Phewas-Multimorbidity Explorer

Usage Manual

Nick Strayer

2019-10-14

1 Data

In order to use the app you will need three tables of data: two sets of individual-level data and one model results table. If using the app in data loading mode these are provided as `.csv` files.

1.1 Individual phenomes

The first set of individual data you will need is a table containing pairs of patient-id to phecode-id (or ICD9/ICD10). This table is just two columns, so a patient `p1` with phecodes `a`, `b`, and `c` would have three lines in the table, `p1, a`, `p1, b`, and `p1, c`.

The patient id column should be titled either `grid`, `id`, or `iid` (case insensitive) and the phenotype id column should be titled `code`.

1.2 Individual SNP info

The next set of individual data comes in the form of presence of the minor allele for each patient. Again the table is two columns, with the first being patient-id and the second being the integer (`0`,`1`, or `2`) corresponding to how many copies of the minor allele the patient had. (Note that it is possible to omit rows for patients with no copies of the minor allele and the application will assume those patients had zero copies. Like the individual phenomes table, the title of the patient id needs to be `grid`, `id`, or `iid`, and the title of the second column should be the rsid of the SNP of interest, which the app uses to infer the SNP id.

1.3 PheWas results

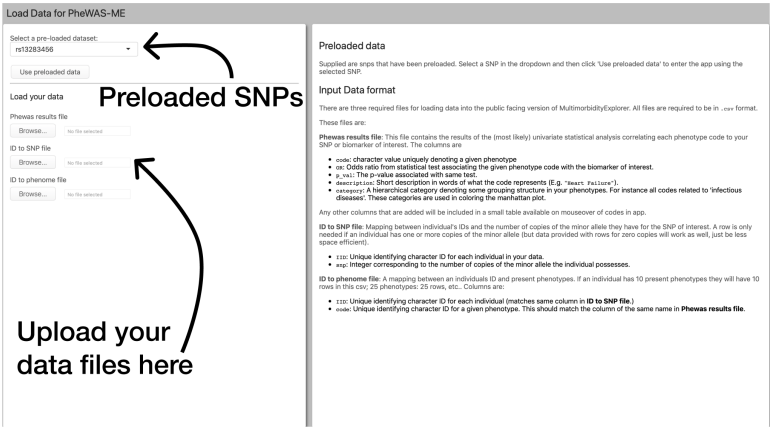
The last set of data needed is the results of the PheWas study you wish to analyze. This takes the form of a table with the following columns. - `code`: PheCode, ICD9, or ICD10 code - `OR`: Odds ratio for test with SNP - `p_val`: P-Value for significance of code's association - `description`: The plain text description of the code - `category`: Plain text description of code's broad-level category.

Any other columns included will be displayed in the app as additional information when investigating a single SNP via tooltips, but is not necessary.

1.4 App API

If you are starting an ME application with preloaded data from the R package API, the format of these files remains almost the same, with the exception being R dataframes are passed directly to the function `build_me_app()`. For more detailed information see the application documentation either in the 'reference' tab of this website or by accessing the package documentation in your R console with `?meToolkit:: build_me_app`.

1.5 Data Loading Screen



When the app is first loaded the user is greeted by the data-loading screen. This screen prompts the user to provide the three previously described datasets as CSV's or choose from available preloaded data (see next section for how to prepare this data.)

As each file is uploaded the app will check to make sure the files match the required format and inform the user of malformed data if it is uploaded.

Once all data is loaded, the user is automatically sent to the main dashboard.

1.6 Preloading data

If a set of results are going to be repeatedly visited in the app, the data for the results can be preloaded to save time. Once data is preloaded it populates a dropdown menu on the data-loading screen that can be used to select the desired dataset.

To do this, the data must be loaded into a path (relative to the main app working directory) provided to `run_data_loader()` in the `preloaded_data_path` argument . The patient phenome file must be stored at `preloaded_data_loc/id_to_code.csv`, and each individual SNP's patient-to-snp and phewas results stored in `preloaded_data_loc/<SNP_ID>/id_to_snp.csv` and `preloaded_data_loc/<SNP_ID>/phewas_results.csv` respectively.

Note that due to sharing the same phenome mappings, all preloaded phewas results must be done on the same population.

2 Accessing app

There are two main ways to access Multimorbidity Explorer.

Hosted The first is to go to the hosted example at prod.tbilab.org/multimorbidity_explorer. Here you can run the app on preloaded simulated data or upload your own data. No uploaded data is saved on the hosting server.

