Puzzles in massively parallel genome biology

await GPU solutions

Oleksiy Karpenko
Bioinformatics Program
University of Illinois at Chicago

Super Computing 2011
Seattle, WA

SC11 Seattle, WA | Connecting communities through HPC

NSF

UIC UNIVERSITY OF ILLINOIS AT CHICAGO COLLEGE OF ENGINEERING

This presentation is funded by Pan-American Advanced Studies Institute of NSF

Human genome for computer scientists

…ACGTTGGATCGAGACATGACGATG…

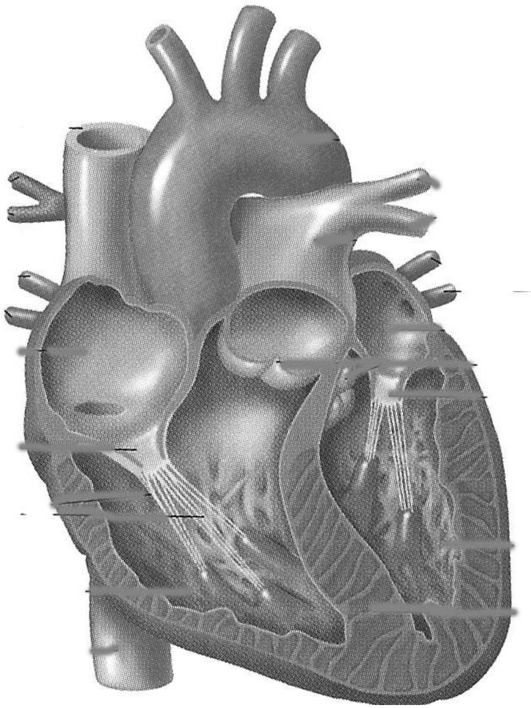
- 4 letter alphabet of DNA: {**A**, **C**, **G**, **T**}
- ~3 GB letters long

Human genome for computer scientists

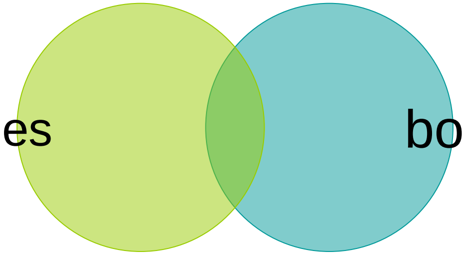
…ACGTTGGATCGAGACATGACGATG…

- 4 letter alphabet of DNA: {**A, C, G, T**}
- ~3 GB letters long
- 2% are genes coding for proteins
- 98% formerly known as “junk DNA”

DNA is the same
but active genes are different



≠



heart genes

bone genes

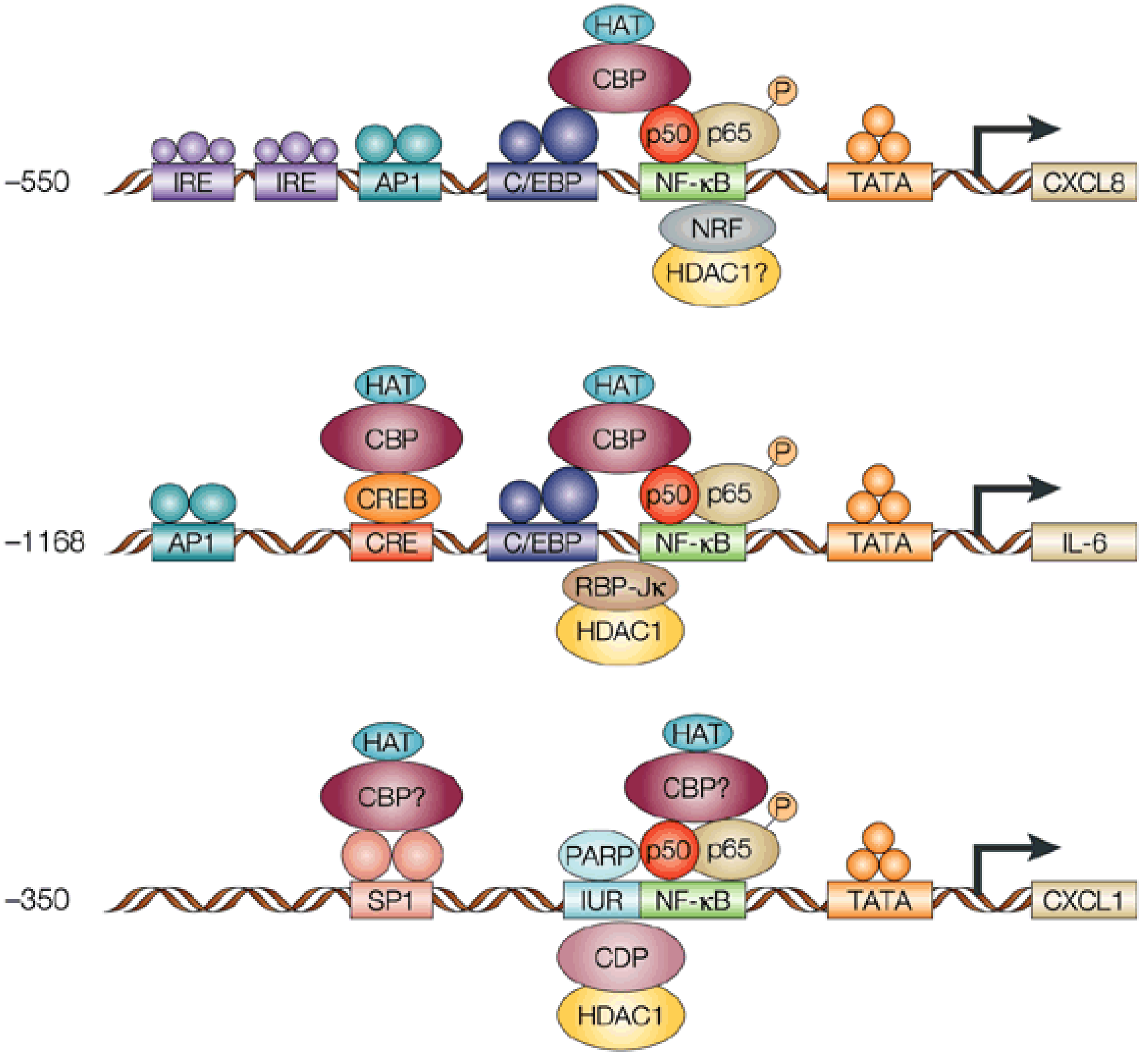
http://www.bio.aps.anl.gov/scihi/img/heart.jpg, http://www.sciencephoto.com/image/190555/large/F0014577-The_femur-SPL.jpg

Human genome for computer scientists

…ACGTTGGATCGAGACATGACGATG…

- 4 letter alphabet of DNA: {**A, C, G, T**}
- ~3 GB letters long
- 2% are genes coding for proteins
- 98% formerly known as “junk DNA”
- the non-coding regions host important regulatory sites

