**Running on Hummingbird**

d-128-95-149-219:bsmap-2.74 sr320$ ./bsmap -w 1000 -a /Volumes/NGS\ Drive/NGS\ Raw\ Data/Oyster\_BSseq/filtered\_Unlabeled\_NoIndex\_L003\_R1.fastq.gz -d /Volumes/web/cnidarian/TJGR\_RepProMask\_TE.fa -o /Volumes/web/cnidarian/BiGill\_BSMAP\_TEonly\_v2.sam -p 16

BSMAP v2.74  
Start at: Fri Jun 14 13:19:06 2013  
  
Input reference file: /Volumes/web/cnidarian/TJGR\_RepProMask\_TE.fa (format: FASTA)  
Load **in** 119787 db seqs, total size 39414569 bp. 2 secs passed  
total\_kmers: 43046721  
Create seed table. 5 secs passed  
max number of mismatches: read\_length \* 8% max gap size: 0  
kmer cut-off ratio: 5e-07  
max multi-hits: 1000 max Ns: 5 seed size: 16 index interval: 4  
quality cutoff: 0 **base** quality **char**: '!'  
min fragment size:28 max fragemt size:500  
start **from** read *#1 end at read #4294967295*  
additional alignment: T **in** reads => C **in** reference  
mapping strand: ++,-+  
Single-**end** alignment(16 threads)  
Input read file: /Volumes/NGS Drive/NGS Raw Data/Oyster\_BSseq/filtered\_Unlabeled\_NoIndex\_L003\_R1.fastq.gz (format: gzipped FASTQ)  
Output file: /Volumes/web/cnidarian/BiGill\_BSMAP\_TEonly\_v2.sam (format: SAM)

BSMAP complete on Hummingbird   
Total number of aligned reads: 2545683 (1.8%) Done. Finished at Sat Jun 15 20:45:07 2013 Total time consumed: 113161 secs

python [methratio.py](http://methratio.py/" \t "_blank) -d /Volumes/web/cnidarian/TJGR\_RepProMask\_TE.fa -z -g -o  /Volumes/web/cnidarian/BiGill\_methratio\_TEonly\_A.txt -s /Volumes/Bay3/Software/BSMAP/bsmap-2.73/samtools /Volumes/web/cnidarian/BiGill\_BSMAP\_TEonly\_v2.sam

[http://eagle.fish.washington.edu/cnidarian/BiGill\_methratio\_TEonly\_A.txt](http://eagle.fish.washington.edu/cnidarian/BiGill_methratio_TEonly_A.txt" \t "_blank)

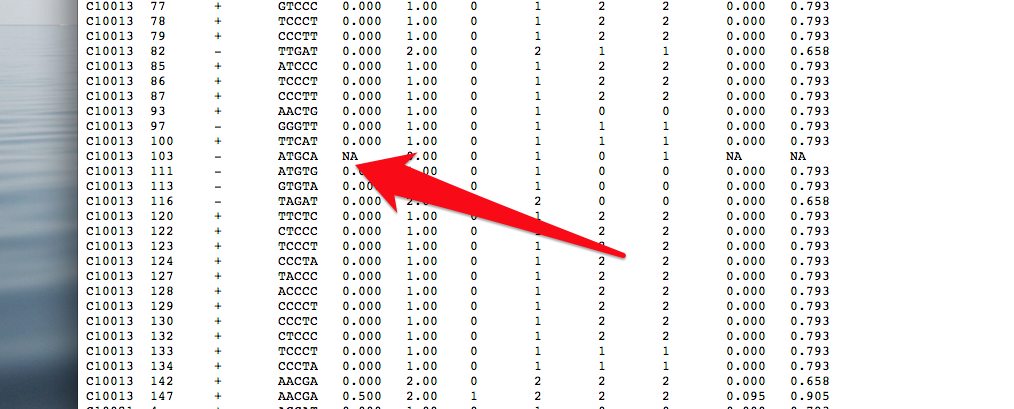
|  |
| --- |
| **BiGill: Figuring out this Effective CT factor** |

methratio file in  SQLShare

[https://sqlshare.escience.washington.edu/sqlshare#s=query/sr320%40washington.edu/BiGill\_methratio\_v9\_A.txt](https://sqlshare.escience.washington.edu/sqlshare" \l "s=query/sr320%40washington.edu/BiGill_methratio_v9_A.txt" \t "_blank)

