glossary and GENRE software download/tutorial

- Download glossaryGENRE_download.zip from http://thebrain.bwh.harvard.edu/glossary-GENRE/download.html
- Unzip the scripts in a directory you can use as the GENRE-glossary home directory. Everything in the subsequent downloads will move to a subdirectory of this home directory.
  - $ unzip glossaryGENRE_download.zip
- Verify all dependencies installed.
  - $ sh testEnv.sh
  - Dependencies are:
    - Python 2.X (tested in 2.6.6 and 2.7.6)
      - argparse
      - os
      - sqlite3
      - re
      - time
      - random
      - csv
      - math
      - hashlib
      - subprocess
      - decimal
    - R (tested in 3.0.2 and 3.2.3)
      - zoo
      - Biostrings
      - methods
      - BiocGenerics
      - parallel
      - IRanges
      - XVector
  - BEDTools (tested in 2.23.0 and 2.25.0)
  - sqlite3 (tested in 3.6.20 and 3.11.0)
  - Download all needed dependencies and try again.
- View all available downloads.
  - $ python getGENRE.py -avail
- Obtain a foreground set.
  - Must be in BED file format.
  - Only the first three columns will be utilized.
- Download an appropriate database for your foreground. At the moment, only the database from Mariani et al., 2017 is available. If you are looking for more genomes/lengths, please contact the Bulyk lab.
  - $ python getGENRE.py -download "db_ID"
  - db_ID: The ID is the second column of the -avail output; choose an ID under the DATABASES heading.
  - First, narrow options by your foreground set's genome.
  - Second, narrow options by your foreground set's length distribution.
    - dflt: "default" length of 200 bp

- o For example, if you've chosen a foreground of GATA peaks from the hg19 genome with lengths of 200 bp, you'd download the hg19_dflt database.
  - ▪ $ python getGENRE.py -download hg19_dflt
- o The program will automatically download the GENRE scripts and hg19 genome as dependencies.
- Run GENRE using the pre-defined database you've just downloaded and your chosen foreground set.
  - o $ bin/GENRE/GENRE "db_ID" "FG set"
  - o db_ID: see above
  - o FG set: your foreground set in BED format
  - o Optional arguments:
    - ▪ -seed seed: seed for randomization, preferably a number; default 123456789
    - ▪ -BG BG: BG BED filename. If not given, prefix will be the same as the FG file; suffix will be the db_ID
    - ▪ -mult mult: multiplicity factor (positive integer); default 1 (mult 1 is needed to run with the glossary)
  - o For the above example, to get 1 background sequence for 1 foreground sequence in an output BED file named hg19_GATA_dfltBG.bed:
    - ▪ $ bin/GENRE/GENRE hg19_dflt hg19_GATA.bed -seed 49472047 -BG hg19_GATA_dfltBG.bed
  - o Output:
    - ▪ Message with GENRE version number, seed, and BG filename.
    - ▪ Foreground FASTA file
    - ▪ Background BED and FASTA file
- Download an appropriate motif set
  - o $ python getGENRE.py -download "motifs_ID"
  - o motifs_ID: The ID is the second column of the -avail output; choose an ID under the MOTIFS heading.
  - o For example, to download the glossary in kmer format,
    - ▪ $ python getGENRE.py -download glossary-kmer
  - o The program will automatically download the glossary scripts as a dependency.
- Run glossary script with foreground and background FASTA files and the motif set you've just downloaded.
  - o $ bin/glossary/glossary "FG FASTA file" "BG FASTA file" "motifs_ID"
  - o FG FASTA file: your foreground set in FASTA file format
  - o BG FASTA file: the background set in FASTA file format
  - o motifs_ID: see above
  - o Both the FASTA files are outputted by GENRE if you don't have them from other sources.
  - o For the above example,
    - ▪ $ bin/glossary/glossary hg19_GATA.fa hg19_GATA_dfltBG.fa glossary-kmer
  - o Output:
    - ▪ Best matches per motif file
      - • PWM – PWM score
      - • kmer – E score
      - • location in sequence
      - • sequence matched

- AUC results
  - Motif: Motif file
  - AUROC: enrichment of motif in foreground over background
  - p-val: p-value of AUROC - significance of enrichment
  - numFG: number of genes with motif match in foreground
  - numBG: number of genes with motif match in background
  - medFG: median distance of best motif match in foreground
  - medBG: median distance of best motif match in background
  - posTest: position test p-value
- All-in-One option: Combine GENRE and the glossary.
  - $ bin/glossary/glossary_GENRE "db_ID" "FG set" "motifs_ID"
  - Optional arguments:
    - -seed seed: seed for randomization, preferably a number; default 123456789
    - -BG BG: BG BED filename. If not given, prefix will be the same as the FG file; suffix will be the db_ID
  - Same output is given as for GENRE and glossary separately.
- Add your own motif sets
  - $ bin/glossary/addMotifSet "motifs" "type"
  - motifs: file of motifs separated by >motif_name or directory of single motif files with the motif_name being the filename (minus extension)
  - type: type of motif; either pwm or kmer
    - pwm specifics
      - rectangular matrix
      - all positions add to 1
      - bases are assumed to be ACGT
      - bases must be labeled if length is 4 or 5bp, though it's always a good idea to label them
      - positions may or may not be labeled
    - kmer specifics
      - two columns separated by non-newline whitespace: kmer and E-score
  - Optional arguments
    - -description description: description to use in the -avail output