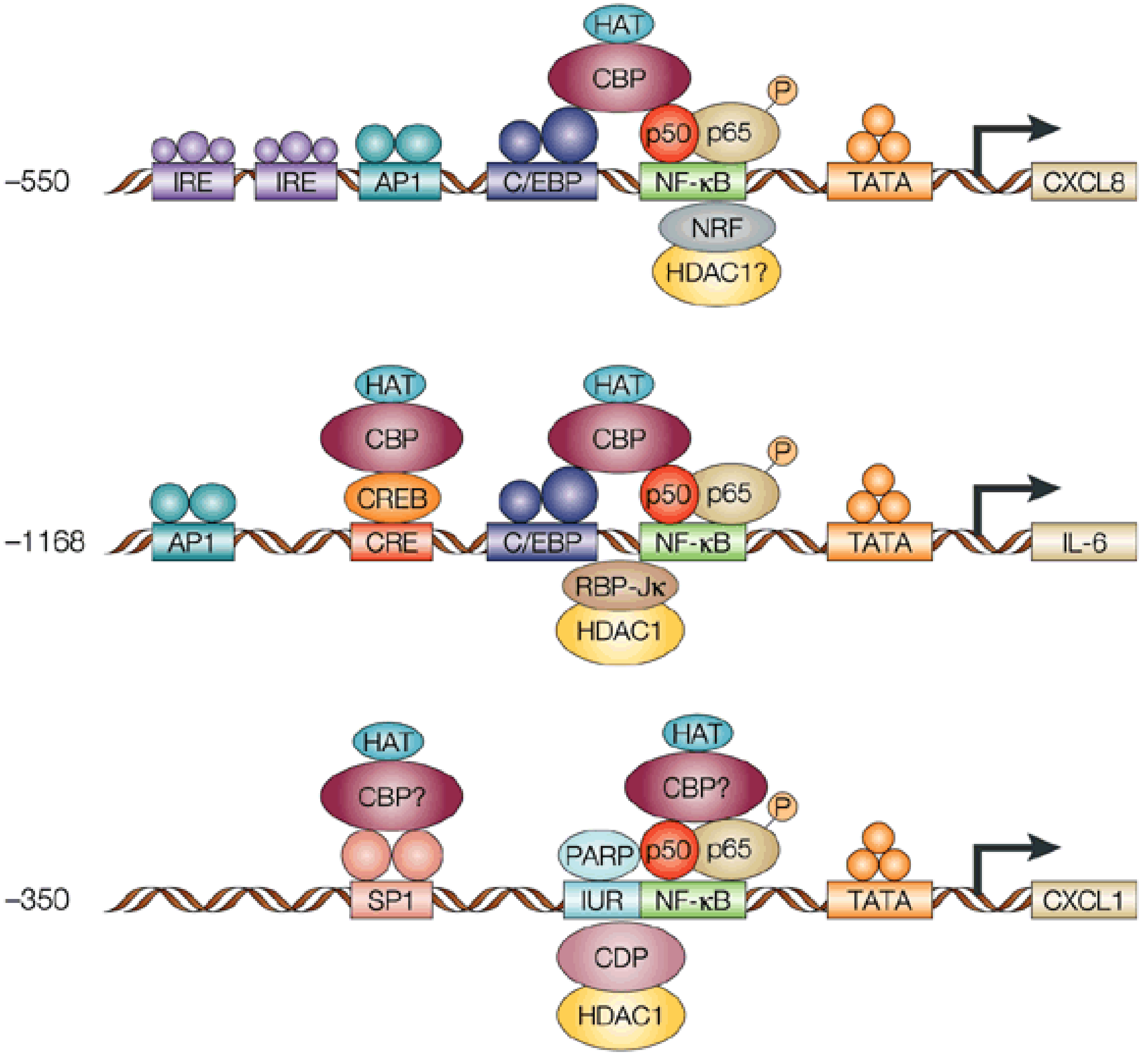
1) Transcription factors



Nature Reviews | Immunology

http://www.nature.com/nri/journal/v2/n9/images/nri887-f2.gif

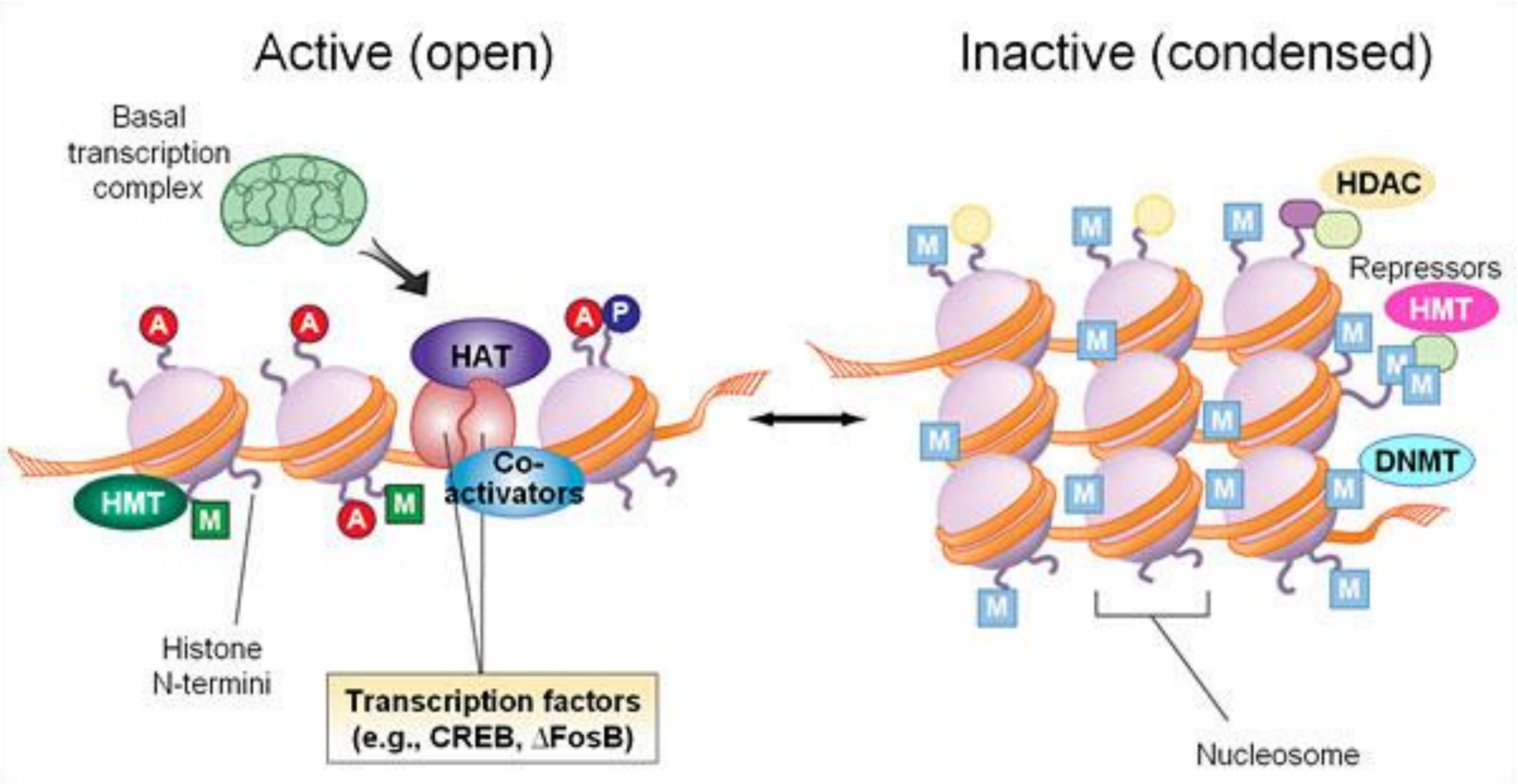
1) Transcription factors



Nature Reviews | Immunology