----

Having look at raw output:



Based on methratio - NA methratio always

C     CT     revG     revGA

0       #     0     #

CI\_lower and CI\_upper also NA

---

7.1 [methratio.py](http://methratio.py/" \t "_blank)

python script to extract methylation ratios from BSMAP mapping results. Require python 2.X.

For human genome, [methratio.py](http://methratio.py/" \t "_blank) needs ~26GB memory.

For systems with limited memory, user can set the -c/—chr option to process specified chromosomes only,

and combine results for all chromosomes afterwards.

Usage: python [methratio.py](http://methratio.py/" \t "_blank) [options] BSMAP\_MAPPING\_FILES

BSMAP\_MAPPING\_FILES could be one or more output files from BSMAP.

The format will be determined by the filename suffix.

(SAM format for \*.sam and \*.bam, BSP format for other filenames.)

Options:

  -h, —help            show this help message and exit

  -o FILE, —out=FILE   output file name. (required)

  -d FILE, —ref=FILE   reference genome fasta file. (required)

  -c CHR, —chr=CHR     process only specified chromosomes. [default: all]

                        example: —chr=chr1,chr2 (this uses ~4.5GB compared with ~26GB for the whole genome)

  -s PATH, —sam-path=PATH

                        path to samtools. [default: none]

  -u, —unique          process only unique mappings/pairs.

  -p, —pair            process only properly paired mappings.

  -z, —zero-meth       report loci with zero methylation ratios.

  -q, —quiet           don’t print progress on stderr.

  -r, —remove-duplicate

                        remove duplicated mappings to reduce PCR bias.

            (This option should not be used on RRBS data. For WGBS, sometimes

            it’s hard to tell if duplicates are caused by PCR due to high seqeuncing depth.)

  -t N, —trim-fillin=N

                        trim N fill-in nucleotides in DNA fragment end-repairing. [default:2]

            (This option is only for pair-end mapping. For RRBS, N could be detetmined by the distance between

                        cuttings sites on forward and reverse strands. For WGBS, N is usually between 0~3.)

  -g, —combine-CpG     combine CpG methylaion ratio from both strands. [default: False]

  -m FOLD, —min-depth=FOLD

                        report loci with sequencing depth>=FOLD. [default: 1]

  -n, —no-header       don’t print a header line

  -i CT\_SNP, —ct-snp=CT\_SNP

                        how to handle CT SNP (“no-action”, “correct”, “skip”),

                        default: “correct”.

                        “correct”:      correct the methylation ratio according to the C/T SNP information

                        estimated by the G/A counts on reverse strand, see the output format below for details.

                        “skip”:         do not report loci with C/T SNP detected (i.e. detected A on reverse strand)

                        “no-action”:    do not consider C/T SNP.

Output format: tab delimited txt file with the following columns:

    1) chromorome

    2) coordinate (1-based)

    3) strand

    4) sequence context (2nt upstream to 2nt downstream in Watson strand direction)

    5) methylation ratio, calculated as #C\_counts / #eff\_CT\_counts

    6) number of effective total C+T counts on this locus (#eff\_CT\_counts)

            CT\_SNP=”no action”, #eff\_CT\_counts = #CT\_counts

            CT\_SNP=”correct”, #eff\_CT\_counts = #CT\_counts \* (#rev\_G\_counts / #rev\_GA\_counts)

    7) number of total C counts on this locus (#C\_counts)

    8) number of total C+T counts on this locuso (#CT\_counts)

    9) number of total G counts on this locus of reverse strand (#rev\_G\_counts)

    10) number of total G+A counts on this locus of reverse strand (#rev\_GA\_counts)

    11) lower bound of 95% confidence interval of methylation ratio, calculated by Wilson score interval for binomial proportion.

