Assessing Noncoding Functional Elements by Experimental and Computational Means

Current Topics in Genome Analysis

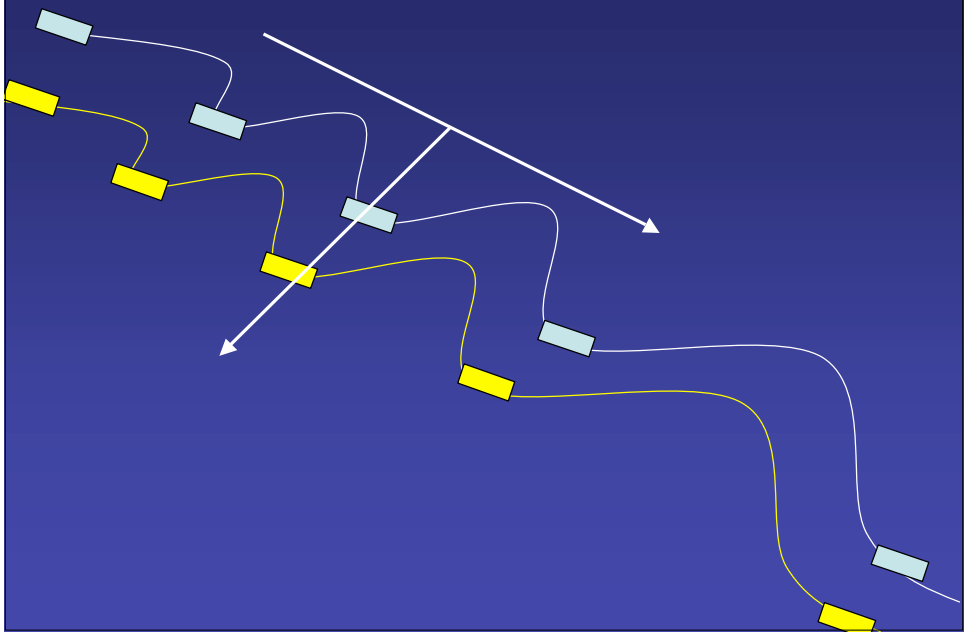
Laura Elnitski, Ph.D.
NHGRI/NIH

Comparative Sequence Analysis

Dr. Elliott Margulies
Current Topics Lecture *3-1-05*

- Tools for aligning genomic sequences
- Electing to use multiple species
- Conservation as an indicator of functional elements
- Sequence similarities reflect evolutionary relationships

Intra- and Interspecies Sequence Comparisons



How do I use genomic databases to find information about functional elements that might regulate expression of my gene of interest?

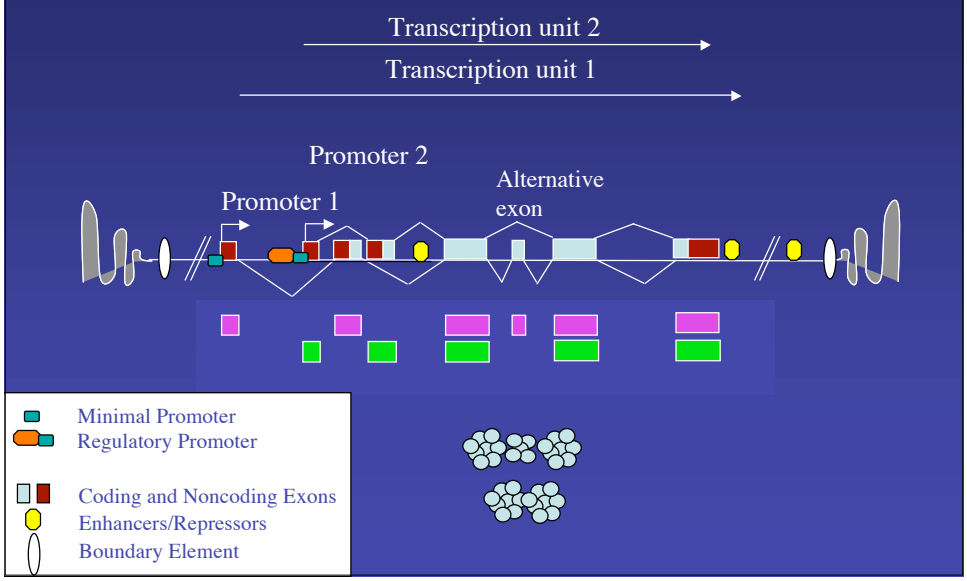
Outline

- Classes of Functional Elements
- Computational Analyses
- Experimental Analyses/Validation
- Preponderance of Evidence

Functional Elements

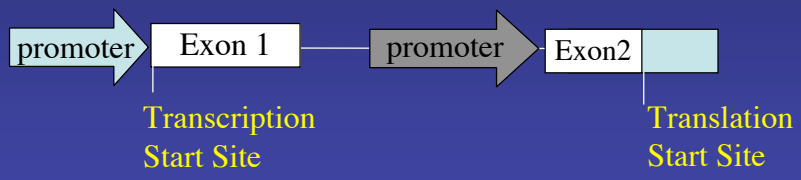
- Intragenic Functional Elements
  - Coding Exons
  - Untranslated Regions (UTRs)
  - Introns
- Intergenic Functional Elements
  - Boundary Elements
- Regulators of Gene Expression
  - Promoters
  - Enhancers/Repressors

A Typical Genomic Locus

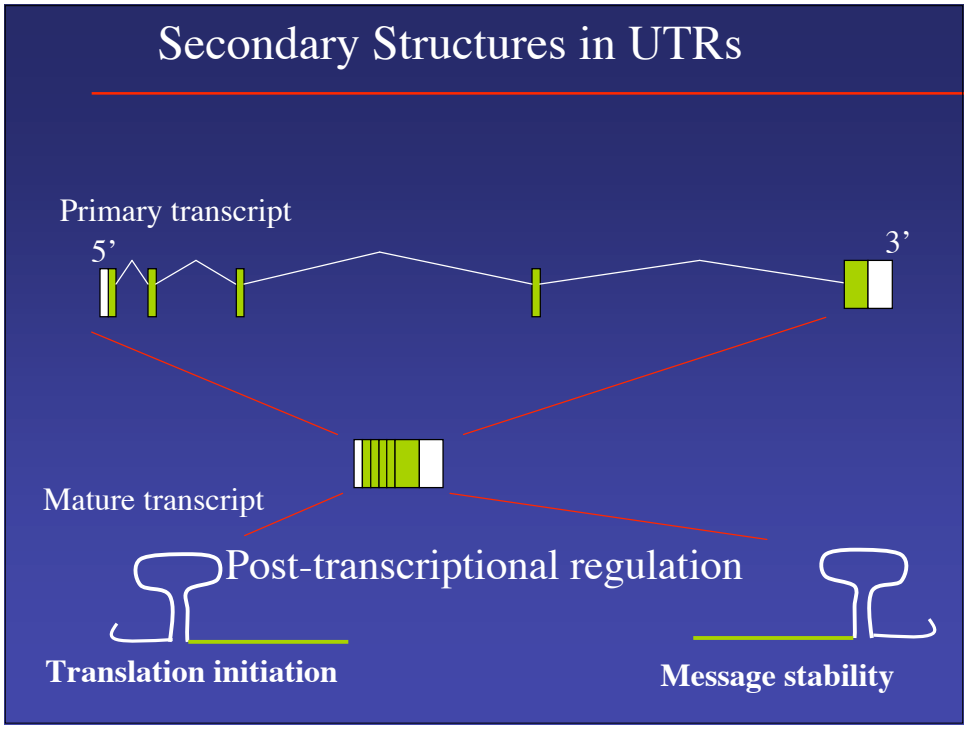


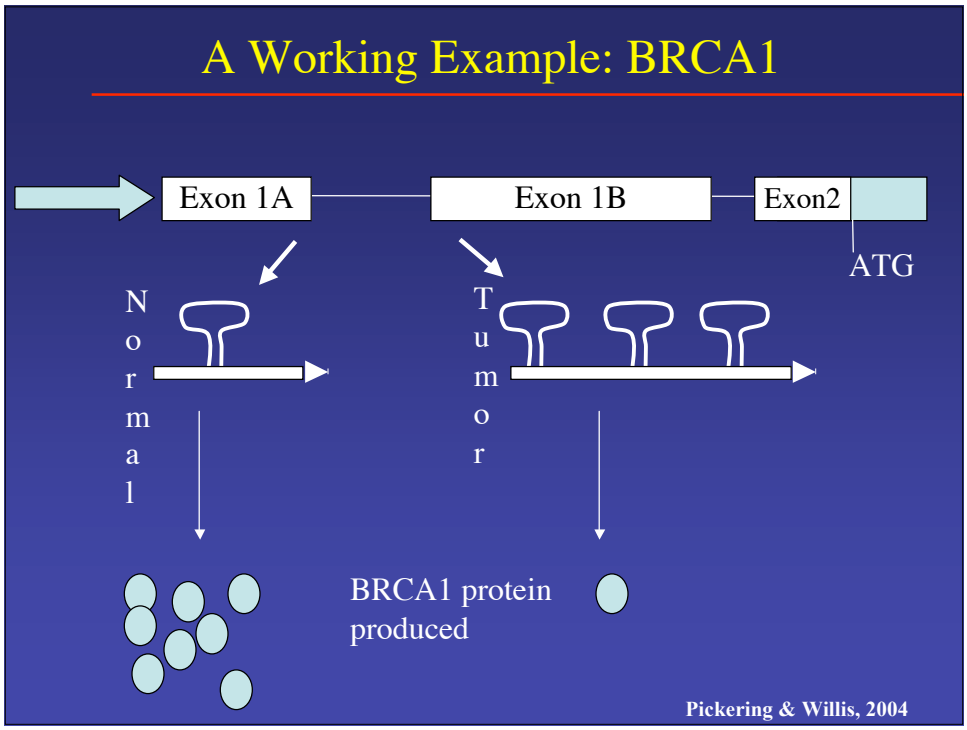
UTRs

- Often omitted from gene annotations
- 5' UTRs indicate transcription start site

