Introduction to GenAMap

For further questions / comments / suggestions, please contact
genamapsupport@cs.cmu.edu

1 What is GenAMap?

GenAMap is software for association mapping and eQTL- and genome-wide association studies. It combines automated data analysis and rich visual exploration tools to enable users to perform an association study end-to-end within a single platform. Once users have imported genome data, transcriptome data and / or phenotypic trait data into GenAMap, they can deploy a variety of methods for analysis, both classical test-based methods such as the wald test and modern structured regression algorithms such as the GFLasso. The results produced can then be explored directly within the software using a range of visualization tools that are designed to enable the users to gain an overview of a potentially large amount of result data, spot interesting patterns easily and then drill down into those patterns and request details. In this way, GenAMap enables users to gain a quality of insight that is difficult to achieve using either machines alone or by database queries.

GenAMap offers the following advantages:

- A range of sophisticated algorithms at the users disposal that can be started with a few clicks from a graphical interface.
- Automatic delegation of algorithm computation to a remote computer cluster. This frees the user from having to provide high-performance hardware and enables him or her to run algorithms on even large data sets in a timely manner.
- Customized visualization tools that both scale to large datasets and allow the user to discover and explore nuances.

- Results are ready for visualization immediately as an algorithm finishes running. The user does not need to reformat data or port information between different stand-alone tools.
- Graphs produced within GenAMap are high-quality and can be embedded in publications.

This tutorial is structured as follows: In section 2, we will introduce the reader to the GenAMap interface using an example. In section 3, we describe the core features of GenAMap in detail. In section 4, we outline the technical architecture underlying GenAMap and describe GenAMap's technical capabilities, limitations and system requirements. Finally, in section 5, we describe the more specialized features of GenAMap.

Throughout this tutorial, I will be referring to the video tutorials that can be accessed via the GenAMap website.

Please visit http://cogito-b.ml.cmu.edu/genamap/tutorials.html

In addition to online resources, further documentation comes packaged with GenAMap. In the 'Documentation' folder, you can find the following tutorials:

- 'Setting up GenAMap.pdf' guides you through the installation, registration and login process.
- 'DataImportingGuide.pdf' shows you how to format data for importing into GenAMap. We also provide examples of correctly formatted data in the 'ExampleData' subfolder.
- 'AlgorithmsGuide.pdf' provides details on each of the algorithms available in GenAMap, together with information about their implementation.

2 A first example

In this section, we will give a flavor of the GenAMap interface using a simple example. Given synthetic marker data (marker = SNP) and phenotypic trait data, we would like to find associations between individual markers and traits using the Lasso algorithm. We assume that GenAMap is already set up on the users

machine and is ready for use. To learn more about this initial setup, please refer to 'Setting up GenAMap.pdf' in the Documentation folder. The data used in this section is available in 'Documentation/ExampleData'. Using this data, you can perform the same steps as described here.

After logging in for the first time using a fresh user account, the interface presents itself as in Figure YYY. The first thing we do is add a project. A project is similar to a folder, used to organize a data collection. We right-click on the XXX field in the top left and choose 'XXX'. This brings up the project creation dialog box as in Figure YYY. We choose name XXX and click on XXX.

Once the project has been created, we can import the marker data. Right-click on XXX, choose XXX and fill out the appearing marker import dialog. The dialog requires you to specify three files: A file containing the actual marker values, a file containing the names and locations of the markers and a file containing the names of the samples / individuals. We choose the following files from the ExampleData folder: 'markerVals.txt', 'markerLabelsReal.txt' and 'sampleLabels.txt'. The completed XXX is shown in Figure YYY. Now, we click on the XXX button. This will open up a separate window as in Figure YYY that displays a progress bar that indicates what proportion of the import has (approximately) been completed. This window is used to keep the user updated on many potentially time-intensive data operations. While the marker data is being imported, the user is free to use any other feature available in GenAMap.

Many features in GenAMap are accessed through right-click menus. It can help to click on different objects within the GenAMap interface to discover specialized functions.

