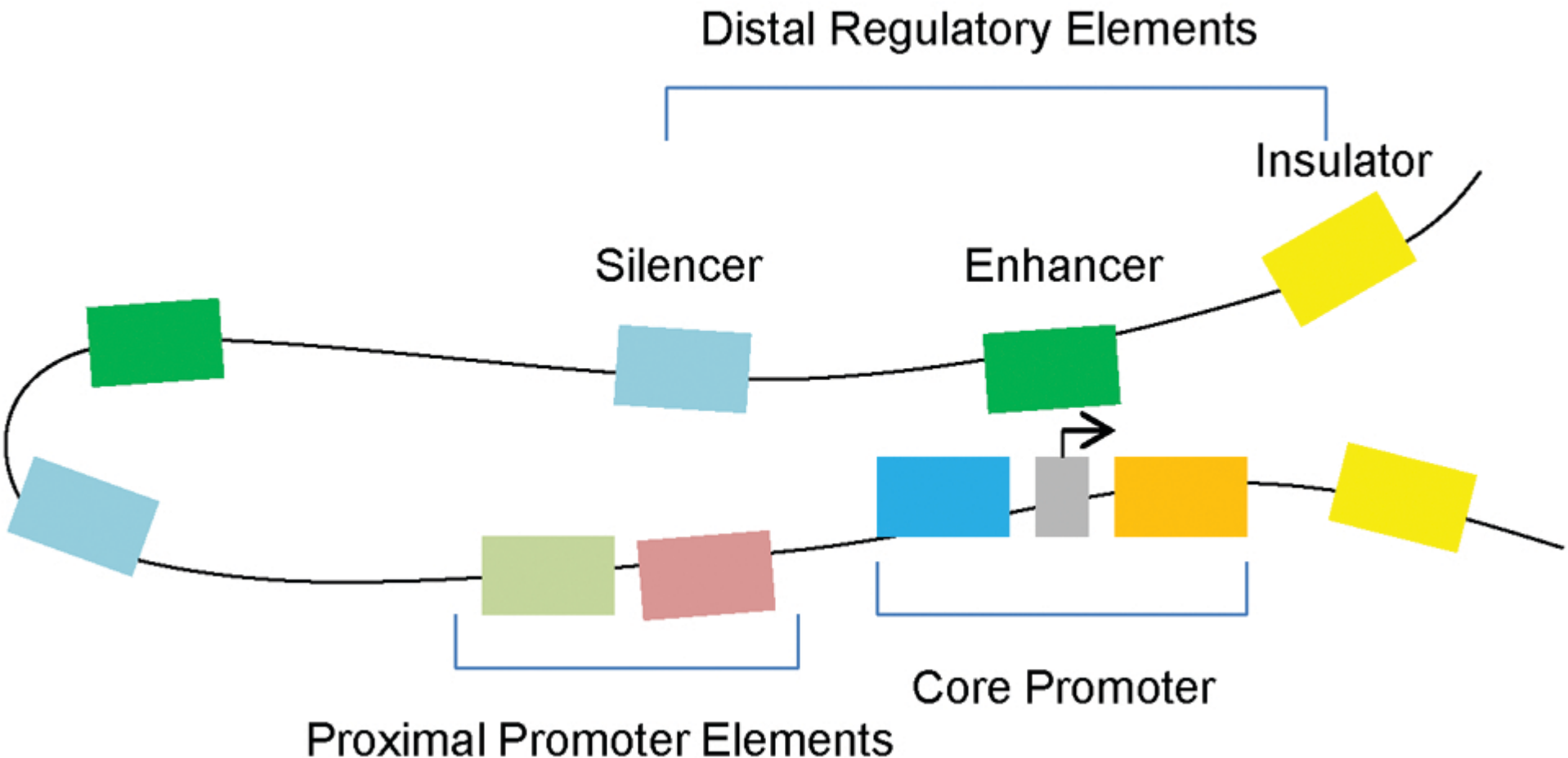
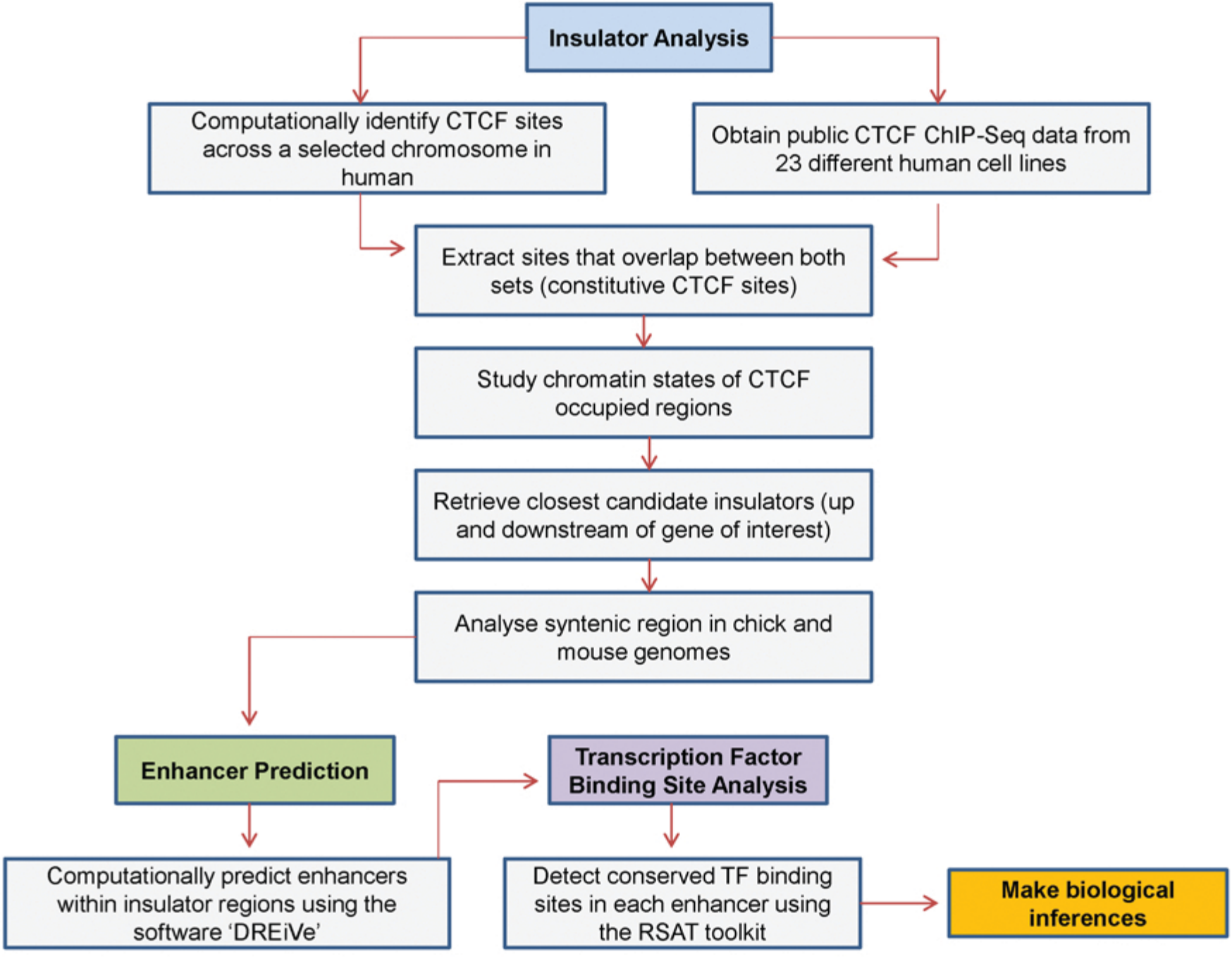
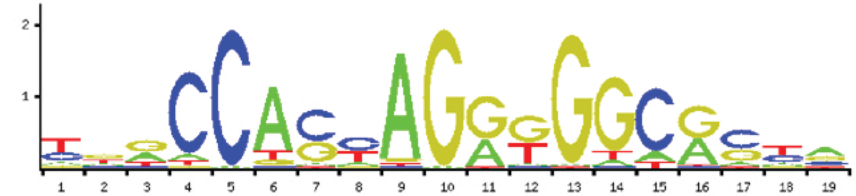