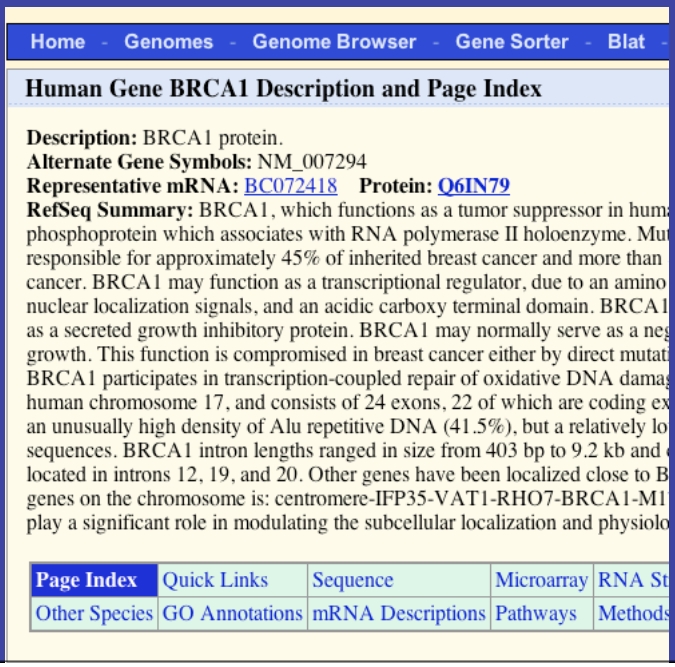


- Can be regulatory in nature

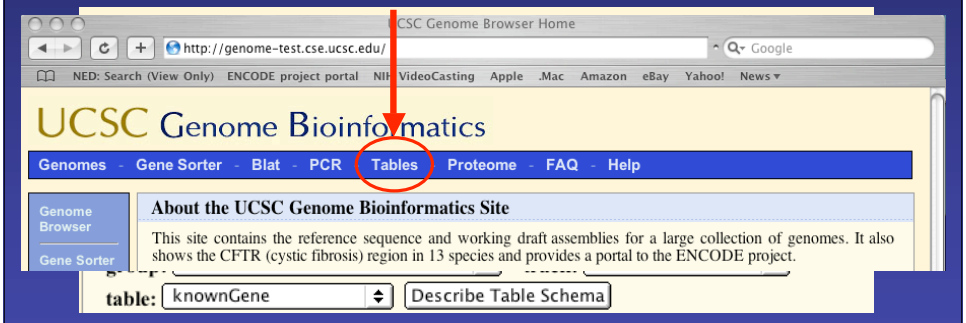

Secondary Structures in UTRs
Primary transcript
5'
3'
Mature transcript
Post-transcriptional regulation
Translation initiation
Message stability


A Working Example: BRCA1
Exon 1A
Exon 1B
Exon2
ATG
Normal
Tumor
BRCA1 protein produced
Pickering & Willis, 2004

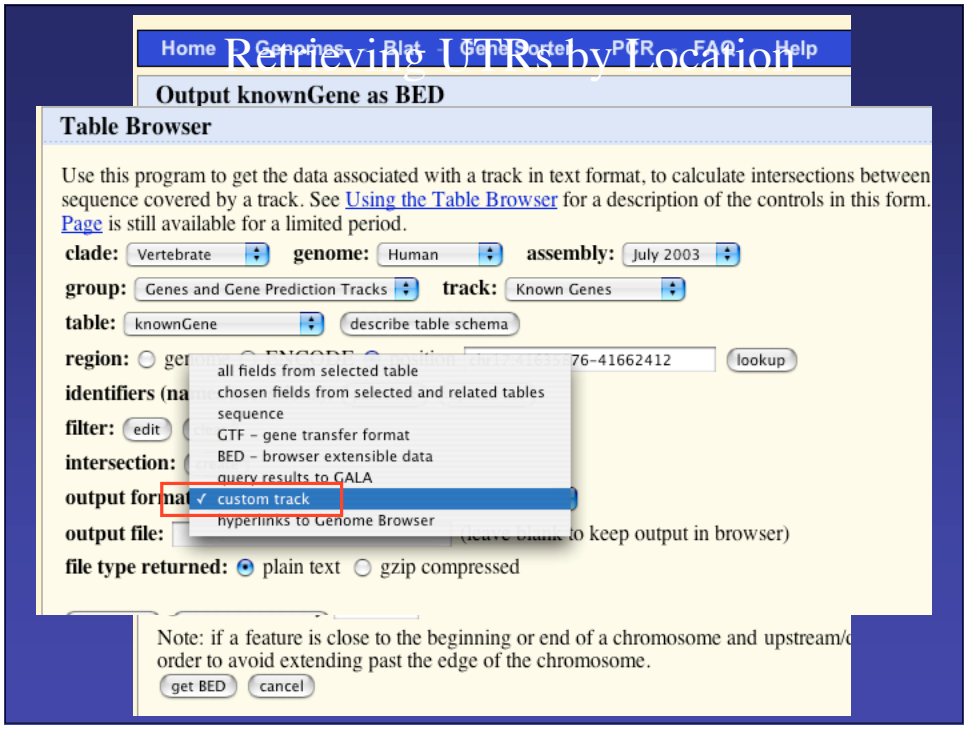
Genome Browser - Description Page



Data Retrieval via Table Browser

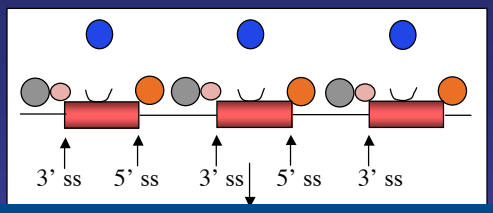


Retrieving UTRs by Location



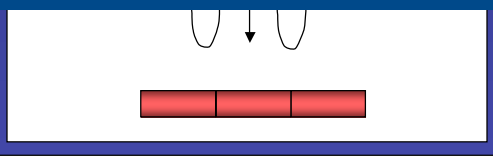
Regulatc Splice Junctions in Introns

Splicing Enhancers (Exonic)



RESCUE-ESE Web Server

An online tool for identifying candidate ESEs in vertebrate exons



http://genes.mit.edu/burgelab/rescue-ese/

RESCUE-ESE Server

Choice of Vertebrate ESEs

Check if **YES**, default is **NO**

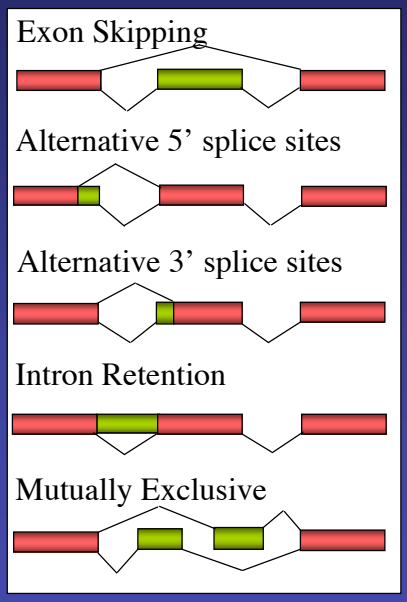
- [x] Human (Homo sapiens)
- [ ] Mouse (Mus musculus)
- [ ] Zebrafish (Danio rerio)
- [ ] Pufferfish (Fugu rubripes)

Find ESEs | Reset

Enter Your Sequence (RNA or DNA) as plain text or multi-FASTA format (maximum 4k bases of sequence data)

```
ACCAAGGTCAGAACATTCACCGAAAGACAACAGCATCCACACGAAAAGTG
TCACTGGCCCCTCAGGCAAACTTGACTGAACTGGATATATATTCAAGAAG
GTTATCTCAAGAAACTGGCTTGGAAATAAGTGAAGAAATTAACGAAGAAG
ACTTAAAGGTAGGTATACATCGCTTG
```