Self-Hosting If data privacy is a concern, the app can be run entirely locally on any computer with R installed and the ability to install packages. To run locally install the `meToolkit` package from github using the `devtools` package...

```
devtools::install_github('tbilab/meToolkit')
```

Once the package is installed the app can be launched with the data loading interface using the command

```
meToolkit::run_me()
```

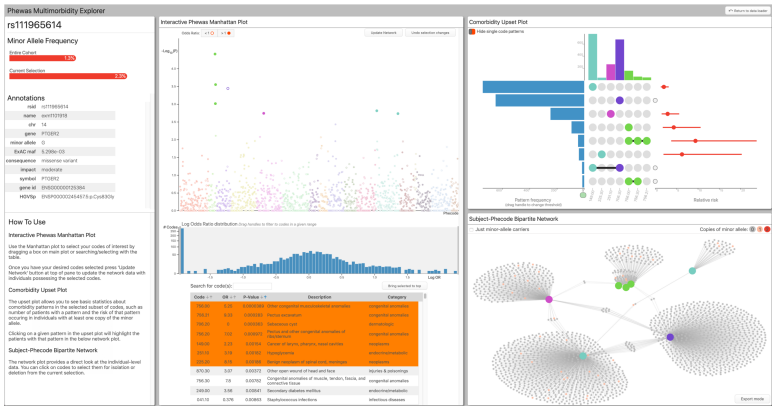
Customizing with R API The `meToolkit` package is setup modularly so the data loading page can be skipped if data is provided directly to the app via the function `meToolkit::build_me_app()`.

```
my_ME_app <- build_me_app(
  phewas_table,
  id_to_snp,
  phenotype_id_pairs
)
```

The data format for `phewas_table`, `id_to_snp`, and `phenotype_id_pairs` follows exactly from the required input files as outlined in the Data section (page 0)

For further information on customizing applications read the packages documentation at prod.tbilab.org/meToolkit.

3 Main App



3.1 Application state and reading this manual

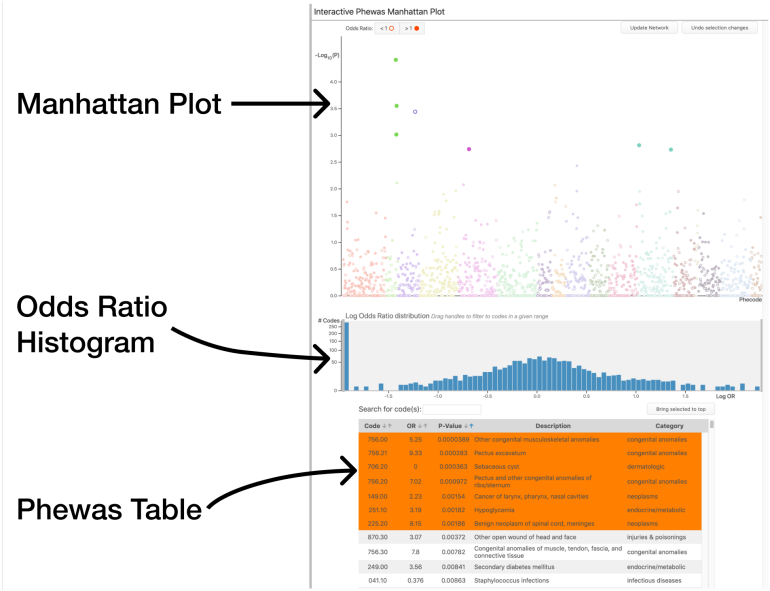
The state of the app is primarily driven by the list of currently selected phenotypes. At load this is small number of the most significant results from the Phewas analysis, but by interacting with the application the user changes this state.

Other state information includes, subjects that are to be highlighted or phenotypes that are to be 'inverted' (see 'Subject-Phecode Bipartite Network' section for details.)

Throughout the following sections the functionality of each panel of the dashboard is broken into two parts: what the panel shows and how interactivity works within the context of the panel, along with how that panel can interact with the main application state (the **App Interaction**) subsection for each panel.

The application state is saved in the URL of the page. To return to a pre-determined view, simply bookmark or copy the current URL. For apps with data loading screen, the state will automatically select the correct snp from preloaded data if it is available.

3.2 Interactive Phewas Manhattan Plot



3.2.1 Purpose

Visualizing results of entire Phewas analysis in graphical and table format and select codes to investigate comorbidities.

