README FILE WITH INSTRUCTIONS FOR USE

This file provides instructions for running the CTCF program locally on your computer.

Developer: Mohsin A.F Khan (ucbtmaf@ucl.ac.uk). For any questions/queries, please send me an email.

VERSION 1.1 UPDATED ON 08/01/2013

CONTENTS

1. MINIMUM SYSTEM REQUIREMENTS/SOFTWARE
2. OBTAINING DNA REGION(S) FOR CTCF ANALYSIS VIA UCSC GENOME BROWSER
3. CHANGING PROPERTIES IN PERL FILES BEFORE RUNNING
4. RUNNING THE PERL FILES
5. SYNTENY ANALYSIS
6. MINIMUM SYSTEM REQUIREMENTS/SOFTWARE

¥ Any version of Windows or UNIX with a Perl installation

¥ Before running the CTCF program, it is required for users to have Perl installed on their systems. If you do not have it installed, please visit <http://www.activestate.com/activeperl/downloads> and click “Download ActivePerl [version number] for Windows (x86)” located on the top right of the page.

¥ By default, Perl should be installed on your C: drive. After installation, to check that you have successfully installed it, open your command prompt (Start 🡪 all programs 🡪 accessories 🡪 command prompt) and then type **cd C:\Perl** in the command window

¥ If you are able to successfully change directory to above, then your Perl installation was successfully



1. OBTAINING DNA REGION(S) FOR CTCF ANALYSIS VIA UCSC GENOME BROWSER

Before performing CTCF analysis, you will first need to obtain the DNA sequence of your region of interest. To do this, follow the instructions below:

1. Go to <http://genome.ucsc.edu/> and click “Genome Browser” located on the left side of the page. That should take you to the page below:



1. In the “search term”, enter the name of the gene you are interested in. i.e. Sox2 (if you get a list of genes from your search, find the one you are interested in and click it), and choose a genome. You should get a page like the one below:



1. Because the purpose here is to find putative CTCF sites harbouring your gene of interest, you would want to cover a wide genomic location. It is recommended that a region of ~4 megabases is covered (~2 megabases up and ~2 downstream of your gene of interest). So in the above example, the genomic coordinates for the Sox2 gene are chr9:17,990,091-17,991,429. To cover a 4 megabase distance with the Sox2 gene approximately in the centre, the genomic coordinates would be chr9:15,990,091-19,991,429 (~ 4 mb). The next step is to then extract the DNA sequence for this 4 mb region.
2. To do this, click anywhere on the conservation profile
3. This will open a page like the one below:



1. Next to the species of choice, click on the blue hyperlink labelled “D”. In this example because the coordinates for the 4 mb sox2 region are for chick, the “D” hyperlink next to “chicken” would be clicked.
2. This will open up another page as shown below:



1. Paste the coordinates for the 4 mb region determined in step c into the text box labelled “Position”. Click “get DNA”. This will generate the DNA sequence for the region.

NOTE: BEFORE PROCEEDING, YOU WILL NEED TO COPY THE CONTENTS OF THE “CODE” FOLDER FROM THE SUPPLEMENTARY INFORMATION INTO YOUR C:/PERL DIRECTORY

1. Paste the generated DNA sequence into the text file called “forward.txt”, which should now be located in your C:/Perl directory. NOTE: The forward.txt file by default contains an example dataset containing a region of human chromosome 3, where the *Sox2* putative insulator region can be identified as explained in the paper. To use your own DNA sequence, replace the contents of this file with your sequence of interest.
2. Because CTCF analysis applies to both strands, you will also need to generate the reverse complement of the 4 mb region. To do this, repeat step f and when performing step g, click the “Reverse complement (get ‘-‘ strand sequence)” check box. Then click “get DNA”. This will generate the reverse complement sequence. Copy it, and paste it into the file called “reverse.txt”, which should be located in your C:/Perl directory. NOTE: Again, replace the contents of this file with the reverse complement DNA sequence of interest.
3. CHANGING PROPERTIES IN PERL FILES BEFORE RUNNING
4. In your C:/Perl directory, you should be able to see a Perl file called “CTCF-forward.txt”, and another called “CTCF-reverse.txt”. Because Perl files have the extension .pl, you will need to rename these files to “CTCF-forward.pl” and “CTCF-reverse.pl” respectively
5. Right click the “CTCF-forward” file and then click “edit”. This should open the code in notepad. Go to the end of the code , where you will see the following lines of code:



Change this to your starting coordinate of the 4 mb region.

Change this to the chromosome number harbouring your gene of interest. If your gene of interest is located on chromosome 9, then change it to chr9

Change this to the starting coordinate of your 4 mb region + 18 (add 18)

1. Save the CTCF-forward.pl file after making the above changes and then repeat the process with “CTCF-reverse.pl”. The end of the CTCF-reverse.pl file, you should see the following code:



Change this to the ending coordinate of your 4 mb region – 18 (subtract 18)

Change this to the ending coordinate of your 4 mb region.

Change this to the chromosome number harbouring your gene of interest. If your gene of interest is located on chromosome 9, then change it to chr9

IV-RUNNING THE PERL FILES

Open your command prompt and go to C:/Perl

Type **perl CTCF-forward.pl** into the command prompt. The program will execute and once the analysis is complete, your results can be accessed in an output file called “output-forward.txt”

Repeat the above with **perl CTCF-reverse.pl**. Results for the reverse strand can be accessed in an output file called “output-reverse.txt”

You can now use these results to look for constitutive CTCF sites by turning on relevant tracks in the UCSC genome browser (to display CTCF ChIP-seq data) and use “Custom Tracks” to display your results into the genome browser. To carry out synteny analysis, you will need to repeat all of the above with equivalent regions in other species, and then look for syntenic sites in the UCSC genome browser

V-SYNTENY ANALYSIS

To perform synteny analysis, you will need to repeat the above analysis for equivalent regions in other species of interest. Once you have this done, you will have CTCF predicted results from all species. The next step is to look at synteny across these species to see whether the same set of genes in different genomes are insulated by CTCF predicted sites. To do this, you will need to feed your results from “output-forward.txt” and “output-reverse.txt” as a track into the UCSC genome browser for a particular species. NOTE: The coordinates in these output files need to be in BED format (e.g. chr3 50096954 50096972) , so you will first need to convert them. To convert, you can simply open the output files in notepad, then go to “edit” 🡪 “replace”. In the “find what” box, type “:” and in the “replace” box, type a space. Then click on “replace all”. Repeat this step, this time replacing “-“ with a space. This will convert the format to BED, which will allow you to feed this data as a track in the UCSC genome browser.

Next, on the UCSC genome page, you should see an option called “add custom tracks”. Click this button.



Then paste your coordinates in BED format into the box shown below:



Click the “Submit” button.

You can now view the distribution of CTCF sites across the genome in the UCSC genome browser by clicking on “go to genome browser”, shown below:



In the page that will open, there are options to zoom in and out so that you can change your resolution to what you like. When you repeat this analysis with a region of interest in different genomes, you will be able to look at synteny by detecting CTCF sites that harbour the same genes across multiple species.

NOTE: You can also perform constitutive CTCF analysis with publically available ChIP-seq data for several cell-lines as described in the paper.