Alternative splicing events



Conserved across species:

38%

18%

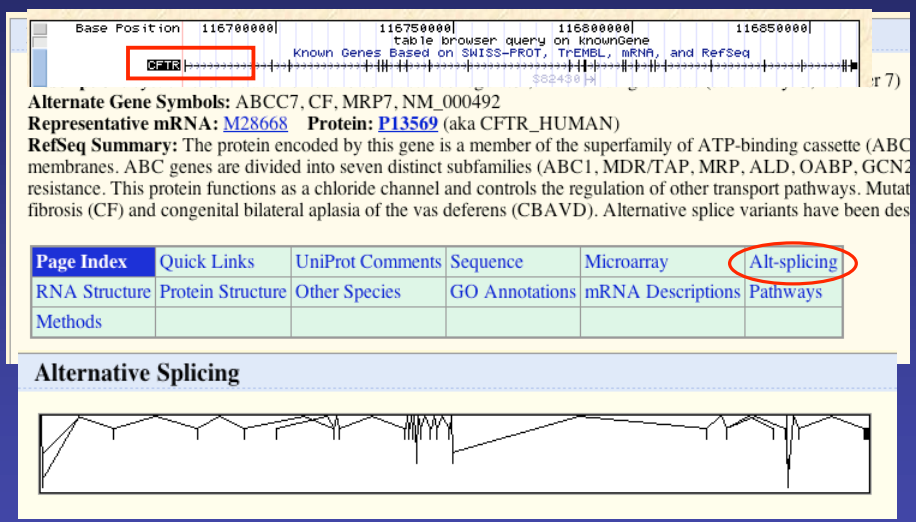
8%

3%

N/A

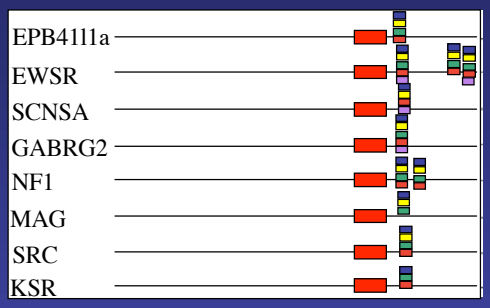
Nature Reviews Genetics
5; 773-782 (2004)

UCSC Alt-splicing Data



http://genome-test.cse.ucsc.edu

Splicing Enhancers (Intronic)



UGCAUG

- Associated with alternatively spliced exons that have brain specific expression pattern.
- Show conservation across mammalian species

Minovitsky et al. NAR 2005

Cis-Acting Regulatory Elements

Defined with respect to location and orientation from the transcription start site

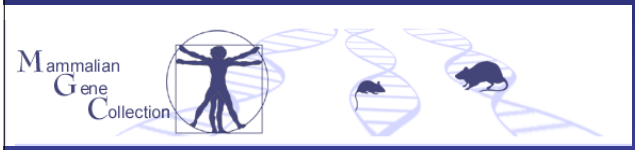
- Promoters
  - Proximal to transcription start site
  - Determine timing of gene expression during development
  - Recruit basal transcription machinery
- Enhancers/(Repressors)
  - Act in a distance and orientation independent manner
  - Augment the level of expression or choice of tissue specificity

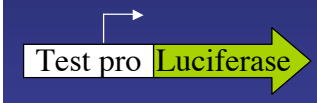
• Core Promoters

Elements located proximal to transcription start sites that determine the timing of gene expression during development and act to recruit the basal transcription machinery.

High Throughput Promoter Analysis

Identification and functional analysis of human transcriptional promoters.
Trinklein et al. (2003) Genome Research 13(2):308-312.



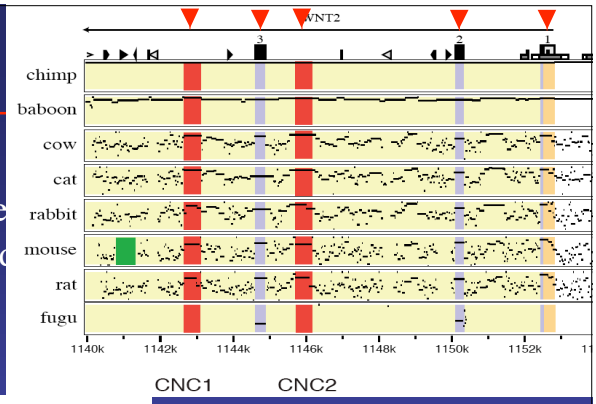


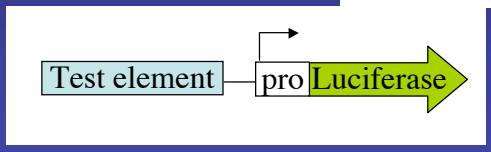
- Upstream regions from the Mammalian Gene Collection.
- 90% of tested predictions are positive for function
- Over 40,000 mapped candidate promoters in the human genome.

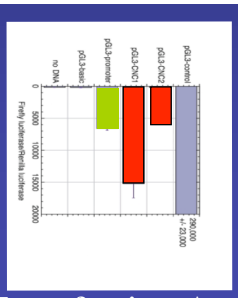
Available at genome-test.cse.ucsc.edu
April 2003

• Enhancers

• Identify candidate e
through multiple sec
alignment

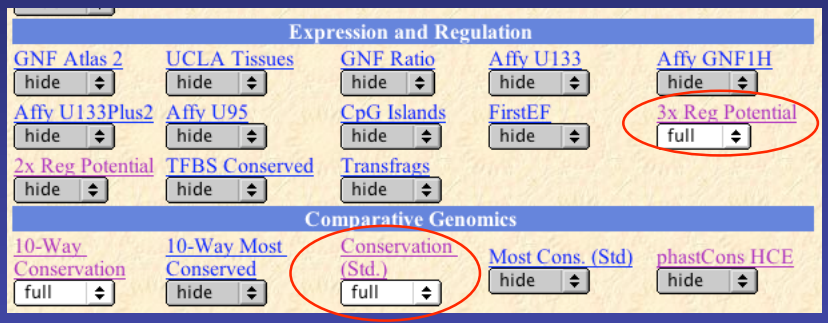






Transfection Assay

Annotations of Putative Regulatory Regions

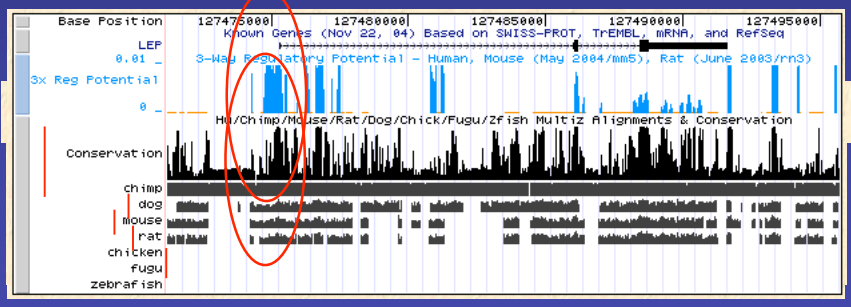


Regulatory Predictions

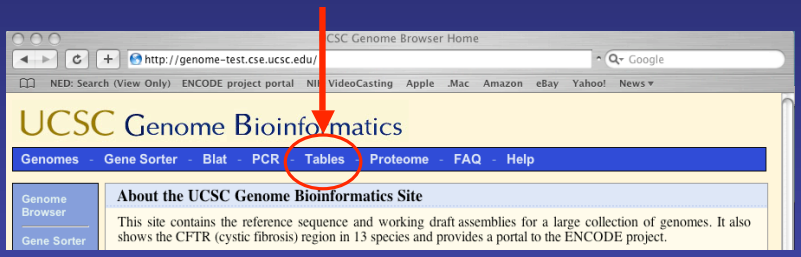
Conservation

PhyloHMM (PhastCons)