Distal Regulatory Elements
Insulator
Silencer
Enhancer
Core Promoter
Proximal Promoter Elements


Insulator Analysis
Computationally identify CTCF sites across a selected chromosome in human
Obtain public CTCF ChIP-Seq data from 23 different human cell lines
Extract sites that overlap between both sets (constitutive CTCF sites)
Study chromatin states of CTCF occupied regions
Retrieve closest candidate insulators (up and downstream of gene of interest)
Analyse syntenic region in chick and mouse genomes
Enhancer Prediction
Transcription Factor Binding Site Analysis
Computationally predict enhancers within insulator regions using the software 'DREiVe'
Detect conserved TF binding sites in each enhancer using the RSAT toolkit
Make biological inferences

A

Obtain CTCF Position Frequency Matrix



Make a SVG logo ?

FREQUENCY MATRIX ?

```
A  [ 87 167 281  56   8 744  40 107 851   5 333  54  12  56 104 372  82 117 402 ]
C  [291 145  49 800 903  13 528 433  11   0   3  12   0   8 733  13 482 322 181 ]
G  [ 76 414 449  21   0  65 334  48  32 903 566 504 890 775   5 507 307  73 266 ]
T  [459 187 134  36   2  91  11 324  18   3   9 341   8  71  67  17  37 396  59 ]
```

Reverse complement ?

B

Shuffle chromosome, scan against CTCF PFM and calculate 'weighted' scores

GCCGTCTGCTCCTCGCTGTGTGCCACAGAGAGGCTGGGGCC..nth base

↓ **Shuffle**

GGCTTTTGCAATTAGCCGCCGCGGGGCGGGGCCCCCTCTTC..nth base

↓ **Calculate weight**

$$w = \log2 ((n + \sqrt{N} * b) / (N + \sqrt{N})) / b)$$

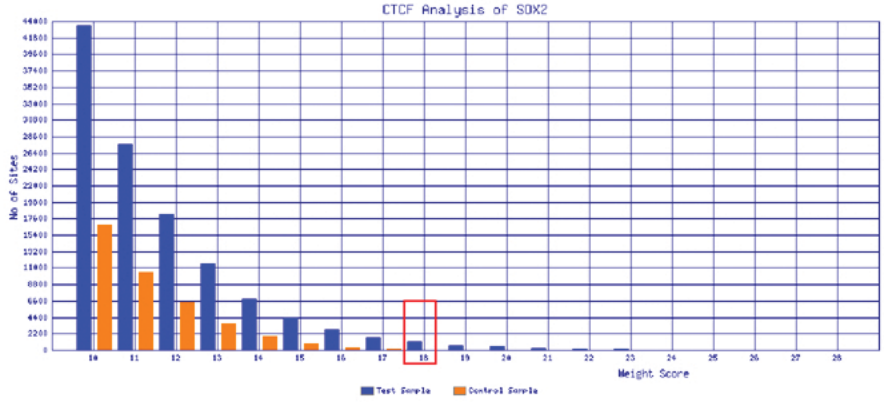
w, weight of current nucleotide
n, no. of occurences of current nucleotide in current column
(e.g. "87" for A in column 1, "414" for G in column 2 etc
N, total no. of observations, sum of all nucleotides in a column
(913 in this case)
b, [background] frequency of current nucleotide

↓ **Report weighted scores and coordinates**

Start	End	Pattern	Score	Strand
1	19	GGCTTTTGCAATTAGCCGC	12.5	+

C

Scan actual chromosome against CTCF matrix, re-calculate 'weight' scores and compare distributions



D

Calculate False Discovery Rate (FDR) and P value

$$FDR = V / V + S$$

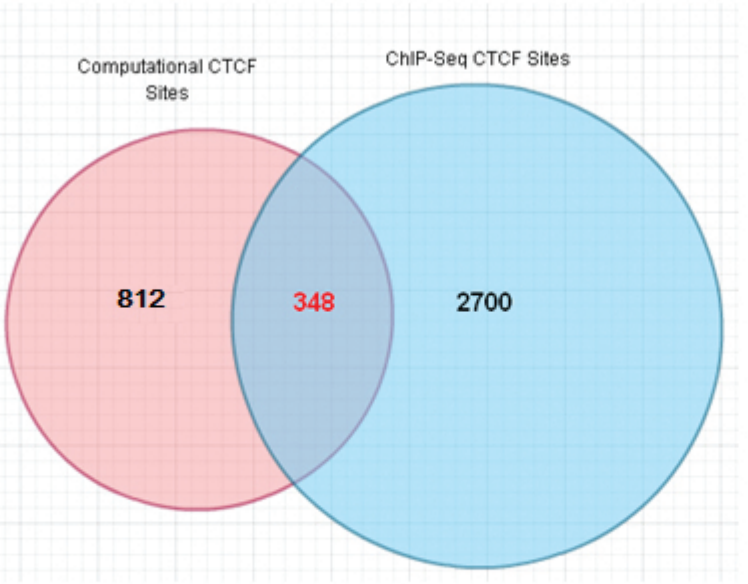
Where,
V = no. of sites in control sample (false positives)
S = no. of sites in test sample (true positives)