    12) upper bound of 95% confidence interval of methylation ratio, calculated by Wilson score interval for binomial proportion.

Example:

    python [methratio.py](http://methratio.py/" \t "_blank) —chr=chr1,chr2 —ref=hg19.fa —out=methratio.txt rrbsmap\_sample\*.sam

    python [methratio.py](http://methratio.py/" \t "_blank) -d mm9.fa -o meth.txt -p bsmap\_sample1.bsp bsmap\_sample2.sam bsmap\_sample3.bam

    python [methratio.py](http://methratio.py/" \t "_blank) -s /home/tools/samtools -t 1 -d arab.fa -o meth.txt bsmap\_sample.sam

Note: For overlapping paired hits, nucleotides in the overlapped part should be counted only once instead of twice.

[methratio.py](http://methratio.py/" \t "_blank) can correctly handle such cases for SAM format output, but for BSP format it will still be counted twice,

because the BSP format does not contain mapping information of the mate.

*[@1 day ago](http://sr320.tumblr.com/post/51475005119" \t "_blank)*

|  |
| --- |
| **BiGill - Gene Specific Methylation (Take 3)** |

Need to grab genes in correct orientation.

robertsmac:bin sr320$ ./ipython notebook

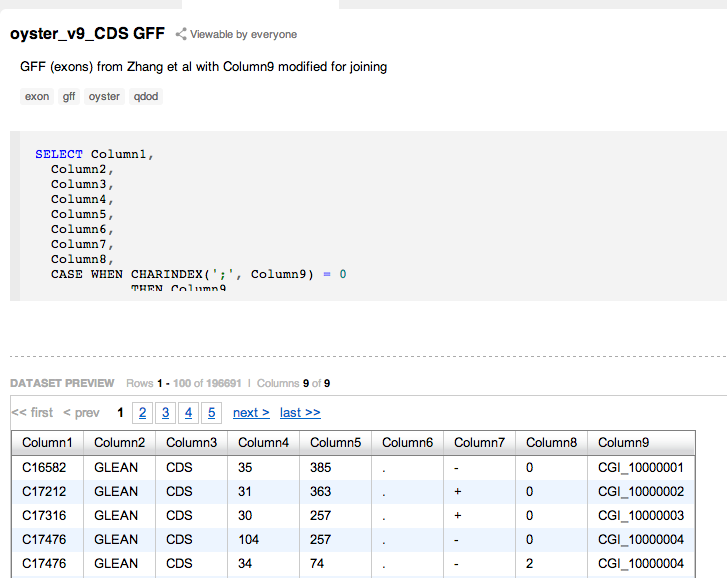
[http://nbviewer.ipython.org/url/eagle.fish.washington.edu/cnidarian/BiGill\_Gene\_Methylation.ipynb](http://nbviewer.ipython.org/url/eagle.fish.washington.edu/cnidarian/BiGill_Gene_Methylation.ipynb" \t "_blank)

^this is getting methylation

Still need to get new Coordinate system (again) for visualization.

SQLShare.

GFF are in with C9 Modified for joining….



Will be modifying CDS gff so that it coordinates are related to genes

Will need to to arithmetic based on strand so that 5' 3' orientation holds true

SELECT cds.\*,

  mRNA.Column4 as mRNA\_start,

  mRNA.Column5 as mRNA\_end

  FROM [sr320@washington.edu].[Snapshot of CDS GFF] cds

  LEFT Join [sr320@washington.edu].[oyster\_v9\_mRNA GFF] mRNA

  on cds.Column9 = mRNA.Column9

  Order by Column1 Desc

Create new Dataset

----

correct stuff

SELECT \*,

 Case when Column7 = '+'

  then Column4 - mRNA\_start + 1

  Else mRNA\_end - Column4 + 1

  END as New\_Start,

  Case when Column7 = '+'

