Building Research Portals Without Coding: The Bring Your Own Results Portal Service (BYOR)

Dongkeun Jang[1], Maria C. Costanzo[1], Jason Flannick[1,2,3], and Noël P. Burtt[1]

Affiliations

1. Programs in Metabolism and Medical & Population Genetics, The Broad Institute of MIT and Harvard, Cambridge, MA 02132, USA
2. Department of Pediatrics, Boston Children's Hospital, Boston, MA 02115, USA
3. Department of Pediatrics, Harvard Medical School, Boston, MA 02115, USA

Summary

Sharing results has always been an essential part of the scientific method and is especially important in current biomedical research, where the translation of large and complex datasets into knowledge requires that data are integrated and shared. Yet the complexity of the results can present a barrier to data sharing for biologists who may lack computational expertise. To help researchers surmount this barrier we have created the open-access Bring Your Own Results (BYOR) service, which facilitates visualization, analysis, and sharing of biomedical results.

Introduction

Our group has developed the HuGeAMP infrastructure that powers the Common Metabolic Diseases Knowledge Portal and other open access portals to integrate genetic, genomic, and epigenomic datasets and facilitate the understanding and treatment of common diseases. We constructed the BYOR service as a light-weight offshoot of the HuGeAMP infrastructure in response to two gaps faced by researchers. First, for scientists who may not be expert in creating data visualizations, it can be challenging to extract meaning from complex data types and to convey that meaning to the larger scientific community. Second, datasets and data visualizations must be shared – privately within a group of collaborators or with manuscript or grant reviewers, and then with the larger scientific community – but not all researchers have avenues for creating websites to share data.

The BYOR concept grew from conversations with scientists who had generated data that were relevant to common diseases but were not suitable for display on the Knowledge Portals, either because the data were unpublished and private, or because the portals did not yet offer a way to incorporate and visualize those data types. In addition to facilitating data sharing, the Research Portals and Research Pages created using BYOR can serve as staging areas for

exploring the best ways to represent certain data types, as a prelude to their eventual incorporation into the Knowledge Portals.

In light of the recently released policy from the National Institutes of Health that will require NIH-funded researchers to share data publicly[1,2], the availability of straightforward ways to share data will become increasingly important. Here we present the BYOR service as one way to address this goal.

Results

The BYOR platform is composed of three parts: the data; the configuration user interface; and the web page presentation (Fig. 1). The configuration specifies the data source, how to render the loaded data for display, and which visualizer(s) to use on the presentation page.

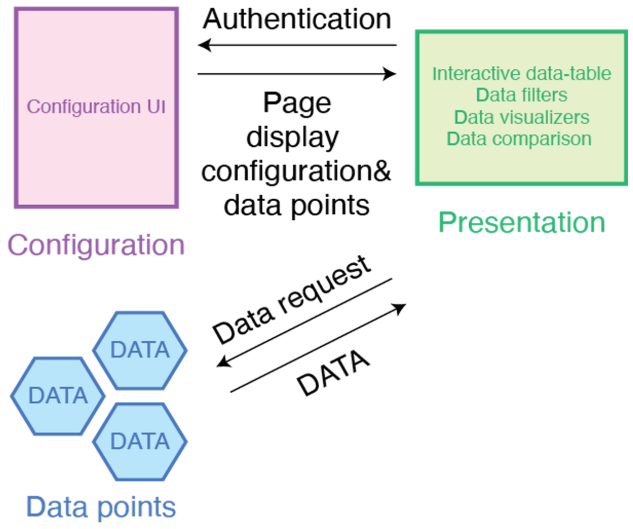


Figure 1. Overview of the BYOR platform. The presentation page (green) requests data (blue) from the data source and loads the data into an interactive table and visualizers. If desired, the presentation page may require authentication (login) in order to be viewed. The configuration (pink), set by the page creator, specifies the format of the data table(s), which visualizers are displayed, and the overall structure of the Research Portal.