$$P = A / B$$

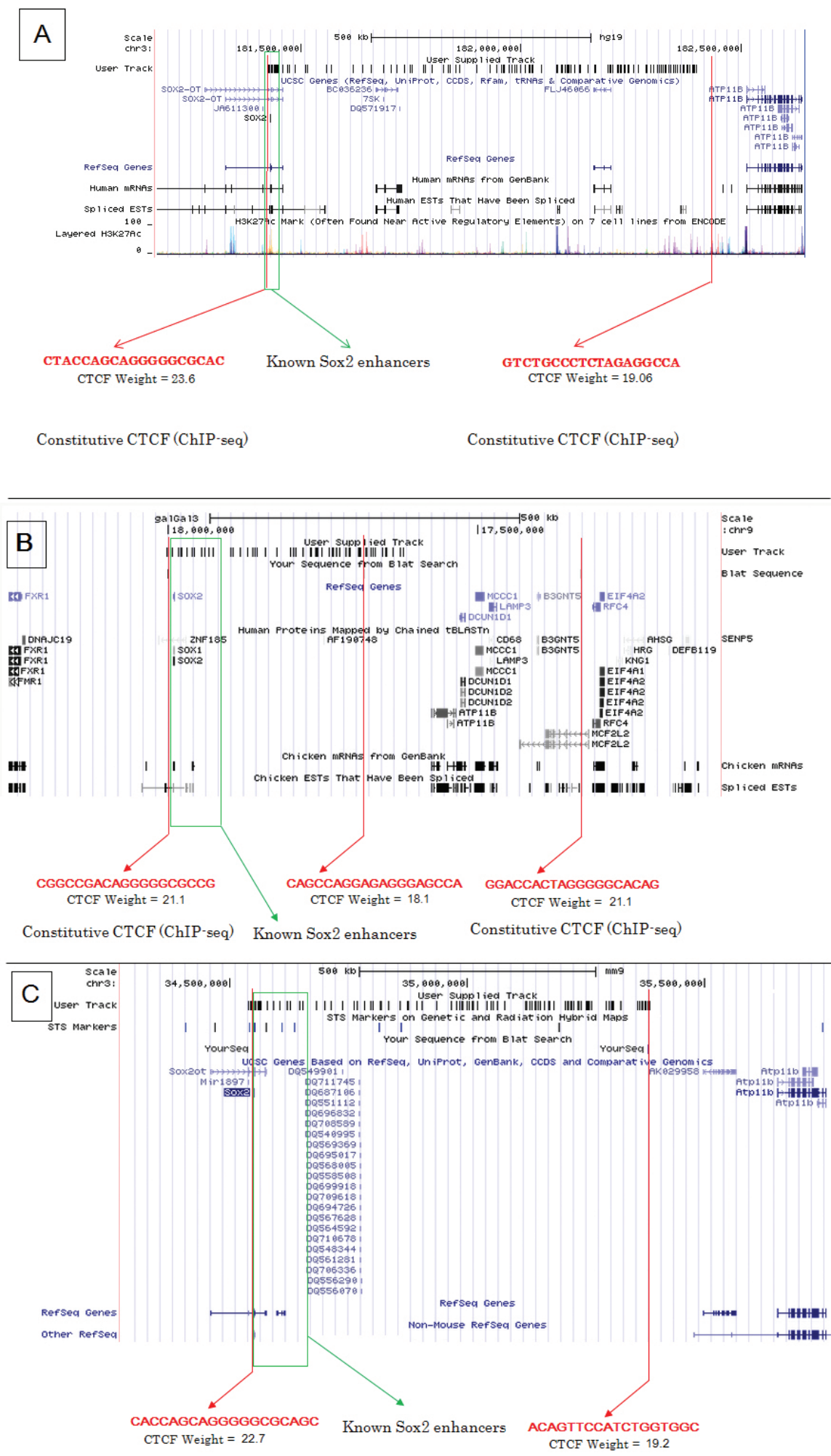
Where,
A = no. of sites of a given weight in control sample
B = total no. of sites in control sample