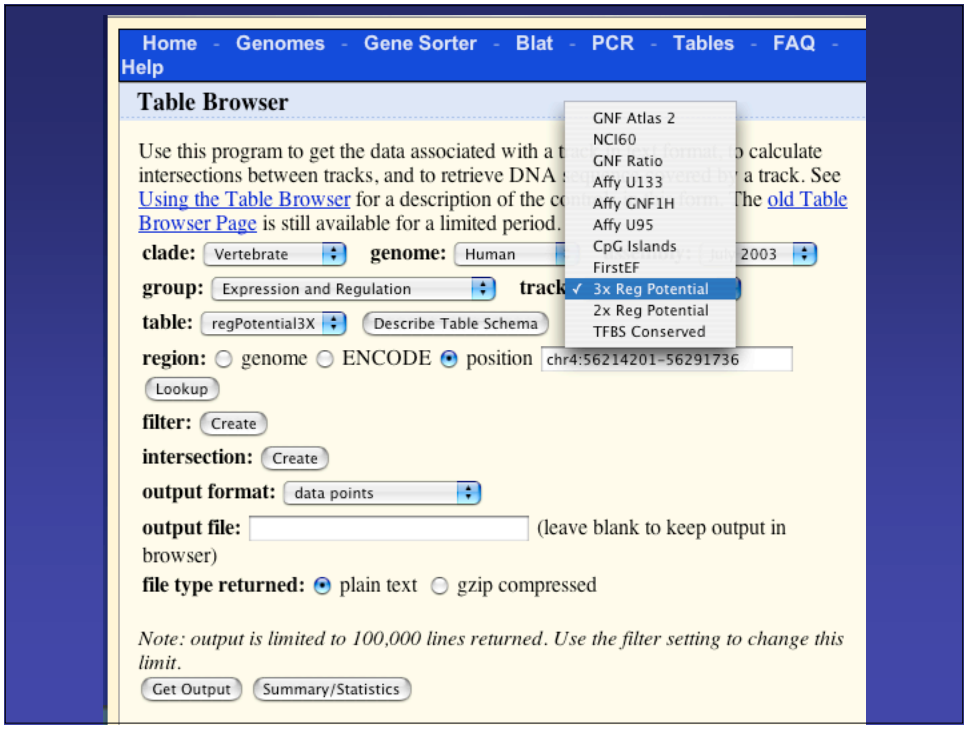
Regulatory Potential Scores

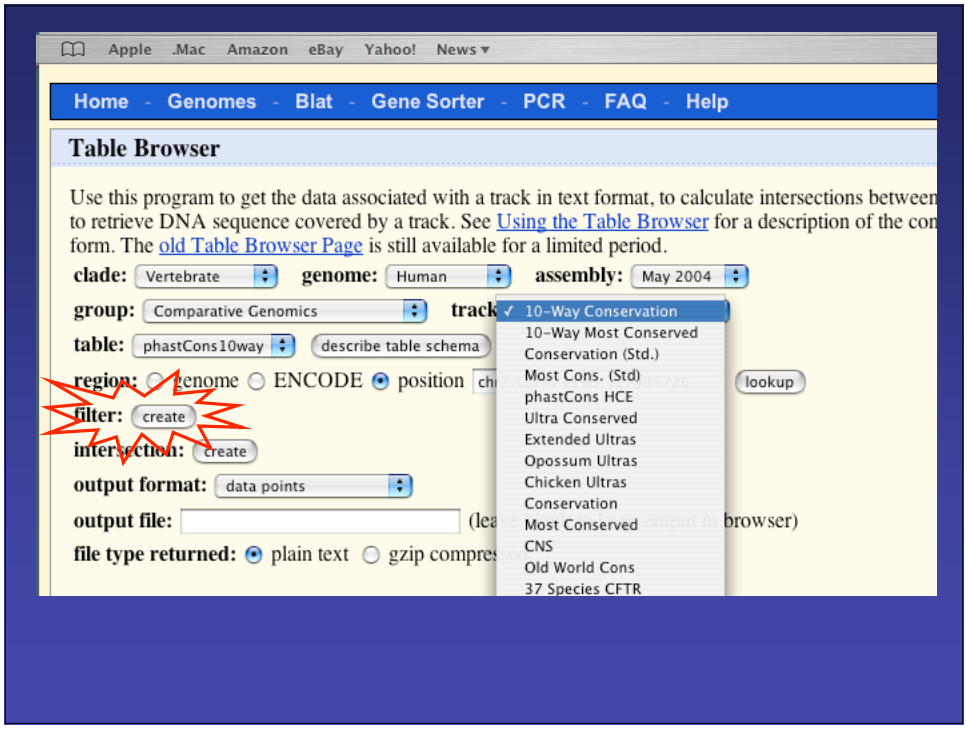


Data Retrieval via Table Browser



http://genome-test.cse.ucsc.edu/


Home - Genomes - Gene Sorter - Blat - PCR - Tables - FAQ - Help
Table Browser
GNF Atlas 2
NCI60
GNF Ratio
Affy U133
Affy GNF1H
Affy U95
CpG Islands
FirstEF
3x Reg Potential
2x Reg Potential
TFBS Conserved
Use this program to get the data associated with a t
intersections between tracks, and to retrieve DNA
a track. See
Using the Table Browser for a description of the co
The old Table
Browser Page is still available for a limited period.
clade: Vertebrate
genome: Human
2003
group: Expression and Regulation
track
table: regPotential3X
Describe Table Schema
region: genome ENCODE position chr4:56214201-56291736
Lookup
filter: Create
intersection: Create
output format: data points
output file:
(leave blank to keep output in browser)
file type returned: plain text gzip compressed
Note: output is limited to 100,000 lines returned. Use the filter setting to change this limit.
Get Output
Summary/Statistics


Apple .Mac Amazon eBay Yahoo! News
Home - Genomes - Blat - Gene Sorter - PCR - FAQ - Help
Table Browser
Use this program to get the data associated with a track in text format, to calculate intersections between
to retrieve DNA sequence covered by a track. See Using the Table Browser for a description of the co
form. The old Table Browser Page is still available for a limited period.
clade: Vertebrate
genome: Human
assembly: May 2004
group: Comparative Genomics
track
table: phastCons10way
describe table schema
region: genome ENCODE position
lookup
filter: create
intersection: create
output format: data points
output file:
browser)
file type returned: plain text gzip compres
10-Way Conservation
10-Way Most Conserved
Conservation (Std.)
Most Cons. (Std)
phastCons HCE
Ultra Conserved
Extended Ultras
Opossum Ultras
Chicken Ultras
Conservation
Most Conserved
CNS
Old World Cons
37 Species CFTR

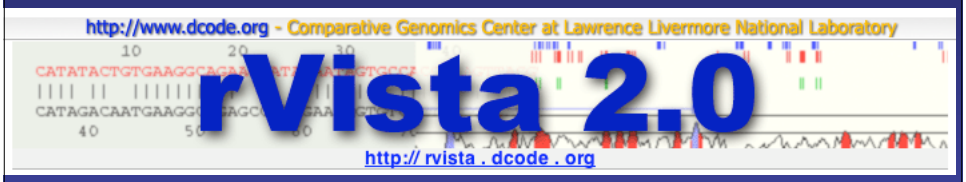
Binding Site Prediction

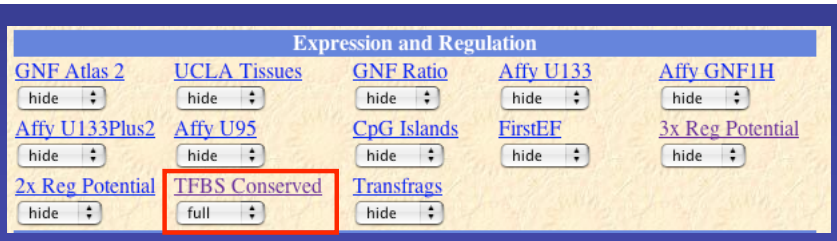
http://zlab.bu.edu/zlab/gene.shtml

- Regulatory Networks
- Clusters of binding sites
- High-througpout approaches

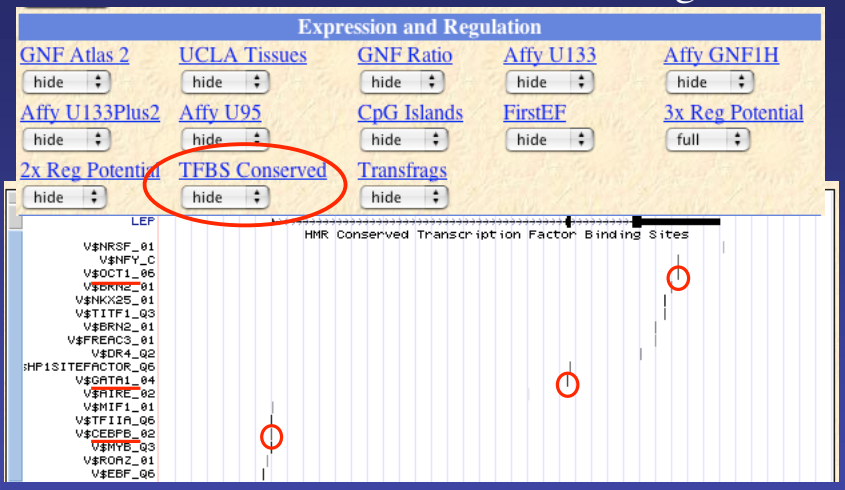
•Precomputed, Genome-Wide Analyses
(rVista/zPicture, Genome Browser, Table-Browser/GALA)