Data types and sources

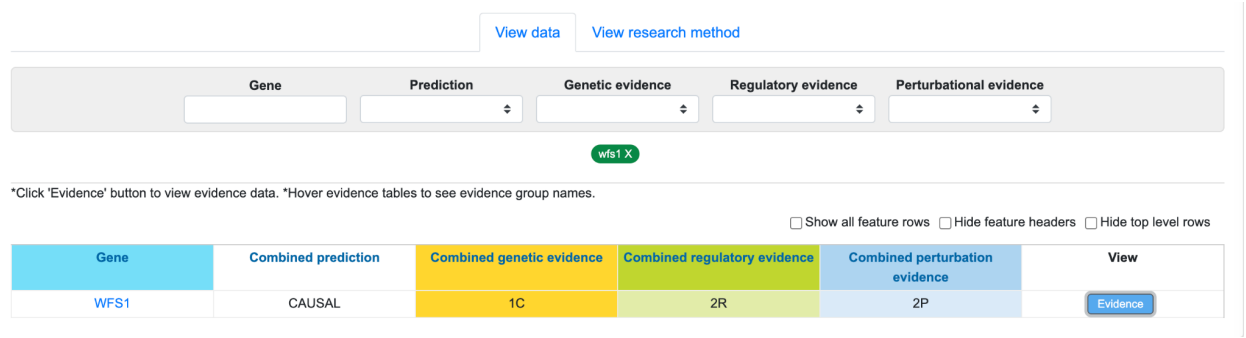
BYOR can accommodate data represented as numeric values, text, or both (see below for examples). Data must be in one of three formats: comma-separated values (.csv), JavaScript Object Notation (JSON), or indexed by our project's custom software, "BioIndex", that enables rapid access to data for specified genomic regions.

Relatively small files, less than 8 MB in size, may be uploaded to our server and stored there. Alternatively, data may be served from a file hosting service, or may be retrieved from a remote source via dynamic APIs. Data transfer from remote sources must be performed using the HTTPS internet communication protocol. Datasets that are relevant to the Knowledge Portals, such as genetic associations for common diseases and related traits, may be served from our project's cloud storage after indexing by the BioIndex software.

Presentation of results

The default view of data in BYOR is in customizable interactive tables (Fig. 2). Filters may be added to allow viewers to filter the data by specific criteria (*e.g.*, gene name or chromosomal location). Table cells may include numeric values, text, or links to web resources such as one of the Knowledge Portals or PubMed. Expandable nested sub-tables may be added to each row to hold additional information.

A.



B.

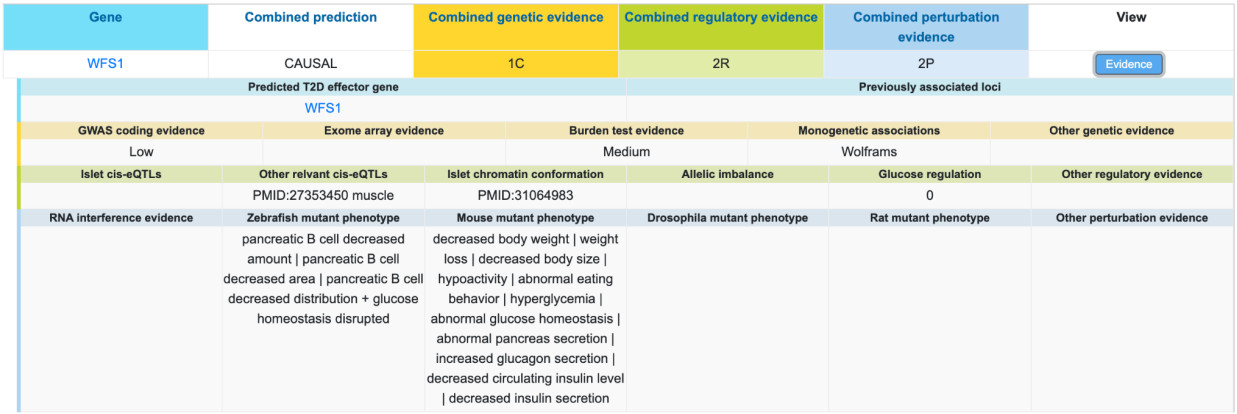


Figure 2. Interactive table structure in BYOR, illustrated with the curated type 2 diabetes (T2D) effector prediction[3] interface on the Type 2 Diabetes Knowledge Portal. A: Above the table, viewers may filter the table by several criteria. The figure shows the table after filtering by the gene name *WFS1*. B: Clicking the "Evidence" button expands the row to several additional rows of information supporting the prediction that *WFS1* is a causal gene for T2D. The headers of the additional rows are color-coded to indicate the type of information that they contain.