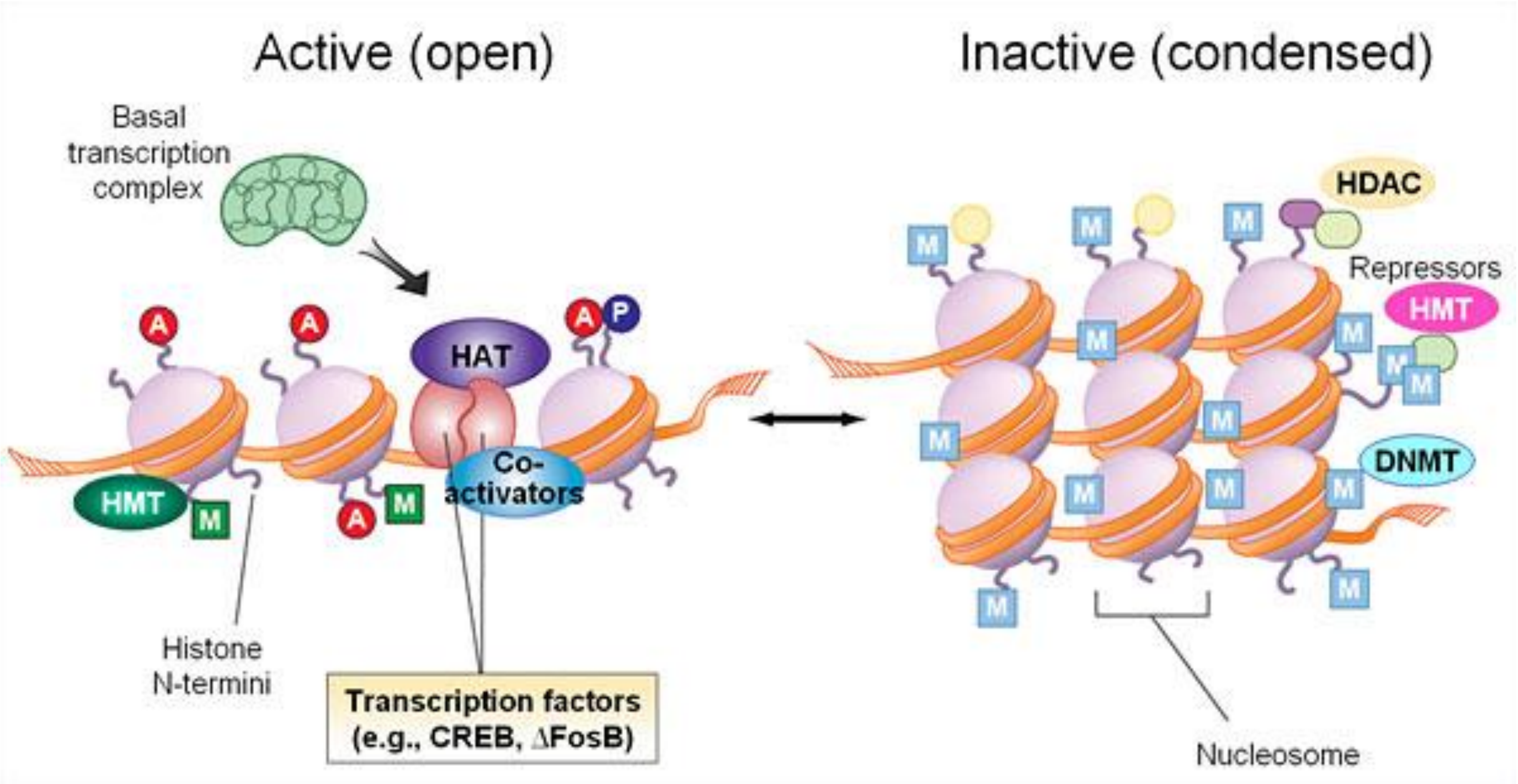
http://www.nature.com/nri/journal/v2/n9/images/nri887-f2.gif

2) Histone modifications



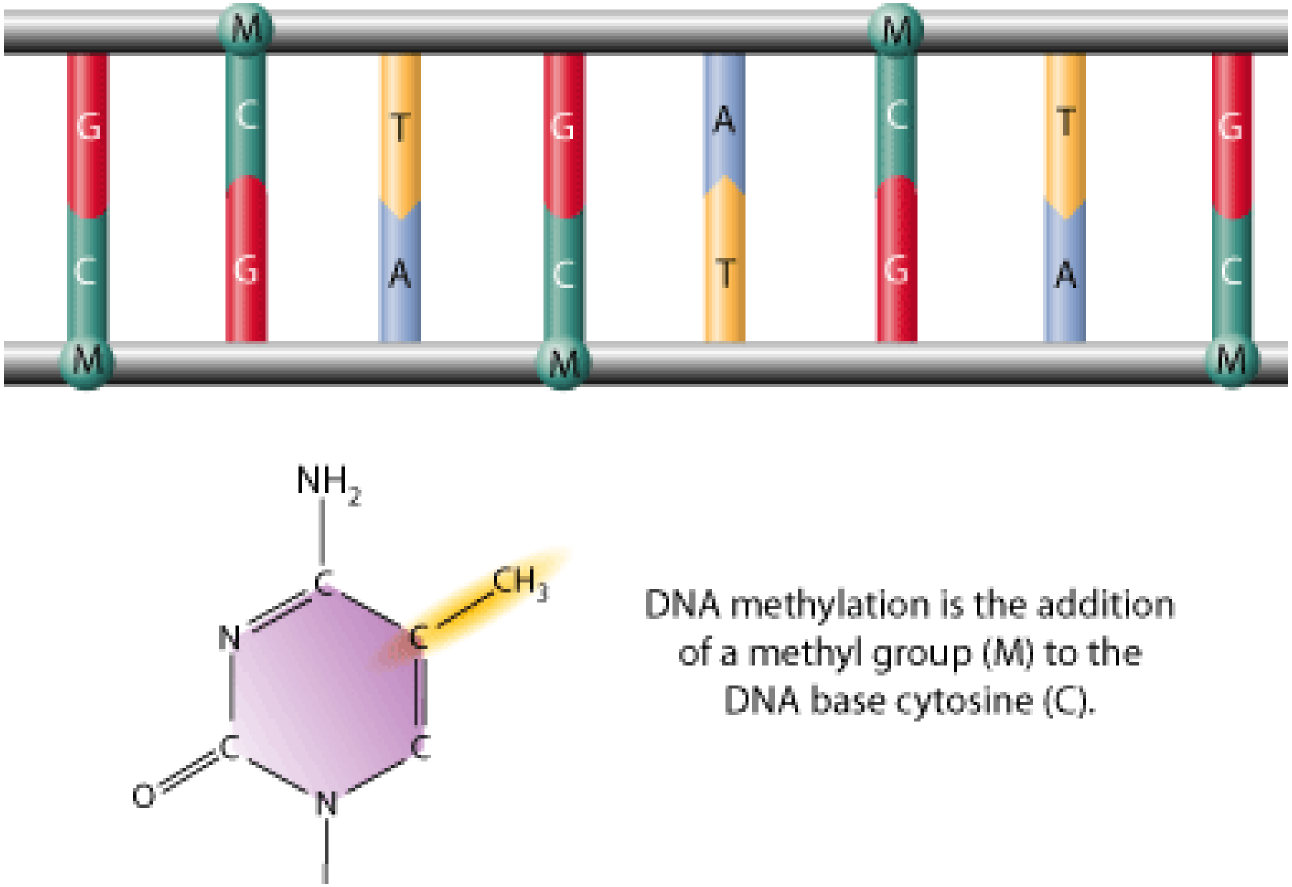
http://www.nyas.org/image.axd?id=a261ff6e-1e42-442c-8dc6-428a60574df4&t=634329511969400000

