Lab06

Introduction

Comparative genomics

We are going to use some of the algorithms used to generate genomic alignments for bacteria chromosomes and use vertebrate pre-computed genome alignments in different browsers:

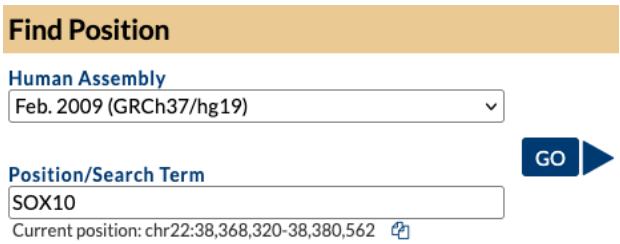
- UCSC Genome Browser http://genome.ucsc.edu
- VISTA: a comprehensive suite of programs and databases for comparative analysis of genomic sequences https://genome.lbl.gov/vista/index.shtml

Document all work during the dry lab for each exercise, defining all your tools and input parameters, data output, and interpretation.

From Lab06, there are three Assignments to be submitted to myCourses:
Discussion 6.1
Discussion 6.2
Activity 6.1

View precomputed genome alignments.

The University of California Santa Cruz (UCSC) Genome Browser can be used to retrieve pairwise alignments between human and selected model organism, as well as a multiple genomic sequence alignment between several organisms.
Let's find out where the human *SOX10* gene is. We better use the Human reference genome 2009 (GRCh37/hg19) in UCSC, since is the latest version VISTA has. All good? OK, let's dive in:
In UCSC Genome Browser:
SOX10 - chr22:38368320-38380562



Discuss in class all tracks and keep only the ones needed (right click "hide"):

- Gene model
- 100 Vertebrate conservation

- Multiz alignments:
    - Chimp
    - Mouse
    - Rat
    - Cow (?)
- RepeatMasker
    - If you click on RepeatMasker it will show the different repeat classes.

The multiple alignment is built by applying a method called MULTIZ to the individual BLASTZ alignments. During the generation of the multiple genomic sequence alignment, conservation scores were calculated using a phylogenetic **Hidden Markov Model** (HMM – not covered in Lecture06, but will be in **Lecture10**).

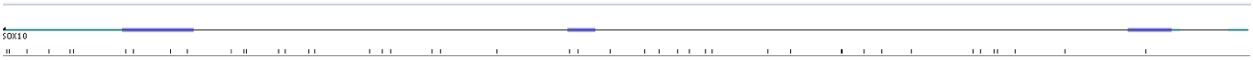
Take a screenshot and save it. For the **Discussion activity**, describe all the components and tracks and what they represent using your screenshot as a reference.

Tip: instead of a screenshot, right-click on the viewer and "View image" to open the image in a different window and save it as a PNG image.

Tip 2: Right-click on each track and select "configure" to see more options and descriptions of what they represent.
What are the greatest overall conservation regions?
Do you notice something odd in the Chimp ***Multiz track*** ?

If you want to know more about the Mutiz track: https://genome.ucsc.edu/cgi-bin/hgTrackUi?db=mm9&g=multiz30way

Now, open a new internet browser tab with VISTA precomputed alignments (Vista-Point).
Disclaimer: The VISTA Browser does not work on my Mac.
Navigate to the SOX10 location (see above or copy the region from UCSC) using the Human Feb. 2009 genome.
Once the human SOX10 is depicted on top as an arrow…What are the colors for each region?



Let's add some genome alignments. The process is a little different than the UCSC browser. But try to get them in the same order as you did in the UCSC genome browser screenshot:

- Chimp
- Mouse
- Rat
- Cow (?)

The graphs (tracks) represent the alignments generated by LAGAN or Multi-LAGAN. Each graph indicates **percent conservation across the region; the value itself is an average identity score using a variable-size window, and the scale goes from 0 to 100%.** The horizontal line represents 70% of Conservation Identity, while the minimum is 50% (such parameters can be adjusted).

Various parts of the curve have been colored red, aquamarine, and purple. What do they represent?
If you click on "alignment in pdf" in one of the genome alignment panels, it will show a more detailed description of all the elements and colors between eg. human-Chimp.

Always try to adjust scaling and thresholds and take a screenshot that will "align" with UCSC screenshot coordinates.