In addition to simply representing data as they appear in the source file, BYOR has the capacity to join, convert, calculate, and score values in the loaded data. For example, the chromosome, genomic coordinate, and reference and alternate alleles for a variant, all contained in separate columns in the input file, can be combined to generate a variant ID. Calculations may be performed on the data, such as generating a score based on the contents of specific table cells. BYOR can also render a bar chart within a table cell to represent the value in that cell. Multiple datasets may be combined in a single table in order to compare them.

Data accessed by BYOR may be visualized in multiple types of plot. Basic plots, applicable to many different data types, include line, bar, and pie charts. More specialized plot types include:

- Manhattan plots, for visualizing associations across the genome;
- Volcano plots, scatterplots of significance versus fold change that are useful for highlighting the most important changes (*e.g.*, differential gene expression) between two conditions (Fig. 3);
- Heat maps, matrices that represent numeric values into color intensity (Fig. 4);
- Region plots, showing genetic associations for specific phenotypes across a defined genomic region (Fig. 5); and
- Score plots, similar to Manhattan plots, for displaying gene-based scores relative to the genomic position of each gene.

A demonstration Research Portal, accessible from the BYOR tutorial, shows examples of each of these visualizations along with the data configuration underlying the plot.

Users of BYOR may create a single Research Page, including a data table and a visualization, or may combine multiple Research Pages into a Research Portal. BYOR offers the ability to add web pages for text and documentation, along with menus allowing the user to navigate through the Research Portal.

Figure 3. A volcano plot on the *APOL1* Portal[4,5] showing differential gene expression between high- and low-risk *APOL1* genotypes.