2) Histone modifications



http://www.nyas.org/image.axd?id=a261ff6e-1e42-442c-8dc6-428a60574df4&t=634329511969400000

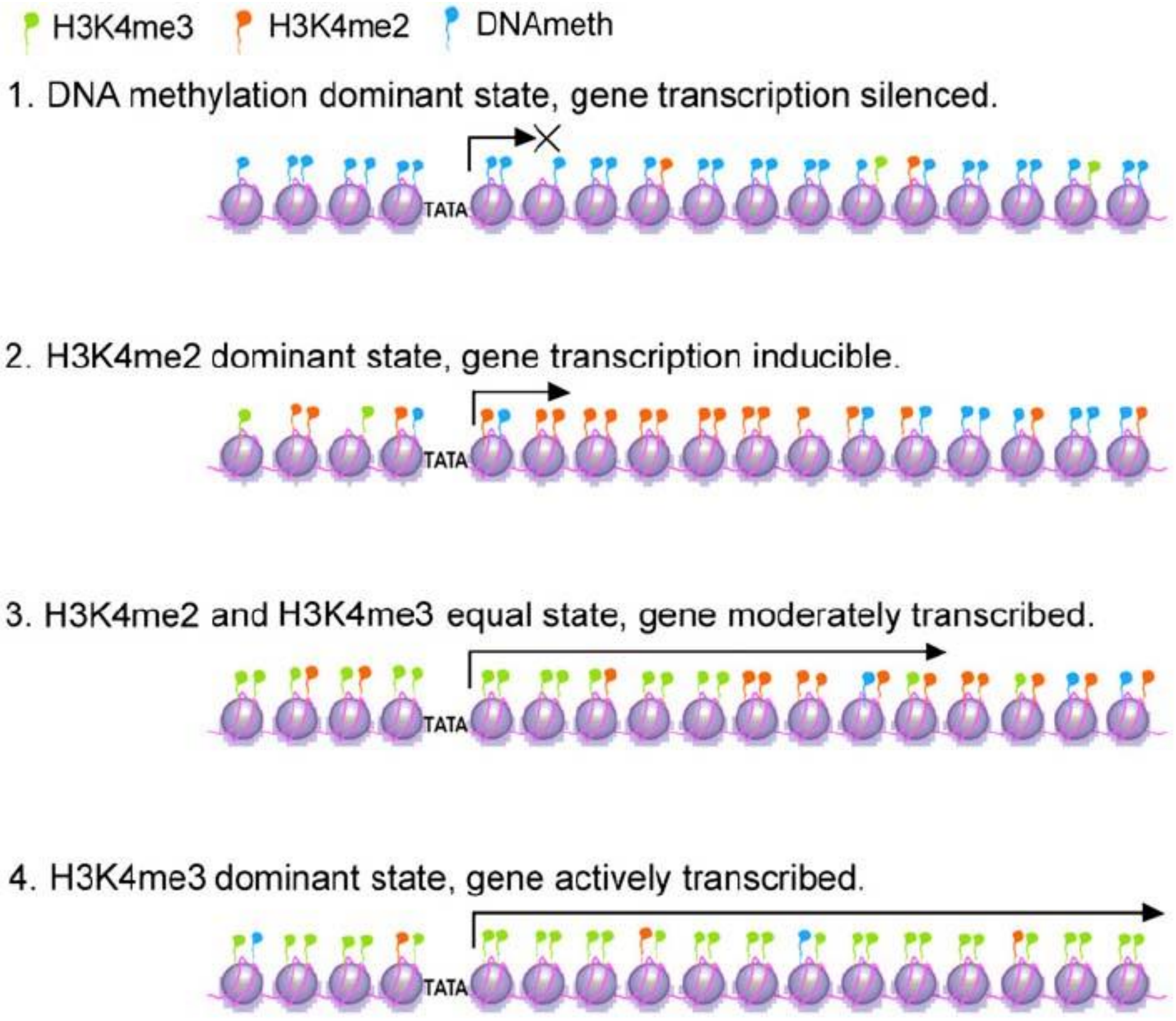
3) DNA methylation



DNA methylation is the addition of a methyl group (M) to the DNA base cytosine (C).

http://www.scq.ubc.ca/wp-content/uploads/2006/08/methylation[1]-GIF.gif

Some known regulatory rules



1. DNA methylation dominant state, gene transcription silenced.

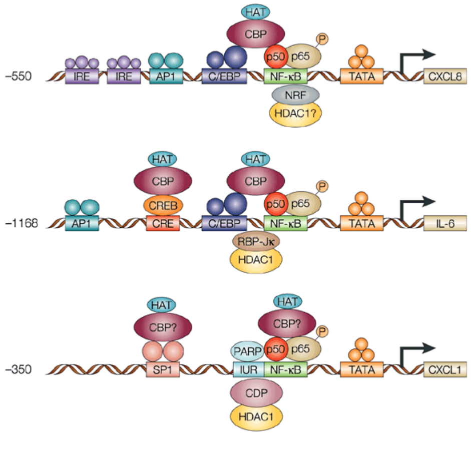
2. H3K4me2 dominant state, gene transcription inducible.

3. H3K4me2 and H3K4me3 equal state, gene moderately transcribed.

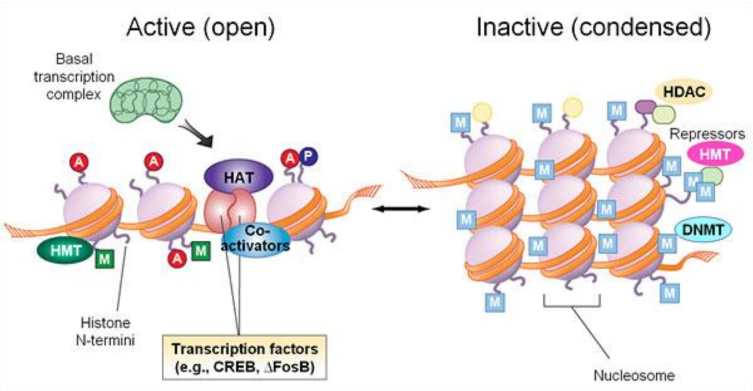
4. H3K4me3 dominant state, gene actively transcribed.