  then Column5 - mRNA\_start + 1

  Else mRNA\_end - Column5 + 1

  END as New\_End,

 Column5 - Column4

  as exon\_size  -- trying to check arithmetic

  FROM

  [sr320@washington.edu].[CDS GFF with Gene start and stop] cd

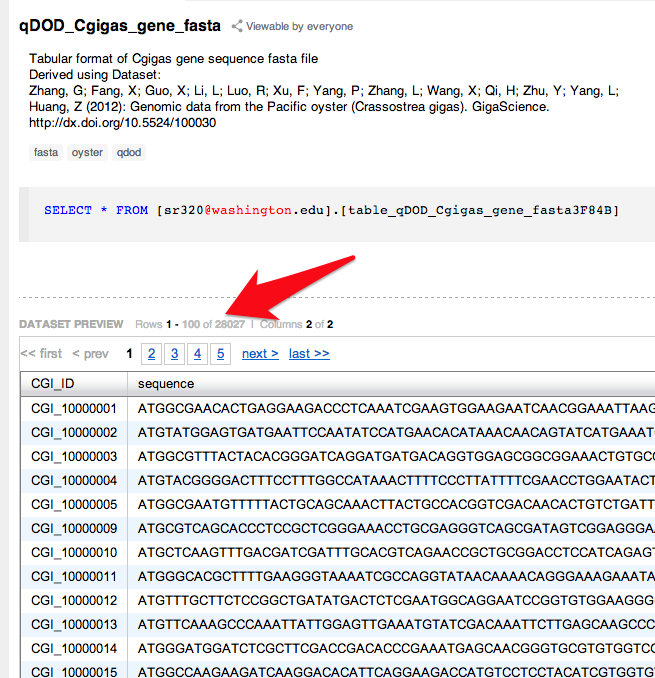
  Order by Column1 Desc, Column9 --makes it easy to see big genes

|  |
| --- |
| **BiGill - Redoing Gene Specific Data** |

There are 28,027 genes

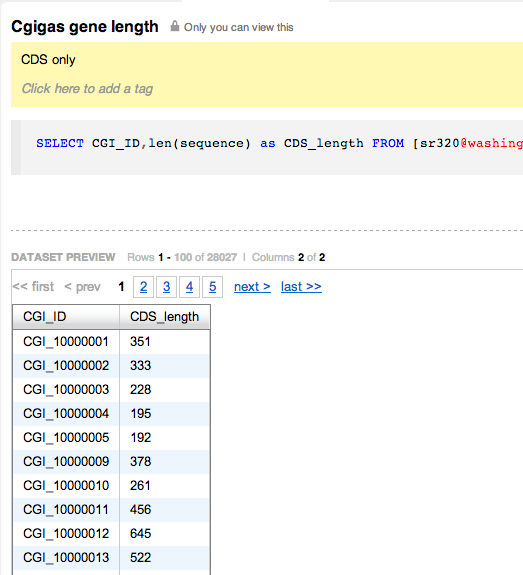
[http://aquacul4.fish.washington.edu/~steven/armina/oyster\_v9\_gene.fasta](http://aquacul4.fish.washington.edu/~steven/armina/oyster_v9_gene.fasta" \t "_blank)