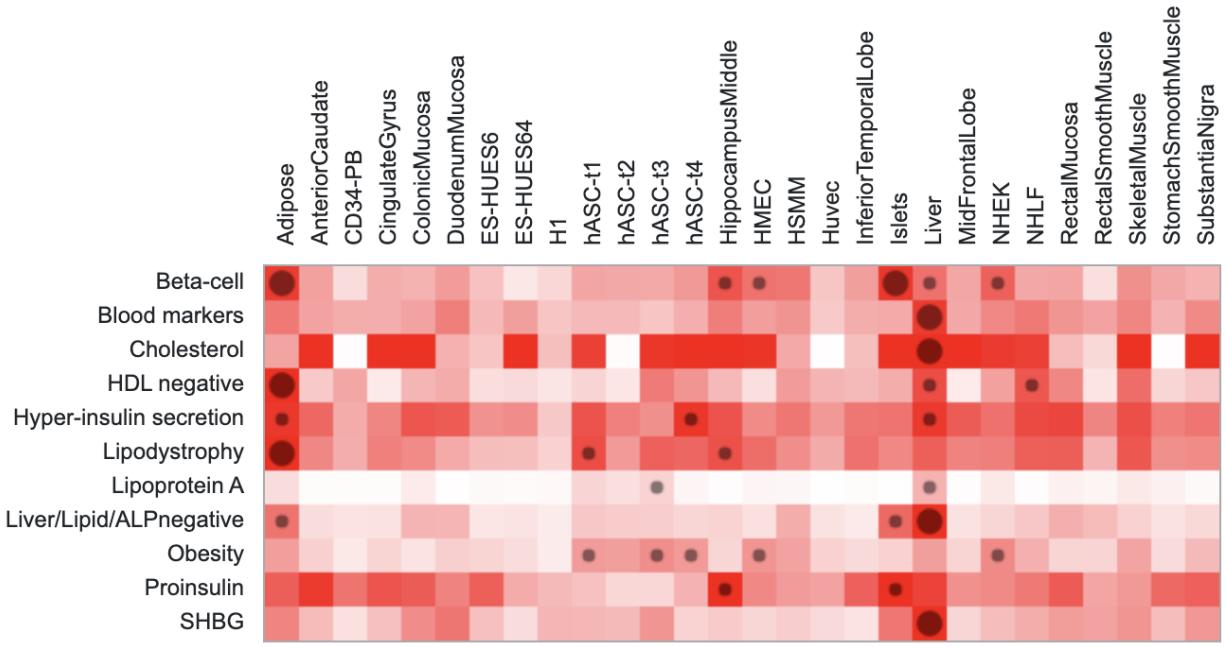


Figure 4. A heat map showing effect size, represented by color intensity, and p-value, represented by the size of the circles within the cells, for tissue-specific epigenomic enrichment of clustered genetic associations for T2D, coronary artery disease, and chronic kidney disease.[6,7]

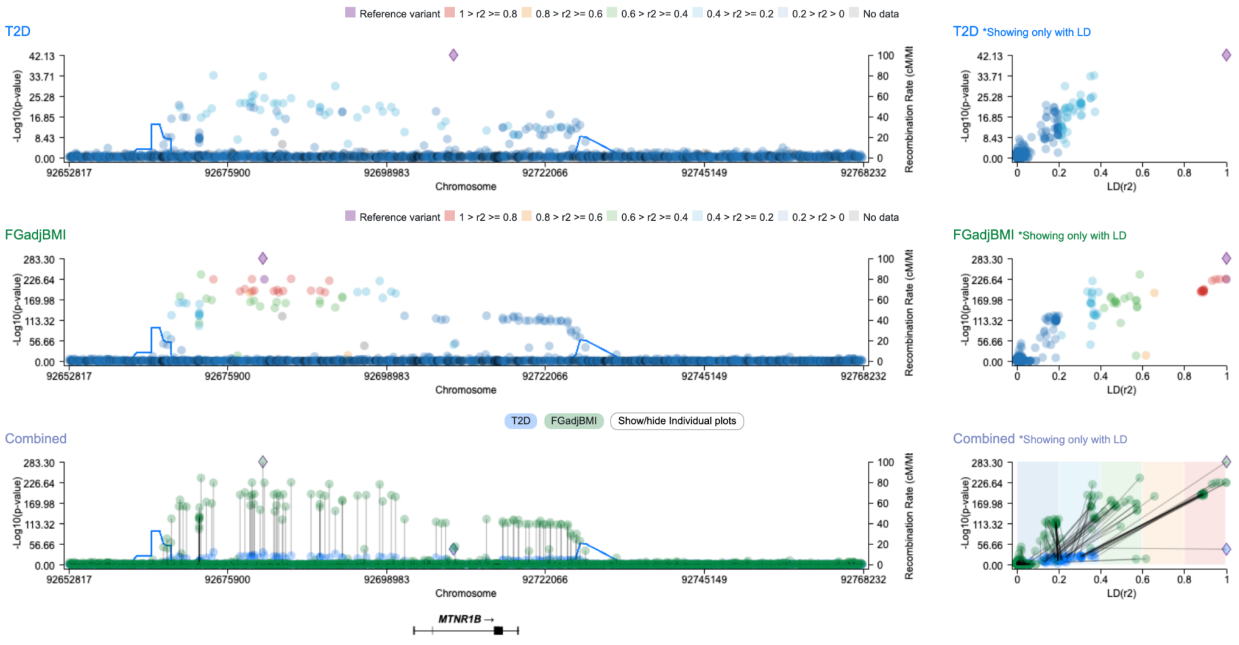


Figure 5. A region plot, drawing genetic association data from the HuGeAMP BioIndex. On the left, top to bottom, plots show associations across the region of the *MTNR1B* gene for T2D, fasting glucose adjusted for BMI, and both phenotypes in a combined plot. On the right, linkage disequilibrium plots display LD relationships between variants associated with each phenotype or with both phenotypes.

Research Pages and Portals exist in one of three access modes. In Unpublished mode, the presentation is accessible only to its creator. In Development/Review mode, access to the Page or Portal is password-protected for secure access by a group of collaborators or reviewers. The Public mode allows open access.