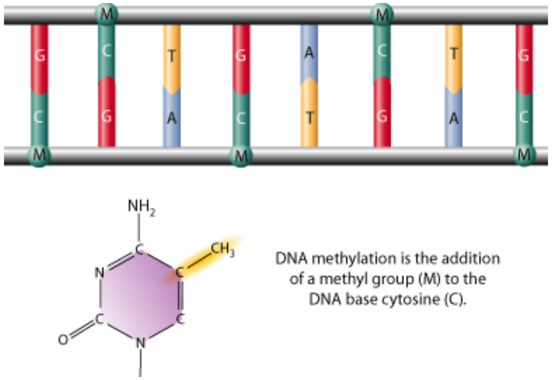
http://ag.arizona.edu/research/xwang/milestones/Rice_Chipchip.jpg

Genome site signals detected individually



Nature Reviews | Immunology



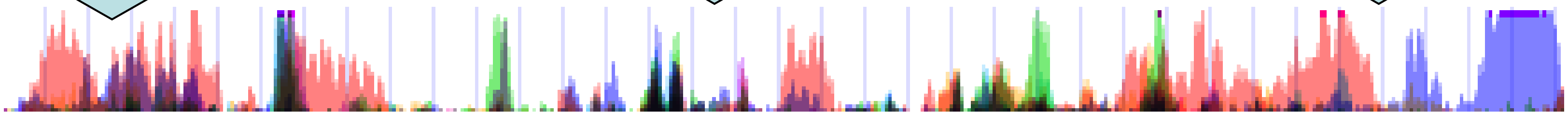


DNA methylation is the addition of a methyl group (M) to the DNA base cytosine (C).

Single transcription factor binding sites

Single histone modifcation sites

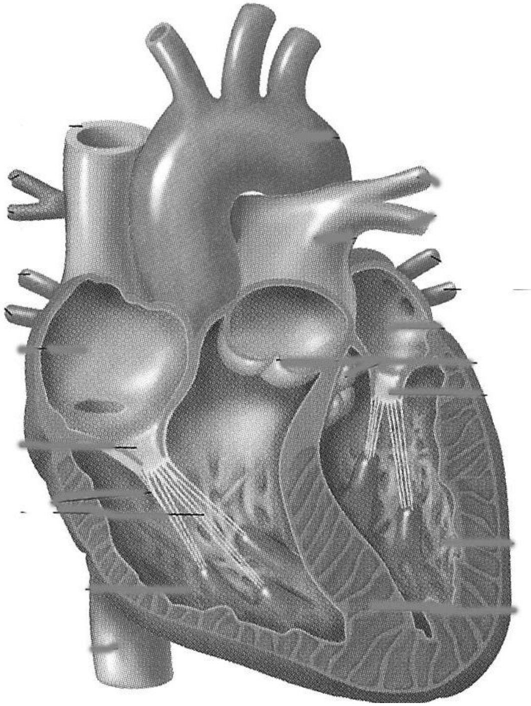
DNA methylation sites

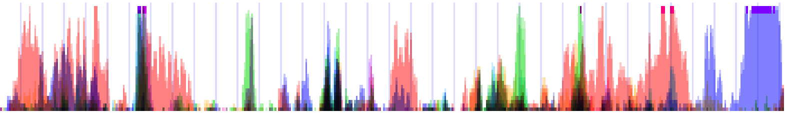


...ACGTTGGATCGAGACATGACGATG...

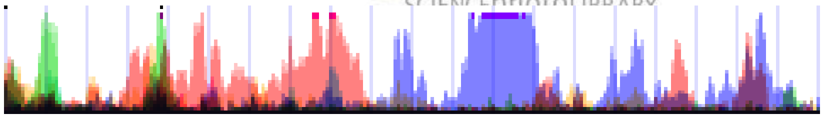
Different signals for different tissues

...ACGTTGGATCGAGACATGACGATG...





≠

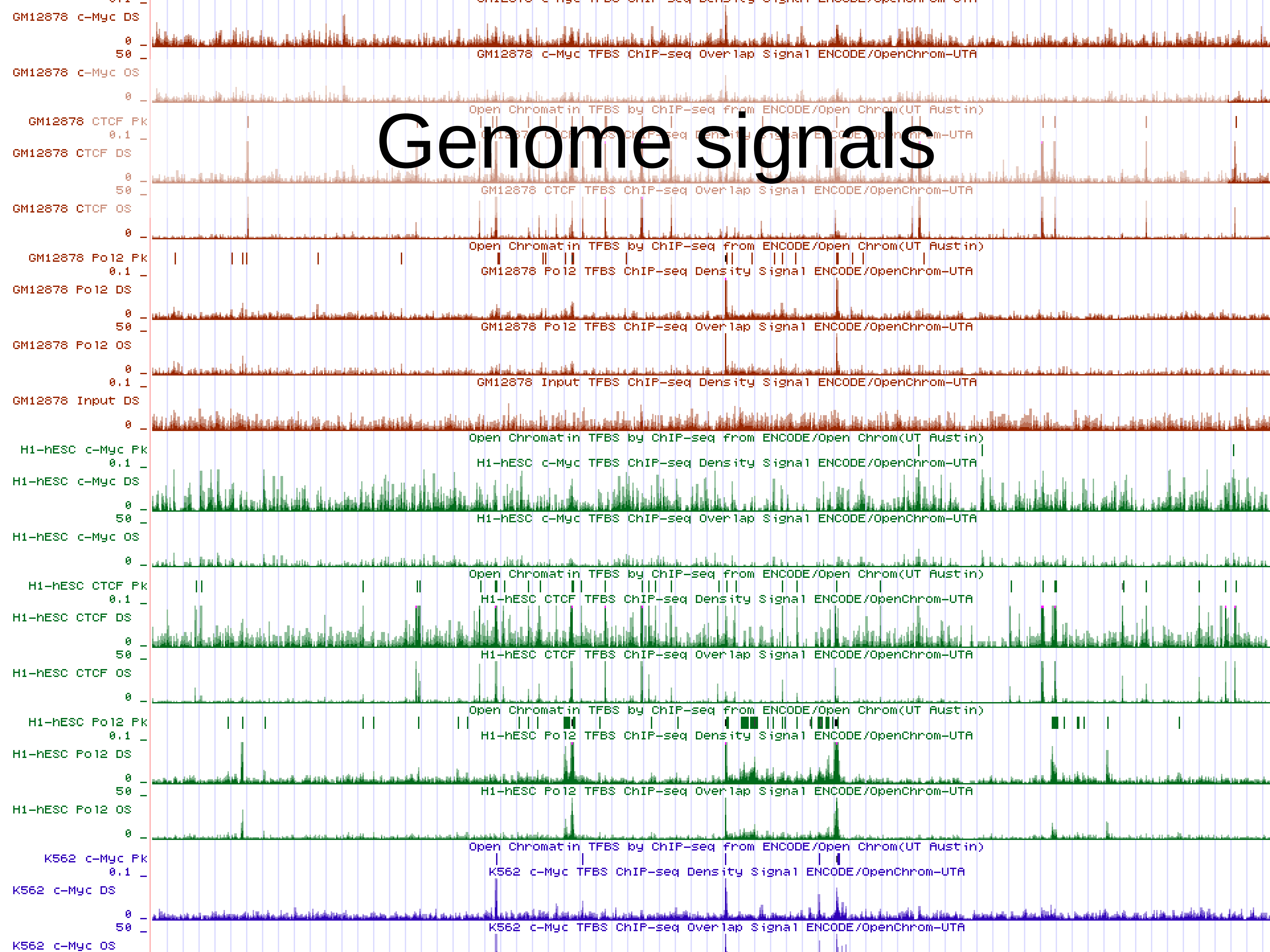


Complexity?

