Generating biased gene lists : pre-processing

Laura Biggins

4 March 2019

Pre-processing of data on cluster

Generating a gene info file for the mouse genome

The script create_gene_info_file_from_gtf.pl takes a gtf file and parses it to create a gene info file that contains all the genes that are annotated in the gtf file. This should be all the genes in that version of the genome.

Downloaded the raw version of the file from github `wget https://raw.githubusercontent.com/s-andrews/GOliath/master gene_info_processing/create_gene_info_file_from_gtf.pl`

Run this script to create the gene info file `perl create_gene_info_file_from_gtf.pl --gtf Mus_musculus.GRCm38.94.gtf.gz --genome GRCm38`

We'll import the gene info file so that we can plot the GC distribution for all genes in the Mus_musculus.GRCm38.94.gtf.gz genome. There are import_GTF and parse_GTF_info functions within the GOcategoryStats package but to get genome information i.e. GC content, the parsing and lookups need to be done with access to genome information, so on the cluster. The import_GTF and parse_GTF_info functions just work with the gtf file itself.

Import the processed gene info file

```
genfo <- read.delim("M:/biased_gene_lists/Mus_musculus.GRCm38.94_gene_info.txt")
head(genfo)
```

```
##              gene_id     gene_name chromosome   start     end strand
## 1 ENSMUSG00000102693 4933401J01Rik          1 3073253 3074322      +
## 2 ENSMUSG00000064842       Gm26206          1 3102016 3102125      +
## 3 ENSMUSG00000051951          Xkr4          1 3205901 3671498      -
## 4 ENSMUSG00000102851       Gm18956          1 3252757 3253236      +
## 5 ENSMUSG00000103377       Gm37180          1 3365731 3368549      -
## 6 ENSMUSG00000104017       Gm37363          1 3375556 3377788      -
##                biotype biotype_family length GC_content no_of_transcripts
## 1                  TEC             NA   1069      0.342                 1
## 2                snRNA             NA    109      0.358                 1
## 3       protein_coding             NA 465597      0.385                 3
## 4 processed_pseudogene             NA    479      0.399                 1
## 5                  TEC             NA   2818      0.408                 1
## 6                  TEC             NA   2232      0.370                 1
```

```
colnames(genfo)
```

```
##  [1] "gene_id"           "gene_name"         "chromosome"
##  [4] "start"             "end"               "strand"
##  [7] "biotype"           "biotype_family"    "length"
## [10] "GC_content"        "no_of_transcripts"
```

Using this gene info file and some extra processing, lists of genes were generated for the following categories:

1. Length biased gene sets
2. High transcript biased gene sets
3. GC biased gene sets

4. Chromosomal biased gene sets
5. Closest genes to random positions
6. Public data gene sets

The processing for categories 1-4 was carried out within an R session and is detailed in the Rmarkdown documents of the same names.

To generate the gene sets for Category 5 - Closest genes to random positions, a python script was written. This generated random locations in the genome and found the closest gene to each position.

Category 6 - the public data required a separate, more extensive workflow.

The processing of each of the 6 categories is detailed in the individual Rmarkdown documents.