Configuring a Research Page or Portal

The interface for creating a Research Page or Portal, based on the Drupal content management system, allows the creator to specify the source of the data, the format of the data table and its filters, and the type of plot desired to visualize the data, and also to add explanatory text and legends. Although the pages are configured using JSON objects and arrays, the creator does not need to be familiar with JSON. A detailed tutorial explains each component and presents example configurations.

Discussion

To date, the BYOR framework has been used both internally in the Knowledge Portal team, to build interfaces that are linked from the Portals, and by external users. Interfaces on the Common Metabolic Diseases Knowledge Portal that were constructed using BYOR include tables of predicted effector genes with supporting evidence[8] and the Genetic Loci Clustering interface[6] based on results from Hyunkyung Kim, Miriam Udler, and colleagues[7]. Researchers outside of the Portal team have used BYOR to construct Research Portals supporting submitted manuscripts. Kenneth Westerman and colleagues constructed a portal cataloging variance QTLs and gene-environment interactions[9] to accompany their preprint[10]. A collaboration between the Portal team and Michelle McNulty, Matthew Sampson, and colleagues led to the creation of the *APOL1* Portal[11] to support a now-published manuscript[5]. The Research Pages and Portals that have been created using BYOR are listed on the HuGeAMP and Knowledge Portal Network home page.

The aim of the Knowledge Portal Network resource is to spark insights into the causes and progression of common diseases by aggregating, analyzing, and integrating shared results. We hope that BYOR, an offshoot of the HuGeAMP infrastructure behind the Knowledge Portals, will further promote the data sharing that is crucial to scientific progress. To get started with BYOR, please see the “URLs” section below.

Methods

In the BYOR framework, the Drupal content management system functions as the configuration management framework (CMF). Vue.js, a JavaScript framework, is the data intake and

rendering engine. Vue.js takes data from the sources defined by the user on the CMF, and loads them to the user's browser. The rendering engine formats the loaded data in a consumable format and builds the user interface, data tables, and plots based on the configuration from the CMF.

URLs

- BYOR entry page: https://kp4cd.org/research_portals
- Indicate your interest in creating a Research Page or Portal: https://forms.gle/mNF8fnxPPZEqRwDP8
- Tutorial: https://hugeampkpncms.org/tutorial
- Demonstration portal: https://hugeamp.org/research.html?pageid=demo_portal_n70

References

1. Final NIH Policy for Data Management and Sharing. Notice Number: NOT-OD-21-013. Accessed February 23, 2022; https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html.

2. Kozlov M. NIH issues a seismic mandate: share data publicly. Nature. 2022 Feb;602(7898):558-559. doi: 10.1038/d41586-022-00402-1. PMID: 35173323.

3. Mahajan A and McCarthy MI, 2019. Unpublished results.

4. The *APOL1* Portal (https://hugeampkpn.org/paper.html?paper=apol1_portal). Accessed February 23, 2022; https://hugeampkpn.org/paperdata.html?trait=null&dataset=apol1_diff_exp.

5. McNulty MT, *et al.* A glomerular transcriptomic landscape of apolipoprotein L1 in Black patients with focal segmental glomerulosclerosis. Kidney Int. 2021 Dec 18:S0085-2538(21)01149-2. doi: 10.1016/j.kint.2021.10.041. PMID: 34929253

6. Common Metabolic Diseases Knowledge Portal (cmdkp.org). Accessed February 23, 2022; https://hugeamp.org/method.html?trait=combined&dataset=epigenomic_clustering.

7. Kim H *et al.*, 2022. Manuscript in preparation.

8. Common Metabolic Diseases Knowledge Portal (cmdkp.org). Accessed February 23, 2022; https://hugeamp.org/effectorgenes.html.

9. Catalog of vQTLs and GxEs for cardiometabolic biomarkers (https://hugeamp.org/research.html?pageid=UKB-vQTL-GxE). Accessed February 23, 2022.

10. Westerman KE *et al*. Variance-quantitative trait loci enable systematic discovery of gene-environment interactions for cardiometabolic serum biomarkers. Submitted for publication, 2021. https://www.medrxiv.org/content/10.1101/2021.11.08.21265930v1