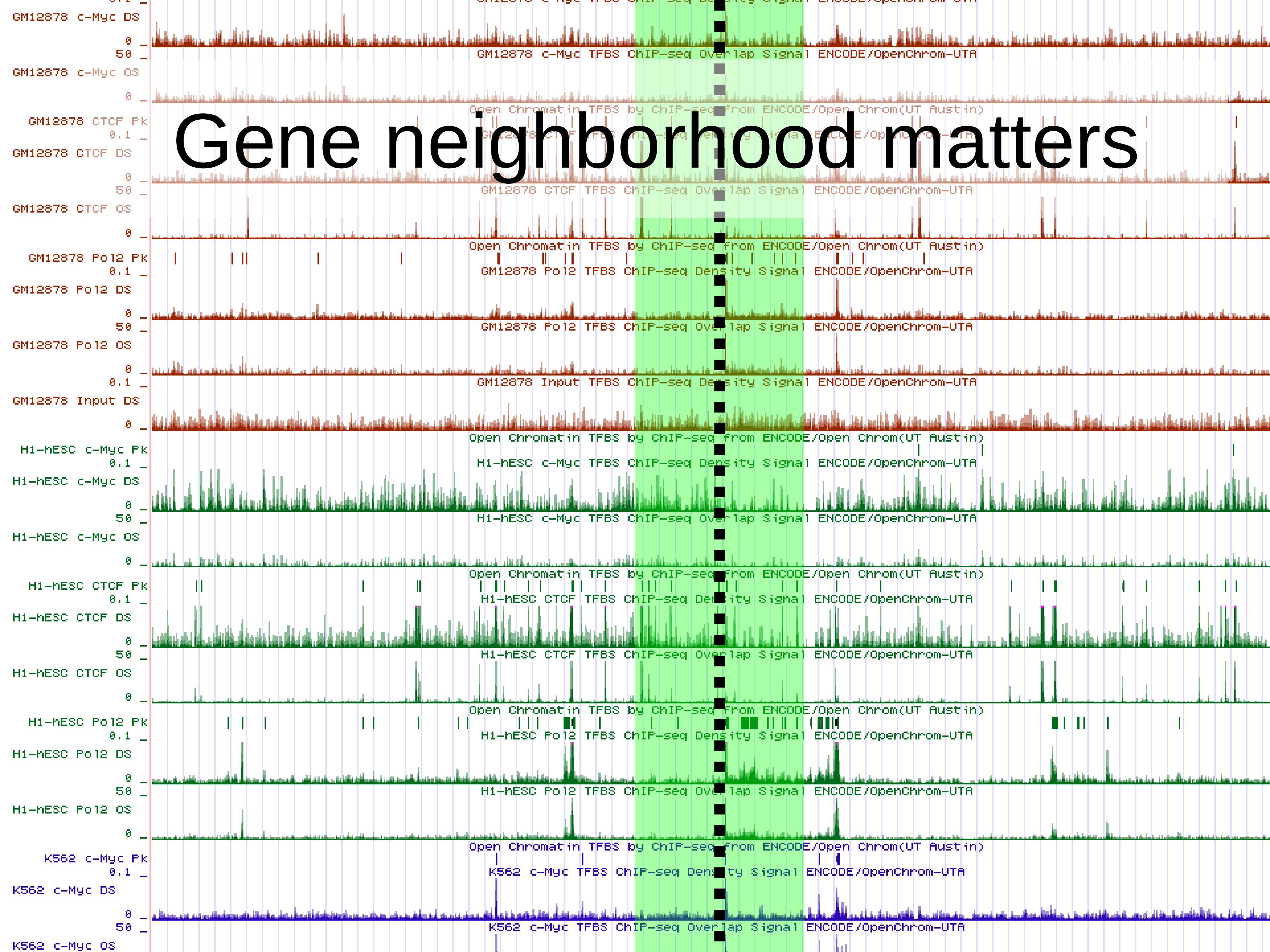
Consider a combination of

- ~2600 transcription factors
- ~210 unique cell types
- dozens of histone modifications