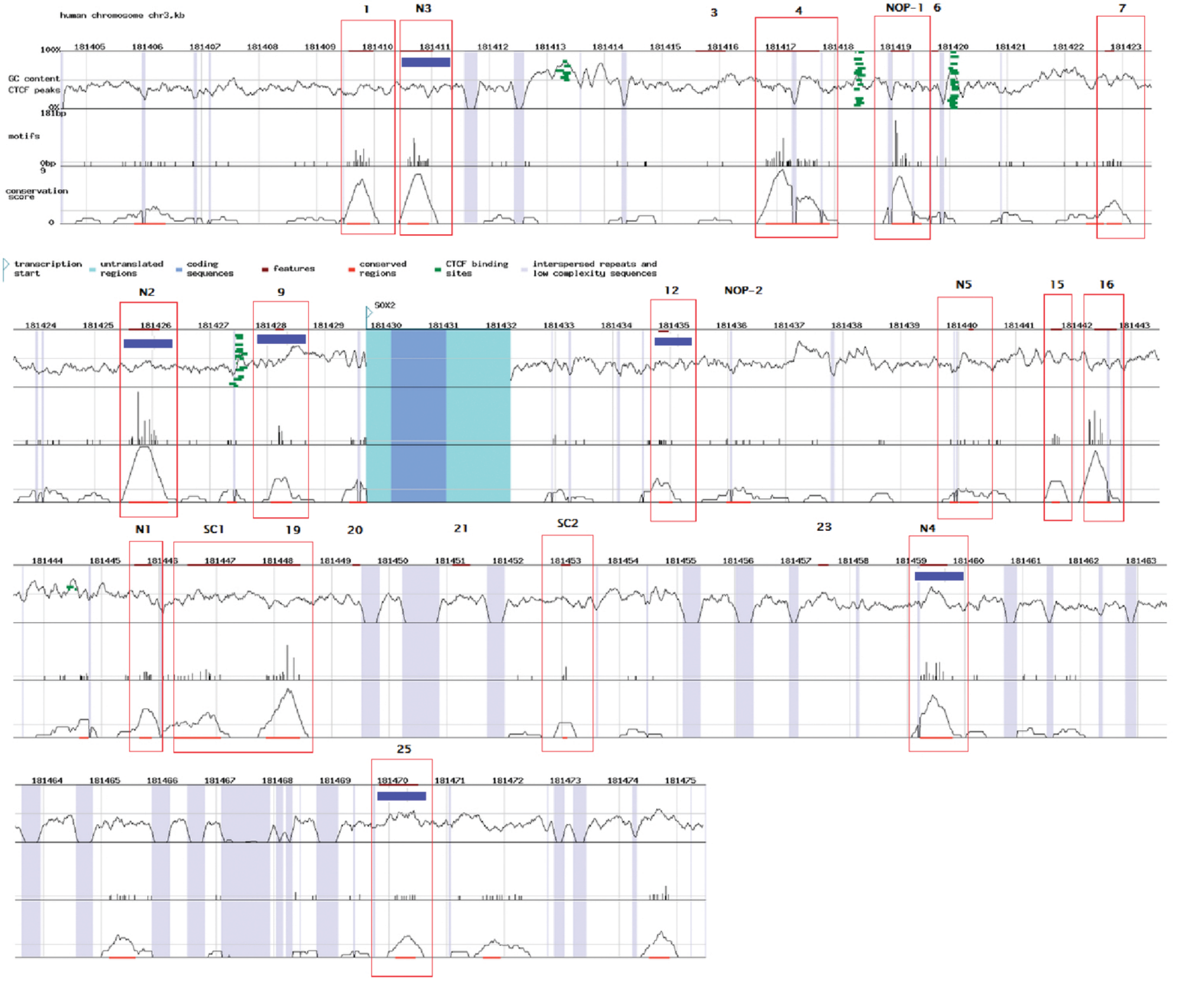
In above example, a weight score of "18" has an FDR and P value of 0 (optimal threshold). Sites of weight >= 18 likely to be real CTCF sites

