Building GeneMANIA

Ruth Isserlin

2021-08-19

Contents

Chapter 1

Attributions:

This book was created using The **bookdown**(Xie, 2015) package and can be installed from CRAN or Github:

```
install.packages("bookdown")
# or the development version
# devtools::install_github("rstudio/bookdown")
```

Icons are from the "Very Basic. Android L Lollipop" set by Ivan Boyko licensed under CC BY 3.0.

Chapter 2

Introduction

When researching with an organism not present in GeneMANIA(Warde-Farley et al., 2010) it can be beneficial to pool together all the data that you have for your organism and create your own instance. There are two ways you can create your own instance of GeneMANIA:

1. Local version of GeneMANIA in cytoscape - only available to you on the device that it is created on. It is quick to set up and offers all the functionality available in the GeneMANIA cytoscape App.
2. Local version of GeneMANIA website - if you expose this website to the internet you can share this resource with your colleagues. Considerably more work to set up but all the code and parts are available in github and docker instances.

For both of the above instances the bulk of the work required is collecting, translating and creating id conversions and networks, for your desired species, that are used in GeneMANIA. Given that this is an organism of interest to you likely you have a lot of the information or know where to get it.

2.1 Overview

There are three major components to the GeneMANIA system. (Only the first two are required for both methods of creating your own instance, identifiers and Network, annotation and attribute data):

1. **Identifiers** - mapping from all the different identifiers available for your species of interest. Ideally the mappping files contain all and any identifiers used for the species. For example, the human dataset in GeneMANIA has hgnc symbol, entrez gene ids, refseq ids, ensembl ids and uniprot ids.

2. **Network, annotation and attribure data** - tables of interactions and associations between your entities (enumerated with one of the above identifiers). Each interaction is associated with a score. If the data in question doesn't have a score then it has the score of 1 (ie. it is present in the dataset)
3. **Indexed data** - Once all the previous data is collected and cleaned the last step involves indexing the data using lucene.

For the purpose of this tutorial we will be creating an instance of GeneMANIA using Tetrahymena.

Chapter 3

Website build Stage 1 - Identifier mapping file

In the main GeneMANIA instance all identifier mapping is extracted from the mySQL data dumps released by ensembl -

There is no requirement to have all the specified identifiers or to get your mappings from ensembl but your identifier mapping file for the website build needs to have the following format:

1. GMID - automatically generated unique genemania identifier. It is specific to the instance of the website and not to any external database.
2. Ensembl Gene ID
3. Protein Coding
4. Gene Name
5. Ensembl Transcript ID
6. Ensembl Protein ID
7. Uniprot ID
8. Entrez Gene ID
9. RefSeq mRNA ID
10. RefSeq Protein ID
11. Synonyms
12. Definition

3.1 How to build your identifier file

In order to build the identifier file we need to create an ensembl mirror for the desired organism, import all the ensembl data and export the desired data into the above specified format.

3.1.1 Create and setup a container of the ensembl mirror docker instance

The docker instance that we are going to use can be found here - gmbuild_ensembl.

1. Install Docker - for instructions see here
2. check out GeneMANIA_build from the Baderlab github (https://github.com/BaderLab/GeneMANIA_build.git)

```
git clone https://github.com/BaderLab/GeneMANIA_build.git
```

Create container of gmbuild_ensembl instance.

