Lab 1: Simple genomic data analysis using R

The main purpose of this lab is to get student familiar with R through analyzing simple genomic data. Before the lab, students should have R installed.

1. UCSC genome browser

Go to UCSC genome browser webpage at http://genome.ucsc.edu/. Click "Genomes" at top left corner. This will bring you to the Genome Browser Gateway. From here you can select genomes for a number of species, the default species is human. Now from the "Human Assembly" pull down menu, select "Mar. 2006 (NCBI36/hg18)". Some information for this assembly will be displayed. Click the "view sequences" button next to "Human Genome Browser – hg18 assembly". Then go to the bottom of the page and click "Summary Statistics" to go to the statistics page for human genome hg18. Briefly go through the statistics and answer following questions based on the statistics:

1. How many chromosomes are there in human genome?
2. What's the longest chromosome and what's its length?
3. What's the total length of human genome for assembled size and sequenced size?

Go back to the previous page, and put "nanog" in the search box at the top of the page. You will get the search results for Nanog gene. Answer following questions based on the first search result under UCSC Genes:

1. What's the short description of Nanog gene?
2. What's the genomic location of the gene, e.g., chromosome, start and end?

Now click the first search result and visualize the gene on the genome browser. Zoom in and out to see nearby genomic features. Now click on the gene on the browser to go to the gene description page. Briefly go through the information there and answer:

1. What's the strand direction, coding region size and exon count of the gene?
2. What's the RefSeq Accession for the gene? Hint: under "Other Names for This Gene".

2. Download and analyze hg18 refseq genes

Go back to the home page of UCSC genome browser. Select "Table Browser" under the "Tools" menu. In the table browser page, select: "Mammal" under clade, "Human" under genome, "Mar. 2006 (NCBI36/hg18)" assembly, "Genes and Gene Prediction Tracks" under group, "RefSeq Genes" under track, "refGene" under table. Then select "genome" under region, which means you want to get data for the whole genome. Note that you can specify chromosome and location to get part of the data. Go down a little bit to select "all fields from selected table". The specify output file name in the textbox by "hg18genes.txt", and select "file type returned" as "plain text", the click "get output"

button. This will take a little time. The downloaded text file is a list of human hg18 refseq genes.

Open the file to take a quick look. Each row is for a gene. Columns are for properties of the genes. For example, “name” gives the refseq gene name (accession number); “chrom” is the chromosome; “strand” is the strand direction (+/-); txStart/txEnd are the transcriptional start/end position on the chromosome, etc.

Now obtain the R code from class website and perform simple exploratory analysis of the human genes. These are typical analysis one want to perform when getting a new genome. Write a short report to summarize the findings from these studies. Your report needs to address following points:

1. Number of genes: total number of genes, genes on different chromosome and strands.
2. Gene length. What’s the distribution of gene length? Are gene lengths different on + and - strands?
3. Short description on longest and shortest genes.
4. Exon counts. Are number of exons for genes different on on + and - strands?
5. Are exon counts and gene length correlated?
6. Gene density. What are the average gene numbers per mega bps on each chromosome? Which chromosome is the most “gene dense”?

Extra R codes are needed in order to answer some of the questions.