Visualizing Binding Sites on Alignments as Genome Annotations

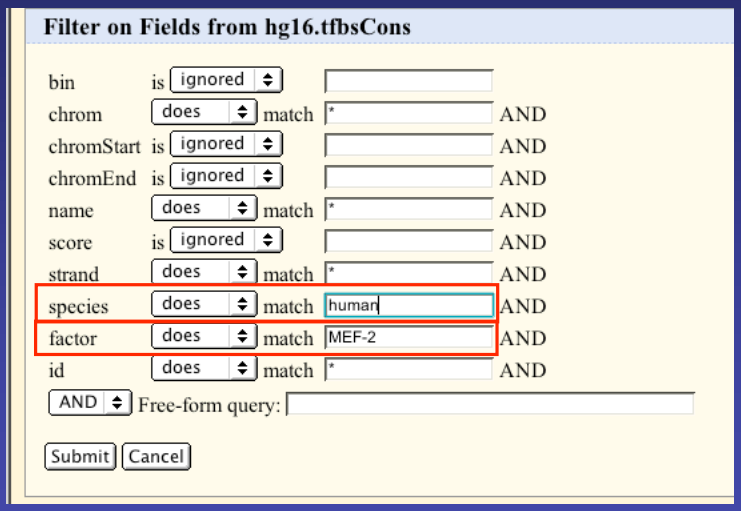




Genome Browser Conserved Binding Sites



Filtering the dataset



Filtering on Binding Sites

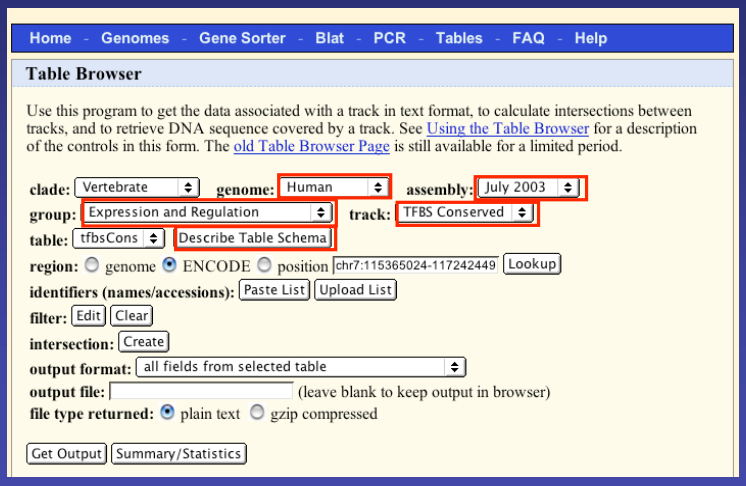
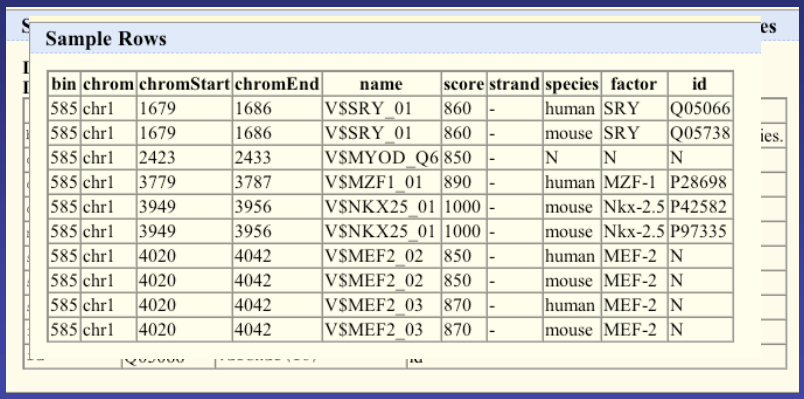


Table Browser - Table Schema



Sample Rows

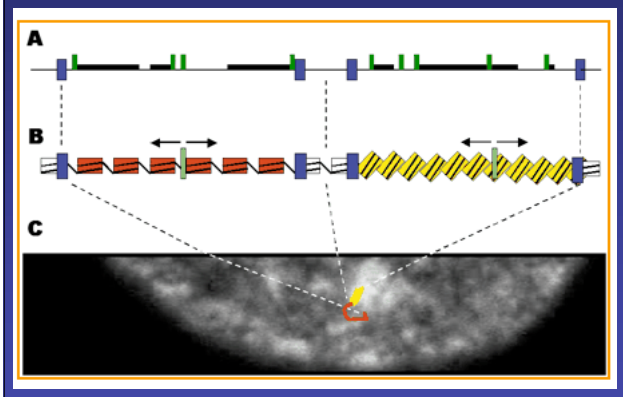
bin	chrom	chromStart	chromEnd	name	score	strand	species	factor	id
585	chr1	1679	1686	V$SRY_01	860	-	human	SRY	Q05066
585	chr1	1679	1686	V$SRY_01	860	-	mouse	SRY	Q05738
585	chr1	2423	2433	V$MYOD_Q6	850	-	N	N	N
585	chr1	3779	3787	V$MZF1_01	890	-	human	MZF-1	P28698
585	chr1	3949	3956	V$NKX25_01	1000	-	mouse	Nkx-2.5	P42582
585	chr1	3949	3956	V$NKX25_01	1000	-	mouse	Nkx-2.5	P97335
585	chr1	4020	4042	V$MEF2_02	850	-	human	MEF-2	N
585	chr1	4020	4042	V$MEF2_02	850	-	mouse	MEF-2	N
585	chr1	4020	4042	V$MEF2_03	870	-	human	MEF-2	N
585	chr1	4020	4042	V$MEF2_03	870	-	mouse	MEF-2	N

- Descriptions of all information relating to the track
- Sample data

Experimental Approaches

- Inference
  - Altered chromatin structure
- Verification
  - Protein binding sites
- Prediction
  - High throughput assays

Levels of Gene Regulation



Linear Sequence

Chromatin Conformation

Nuclear Location

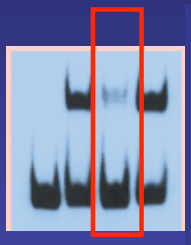
Roel van Driel Journal of Cell Science: 4067-4075 (2003)

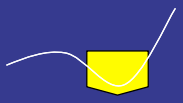
Detecting Protein Binding

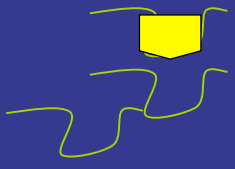
Gel Shift / EMSA (Electrophoretic Mobility Shift Assay)

Purified Proteins
Nuclear Extracts

Free probe





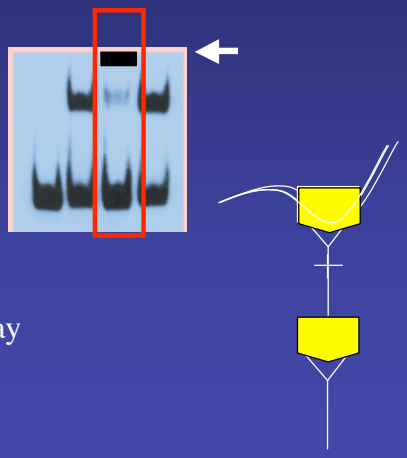


Nuclear Extracts

- Competition Assay
- Supershift assay

Detecting Protein Binding

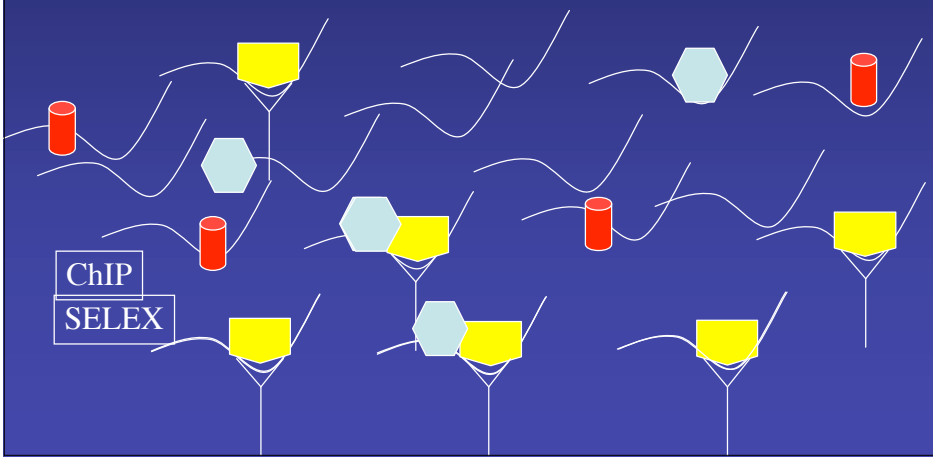
Gel Shift / EMSA (Electrophoretic Mobility Shift Assay)



- Supershift assay

Identifying Preferred Motifs

- Chromatin immunoprecipitation "ChIP"
- SELEX (Casting)



