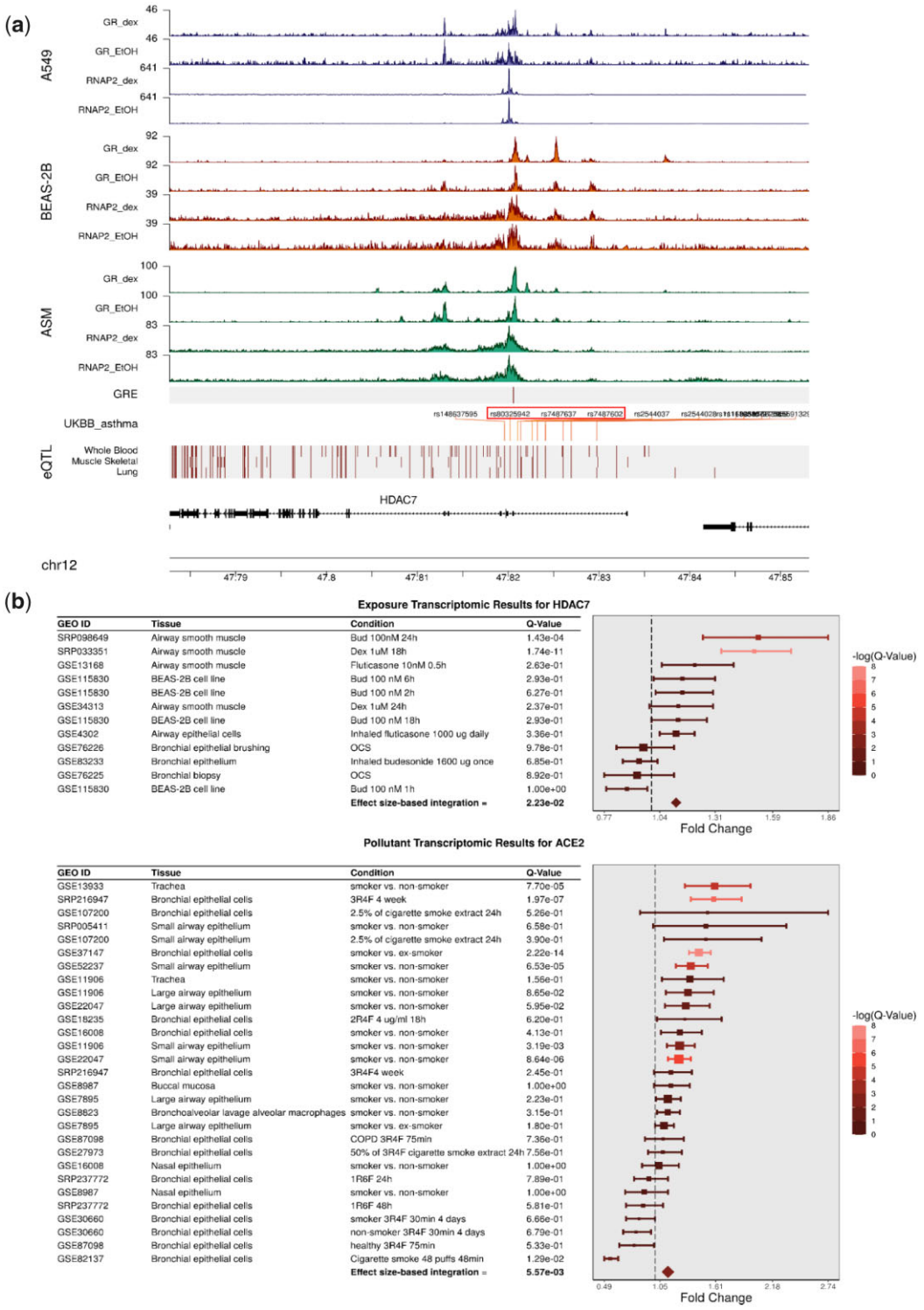

Fig. 1. REALGAR applications. (**a**) Prioritization of HDAC7 variants nominally associated with asthma based on multiomics evidence. Three GR-binding site-overlapping SNPs in the red box are prioritized. (**b**) Differential expression results of ACE2 gene in response to smoking exposure in transcriptomic studies of various airway cell types

our work with experimental researchers who sought to validate asthma gene-association results, namely, that information on SNP function from existing databases was not adequate to effectively design experiments. In ongoing work, we are expanding the scope of REALGAR to include datasets related to COPD and sepsis, as well as additional exposures (e.g. pro-inflammatory reagents). Future efforts will incorporate datasets related to other complex diseases and explore use of automated tools to maintain the app up to date. As single-cell RNA-Seq data become more widely available, we will include it to provide information related to cell population-specific gene expression. We will also provide an option for users to upload their own omics datasets so that these data can be analyzed and corresponding results can be visualized in the context of the integrated data stored in REALGAR. We currently provide integrative scores

on-the-fly for transcriptomic data selected, which is helpful to conduct gene expression meta-analyses. Inclusion of scoring metrics that span omics modalities would be additionally helpful even though such scores remain biased and arbitrary from a biological perspective. Future versions of REALGAR will include multiomics scoring metrics that provide overall rankings for user-selected datasets based on weighted ranks of individual datasets. The goal of such scores will be to increase the ability of users to obtain evidence-based hypotheses for the design of experimental studies.

