There are two types of data we can use: GEO and dbGaP. GEO is gene expression data and dbGaP maps genotypes and phenotypes.

There has been Neena, Tiffany as well as Vinnie who used GEO data for their projects; on the other hand, Skanda and Vinnie used dbGaP.

Git/Kiarash/related papers/GEO\_TBIpaper8\_V2.pdf is a good start to get familiar with GEO data and Neena’s pipeline. Her main code can be found here: Git\Kiarash\Neena pipeline\GEOpipeline.R . As you notice, it is written in R language which is nice for getting the data from the server. Alternatively this can be written in any other programming language and use the raw geo data that can be downloaded from here: <ftp://ftp.ncbi.nlm.nih.gov/pub/geo/> (Read the readme file and it is straightforward from there)

Neena’s pipeline produces arff files as the output. They can be given to Weka to produce Bayesian networks. Other preferable but a little more complicated method is to make Git/Kiarash/Neena pipeline/Multinet.zip to produce the networks instead of weka.

Neena explains that the R file structure is in this way:

1. Determines which experiments are "interesting" based on their contents

2. Finds intersection of genes across all experiments

3. Determines which samples correspond to control vs. noncontrol

4. Finds top differentially expressed genes for each experiment

5. Finds intersection of genes among these top differentially expressed genes

6. Creates "interesting reduced experiments" (IREs), which are essentially data tables that represent each of the interesting experiments and their expression data from each GSM sample. They are "reduced" because only the data from the common genes is included in each data table.

7. Normalizes the IREs by using a reference IRE, finding its median, and subtracting the difference between the reference median and each IRE's median from all values in each IRE

8. Creates ARFF files with the data from these genes

Tiffany has used very similar pipeline to Neena’s that can be found here: Git/Tiffany/geopipelineTiffany.R and her thesis paper is very informative: Tiffany/Tiffany Lin Thesis.pdf One thing to keep in mind is that from my understanding in Tiffany's research Weka is used for cross-validation (She didn't use code to make the actual Bayesian networks). To do similar thing we need the Weka software. But if we want to build the Bayesian networks, we might want to use similar code to HIVMultinetBuilder.java or Multinet.zip; in this case we still need two jar files - to be downloaded from Weka website.

Vinnie’s code is also on the Git. He recently wrote a readme file explaining how the code works. vinnie/code/README.txt . Even if one uses dbGaP data, Vinnie’s code have parts that can be useful.

dbGaP data has been used by Skanda (skoppula) and Yixin. Their codes are also on the Git repository. Skanda’s code has very good documentations there. His code can be used both to just get an understanding how to use dbGaP data as well as to understand pathway analysis. An some Yixin codes are in Kiarash/dbGaP to arff files/

A good read is this document on how to convert data type: skoppula/guide-converting-data-types.txt and the related code is here \skoppula\quick-start-guide\predictive-pathway-analysis\pipeline-with-src-april-2012\src\main\ExtractInstances.java

After all, the result of a project can be messy and unclear. A good idea is to have a piece of code to produce a nice output document. Aaron has had similar thing, and his code is on dropbox. This particular piece of code can be found here: bnGui0/src/bnGui/writereport.java