FAQ for Segtor

February 28, 2011

1 Frequently asked questions

1.1 How come the numbers do not tally up in the all.stat or single.stat ?

If you use those files, please note that Segtor will tally up all the unique positions for a given coordinate/interval.
For instance, if a coordinate is both upstream of a transcript and downstream of another, Segtor will tally both positions rather than pick one in all.stat. The file single.stat will report both if they belong to different genes but will pick the best one if the transcripts belong to the same gene.

1.2 Is downloading the genome needed ?

If the user seeks to build a new segment tree index, the genome is required. However, for annotating coordinates/intervals and finding the closest TSS, only the chromosome index file is required.

1.3 I already have the genome (chromosome files) for the species I wish to use, do I have to download it again ?

No, but make sure the genome was downloaded from UCSC, that you have all the files and that all the files were downloaded without errors. There are two options:

1) If you have write permission in the directory with the chromosomes, you can use the createChromosomeIndex.pl script to create an index using the following command :

./createChromosomeIndex.pl [directory with the chromosomes]/ > [directory with the chromosomes]/index.txt
Make sure you do not have any extra files besides all of the chromosome files in fasta format downloaded from UCSC in that directory. Then create a symbolic link for the directory in your Segtor directory (see quick start for more details) called "chromosomes" in the directory for the species you wish to use. (e.g. ln -s /usr/share/data/ucsc/hg19/ /home/user/.segtor/hg19/chromosomes/)

2) If you do not have write permissions, create a directory in Segtor's directory (see quick start for more details) for the species you wish to use. Create a directory called "chromosomes" and create a symbolic links for each chromosomes files. Run the command:

./createChromosomeIndex.pl [Segtor's directory with the symlinks to chromosomes]/ > [Segtor's directory with the symlinks to chromosomes]/index.txt

1.4 Why is the output file empty ?

It is likely that no genes overlap the inputs you had. Make sure some of your inputs are within genes or increase the range. Make sure that the chromosome you are using has annotated genes. Also make sure that you are using the correct genome assembly since coordinates and chromosome names can vary greatly.

2 Common problems

2.1 Problem:

The program stalls at "testing internet connection..."

2.2 Solution:

Check to see if you have access to the internet. Check to see if you are behind a proxy and if so, enter the ip using the '-p' option

2.3 Problem:

I get either: Unable to retrieve ftp://hgdownload.cse.ucsc.edu/goldenPath/sasda12/chromosome
Status line 500 Server closed connection without sending any data back or
503 Bad Gateway or Downloading stalls

2.4 Solution:

Make sure the organism exists on the UCSC Genome Browser. Make sure you have the correct code and the organism's genome exists in:
ftp://hgdownload.cse.ucsc.edu/goldenPath/[organism]/chromosomes/

2.5 Problem:

The genome downloads correctly but does not download the database files

2.6 Solution:

Check to see if the database you wish to use exists in:
ftp://hgdownload.cse.ucsc.edu/goldenPath/[organism]/chromosomes/

2.7 Problem:

I know that my input coordinate is in a gene but Segtor does not detect it

2.8 Solution:

Is it possible that either your gene was newly added/modified in the database and was not the index. Are you sure you used the correct assembly for the species ? If you believe this is a bug, please contact us.

2.9 Problem:

My input coordinate is in a gene according to Segtor but I do not see that gene in the UCSC Genome Browser

2.10 Solution:

Segtor's indices were built using UCSC flat files. Genes/transcripts routinely get removed or added. Hence, it is possible that the gene was removed. If you build a new index and the gene is not present in the flat file yet Segtor detects it, please contact us.

2.11 Problem:

The program has problems upon building the data structure.

2.12 Solution:

Perhaps there is a parsing error in the gene files, please email your command line to passetti[at]inca.gov.br

2.13 Problem:

Upon using ESTs, the program stalls

2.14 Solution:

ESTs use a lot of RAM since there are a lot of ESTs especially for human builds. Try using a machine with more RAM.

2.15 Problem:

The RAM footprint increases and reaches 100

2.16 Solution:

Do not use the -n option to avoid having the per gene report or, if you need the report, breakdown your files into chunks and merge the results.

3 For any other problem

please contact: passetti[at]inca.gov.br