A limitation of REALGAR is that SNPs are linked to genes based on physical proximity according to a 20-kb distance to gene borders. Use of physical proximity along a chromosome alone may not identify the causal gene that underlies a GWAS signal (Ragvin *et al.*, 2010) as relevant SNP–gene relationships may occur via cell-type-specific enhancer-gene regulation (Nasser *et al.*, 2021) or chromosome conformation (Kong and Jung, 2020). While analysis of colocalized eQTL data can be helpful to prioritize SNP–gene relationships, causal relationships may not have the strongest observed statistical associations (Liu *et al.*, 2019). REALGAR includes all possible colocalizations within a gene region by presenting GWAS SNPs and cis-eQTLs with *P*-values <0.05 for asthma-relevant tissues, along with other layers of omics data that may support SNP–gene relationships as being related to disease. However, some SNPs may be biologically linked to more distant genes. In future versions of the app, we will include candidate genes for SNPs based on varying distances to gene borders, trans-eQTL data, putative enhancer data and chromosome conformation capture data. As these efforts to link variants to genes improve, users are encouraged to query all genes that they hypothesize are related to a SNP signal to find evidence that the gene is relevant based on the other omics data in REALGAR.

Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 40375.

Funding

This work was supported by the National Institutes of Health (NIH) [R01-HL133433 and R01-HL141992] and the Center of Excellence in Environmental Toxicology [P30 ES013508].

Conflict of Interest: none declared.

Data availability

Results of GWAS studies can be downloaded from links provided in the original publications, which are listed in the Supplementary Note. UK Biobank GWAS results can be provided upon request. Raw data from transcriptomic and ChIP-Seq studies are available in GEO with accession numbers displayed in REAGLAR website.

References

Cai,G. *et al.* (2020) Tobacco smoking increases the lung gene expression of ACE2, the receptor of SARS-CoV-2. *Am. J. Respir. Crit. Care Med.*, **201**, 1557–1559.

Diwadkar,A.R. *et al.* (2019) Facilitating analysis of publicly available ChIP-Seq data for integrative studies. *AMIA Annu. Symp. Proc.*, **2019**, 371–379.

Ferreira,M.A.R. *et al.*; BIOS Consortium. (2019) Genetic architectures of childhood- and adult-onset asthma are partly distinct. *Am. J. Hum. Genet.*, **104**, 665–684.

Grant,C.E. *et al.* (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.

Hoffmann,M. *et al.* (2020) SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell*, **181**, 271–280.e8.

Kan,M. *et al.* (2019) Airway smooth muscle-specific transcriptomic signatures of glucocorticoid exposure. *Am. J. Respir. Cell Mol. Biol.*, **61**, 110–120.

Kan,M. *et al.* (2018) Integration of transcriptomic data identifies global and cell-specific asthma-related gene expression signatures. *AMIA Annu. Symp. Proc.*, **2018**, 1338–1347.

Kan,M. *et al.* (2021) Multiomics analysis identifies BIRC3 as a novel glucocorticoid response-associated gene. *J Allergy Clin. Immunol.*, **149**, 1981–1991.

Kan,M. and Himes,B. E. (2020) Insights into glucocorticoid responses derived from omics studies. *Pharmacol. Ther.*, **218**, 107674.

Killerby,M.E. *et al.*; CDC COVID-19 Response Clinical Team. (2020) Characteristics associated with hospitalization among patients with COVID-19—metropolitan Atlanta, Georgia, March–April 2020. *MMWR. Morb. Mortal. Wkly. Rep.*, **69**, 790–794.

Kong,N. and Jung,I. (2020) Long-range chromatin interactions in pathogenic gene expression control. *Transcription*, **11**, 211–216.

Liu,B. *et al.* (2019) Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.*, **51**, 768–769.

Nasser,J. *et al.* (2021) Genome-wide enhancer maps link risk variants to disease genes. *Nature*, **593**, 238–243.

Pividori,M. *et al.* (2019) Shared and distinct genetic risk factors for childhood-onset and adult-onset asthma: genome-wide and transcriptome-wide studies. *Lancet Respir. Med.*, **7**, 509–522.

Ragvin,A. *et al.* (2010) Long-range gene regulation links genomic type 2 diabetes and obesity risk regions to HHEX, SOX4, and IRX3. *Proc. Natl. Acad. Sci. USA*, **107**, 775–780.

Shumyatcher,M. *et al.* (2017) Disease-specific integration of omics data to guide functional validation of genetic associations. *AMIA Annu. Symp. Proc.*, **2017**, 1589–1596.

Zhang,H. *et al.* (2020) Expression of the SARS-CoV-2 ACE2 receptor in the human airway epithelium. *Am. J. Respir. Crit. Care Med.*, **202**, 219–229.

Zwinderman,M.R.H. *et al.* (2019) Targeting HDAC complexes in asthma and COPD. *Epigenomes*, **3**, 19.