Annotation - Overview

Kasper D Hansen

@KasperDHansen

www.hansenlab.org

Annotation is the process of giving context to your data using external information.

This can be done in many ways, but it usually involves linking your data to online databases.

Example: annotating a probeset on the Affymetrix HGU95av2 microarray.

Example: annotating a genomic interval.

Genome coordinates
chr7:27,136,000-27,137,000

Some annotation is just experimental data from other groups.

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

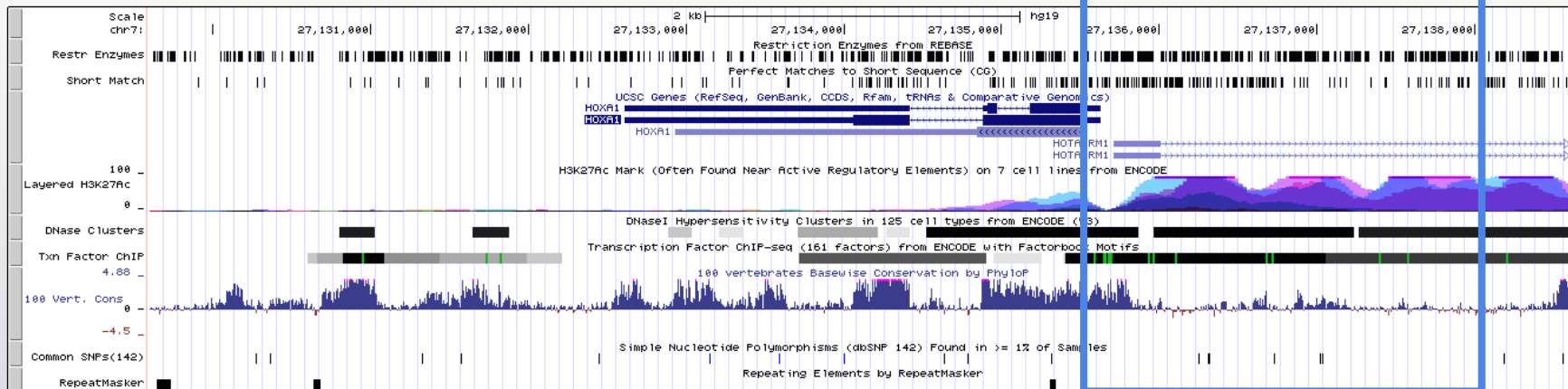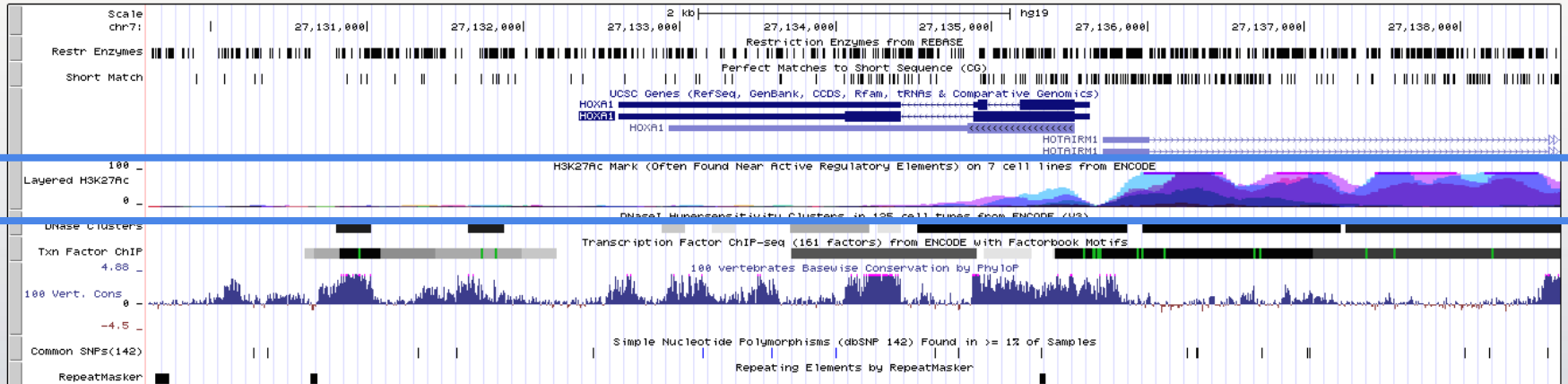move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr7:27,129,602-27,138,637 9,036 bp. [enter position, gene symbol or search terms] go *More on-site workshops available!*

chr7 (p15.2) 7p21.3 21.1 15.3 7p14.3 7p14.1 13 11.2 11.21 11.22 11.23 7q21.11 q21.3 7q22.1 7q31.1 7q33 7q34 7q35 36.1 36.3

Scale   2 kb   hg19
chr7:   27,131,000| 27,132,000| 27,133,000| 27,134,000| 27,135,000| 27,136,000| 27,137,000| 27,138,000|
Restr Enzymes   Restriction Enzymes from REBASE
Short Match   Perfect Matches to Short Sequence (CG)
                  UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
              HOXA1
              HOXA1
                  HOXA1   <<<<<<<<<<<<<<<<
                                                              HOTAIRM1
                                                              HOTAIRM1
       100   H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
        0
                                      DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
DNase Clusters
Txn Factor ChIP   Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
      4.88   100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons 0
     -4.5
                              Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Common SNPs(142)
RepeatMasker   Repeating Elements by RepeatMasker

How to annotate 100-10,000 items?
Need to do this programmatically.
And easily!

Two main approaches in Bioconductor:

Annotation packages: (ex. hgu95av2.db) This is "preprocessed" and "packaged" annotation.

Query online resources: like UCSC or ENSEMBL.