- each -v parameter specifies a local directory, or volume, that is mapped to a directory on the docker. For example, the directory /home/gmbuild/ensembl data on your machine gets mapped to the location /home/gmbuild/ensembl_data in the docker container. Any file that is put into that directory on the docker will show up on the corresponding directory on your machine.
- There are multiple volumes mapped to the ensembl_mirror
  - /home/gmbuild/ensembl_data –> /home/gmbuild/ensembl_data
  - /home/gmbuild/gmbuild_code_dir/genemania-private –> /home/gmbuild/ensembl_code - This is the directory where you checked out the code in the previous step and contains all the code we will need to build the identifier mapping files.
  - /home/gmbuild/gmbuild_code_dir/genemania-private/Docker_contatiners/Ensembl_docke –> /etc/mysql/conf.d - configuration files needed for the set up of mySQL instance on the docker.
  - /home/gmbuild/db_files –> /var/lib/mysql - directory where the mySQL database will store its data files.
  - With the –name tag you can specify what you would like to call your instance. This name can be used when logging into the instance. For this example we have called this instance ensembl_mirror.

```
docker run -d
-v /home/gmbuild/ensembl_data:/home/gmbuild/ensembl_data
-v /home/gmbuild/gmbuild_code_dir/genemania-private:/home/gmbuild/ensembl_code
-v /home/gmbuild/gmbuild_code_dir/genemania-private/Docker_containers/Ensembl_docker/cu
-v /home/gmbuild/db_files:/var/lib/mysql
```

```
--name ensembl_mirror
baderlab/gmbuild_ensembl
```

Once the instance is created, log into it.

```
docker exec -it ensembl_mirror /bin/bash
```

Download the ensembl data.

- On the docker you need to change into the directory ensembl_code/ensembl_mirror (remember that actually points to /home/gmbuild/gmbuild_code_dir/genemania-private on your main computer which is the directory containing the code that we downloaded from github.)

```
cd ensembl_code
cd ensembl_mirror
```

- Download the data - There are two separate scrits in the ensembl_code/ensembl_mirror that you can use to download the data. For a new species you will need to modify the scripts to make them specific for your organism.

1. get_ensembl.sh - this script demonstrates how to download all the species that are currently available in the public GeneMANIA server. A selection of them are available from the main ensembl ftp site (including human, mouse, fly ….) but some are not (including e-coli and arabidopsis). This script shows how you need to specify the ftp site depending on the data you are grabbing.
2. get_ensembl_indiv_species.sh - this script demonstrated how to donwload one example species. For this example we are using Tetrahymena which is available in ensemblgenomes protists section.

- Open get_ensembl_indiv_species.sh
- update script to have the following variables:
  - SPECIES=‘tetrahymena_thermophila_core’
  - FTP_SITE=‘ftp.ebi.ac.uk/ensemblgenomes/pub/current/protists/mysql/’

Depending on the species that you are using the species and the ftp_site variables will be different. Not all species are available on the ensembl servers (or available to RSync.). For the tetrahymena example, although the files are listed on the ensemblgenomes ftp site, the files failed to download using rsync. Changing to the ebi mirror fixed that issue.
There are many ftp sites you can check to see where your organism specific files are located. Ultimately, it depends which division the species falls into.

1. ‘ftp.ensembl.org/pub/current_mysql/’
2. ‘ftp.ensemblgenomes.org/pub/current/plants/mysql/’
3. ‘ftp.ensemblgenomes.org/pub/current/bacteria/mysql/’
4. ‘ftp.ensemblgenomes.org/pub/current/fungi/mysql/’

5. 'ftp.ensemblgenomes.org/pub/current/protists/mysql/'
6. 'ftp.ensemblgenomes.org/pub/current/metazoa/mysql/'

Navigate to the right division and find your species of interest. There will be additional numbers after the species name associated with the directory name but when setting the species variable just include the species name. The additional numbers indicate which release these files are associated with. **Given that you want to get the current release, make sure that you don't include those numbers.**

For the above directory found in the protist division we set the following variables:

- SPECIES='tetrahymena_thermophila_core'
- FTP_SITE='ftp.ensemblgenomes.org/pub/current/protists/mysql/'

```
./get_ensembl_indiv_species.sh
```

Problem: When running *./get_ensembl_indiv_species.sh* nothing happens, the script finishes right away with no output.

Solution: Check to see that you have defined the SPECIES variable correctly. * Go to ftp. ensembl.org/pub/current_mysql and check the spelling of your organism's directory.

Problem: ==> Rsync tetrahymena_thermophila_core_51_104_1 FROM ftp.ensemblgenomes.org/pub/current/protists/mysql/: @ERROR: Unknown module 'pub' rsync error: error starting client-server protocol (code 5) at main.c(1666) [Receiver=3.1.2]

[[ERROR]] : tetrahymena_thermophila_core_51_104_1 – Trying again...

