Gene Score Gene Definition Files for Standardized Data: Provenance
Tod Casasent
2019-12-30-1111

Background

For the Gene Score Algorithm, we needed a reproducible way of getting gene definitions. (For details, about sources not to use, see the Archive section at the end of this document.)

GTF Files

GDC references files are listed at: https://gdc.cancer.gov/about-data/data-harmonization-and-generation/gdc-reference-files Current files in data were downloaded 2018-08-03-1500

For HG38 genes, we use this file: gencode.v22.annotation.gtf.zip. This contains a reference genome, "used in RNA-Seq alignment and by HTSeq".

For HG19 genes, we use this file: TCGA.hg19.June2011.gaf.zip.

GTF Gene Conversion

For processing the GTF files, we ignore any line beginning with #, which indicates a comment line. The first line is a header line. In the table below the entry column indicates from which column the data is derived based on a zero index, with an additional string for column 8, which takes the form:

key1 "value 1"; key 2 "value 2";

Description	Entry	Notes
Value	2	Type of entry, such as gene or exon. For gene conversion, we u
Chromosome	0	Chromosome entry, which can include mitochondrial entries. F
Start	3	
End	4	
Strand	6	
Gene Source	8 (gene_source)	Gene calling entities are "ensembl_havana", "ensembl", "havana
Gene Name	8 (gene_name and gene_id)	Result is the gene name, a pipe "|", and the gene id. The gene_
Gene Type	8 (gene_biotype)	Genes of type "rna" may have multiple entries for locations. O

Archive: DCC Background

The original HG18 and HG19 mapping/annotation files for Standardized Data had no provenance. The files were provided from non-reproducible sources. I decided that we needed a reproducible, documented method of creating HG19 and HG19 annotation files.

I first considered the stand-by most people go to: the UCSC Genome Browser known.Genes.txt file. The issue with the knownGenes file is that it is not a list of genes. It is a list of transcripts and their mapping to UCSC ids. The difference is important. A gene is defined as existing along a particular length of the genome, on a particular chromosome and strand. The UCSC knownGenes file assigns a UCSC id to each line in the knownGenes files. However, there does not exist a complete, reliable one-to-one mapping from UCSC ids to HGNC gene symbols, which is what most researchers are familiar with and want to use.

So, I created a one to one mapping file for UCSC ids to HGNC gene symbols. I used two files. First, I used a file which contains one-to-one HGNC to UCSC mappings from http://ftp.ebi.ac.uk/pub/databases/genenames/hgnc_complete_set.txt.gz. This generates a handful of name collisions (multiple UCSC ids mapping to a single HGNC gene symbol for eight pseudogenes). I then used the UCSC kgXref file which maps UCSC ids to HGNC gene symbols, and is retrieved from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/kgXref.txt.gz for HG19 and http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/kgXref.txt.gz for HG18. However, only UCSC ids which map to one HGNC symbol are processed.

Using these "one-to-one mappings" and the knownGene files from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/knownGene.txt.gz for HG19 and http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/knownGene.txt.gz for HG18, gives disappointing results. The HG18 file produces 1975 gene symbols and the HG19 file produces 7973 gene symbols.

This reduced output led to a search for alternate solutions.

I discovered that NCBI has a peculiar text format for gene annotations which seemed to contain all the information of interest. However, NCBI no longer provides HG18 and HG19 annotation data. The other online sources I check either provide nothing but the gene annotations for the newest reference genome GRCh38.p3 (NCBI/Entrez does this) or seem to have stopped creating gene annotation files before GRCh37.p13 was released. This led me to the Ensembl site.

Converting Ensembl Data to Gene Annotations

HG18 data from Ensembl consists of four dat files. HG19 data from Ensembl consists of two sets of dat files. The first HG19 set is the 24 chromosome files. The second HG19 set is the patch files.

Each file is processed the same.

1. The file is parsed looking for a line that starts with "LOCUS". This line is used to determine the chromosome.

a. The word after LOCUS is extracted, to determine the chromosome.

b. If the word starts with c, HG, GL, or NT, we cannot determine a chromosome, so the entry is skipped.

c. If the word starts with HSCHR, we remove that portion, leaving a chromosome.

2. If we are within an entry which gave us a good chromosome value, we look for additional data.

d. If the line starts with "gene ", we remove that string.

i. Check if the substring starts with "join", which indicates a disjoint "gene", which we skip. (Disjoint genes are those whose mapping indicates two or more widely separated genomic locations.)

ii. If the substring does not start with "join", we extract the genomic location data.

1. If the line starts with complement, then the gene is on the minus strand, otherwise, it is on the plus strand.

2. Remove "complement", "(", ")", and ".." from the substring, and split the string on the remaining space. This gives us two long values with the smaller used as the start coordinate, and the larger and the ending coordinate for the gene.

3. This genomic location is used for the locus determined next.

e. If the line did not start with "gene", we check for locus symbols (such as, gene symbols, locus tags, and ribosomal RNA) or to see if we are done with this LOCUS entry.

iii. If the line starts with "/locus_tag=", extract the tag name in double quotes after the locus_tag string, and use that as the locus name.

iv. If the line starts with "/gene=", extract the gene identifier after the equal sign, and use that as an alternate gene id.

v. If the line does not start with a "/", then you have reached the end of data for this "gene". Combine the current locus, a pipe "|", the chromosome, another pipe, and the alternate gene id into a temporary "gene symbol". These symbols will later be reduced to the smallest portion (between pipes) necessary to identify a "gene".

All files for HG19 (patch and regular) or HG18 are processed as described above. This gives a hashmap of locus|chromosome|gene-symbol and the corresponding genomic location data.

The locus token from the hashmap key corresponds to HGNC gene symbols (which include non-traditional genes that have products such as miRNA or ribosomal RNA, rather than mRNA and therefore protein). The keys are reduced to the smallest possible component that uniquely identifies a gene. One caveat to that is that any name collisions at a lower level (such as "foo|X" and "foo|X") mean that the higher level name/token is included (meaning "foo|X|X2000" and "foo|X|X1000", rather than "foo|X" and "foo|X|X1000").

For HG18, the original Ensembl files cover approximately 43,779 loci. The gene annotation file we generate contains 36,725 gene entries.

For HG19, the original Ensembl files cover approximately 57,736 loci. The gene annotation file we generate also contains 57,736 gene entries.