Solution - Week 10

Genomics and bioinformatics

1 General Notes

To completely understand the exercises and solutions, you need to understand the paper first. The exercise does not cover all the points in the paper, but guides you through important ones. After completing it, you should be able to reproduce all the figures in this paper and others of its kind.

2 Getting the data

The GEO accession code for this paper's dataset is GSE25107. Since all the NCBI databases are linked (they are actually on the same ftp server), all you need is one accession number for one database (in this case GEO).

1. From which database would you download the raw sequencing files?

   The raw sequencing files would be found in the SRA database (accession code: SRP004431)

2. From which database would you get general information about a specific sample?

   You can find that in GEO or in BioSample.

3. What is the difference between the datasets in SRA and those in GEO?

   The "wig" format you would find in GEO is a processed file, whereas the SRA is a compressed format for storing raw reads. The SRA database contains raw sequencing files as obtained from the sequencing machines. In this case, they are the output of the Illumina Genome Analyzer II and are text files with all the short reads that have been sequenced. On the other hand, the GEO "wig" files are obtained after mapping those reads to the yeast genome and performing normalization to obtain a quantification of reads throughout the yeast genome. See GSE25107_RAW.tar (custom).

3 Viewing the data in UCSC

1. In figure 1b in the paper, the authors show a proof of principle of their NET-seq protocol by comparing it to RNA-seq. Upload the appropriate tracks to the genome browser and zoom to the RPL30 gene (YGL030W). Do you get the same result?

   Since the RPL30 gene is on the + strand, the tracks that you need to upload are: WT_NC_plus (for the NET-seq data in the wild-type) and WT_mRNA_plus (for the RNA-seq in the wild-type). Notice that there are many reads mapping to the intron in the NET-seq data, but not the RNA-seq data (figure 1)

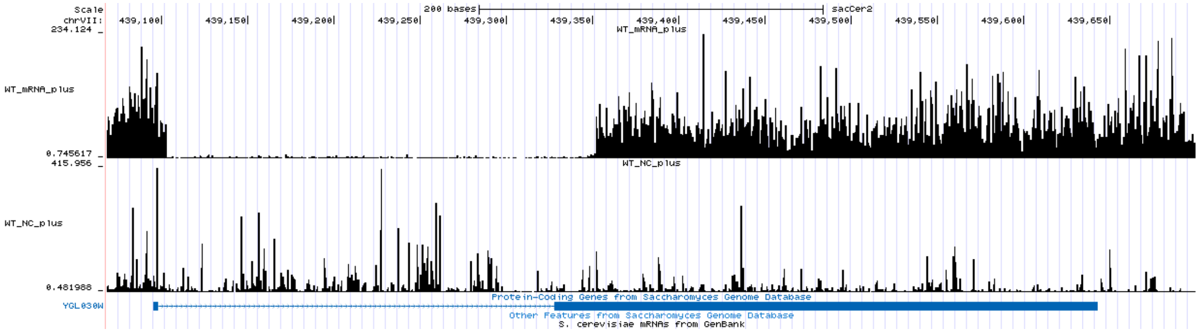

Figure 1: UCSC view of of the RPL30 gene

2. The authors talk about CUTs in the paper. What are they?

   Cryptic unstable transcripts (CUTs) are a class of short transcripts that are upstream and antisense to an annotated gene. They are degraded quickly, so they will always be undetected by RNA-seq. The biological function of these transcripts, if it exists, is not yet known.

3. Visualize the CUTs in the cluster of genes in figure 3a (make sure you upload the appropriate tracks). What do you notice when you compare *RCO1*Δ to the wild-type?

   For this exercise, you will need to upload the following tracks:

   - WT_NC_plus
   - WT_NC_minus
   - RCO1D_plus
   - RCO1D_minus

   Then you will need to navigate to a window where you could see all those genes like: chrVII:662,257-675,621 (figure 2). Notice the CUT sites in the "minus" tracks. Also notice that the RCO1D deletion mutant has a higher CUT expression (up to 533.394 reads per 10^7)

4. For the next exercise, you will need to extract gene annotation information (like strand and transcription start) for all the genes on chromosome VII. This is easy to do in UCSC. Go to Tools>Table Browser. Under group, select Genes and Gene Prediction Tracks, then select SGD Genes in the track drop down list. Define the position as "chrVII" and then click on get output. Observe the resulting table, for this will be useful for the next exercise.