Solution: The ensembl sites are not always consistent. Verify that you have got the address right but if the address right and you still get this error try using the ebi mirror instead:

When the script is done running you will find a directory in your ~/ensembl_data directory with today's data. In that directory you will find all of the ensembl files that were just downloaded.

```
ls -r ~/ensembl_data/*/tetrahymena_thermophila_core*
```

Load ensembl data into local database.

The next script will take all the files downloaded from ensembl and load them into a local mySQL database.

Using the gmbuild_ensembl docker will help with this step because there is no requirement to install mySQL. The docker instance comes with a compatible mySQL server.

```
./create_ensembl.sh
```

mySQL 8 or greater no longer has INFORMATION_SCHEMA. If a database was exported from an older version of mySQL there might be references to INFORMATION_SCHEMA and script will crash with error. You can:

- Update any file containing it from INFORMATION_SCHEMA to PERFORMANCE_SCHEMA
- Of alternately, Easy fix for this issue. Open mySQL and run the following command: set @@global.show_compatibility_56=ON;

Process ensembl data. Create summary files needed for GeneMANIA

This step creates identifier mapping files as well as shared domain information present in ensembl associated with your species. Shared domains is not required for the build but they will be automatically created during this step.

- Change into the identifier mapping directory.

```
cd ../identifier-mapper-perl/
```

- modify the runall_indiv_species.sh script to use your newly downloaded species data. - the runall.sh script shows how the main GeneMANIA build processes multiple species. The runall_indiv_species.sh runs the exact same process but only for one species.

Update the line 34 of *runall_indiv_species.sh* to reflect your species of interest:

- ./idmapper.pl $DATADIR/Work **tetrahymena_thermophila Tt** 19
- This line will call the idmapper perl script with parameters (in the following order, order is important):
- output directory $DATADIR/Work - don't change. The script automatically detects the newest ensembl data download directory and places the output files there.
- species name - Change to species of interest. For this example 'tetrahymena_thermophila'.
- species two letter code - change to two letters that best represent your species.

- random number - this is used when generating GeneMANIA unique identifers. If you have multiple species in your instance make sure that this number is different. In the above example, all random GeneMANIA identifiers will start with 19 for tetrahymena.

  Also, before running the script you need to update the configuration file. For this example the configuration file is spd_tetra.cfg. the only thing that needs to be updated in this file is the **spd_org** variable.

You can choose to modify the spd_tetra.cfg file or create a new file specific for you organims. If you create a new file for your organism make sure to update runall_indiv_species.sh line 37 to reflect the new file name.

./1.export_spd_from_ensembl.sh $ensembl_version spd_tetra.cfg

to

./1.export_spd_from_ensembl.sh $ensembl_version <new_spd_config_filename.cfg>

```
[BuildScriptsConfig]
# this section is for the configuration for the build process script
# ___[revision]_____________________________________________
revision = R2

# ensembl releases to use
ensembl_core_release =104
ensembl_plants_release =51
ensembl_metazoa_release =51
ensembl_bacteria_release =51

#had to move this because of issues with get_spd.pl script
# mysql host, username, password
mysql_h = localhost
mysql_u = root
mysql_p = gm.build

# shared protein domain organisms
spd_org = tetrahymena_thermophila
```

```
./runall_indiv_species.sh
```

Once the script is finished running you will have the following files and directories (It will exist both on the docker and on the computer that the docker is running on. Its location on main computer depends on what you set *-v /home/gmbuild/ensembl_data:/home/gmbuild/ensembl_data* in the docker

run command) :

```
ls ~/ensembl_data

ensembl_data/
  July_23_2021/
    current_build.log
    tetrahymena_thermophila_core_51_104_1/
    Work/
      ENSEMBL_ENTREZ_Tt
      Tt_done.txt
      spd/
        interpro/
        pfam/
```

The *ENSEMBL_ENTREZ_Tt* file is the identifier mapping file. If you have additional identifier mapping data that is not present in Ensembl you can modify this file directly (make sure to keep the overall structure) but, for example, Tetrahymena maintains gene names that are not incorporated into Ensembl. (table of gene names can be found here - http://ciliate.org/index.php/show/namedgenes) Through Scripting or Excel you can add these gene symbols to the *ENSEMBL_ENTREZ_Tt*.