3.2.2 What's Shown

Manhattan Plot The Interactive Phewas Manhattan plot contains a standard manhattan plot, which is a scatter plot with the x-axis containing each phecode tested for association with the SNP of interest, and the y-axis representing the p-value of the association (on the negative log base-10 scale.)

The points are colored according to the broadly provided code `category` in the Phewas results table.

The plot contains a few features not normally found on manhattan plots. First, the points themselves are either hollow or solid, corre-

sponding to a negative association ($\texttt{OR} < 1$) or positive association ($\texttt{OR} > 1$) respectively. This allows the viewer to get an idea of general trends in association by category and subcodes.

Hovering over a given Phecode in the plot will show a tooltip containing all the information passed to the app in the Phewas results data (Odds ratio, p-value, ect.).

Odds Ratio Histogram In addition to showing positive and negative association in point shape, a histogram of the log-odds ratio of all tests is provided below the manhattan plot. This plot allows the viewer to see the broad distribution of associations with the SNP of interest.

A range selector is overlaid. This can be dragged to control which points are shown in the manhattan plot based upon their odds ratio. E.g. if the histogram upper range is reduced to a log odds ratio of 0, only phecodes that had a negative association would be plotted on the manhattan plot. This allows the viewer to target phenotypes based on effect size.

Phewas Table Below the manhattan plot and odds ratio histogram is a table corresponding to all the provided results. This table contains columns for the code name, odds ratio, p-value, code description, and code category. The table can be sorted by all column values in addition to being searched for both code name and description.

3.2.3 Filtering codes

Manhattan Plot Codes can be selected in a variety of ways. The main method of interaction involves dragging a box around a region of codes on the manhattan plot. A common case is to select the top most significant codes by dragging a box across the upper portion of the plot.

There are two keyboard shortcuts that can modify drag-selection behavior. First, if the ‘a’ key is held down (for __a__dd) while the selection box is drawn, the selection is added to any previously selected codes. This can be useful if two distinct regions of the manhattan plot are desired, potentially corresponding to two categories of interest. Second, if the ‘d’ key is held down (for __d__elete), any newly

selected codes will be removed from the selection. This can help fine tune selections without needing to redraw the entire region.

Individual codes can be selected and unselected with a click.

Any codes outside of the current odds ratio bounds set by the histogram range-slider are not able to be selected.

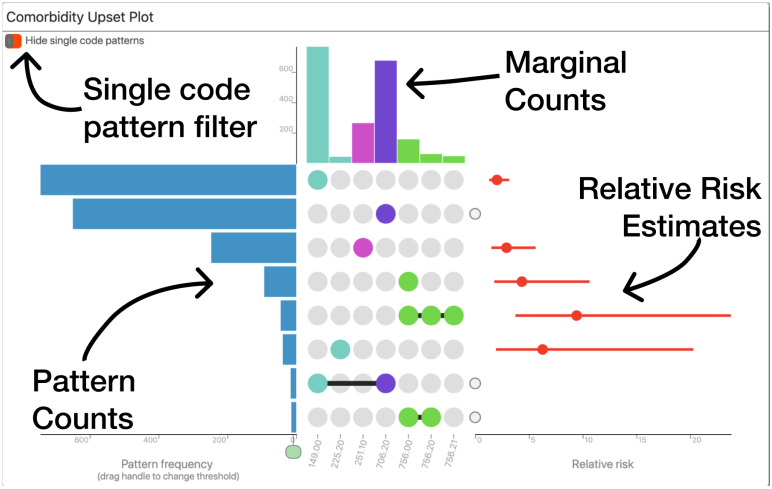
Phewas Table Codes can be either added or removed from the selection by clicking on their row in the Phewas table. Codes that are currently selected are colored orange.

3.2.4 App Interaction

Once a selection of codes has been made, pressing the 'Update Network' button in the upper right of the panel will update the rest of the app with the current selection of codes.

Reset Button The reset button allows the user undo all code changes they have made since the last app-update. I.e. loading state or the last time the app-wide state of selected codes was changed.

3.3 Comorbidity Upset Plot



3.3.1 Purpose

To visualize unique comorbidity patterns in currently selected Phecodes, along with those pattern's associations with the SNP of interest.

3.3.2 What's Shown

The Comorbidity Upset Plot is an Upset plot (https://caleydo.org/tools/upset/) that visualizes unique combinations of Phecodes seen in the subjects.