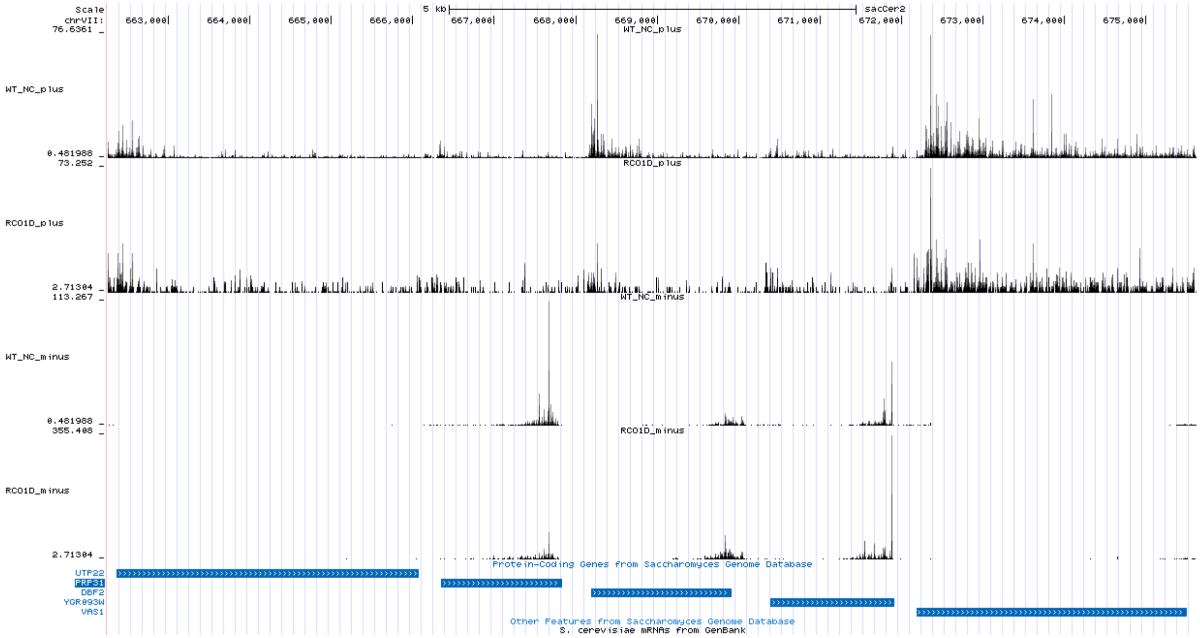

Figure 2: UCSC view of of the genes in the paper's figure 3a

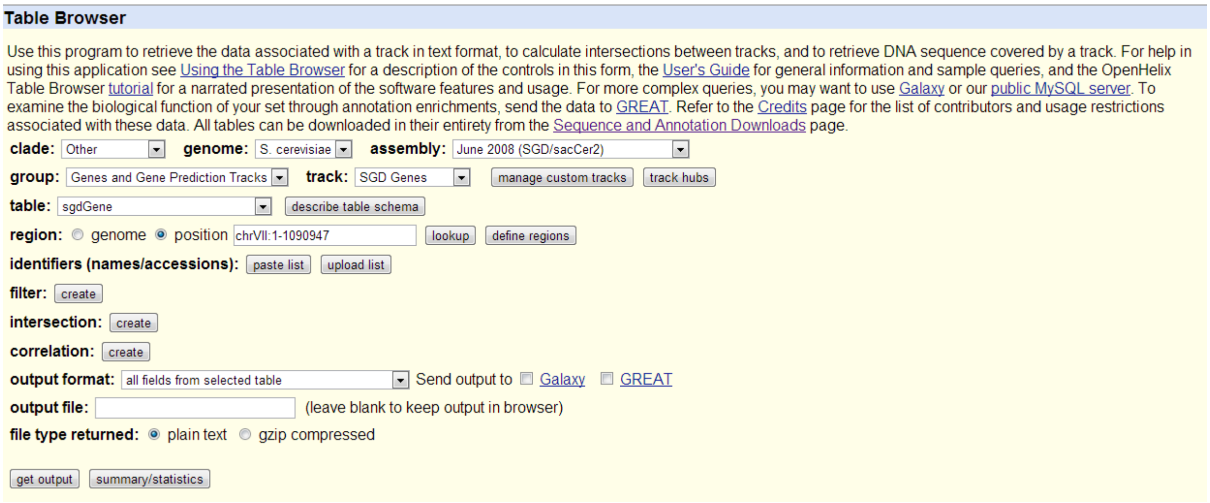

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Other **genome:** S. cerevisiae **assembly:** June 2008 (SGD/sacCer2)

group: Genes and Gene Prediction Tracks **track:** SGD Genes

table: sgdGene

region: ○ genome ● position chrVII:1-1090947

identifiers (names/accessions):

filter:

intersection:

correlation:

output format: all fields from selected table Send output to □ Galaxy □ GREAT

output file: (leave blank to keep output in browser)

file type returned: ● plain text ○ gzip compressed

Figure 3: UCSC table browser snapshot

4 Reproducing figures in R or Python

1. Try to reproduce figure 1b in R or Python. You will need to load the appropriate ".wig" files and to extract the region corresponding to the RPL30 gene. You can find a text file on moodle named `chrVII_UCSC.txt`, which contains annotation information of all genes on chromosome VII.

   You can find two solutions for this exercise in R. One is more complicated than the other.

2. Explain the significance of figure 3b and try to reproduce it.

   In figure 2d, the authors show that there is a strong correlation between antisense transcription and histone H4 acetylation. So, to test if there is a causal role of this acetylation, they perform NET-seq on a yeast strain that has no RCO1 (a required subunit of the Rpd3S histone 4 deacetylation complex). And indeed, they observe a higher antisense/sense ratio in the mutant when compared to the control (figure 3b).

   The R solution is in the same file as that of the previous exercise.