PWM

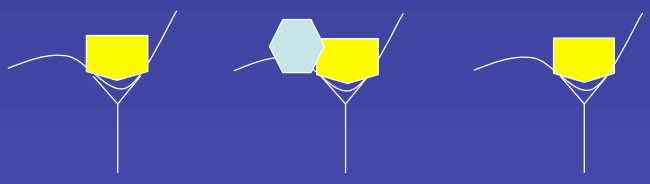
BIOBASE Biological Databases/Biologische Datenbanken
TRANSFAC - PathoDB
TRANSPATH - S/MARt DB
TRANSCompel - Cytomer

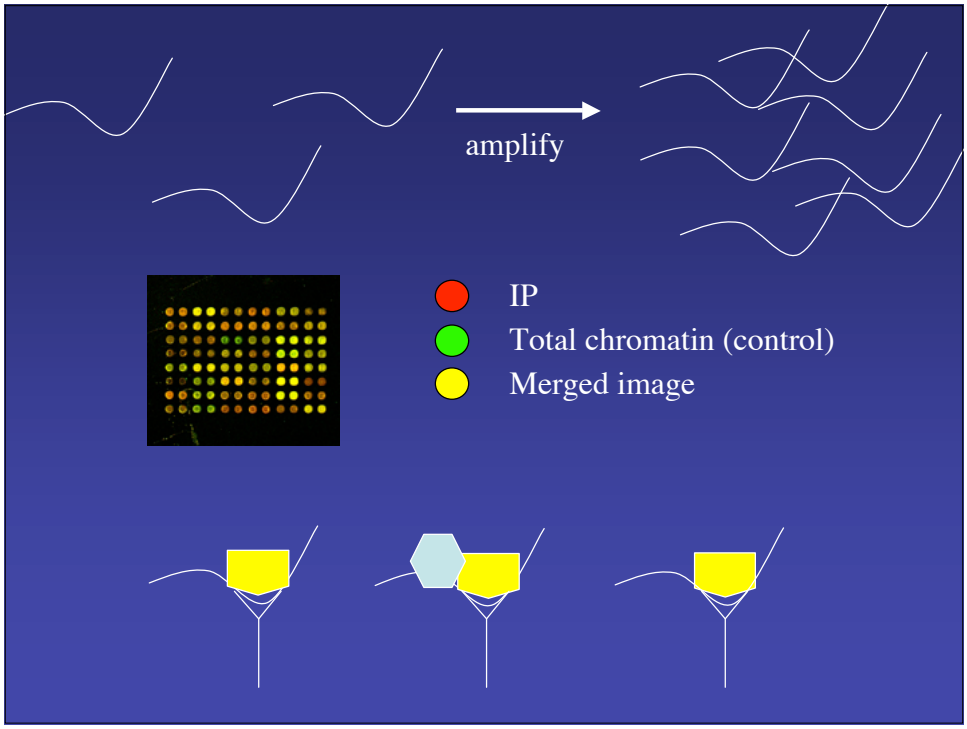
Gene Regulation http://www.gene-regulation.com/

PO	A	C	G	T	
01	6	0	16	3	G
02	2	0	24	0	G
03	1	0	0	25	T
04	2	0	0	24	T
05	25	0	0	1	A
06	20	2	1	3	A
07	2	0	0	24	T
08	8	3	8	7	N
09	16	0	1	9	W
10	3	1	1	21	T
11	1	4	0	21	T
12	16	1	6	3	A
13	9	11	3	3	M
14	7	14	0	5	M
15	11	7	4	3	N
XX					
BA	26 binding sites from 20 genes				

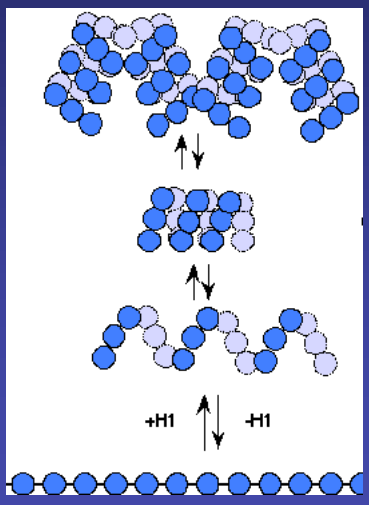
High Throughput Binding Assays

- ChIP-chip
  chromatin immunoprecipitation using DNA microarray chips
  - Detects Protein-DNA Interactions (Direct)
  - Protein-Protein Interactions (Indirect)





Chromatin Compaction



solenoid

- Transcriptionally inactive
- Nuclease insensitive

~ 30 nm fiber

- Decreasing levels of compaction

Beads on a string

10 nm fiber, with exposed linker region

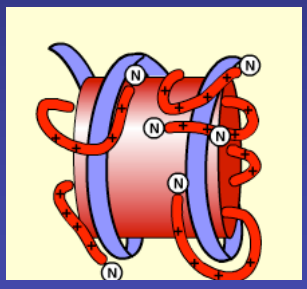
Detectable Histone Modifications

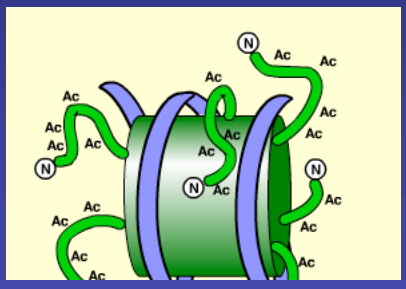
Gross rearrangements : movement or removal

Detected by DNAse I sensitivity

Subtle alterations: methylation, actylation, phosphorylation

Detected by ChIP assays

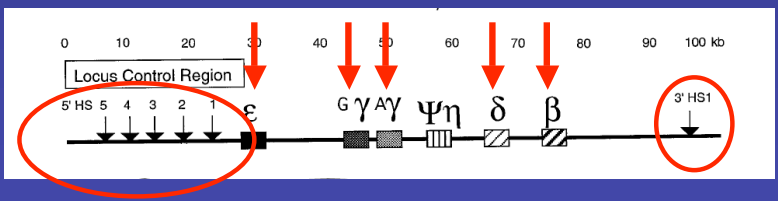




http://www-schreiber.chem.harvard.edu/home/animation.html

Chromatin and Gene Expression

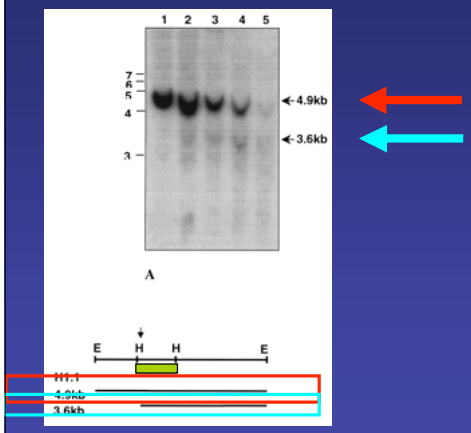
- Changes in chromatin conformation can be measured by accessibility of DNA to cleavage by DNase I
- DNase I *hypersensitivity* is an indicator of relaxed chromatin
- Often seen prior to detectable gene expression

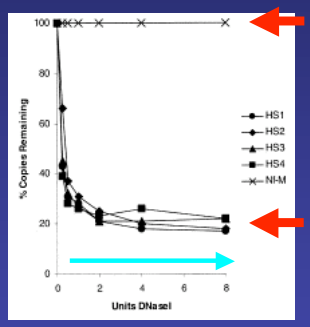


DNAse I Assays

- Targeted Loci
  - *In vitro* and *in vivo*
    - Indirect end-labeling/ Southern Blot
    - Quantitative PCR
- High-Throughput Analyses
  - Cloning endpoints of enzymatic cleavage

DNAse I Assays

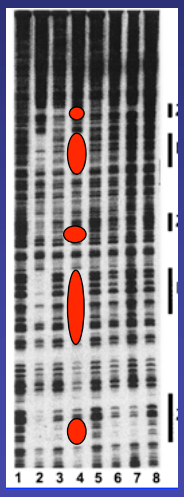




Indirect end-labeling/ Southern Blot

Quantitative PCR

DNAse I Footprinting

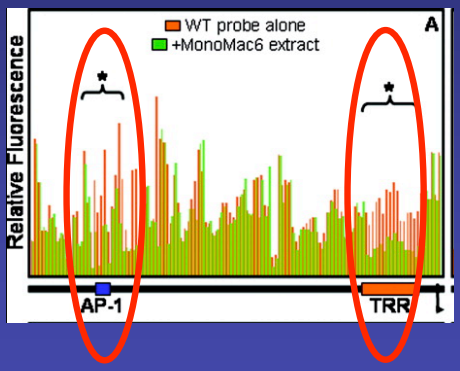


Identifies regions inaccessible to the enzyme that cleaves the DNA

"DNAse Protection"