also in SQLShare



Can get lengths

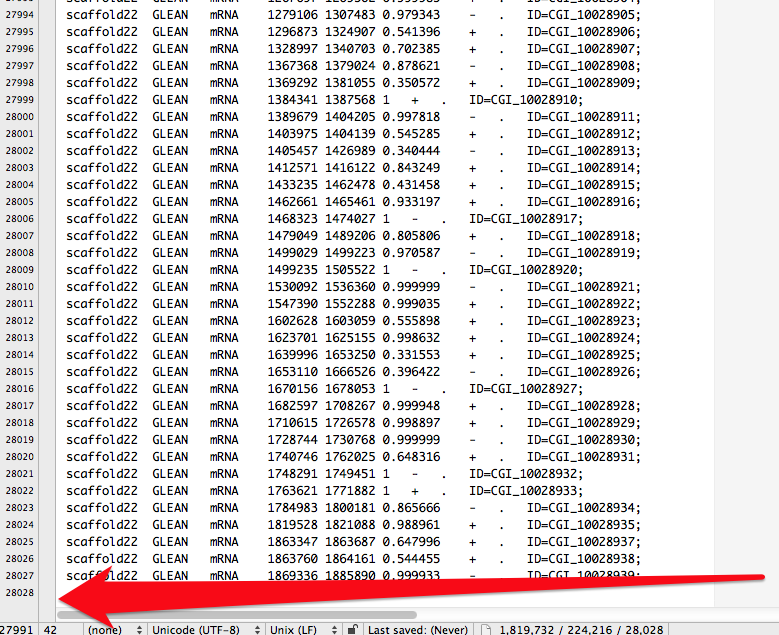
SELECT CGI\_ID,len(sequence) as CDS\_length FROM [sr320@washington.edu].[qDOD\_Cgigas\_gene\_fasta]



--------------------

Now want to get genomic structure of gene..

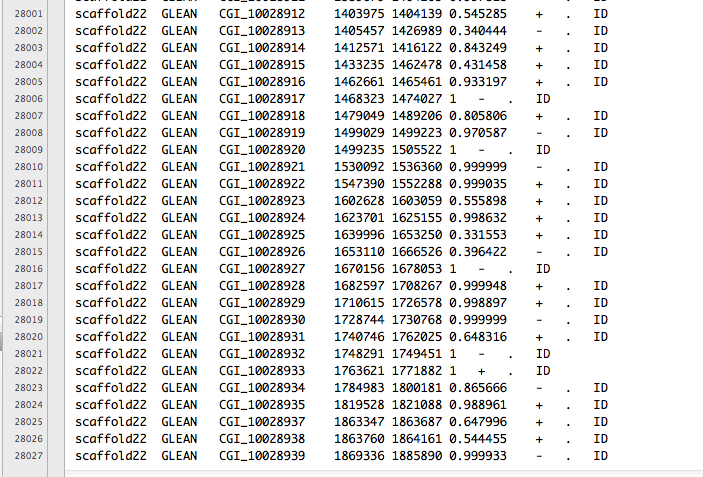
[http://aquacul4.fish.washington.edu/~steven/armina/oyster.v9.glean.final.rename.mRNA.gff](http://aquacul4.fish.washington.edu/~steven/armina/oyster.v9.glean.final.rename.mRNA.gff" \t "_blank)



GFF has Start on Stop and presumably includes introns….

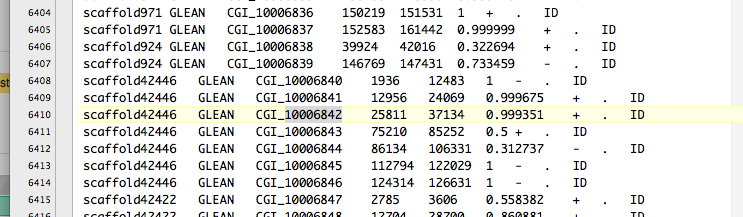
Reconfigured to get ID out

[http://eagle.fish.washington.edu/cnidarian/TJGR\_Gene\_28027\_column\_mod.gff](http://eagle.fish.washington.edu/cnidarian/TJGR_Gene_28027_column_mod.gff" \t "_blank)



Now lets get the corresponding fasta (again)

to avoid [http://genetwit.tumblr.com/image/51023089882](http://genetwit.tumblr.com/image/51023089882" \t "_blank)

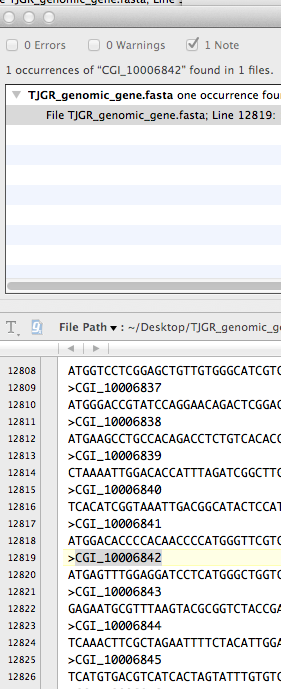


missing ID is in GFF.