Finally, try overlaying the information displayed within the VISTA and UCSC browsers, providing a unified view of all the alignments.

Tip: if you make the VISTA browser work, there is a function to get both browsers in one nice picture.

Discussion 6.1
Superimposed both browser's screenshots (UCSC and VISTA). Write down the steps and parameters (methods) you used, describing all the components and tracks and what they represent. (Once you finish, submit the report with explanations to myCourses (in Assignments).

When you move to the next activity in a new window, keep the VISTA-point browser open (do not close) with your results.

Align and compare two large-scale sequences (mVISTA)

mVISTA allows for two or more genomes to be aligned in a straightforward fashion, using the AVID or LAGAN algorithms. We will give the genome inputs as fasta format and an optional gene annotation file (gff).
Here are some runtimes for the alignment of whole-genome assemblies (Frazer *et al.*, 2004). Of course, due to time constraints, we are using bacteria genomes for comparisons. Right now, the VISTA server seems to run fast.

Microbial genome:	Size	Microbial genome:	Size	Runtime
Bartonella bacilliformis KC583	1.4Mb	Bartonella Quintana st r. Toulouse	1.6Mb	7 min
Bartonella bacilliformis KC583	1.4Mb	Bartonella henselae str. Houston-1	1.9Mb	8 min
Bartonella henselae str. Houston-1	1.9Mb	Bartonella - tribocorum CIP 105476 plasmidBtr	2.6Mb	31 min

Yeast genome: Candida tenuis	10.7Mb	**Yeast genome:** Spathaspora assalidarum	13.3Mb	1 h. 52 min
Fungal genome: Batrachochytrium dendrobatidis JAM81	30Mb	**Fungal genome:** Piromyces sp. E2	71Mb	7 h. 13 min
Fungal genome: Batrachochytrium dendrobatidis JAM81	30Mb	**Fungal genome:** Batrachochytrium dendrobatidis JEL423	23.7Mb	7 h. 30 min

Find two strains in *Bartonella* using a standard NCBI search. Download both fasta assembly file and GFF annotation.
Eg.
Search in the NCBI Genome portal: https://www.ncbi.nlm.nih.gov/datasets/genome/

Bartonella bacilliformis KC583
Bartonella henselae strain Houston-1

Click on Download and select Genome sequences (FASTA) and Annotation Features (GFF). Save each entry and navigate through the files. Once you have all files needs go to mVISTA and select two sequences.

Select your input FASTA sequences (Inquiry).
Under additional options, upload the annotation file (GFF). You can rename the sequences; instead of sequence#2, use something like “Bhenselae”.

Submit and…. wait. Within a few minutes, you will receive a link with the results in your mail.

While we wait, let’s go back to the VISTA-point screen and check the “Synteny” window from the previous exercise.

Once VISTApoint is launched with the link sent to your email, explore the genomes alignment. Can you spot highly conserved regions? Can you spot regions poorly conserved? Take a couple of screenshots (conserved vs. low conservation) and comment on what regions (genes / intergenic regions) you see.

Some additional questions:
-In bacteria genomes most of the space is composed by_____.
-What is the genome length for each strain?
-%GC content in each strain?

Discussion 6.2
Superimposed browser’s screenshots (conserver vs. no conserved). Write down the steps and parameters (methods) you used, describing all the components and tracks and what

they represent. (Once you finish, submit the report with explanations to myCourses (in Assignments).

Another precomputed genome alignment.

Compare a discrete human genome segment by aligning homologs from other vertebrate species. Find the human gene pyruvate kinase (PKLR) and the corresponding PKLR homologs from chimp, gorilla, mouse, rat, and cow using precompute genome alignments. Find the regions of high homology across the 12-Kilobases region of the human PKLR gene. What do you notice in the alignments? What regions are being conserved across the taxa? What regions are losing homology?
Since they are loaded only a specific number of precomputed alignments for the human genome, could you find homologs in other species (chicken, zebrafish) with other precomputed alignments? You will have to "jump" from one alignment to another.
Tip: check the mouse genome.
Build another panel with similar boundaries for PKRL with chicken and zebrafish genomes, and superimpose it to the previous panel with human homologs.

Activity 6.1
Once you finish, submit the report to myCourses (in Assignments).