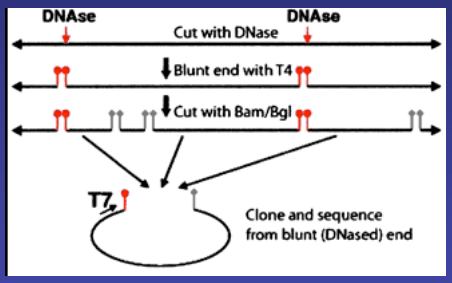
Advancing Technologies

Automated sequencing machines for hypersensitivity assays



Base pair resolution

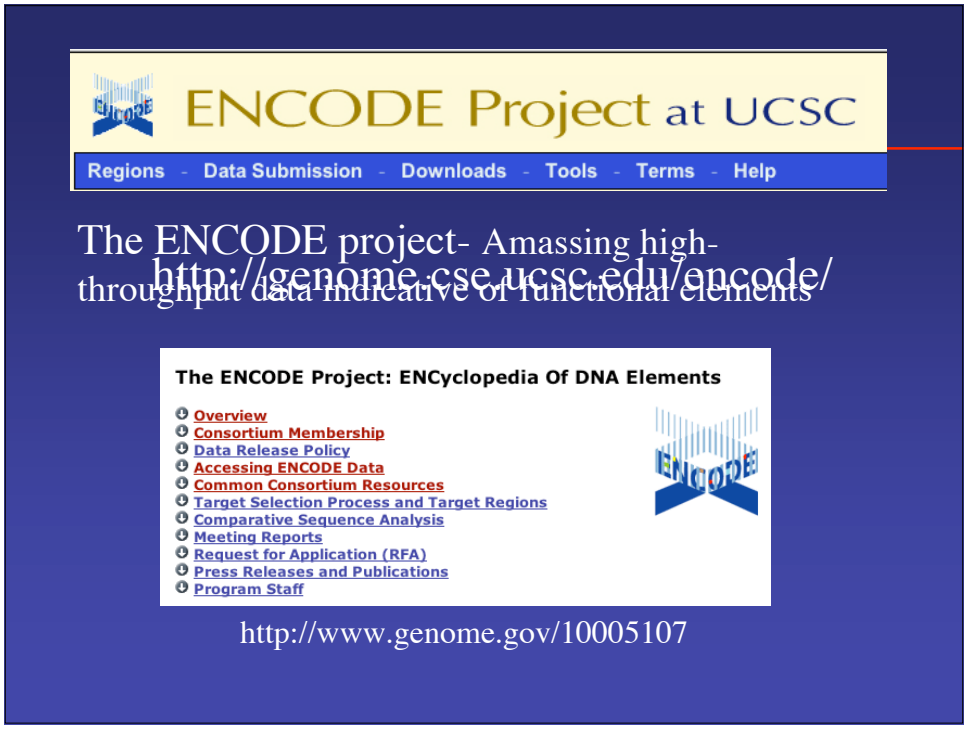
High Throughput DNAse I Hypersensitivity



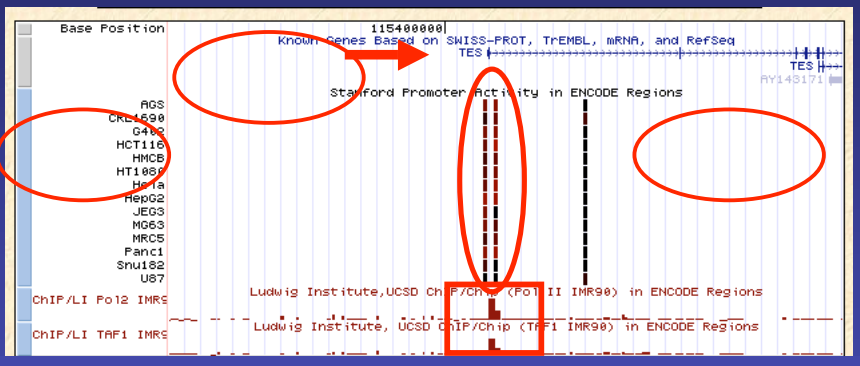
Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites.
Crawford et al. (2004) Proc Natl Acad Sci U S A. 101:992-7.

Characteristics Implicating Function

- Location in genome
- Conservation in a multiple sequence alignment
- Predictive tracks
- Collections of Protein binding sites



Features that identify promoters



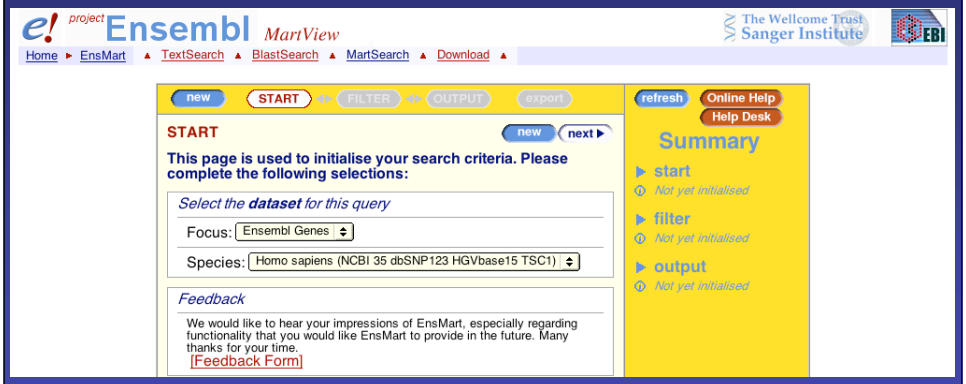
DNAse I hypersensitivity ?

Integrative Data Management

Servers providing tools to:

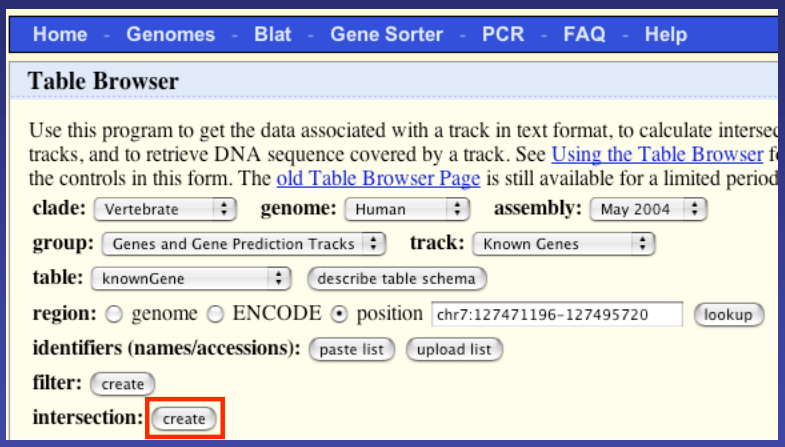
- Compare and contrast annotation tracks
- Include or exclude features on command
- Progressively refine the search criteria