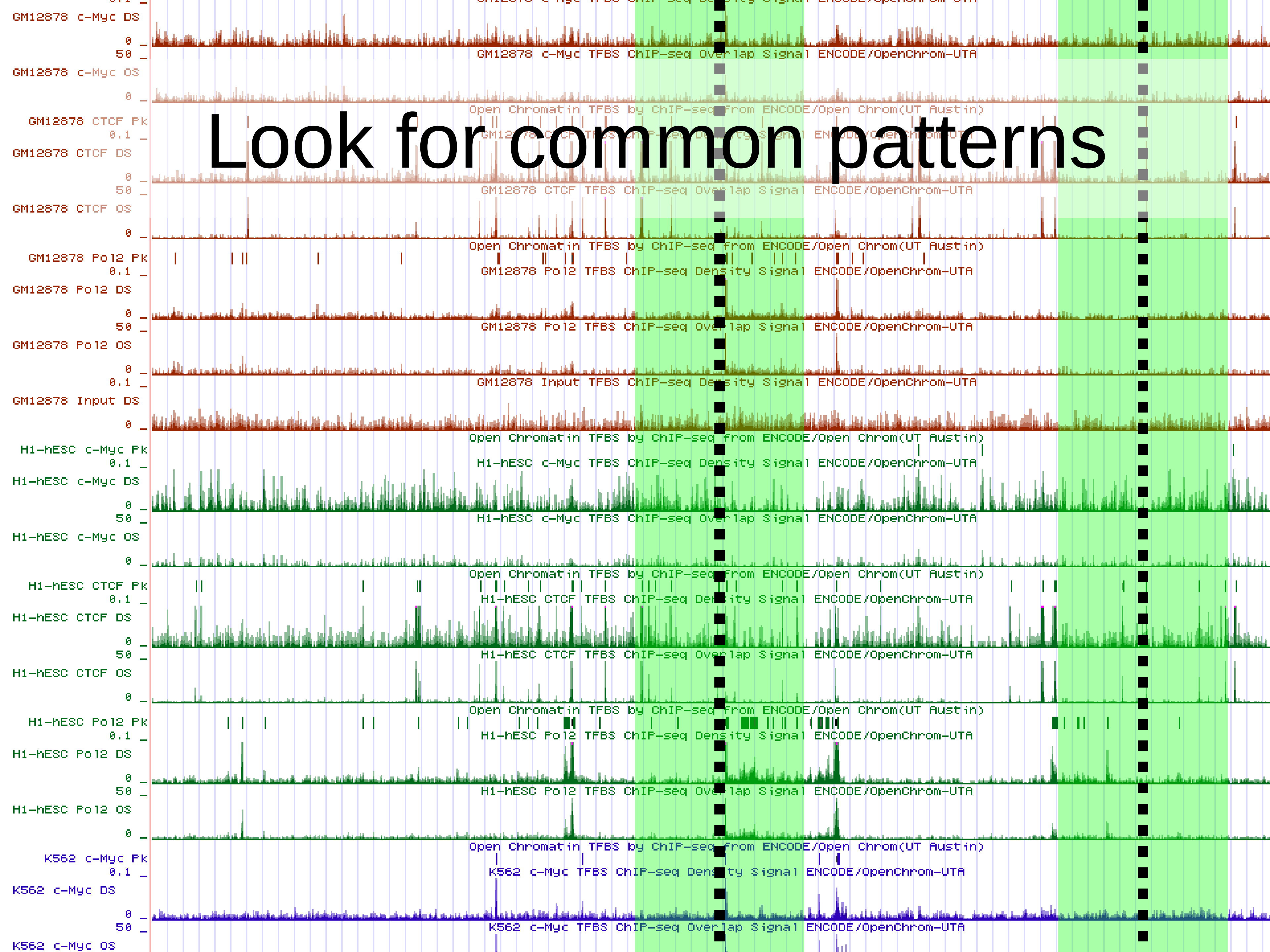
Genome signals


GM12878 c-Myc TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
GM12878 c-Myc DS
0
50
GM12878 c-Myc TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
GM12878 c-Myc OS
0
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
GM12878 CTCF Pk
0.1
GM12878 CTCF TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
GM12878 CTCF DS
0
50
GM12878 CTCF TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
GM12878 CTCF OS
0
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
GM12878 Pol2 Pk
0.1
GM12878 Pol2 TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
GM12878 Pol2 DS
0
50
GM12878 Pol2 TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
GM12878 Pol2 OS
0
0.1
GM12878 Input TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
GM12878 Input DS
0
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
H1-hESC c-Myc Pk
0.1
H1-hESC c-Myc TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
H1-hESC c-Myc DS
0
50
H1-hESC c-Myc TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
H1-hESC c-Myc OS
0
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
H1-hESC CTCF Pk
0.1
H1-hESC CTCF TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
H1-hESC CTCF DS
0
50
H1-hESC CTCF TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
H1-hESC CTCF OS
0
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
H1-hESC Pol2 Pk
0.1
H1-hESC Pol2 TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
H1-hESC Pol2 DS
0
50
H1-hESC Pol2 TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
H1-hESC Pol2 OS
0
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
K562 c-Myc Pk
0.1
K562 c-Myc TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
K562 c-Myc DS
0
50
K562 c-Myc TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
K562 c-Myc OS

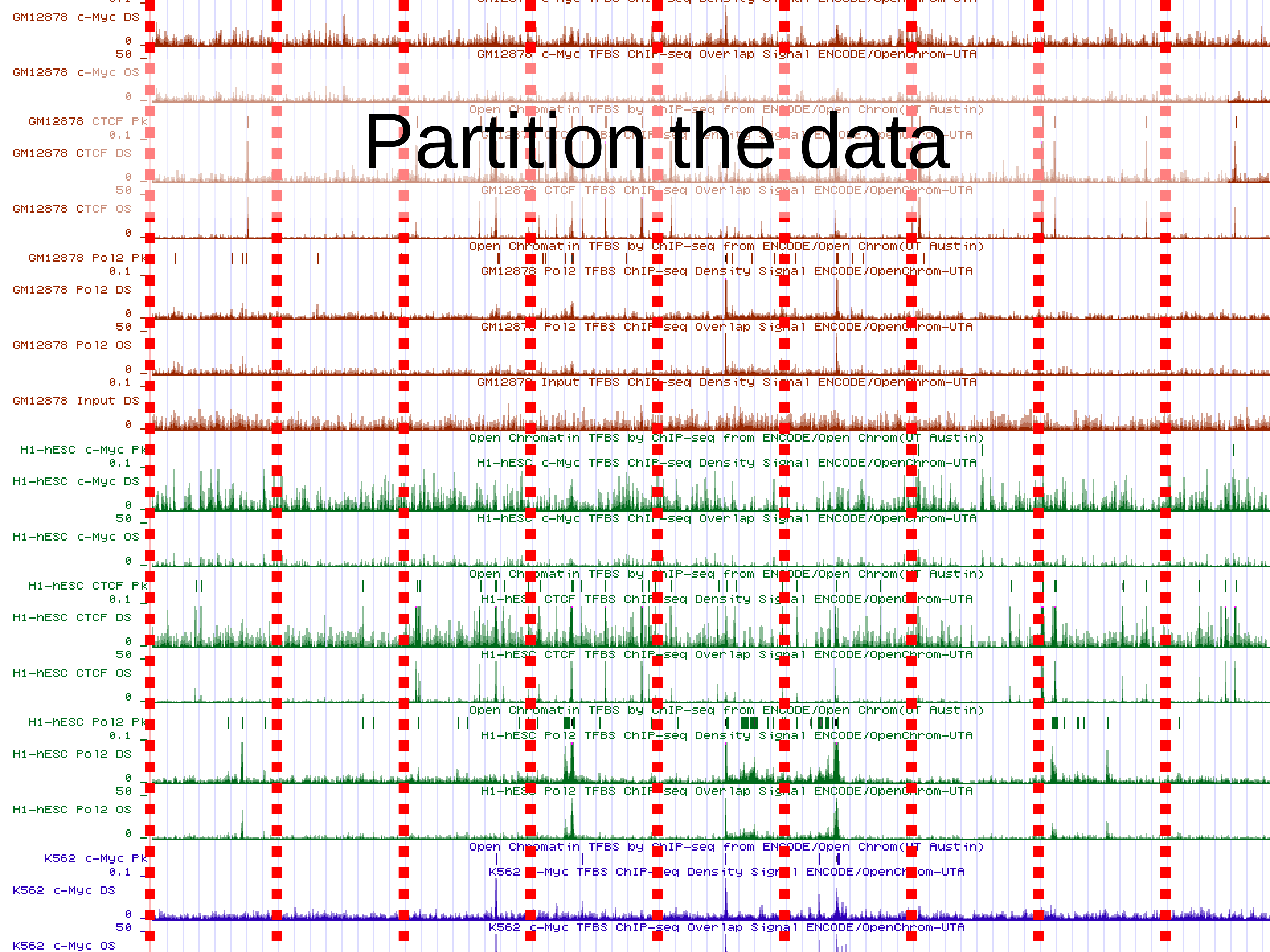
Gene neighborhood matters


GM12878 c-Myc DS
GM12878 c-Myc TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
GM12878 c-Myc OS
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
GM12878 CTCF Pk
GM12878 CTCF TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
GM12878 CTCF DS
GM12878 CTCF TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
GM12878 CTCF OS
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
GM12878 Pol2 Pk
GM12878 Pol2 TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
GM12878 Pol2 DS
GM12878 Pol2 TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
GM12878 Pol2 OS
GM12878 Input TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
GM12878 Input DS
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
H1-hESC c-Myc Pk
H1-hESC c-Myc TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
H1-hESC c-Myc DS
H1-hESC c-Myc TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
H1-hESC c-Myc OS
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
H1-hESC CTCF Pk
H1-hESC CTCF TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
H1-hESC CTCF DS
H1-hESC CTCF TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
H1-hESC CTCF OS
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
H1-hESC Pol2 Pk
H1-hESC Pol2 TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
H1-hESC Pol2 DS
H1-hESC Pol2 TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
H1-hESC Pol2 OS
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
K562 c-Myc Pk
K562 c-Myc TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
K562 c-Myc DS
K562 c-Myc TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
K562 c-Myc OS
0.1
50
0

