

Module 3: Lab Practical

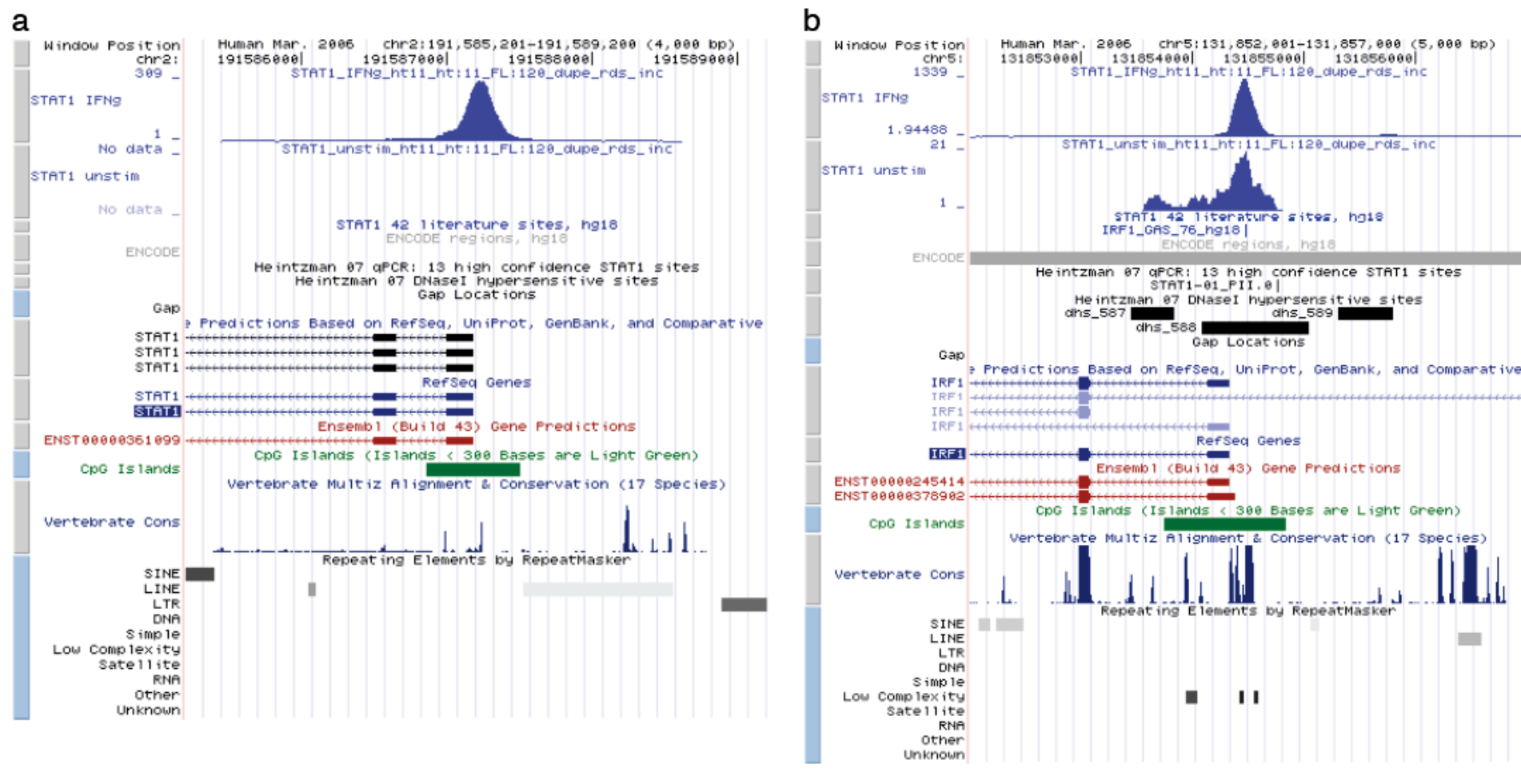
Asim Siddiqui

Estimated completion time: 1-2 hours

Extra credit: an additional hour

Objectives

- The practical will lead you through an analysis of ChIP-seq data.
- Since it is not possible to complete a whole genome analysis in the allotted time, we will focus on aligning reads to a small region of chromosome 1.



Supplementary Fig. 1 FDR-thresholded XSET peaks. Stimulated and unstimulated ChIP-seq peaks for: (a) a 5 kb region around the 5' end of the IRF1 gene, and (b) a 4 kb region around the 5' end of the STAT1 gene. Peaks overlapped CpG islands and sequence conservation peaks. The IRF1 region had both a stimulated and an unstimulated peak (heights 1339 and 21, respectively). These peaks overlapped HeLa DNaseI hypersensitive sites, and a site known to be a functional GAS sequence (Supplemental XLS) that was ~275 bp from a QPCR-validated STAT1 site.¹ The displays are from the UCSC hg18 genome browser² (<http://genome.ucsc.edu>).

Regions of interest

- Using the UCSC genome browser, extract
 - approx. 5kb of sequence around the 5' end of the IRF1 gene
 - Approx. 7kb of sequence around the 5' end of the STAT1 gene

Illumina data

- Extract the sequences from the stimulated files and create a fasta file:
e.g. for file 1
 - `cat stimulated_1_seq.txt | awk ' { print $5 } ' | awk ' { print ">" } { print $0 } ' > output1.fasta`
 - Note the created file is missing the first sequence from the file
- Use blat (with default parameters) to find the location of query sequences against each of the target regions
 - Start with stimulated 1, 2, 3 – will together take ~ 1 minute of CPU time
 - 4-8 will together take ~ 20 minutes to run
 - You can time the execution by prepending the blat command with “time”
- While waiting for 4-8 to complete, review the output from 1-3 and determine the start and end location of each hit
 - `cat output.psl | awk ' { print $15 } ' > start`
 - `cat output.psl | awk ' { print $16 } ' > end`

Plot

- Plot a histogram of the data using a suitable plotting program e.g. R
 - R> start = scan("start")
 - R> hist(start, breaks=300)

Review

- Does the data look like the graphs from the paper?
- Why might there be differences?
- How did the authors perform the alignment?
- How might a different aligner perform?
- What pitfalls do we need to be aware of when focussing on a small regions of the genome in this manner?

Extra credit

- Use the find peaks program on the data.
- How does find peaks work and how might it be improved upon?
- Run files 1,2,3 on a 10kb dna sequence and 20b genome sequences. Using these timings together with the original estimate how long it will take Blat to run the entire dataset on the entire genome