Unix, Perl, and Python

Session 1: Introduction to Unix and LSF

Exercise 4

Goal: Manipulate genomic features with bedtools (http://code.google.com/p/bedtools/).

Given a file of genome coordinates describing regions bound by H3K36me3 in human embryonic carcinoma cells (as determined by ChIP-Seq) we want to identify RefSeq genes with exons that overlap these bound regions.

Note: Commands are in the `Courier` font and sometimes appear over multiple lines. Each command should be executed as a single line on tak.

1	The X Windows System provides a graphical user interface (GUI) for computers. Mac comes with the program X11, an X windows program. If you are using Mac, open terminal with Go->Utilities->Terminal, and login to tak with `ssh -Y username@tak` -Y option will allow you to use X windows. For the PC, we use two software products: Xming and PuTTY. Xming is an X Window Server, which can display tak GUI programs on your local desktop. PuTTY is a free SSH client, which allows you to connect to tak. X server (Xming) must be started before setting up a SSH connection with PuTTY. Refer to the website on how to connect your PC to tak: http://jura.wi.mit.edu/bio/education/docs/ssh-sftp.php
2	Gene structures can be described in gene transfer format (gtf, a variant of gff). See http://mblab.wustl.edu/GTF2.html for a description of the format. tak has a gtf-format file of all human Refseq genes at /nfs/genomes/human_gp_feb_09/gtf/hg19.refgene.gtf. Can you display the last 10 lines of the refgene gtf file? `tail /nfs/genomes/human_gp_feb_09/gtf/hg19.refgene.gtf`
3	Show the hidden characters of the last 10 lines of refgene.gtf? `tail /nfs/genomes/human_gp_feb_09/gtf/hg19.refgene.gtf | cat -A`
4	Extract gene names from the last 10 lines of refgene.gtf? `tail /nfs/genomes/human_gp_feb_09/gtf/hg19.refgene.gtf | cut -d " " -f2`
5	How many genes are described in refseq.gtf? `cut -d " " -f2 /nfs/genomes/human_gp_feb_09/gtf/hg19.refgene.gtf | sort -u | wc -l`

6	ChIP-seq Illumina reads (GSE25882) were mapped to the human genome (hg19) and H3K36me3-enriched regions were identified with Model-based Analysis of ChIP-Seq (MACS). Download the H3K36me3 enriched regions H3K36me3_peaks.bed from our website. `wget http://jura.wi.mit.edu/bio/education/hot_topics/Unix_Perl_Python/ex/H3K36me3_peaks.bed`
7	Find out what RefSeq genes overlap these H3K36me3-enriched regions with the *intersectBed* command from bedtools. Please only show the top 10 lines. `intersectBed -a /nfs/genomes/human_gp_feb_09/gtf/hg19.refgene.gtf -b H3K36me3_peaks.bed|head`
8	How many RefSeq genes have exons that overlap these H3K36me3-enriched regions? `intersectBed -a /nfs/genomes/human_gp_feb_09/gtf/hg19.refgene.gtf -b H3K36me3_peaks.bed|cut -d " " -f2 |sort -u|wc -l`