Look for common patterns

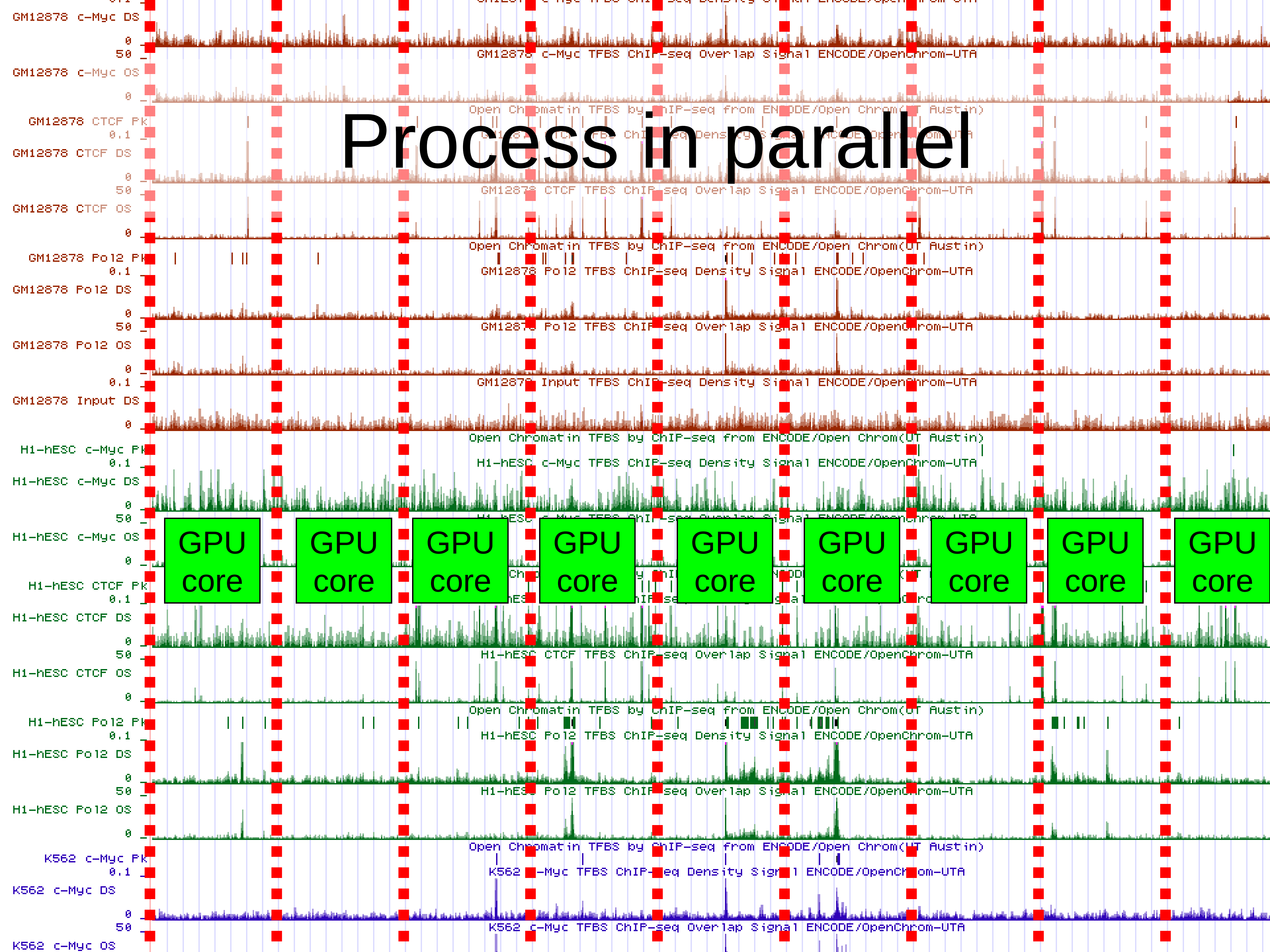

GM12878 c-Myc TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
GM12878 c-Myc DS
0
50
GM12878 c-Myc TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
GM12878 c-Myc OS
0
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
GM12878 CTCF Pk
0.1
GM12878 CTCF TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
GM12878 CTCF DS
0
50
GM12878 CTCF TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
GM12878 CTCF OS
0
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
GM12878 Pol2 Pk
0.1
GM12878 Pol2 TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
GM12878 Pol2 DS
0
50
GM12878 Pol2 TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
GM12878 Pol2 OS
0
0.1
GM12878 Input TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
GM12878 Input DS
0
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
H1-hESC c-Myc Pk
0.1
H1-hESC c-Myc TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
H1-hESC c-Myc DS
0
50
H1-hESC c-Myc TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
H1-hESC c-Myc OS
0
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
H1-hESC CTCF Pk
0.1
H1-hESC CTCF TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
H1-hESC CTCF DS
0
50
H1-hESC CTCF TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
H1-hESC CTCF OS
0
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
H1-hESC Pol2 Pk
0.1
H1-hESC Pol2 TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
H1-hESC Pol2 DS
0
50
H1-hESC Pol2 TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
H1-hESC Pol2 OS
0
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
K562 c-Myc Pk
0.1
K562 c-Myc TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
K562 c-Myc DS
0
50
K562 c-Myc TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
K562 c-Myc OS

Partition the data



GM12878 c-Myc DS
0
50
GM12878 c-Myc TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
GM12878 c-Myc OS
0
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
GM12878 CTCF Pk
0.1
GM12878 CTCF TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
GM12878 CTCF DS
0
50
GM12878 CTCF TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
GM12878 CTCF OS
0
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
GM12878 Pol2 Pk
0.1
GM12878 Pol2 TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
GM12878 Pol2 DS
0
50
GM12878 Pol2 TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
GM12878 Pol2 OS
0
0.1
GM12878 Input TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
GM12878 Input DS
0
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
H1-hESC c-Myc Pk
0.1
H1-hESC c-Myc TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
H1-hESC c-Myc DS
0
50
H1-hESC c-Myc TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
H1-hESC c-Myc OS
0
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
H1-hESC CTCF Pk
0.1
H1-hESC CTCF TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
H1-hESC CTCF DS
0
50
H1-hESC CTCF TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
H1-hESC CTCF OS
0
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
H1-hESC Pol2 Pk
0.1
H1-hESC Pol2 TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
H1-hESC Pol2 DS
0
50
H1-hESC Pol2 TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
H1-hESC Pol2 OS
0
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
K562 c-Myc Pk
0.1
K562 c-Myc TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
K562 c-Myc DS
0
50
K562 c-Myc TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
K562 c-Myc OS

Process in parallel


GM12878 c-Myc DS
GM12878 c-Myc TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
GM12878 c-Myc OS
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
GM12878 CTCF Pk
GM12878 CTCF TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
GM12878 CTCF DS
GM12878 CTCF TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
GM12878 CTCF OS
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
GM12878 Pol2 Pk
GM12878 Pol2 TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
GM12878 Pol2 DS
GM12878 Pol2 TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
GM12878 Pol2 OS
GM12878 Input TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
GM12878 Input DS
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
H1-hESC c-Myc Pk
H1-hESC c-Myc TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
H1-hESC c-Myc DS
H1-hESC c-Myc OS
H1-hESC CTCF Pk
H1-hESC CTCF DS
H1-hESC CTCF TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
H1-hESC CTCF OS
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
H1-hESC Pol2 Pk
H1-hESC Pol2 TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
H1-hESC Pol2 DS
H1-hESC Pol2 TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
H1-hESC Pol2 OS
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
K562 c-Myc Pk
K562 c-Myc TFBS ChIP-seq Density Signal ENCODE/OpenChrom-UTA
K562 c-Myc DS
K562 c-Myc TFBS ChIP-seq Overlap Signal ENCODE/OpenChrom-UTA
K562 c-Myc OS
GPU core
GPU core
GPU core
GPU core
GPU core
GPU core
GPU core
GPU core
GPU core

Take home message

- Biology as a science is becoming more computationally oriented
- Genome biology is massively parallel by its nature
- GPUs fit perfectly for solving problems in genome biology