Pattern Matrix The center of the visualization contains a matrix with the currently selected phecodes as columns and unique patterns of those phecodes seen in the individual-level data as the rows. A pattern is represented by a dot being filled in at each column corresponding to a code in the pattern. These dots are colored by the category of the phenotype.

A toggle is provided to hide patterns that consist of only a single phecode. Singleton patterns can dominate the plot due to how common they are, so filtering them out can allow for closer inspection of comorbidity patterns.

Marginal Counts At the top of each Phecode column in the plot is a bar representing how many subjects had that code in the data. These bars can be hovered over to get a text-summary.

Pattern Counts At the left side of each pattern row, another bar is drawn corresponding to how many patients had that specific pattern in the data. (Note that if a pattern is a subset of another, the bar represents that subjects that *only* had exactly the smaller pattern.) Below the x-axis of the pattern-counts bars is a handle that can be moved to change the threshold for minimum number of times seen needed for a pattern to be plotted. This is helpful when there are a large number of unique patterns but only ones with large sample sizes are of interest. By increasing the threshold needed for inclusion the plot is 'zoomed' in on the more common patterns. Like with the marginal bars, more info is available in text form on hover over the patterns.

Relative Risk Estimates To the right of each pattern's row is a point estimate and 95% confidence band of the relative risk of that pattern occurring given presence of at least one copy of the SNP of interest. This is estimated using the small sample adjustment for both point estimate and confidence interval, via the function `riskratio.small` in the `epitools` package (https://cran.r-project.org/web/packages/epitools/epitools.pdf).

Again, hovering over the interval shows details of plot in text format.

3.3.3 App Interaction

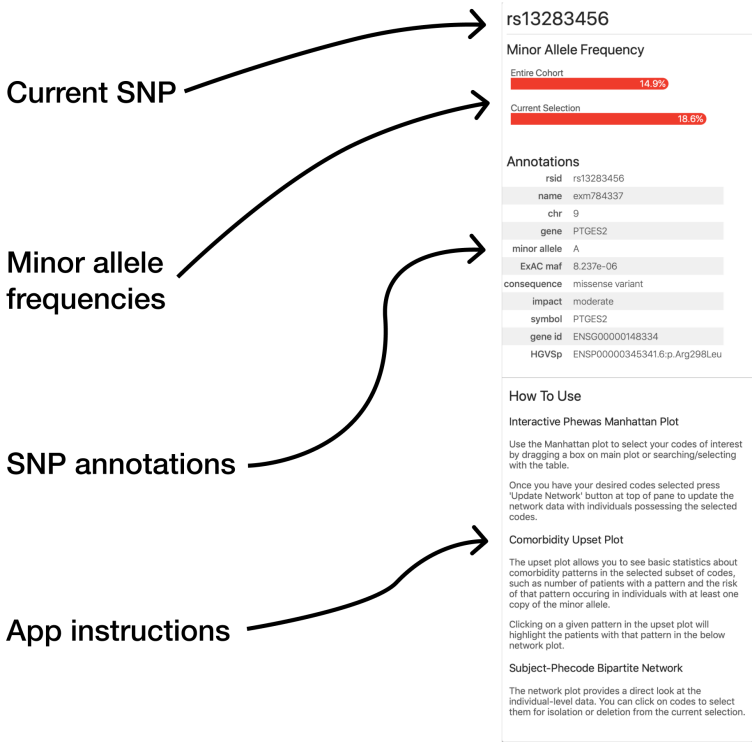
The upset plot can highlight individual-level data in the network plot. There are two forms this highlighting takes.

Code Highlighting By clicking on a column corresponding to a given code, all subjects that have that code present in their phenomes will be highlighted on the network plot. This means if subject A had a phenome with codes *i* and *j*, and subject B had a phenome with codes *i*, *j*, and *k*, both would be highlighted when code i or j's columns were selected.

Pattern Highlighting By clicking on a row corresponding to a given comorbidity pattern of codes, all subjects who possess that *exact* pattern in their phenomes will be selected. Returning to the example in the 'Code Highlighting' section, this means if the pattern of phecodes (*i*, *j*) was selected *only* subject A would be highlighted.

Highlights for both codes and patterns are reset when either another pattern is selected of the column/row is selected again.

3.4 Info Panel



3.4.1 Purpose

Show information about the current SNP of Interest including minor allele frequency and basic annotations, along with usage instructions for application.

