Submission of High-Seq data to GEO (before submission of paper for publication)

Most journals have now implemented a rule that requires submission of all raw data to a public access database prior to publication. In practice, this means that you need to have the raw data deposited before sending paper for review, as you need to provide the reviewers with access the these data for evaluation. For High throughput sequencing (High-Seq) the most commonly accepted database is the NCBI GEO database, and in the following is a simplified instruction of how to upload your raw data and processed data to GEO. Further details are provided at the NCBI GEO website, and at the GNU website (for information regarding the tar command: www.gnu.org/software/tar/). In order to upload large datasets, the files needs to be compressed. This can be done by the command gzip on all txt files, and by packaging and compressing the folder containing all data into a tar archive. Finally, you need a program for transferring the files from your computer to the GEO server (Filezilla is suggested below).

Goto NCBI GEO: http://www.ncbi.nlm.nih.gov/geo/
Create User ID and password : Will assume "YOUR_NAME" as user id below

Submission of High-Seq data-sets:

https://www.ncbi.nlm.nih.gov/geo/info/seq.html

You Need:

1. Fill in excel form found on their website (this is updated frequently, so download latest version of this form on lower end of above web page!!)
2. You need title of study, abstract, as well as descriptions of protocol and samples used for the RNA-Seq (or ChIP-Seq) to describe in submission form above.
3. Need the following files ready (gzip'ed) for uploading (instructions per June 2012):
    * Rawdata (e.g. fastq.txt) and
    * processed files (not bam, but bedgraph or bigwig), and
    * Final excel (.txt) list of calculated RPKM values (and ratios if referred to), or peak-list (for ChIP-Seq)
    * access to command line (unix/linux).
        * Mac computers has "terminal" which is a unix based system.
        * for windows: see additional help: There might be solutions out there, but as of now, my best suggestion is to use a "Virtual machine" and install a linux-based operating system to perform these tasks. This is also useful (or required) for bioinformatic manipulations using for instance BEDTols.

Create .tar archive file for uploading to GEO: in command line window

1. Make new folder for This GEO submission (on computer with unix/linux access)
    a. Create a new folder: for example "YOUR_NAME_exp1"
    b. Goto your new older: > cd YOUR_NAME_exp1
2. Download the excel sheet: **GEOarchive spreadsheet format (latest version)**
    a. **From: https://www.ncbi.nlm.nih.gov/geo/info/seq.html**
    b. **Under: Metadata spreadsheet (scroll down)**

3. Fill in with your info & save in above directory
    a. Save rawdata files (one for each data-set) in YOUR_NAME_exp1 folder
    b. Gzip txt files (like fastq files)
4. Also save processed data files (one for each data-set): .bedg or .bigwig
    a. Do not Gzip .bigwig
    b. include excel files with RPKM (and ratio's if desired) or ChIP-Seq Peak list as tab delimited .txt
5. Then create a .tar archive named with your GEO UserID (eg. YOUR_NAME_exp1.tar) containing all files from above (excel sheet, fastq, bigwig and .txt files):
    a. Go up, out of your directory: cd ..
    b. Assuming Command line prompt ">"
    c. Write command: > tar --create --verbose --file= YOUR_NAME_exp1.tar YOUR_NAME_exp1
    d. (or – alternatively: normal files into .tar, then gzip entire .tar archive.)
6. Use Filezilla to ftp transfer the .tar archive to GEO (or .tar.gz)
    a. Adr: ftp://geo:D0gDAzr0va@ftp-private.ncbi.nih.gov/fasp
    b. **GEO FTP login information:**
        i. **Host:** ftp-private.ncbi.nih.gov
        ii. (Please use the 'fasp' directory)
        iii. **Username: geo**
        iv. **Password as of June 2012:** D0gDAzr0va
    c. **(Check if this info is current on website listed above!**)
7. After transferring files, please send an e-mail to geo@ncbi.nlm.nih.gov with the following information:
    a. GEO account user name (YOUR_NAME);
    b. Name(s) of the archive file(s) deposited; YOUR_NAME_exp1
    c. Public release date (up to 3 years from now).
8. Processing by GEO is approximately 5 business days: Then you will receive an email with accession number and can log into your GEO account and:
    a. Create specific login for reviewers:
    b. Generate the reviewer URL - click the *'Click here to create a reviewer access link'* button located at the top of your Series record
9. Remember to update your GEO records with a reference to your publication as soon as your manuscript is accepted/published.
10. Website info tar:
    a. http://www.gnu.org/software/tar/
    b. http://www.gnu.org/software/tar/manual/tar.pdf
11. Website info file format: http://www.ncbi.nlm.nih.gov/books/NBK47537/