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 home page (with hg18 selected as the assembly), and put "nanog" in the search box at the top of the page. You will get the search results for Nanog gene. There are multiple matches, from different gene annotations (such as UCSC Genes, RefSeq Genes, etc.). Let's focus on the **first search result under UCSC Genes.** 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 (the top track) to go to the gene description page. Briefly go through the information there and answer following questions:

1. What's the (short) description of Nanog gene?
2. What's the RefSeq summary and RefSeq Accession number (the NM_ number) for the gene?
3. Provide following information for the coding region:
    - What's the genomic location of the gene, e.g., chromosome, start and end?
    - What's the strand direction, coding region size and exon count of the 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)" under **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 for human hg18 assembly. 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" for **output format**. Provide an **output filename** in the textbox as "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" column gives the RefSeq gene name (accession number); "chrom" is the chromosome number; "strand" is the strand direction (+/-); txStart/txEnd are the transcriptional start/end position on the chromosome, etc.

Now perform some simple exploratory analysis of the human genes from hg18 assembly. A template R code is provided in the class website. Write a short report to summarize the findings from these studies. Use tables and figures if necessary. Your report needs to address following points:

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