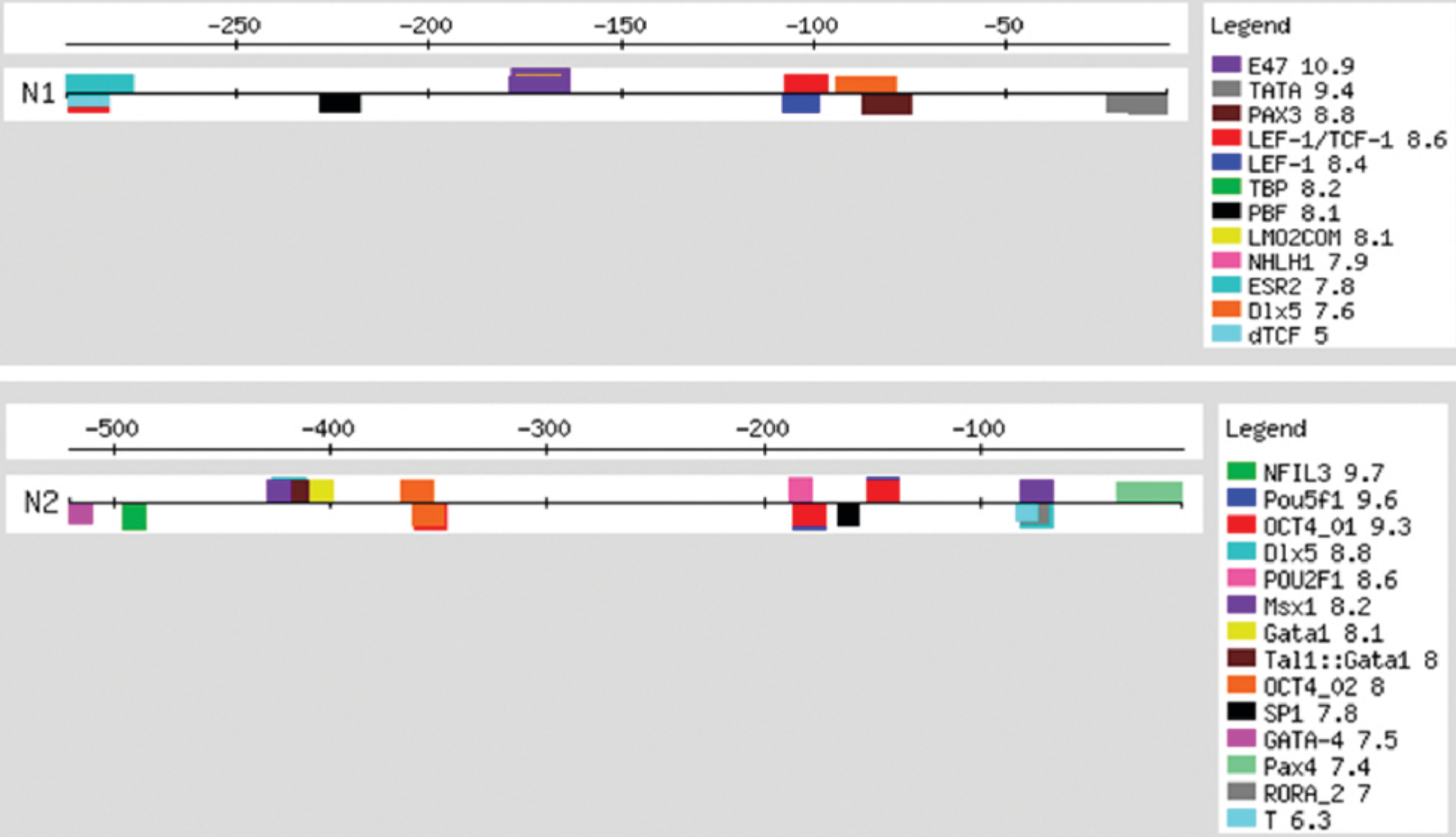
Weight Score	Count in Test Sample
>10	43447
>11	27510
>12	18127
>13	11611
>14	6891
>15	4362
>16	2792
>17	1749
>18	1160
>19	717
>20	473
>21	314
>22	181
>23	106
>24	56
> 25	29
>26	16
>27	3
>28	1