Main output of this step is the directory with all its files. This directory will be mapped onto the main genemania data build docker and used in subsequent steps.

```
~/ensembl_data/July_23_2021/Work
```

The ~/ensembl_data/July_23_2021/Work/spd directory contains shared domains networks computed from the ensembl data. Under the *spd* directory there are two directories (interpro and pfam) each containing a directory for the organism(s) being analyzed with the shared domain interactions.

Chapter 4

Website build Stage 2 - Gather and clean network and attribute data

The objective of this stage is collect all the network and attribute data that we are going to use in this instance of GeneMANIA. Depending on the species there will be varying sources that the data can come from. For the main GeneMANIA there are scripts to download and format data from:

1. Identifiers - extracted from Ensembl in the previous stage.
2. Gene Ontology annotation - downloaded from GO in GAF format
3. Biogrid
4. GEO - imports expression datasets from specified platform series identifiers specified in configuration file.
5. I2D
6. iRef - static resource
7. Pathway Commons - static resources from 2011
8. Shared Protein Domains - as calculated from the files created from the Ensembl export in the previous step.
9. Static Networks - networks created and curated manually.

All of the configuration happens in the genemania config file. For the above resources you need to specify what organisms you wish to download. If there are no tag specified in the config file then nothing will be downloaded and that step will be skipped. Each of the above data sources will be expanded on below. They each have their own script to download and process them.

4.1 How to build your data

In order to build the data we need to create an genemania_databuild docker container. It will need your identifier file that you created in the previous step.

4.1.1 Create and setup a container of the gmbuild data docker instance

The docker instance that we are going to use can be found here - genemania_databuild_base.

1. Install Docker - for instructions see here
2. check out GeneMANIA_build from the Baderlab github (https://github.com/BaderLab/GeneMANIA_build.git)

```
git clone https://github.com/BaderLab/GeneMANIA_build.git
```

Create container of genemania_databuild_base instance.

- each -v parameter specifies a local directory, or volume, that is mapped to a directory on the docker. For example, the directory /home/gmbuild/ensembl data on your machine gets mapped to the location /home/gmbuild/ensembl_data in the docker container. Any file that is put into that directory on the docker will show up on the corresponding directory on your machine.
- There are multiple volumes mapped to the ensembl_mirror
  - /home/gmbuild/Tetrahymena/ensembl_data/July_23_2021/Work –> /home/gmbuild/ensembl_data
  - /home/gmbuild/Tetrahymena/gm_data –> /home/gmbuild/dev - This is the directory where all the code and data is going to be built. When the container is created it will create a directory on the docker in /home/gmbuild/dev/r# (r# is specified in the next variable). This directory will contain the following structure:
  - bp
  - data
  - db
  - src
- -e VERSION=r1 - specifies an environment variable that is used when the container is first created. On creation a direcotry will be created with the revision number and the config file will be updated to reflect this version

Problem: If you want to use the same docker for multiple builds of the data you will have to create the directory structure and update config file manually.

> **Solution**: Recommendation - create a new docker container for every revision to initialized everything correctly

- With the –name tag you can specify what you would like to call your instance. This name can be used when logging into the instance. For this example we have called this instance genemania_build_tetrahymena.

```
docker run -dit
-v /home/gmbuild/Tetrahymena/ensembl_data/July_23_2021/Work:/home/gmbuild/ensembl_data
-v /home/gmbuild/Tetrahymena/gm_data:/home/gmbuild/dev
-v /home/gmbuild/Tetrahymena/db_build:/gm/db_build
-e VERSION=r1
--name genemania_build_tetrahymena
baderlab/genemania_databuild_base /bin/bash
```

Once the instance is created, log into it.

```
docker exec -it genemania_build_tetrahymena /bin/bash
```

4.2 Gene Ontology annotation

4.3 Static Networks

Bibliography

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., et al. (2010). The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research*, 38(suppl_2):W214–W220.

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.