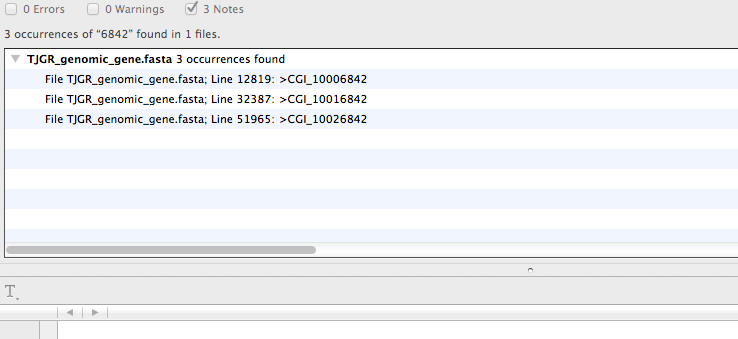
---

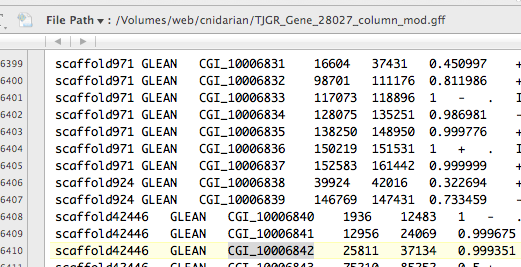
no idea what is going on

CGI\_10006842

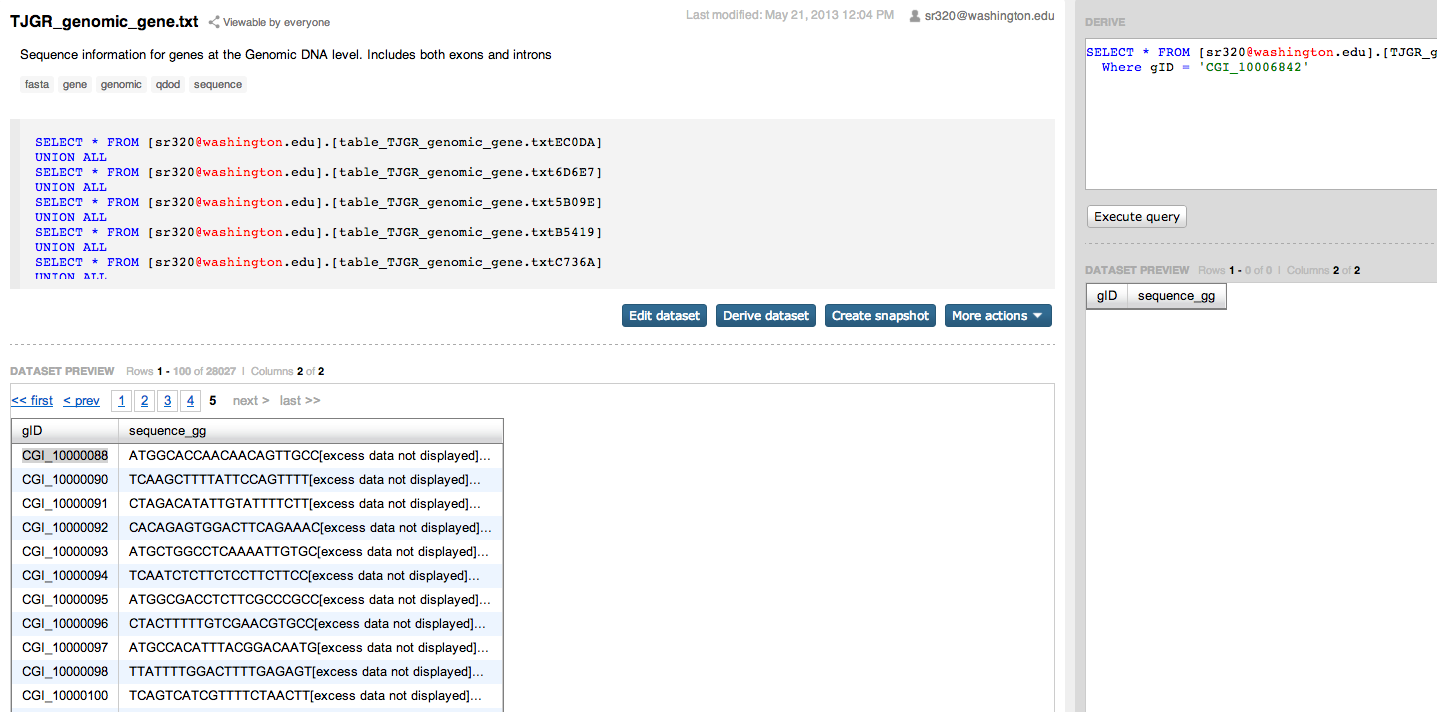
is in fasta

and gff

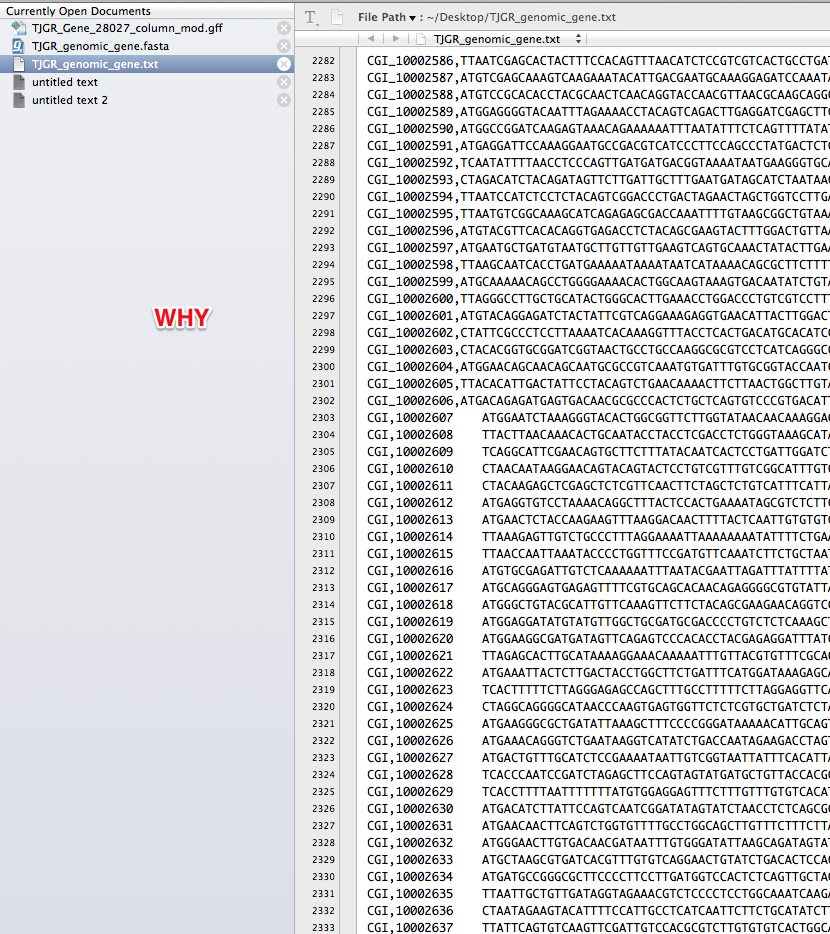




Missing in SQLShare



Downloading to desktop and look at in TextWrangler.



In Short

[https://sqlshare.escience.washington.edu/sqlshare#s=query/sr320%40washington.edu/TJGR\_genomic\_gene.txt](https://sqlshare.escience.washington.edu/sqlshare" \l "s=query/sr320%40washington.edu/TJGR_genomic_gene.txt" \t "_blank)

CGI\_10006842

---