3.4.2 What's Shown

The info panel is broken into three main sections...

Minor Allele Frequencies The top of the panel shows two measures of the minor allele frequency (MAF) for the current SNP of interest. The first of two bars shows the MAF for the entire dataset provided for the app. I.e. what proportion of all subjects provided have at

least one copy of the minor allele. The second bar shows the MAF for the subjects who have at least one of the currently selected phecodes.

SNP Annotations

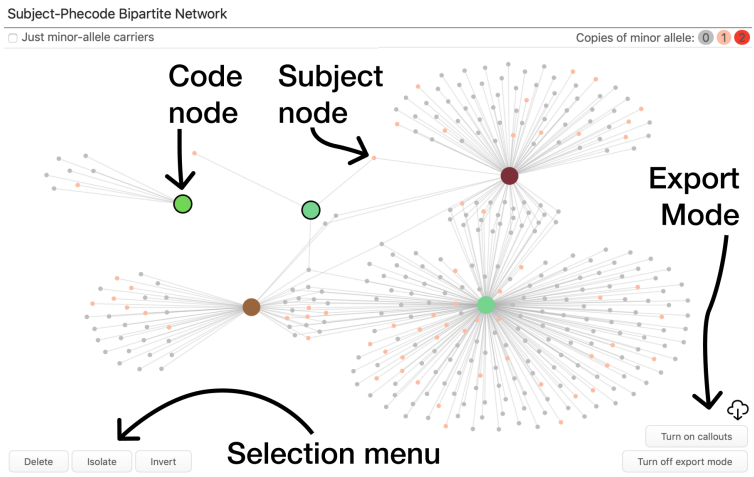
Below the MAF chart is a table containing annotations about the current SNP. Currently these include SNP name, chromosome, gene SNP resides in, minor allele, Exome Aggregation Consortium (ExAC) minor allele frequency, clinical significance, consequence, impact, symbol, gene id, and HGVSp for SNPS found on the Illumina Exome chip (https://www.illumina.com/products/by-type/microarray-kits/infinium-exome.html).

Instructions This panel contains basic instructions on how to use the application.

3.4.3 App Interaction

The info panel has no app-level state interaction capabilities.

3.5 Subject-Phecode Bipartite Network



3.5.1 Purpose

Provide a direct look at individual-level data, showing connections between subjects and their phecodes along with the status for the SNP of interest.

3.5.2 What's shown

The Network Components The network is bipartite, with two different node types. The larger nodes correspond to the currently selected Phecodes, colored by their category as in the Manhattan and Upset plots. The smaller nodes represent individual subjects, colored by their number of copies of the minor allele for the SNP of interest. Edges are drawn from each subject to all the Phecodes present in their Phenome.

Layout The layout of the network is calculated in real time using a basic physics simulation that treats the edges of the network as springs and tries to find the layout with the lowest total tension in the system. This serves to place similar patients and phecodes near each other and acts as a pseudo-dimensionality-reduction technique.

Interaction The network can be panned and zoomed to focus in on subsections or to zoom out for a broader picture. A toggle is provided draw the network with only subjects possessing at least one copy of the minor allele, as a way to investigate differences in overall vs SNP-driven structure.

Like in the Interactive Phewas Manhattan plot, by mousing over a phecode's node a tooltip is revealed with all the supplied information from supplied Phewas results data.

Export Mode The network plot can be exported in SVG format for use in vector-editing software such as Adobe Illustrator or Inkscape to prepare for publication.

Export is enabled by clicking the 'Export mode' button in the bottom right of the plot. This re-renders the plot in vector format and also provides the option to add callouts labeling each Phecode by ID. These callouts are able to be positioned by dragging.

Once the plot has been customized as desired and the download icon is pressed, an SVG of the current view is downloaded to the user's local computer.

Due to the computational overhead of rendering the network in vector format, it is recommended that export mode is kept off until needed to avoid slowing down the app.

3.5.3 App Interaction

The network plot can be used to fine-tune the app-wide selected Phecodes. This is done by clicking or tapping on a Phecode node in the network to select it. Once at least one node is selected an action menu appears in the lower left of the plot. Here the user can choose to remove the selected code(s), to isolate a pattern of selected codes, or to 'invert' a code.

Code Inversion When a code is inverted a subject is considered to 'have' the code if it is absent in their phenome. Visually in the network this is represented as a hollow node. By inverting a code you can frame questions in terms of negative relationships, potentially uncovering previously unseen relationships.