Human ChIP-Seq Cell Lines
retinoblastoma
small airway epithelial cells
epidermal keratinocytes
leukemia
umbilical vein endothelial cells
mammary epithelial cells
cervical carcinoma
embryonic kidney
B-lymphocyte X8
colorectal adenocarcinoma
skin fibroblast
renal epithelial cells
promyelocytic leukemia cells
hepatocellular carcinoma
embryonic stem cells
neuroblastoma


A
Scale
chr3:
500 kb
hg19
181,500,000
182,000,000
182,500,000
User Supplied Track
User Track
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
SOX2-OT
SOX2-OT
JA611300
SOX2
BC036236
7SK
DQ571917
FLJ46066
ATP11B
RefSeq Genes
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
H3K27ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
100
0
CTACCAGCAGGGGGCGCAC
CTCF Weight = 23.6
Known Sox2 enhancers
GTCTGCCCTCTAGAGGCCA
CTCF Weight = 19.06
Constitutive CTCF (ChIP-seq)
Constitutive CTCF (ChIP-seq)
B
galGal3
500 kb
18,000,000
17,500,000
Scale
chr9
User Supplied Track
User Track
Your Sequence from Blat Search
Blat Sequence
RefSeq Genes
FXR1
SOX2
MCCC1
B3GNT5
EIF4A2
LAMP3
RFC4
DCUN1D1
Human Proteins Mapped by Chained tBLASTn
DNAJC19
ZNF185
AF190748
CD68
AHSG
SENP5
SOX1
HRG
DEFB119
KNG1
FMR1
EIF4A1
DCUN1D2
ATP11B
MCF2L2
Chicken mRNAs from GenBank
Chicken ESTs That Have Been Spliced
Chicken mRNAs
Spliced ESTs
CGGCCGACAGGGGGCGCCG
CTCF Weight = 21.1
CAGCCAGGAGAGGGAGCCA
CTCF Weight = 18.1
GGACCACTAGGGGGCACAG
CTCF Weight = 21.1
Constitutive CTCF (ChIP-seq)
Known Sox2 enhancers
Constitutive CTCF (ChIP-seq)
C
Scale
chr3:
500 kb
mm9
34,500,000
35,000,000
35,500,000
User Supplied Track
User Track
STS Markers on Genetic and Radiation Hybrid Maps
STS Markers
Your Sequence from Blat Search
YourSeq
YourSeq
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
Sox2ot
DQ549901
AK029958
Atp11b
Mir1897
DQ711745
SOX2
DQ687106
DQ551112
DQ696832
DQ708589
DQ540995
DQ569369
DQ695017
DQ568005
DQ558508
DQ699918
DQ709618
DQ694726
DQ567628
DQ564592
DQ710678
DQ548344
DQ561281
DQ706336
DQ556290
DQ556070
RefSeq Genes
Non-Mouse RefSeq Genes
Other RefSeq
CACCAGCAGGGGGCGCAGC
CTCF Weight = 22.7
Known Sox2 enhancers
ACAGTTCCATCTGGTGGC
CTCF Weight = 19.2


human chromosome chr3,kb
1
N3
3
4
NOP-1
6
7
100%
GC content
CTCF peaks
0%
181bp
motifs
0bp
9
conservation score
0
transcription start
untranslated regions
coding sequences
features
conserved regions
CTCF binding sites
interspersed repeats and low complexity sequences
N2
9
SOX2
12
NOP-2
N5
15
16
N1
SC1
19
20
21
SC2
23
N4
25


-250
-200
-150
-100
-50
N1
Legend
E47 10.9
TATA 9.4
PAX3 8.8
LEF-1/TCF-1 8.6
LEF-1 8.4
TBP 8.2
PBF 8.1
LMO2COM 8.1
NHLH1 7.9
ESR2 7.8
Dlx5 7.6
dTCF 5
-500
-400
-300
-200
-100
N2
Legend
NFIL3 9.7
Pou5f1 9.6
OCT4_01 9.3
Dlx5 8.8
POU2F1 8.6
Msx1 8.2
Gata1 8.1
Tal1::Gata1 8
OCT4_02 8
SP1 7.8
GATA-4 7.5
Pax4 7.4
RORA_2 7
T 6.3