Once the marker data is imported, the data set is visible under our project in the marker tab as shown in Figure YYY. When we click on the marker data set, the chromosome browser appears at the bottom of the visualization area. This is one of our visualization tools. It allows us to view the location of individual markers along the chromosome, zooming in and out.

Now, it is time to upload the phenotypic trait data. We right-click on the project and select "Add Trait Data". This opens up the trait importing dialog. We have various options for the trait file format. For more information, see 'DataIm. Clicking on "importing" again brings up a status bar that lets us monitor the progress

of the importing process. When the trait data has finished uploading, we can see it under our project when we browse to the trait tab in the data management section of the interface. At this time, we have all the information we need (marker data and trait data) to tun the Lasso algorithm to find associations between individual markers and traits. Right-click on the project and then click on "Add Association Data". This brings up the Association Data creation box. There, we have the option to choose between uploading association data has already been computed or to run an algorithm to compute a new data set. We choose Lasso from the drop down menu and click on import.

3 Core Features

Actions users can take in GenAMap fall roughly into three categories: importing data, running algorithms and visualizing results. Data in GenAMap is represented in the form of *data sets* that can take different *data types*. In the last section, we have already encountered *Marker Data*, *Trait Data* and *Association Data*. These are three data types. We imported one instance of Marker Data, one instance of Trait Data and then ran Lasso to generate an instance of Association Data. All of the data types contain specific kinds of information. For example, Marker Data contains the SNP values of individuals for certain SNPs. Most data types can be imported. To do this, the information contained in the data type has to be presented in one or more files whose format can be recognized by GenAMap. For example, we can present marker data in the form a tab-delimiter matrix file where each row corresponds to a sample and each columns corresponds to a marker, with two additional files for sample labels and marker labels / locations.

In addition to importing, we can create data set instances by running algorithms. An *algorithm* is a program that consumes data set instances of specific types and produces an output data set of a specific type. For example, the Lasso requires a Marker Data set and a Trait Data set as input, and produces an Association Data set as output. Once we have the required input data, running an algorithm in GenAMap is very simple. In fact, both importing and starting an algorithm is achieved through bringing up a dialog as shown in Figure X. Just provide the necessary information and click on "import". Every data type has its own dialog box that outlines how data sets of this type can be created. Most give the user the option to import the data or to run an algorithm that produces a data set of this type.

Starting an algorithm will cause a notification to be sent to a remote computer cluster to execute the computation. In fact, no heavy computation is performed locally. All remote computation runs and completes without any action necessary by the user. The user may close GenAMap or even power down his machine without interrupting the computation. When the user logs on for the first time after the computation is complete, the result is ready for visualization in GenAMap. The progress of the algorithm computation is shown in the algorithm control center, which is in the bottom left corner of the interface. Note that if the datasets involved are large and there is unusually heavy demand on the cluster, algorithms may still take days to complete.

Available data sets are shown in the data management section of the screen, in the top left. This section consists of three tabs: the marker tab, the trait tab and association tab. Switching between tabs will show different data sets. These are organized in a tree structure as shown in Figure X. The top level of this tree is the *project*, which is essentially a storage bin for data. Any data set we import or create is associated to a project. To create a new project, right-click on the X and choose 'Add new project'. The next level in the tree are the marker / trait / association data sets. The next level down are the data sets that are associated to these top-level datasets and so on.

GenAMap has four core data types. Marker Data, Trait Data, Association Data and Network Data. Marker Data is
Once a data set has been imported or creating by an algorithm, it is immediately available for visualization and exploration. All of the major data sets can be selected on the top left hand side in the data management section of the user interface.
It can be expressed as a square matrix where rows represent samples, columns represent SNP locations, and entries represent the value of the SNP at the location specified by the column and the sample specified by the row.

Apart from importing, to obtain an instance of a particular data type, we can run an algorithm that creates this instance. In the last section,
The most important data set types are *marker data*, *trait data*, *network data* and *association data*.

4 The GenAMap Architecture

5 Specialized features