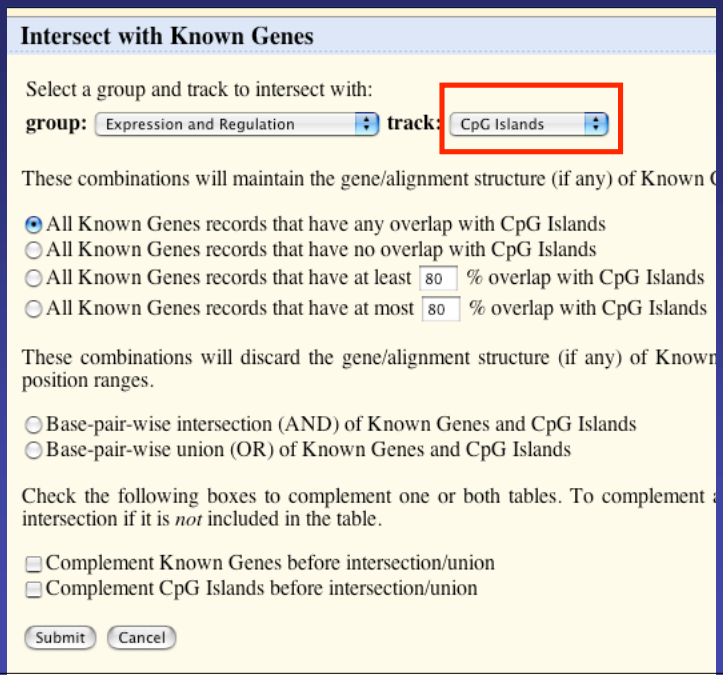
EnsMart Genome Data Mining Tool



http://www.ensembl.org/Multi/martview

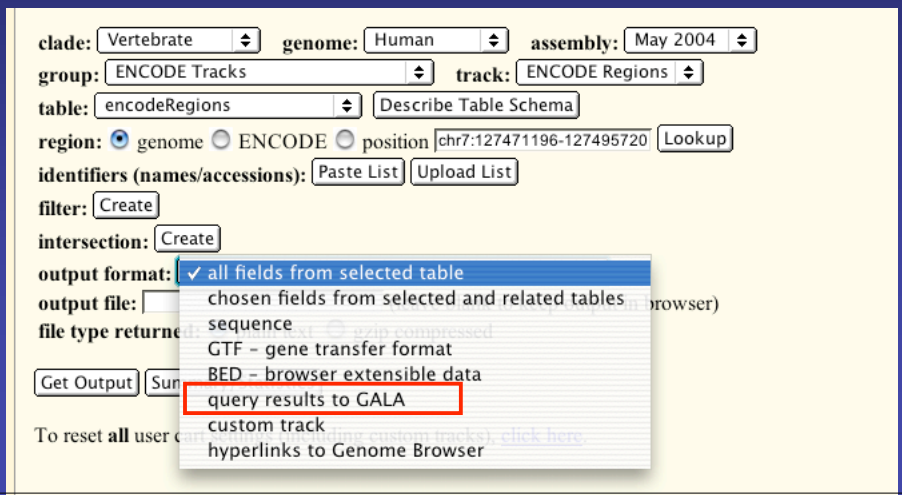
Compound Queries with Table Browser

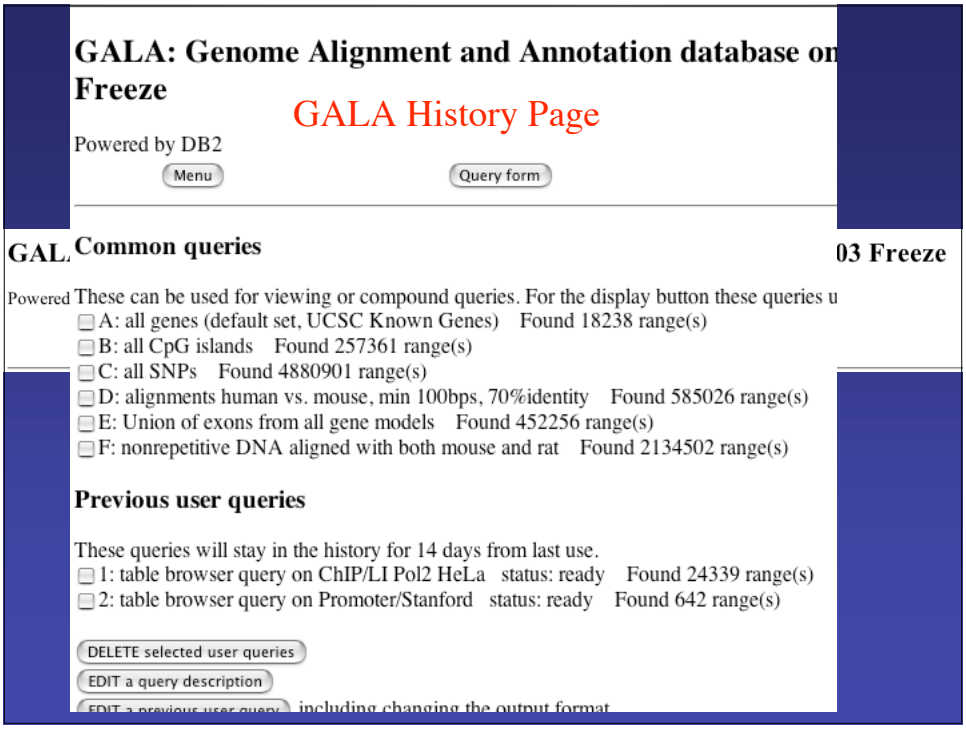




Additional Query/Analysis Tools

GALA Database





GALA Compound Queries

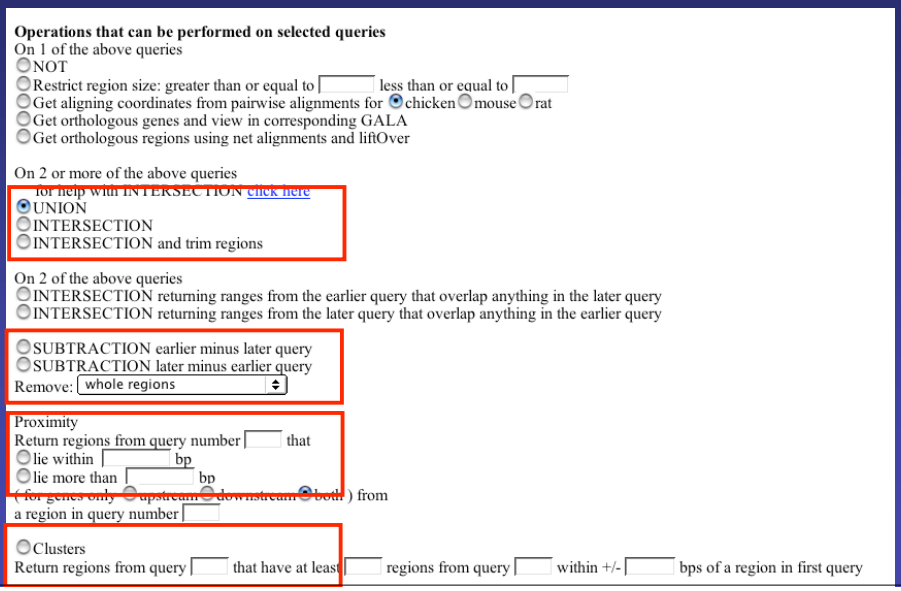
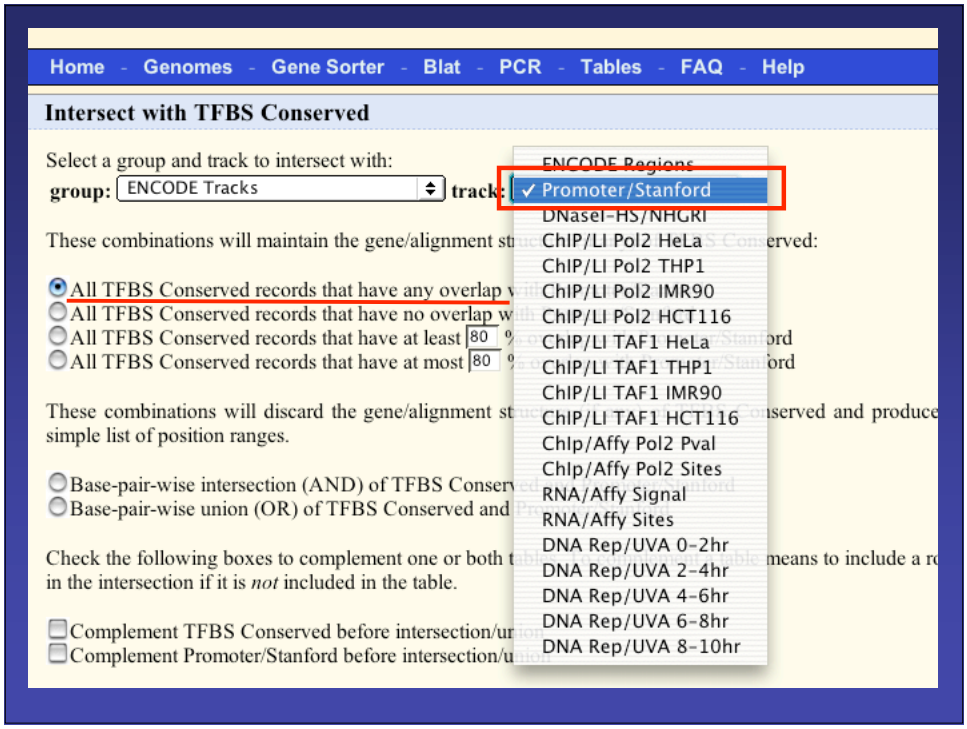


Table Browser Query

Request:

Identify promoters that are regulated by muscle-specific factors

- Of all functional ENCODE promoters (Stanford Promoters Track)
- *How many correspond to conserved regions in mammals?*
- *How many have a conserved MEF-2 site?*


Home - Genomes - Gene Sorter - Blat - PCR - Tables - FAQ - Help
Intersect with TFBS Conserved
Select a group and track to intersect with:
group: ENCODE Tracks
track:
ENCODE Regions
Promoter/Stanford
DNaseI-HS/NHGRI
ChIP/LI Pol2 HeLa
ChIP/LI Pol2 THP1
ChIP/LI Pol2 IMR90
ChIP/LI Pol2 HCT116
ChIP/LI TAF1 HeLa
ChIP/LI TAF1 THP1
ChIP/LI TAF1 IMR90
ChIP/LI TAF1 HCT116
Chip/Affy Pol2 Pval
Chip/Affy Pol2 Sites
RNA/Affy Signal
RNA/Affy Sites
DNA Rep/UVA 0-2hr
DNA Rep/UVA 2-4hr
DNA Rep/UVA 4-6hr
DNA Rep/UVA 6-8hr
DNA Rep/UVA 8-10hr
These combinations will maintain the gene/alignment st
All TFBS Conserved records that have any overlap
All TFBS Conserved records that have no overlap
All TFBS Conserved records that have at least 80
All TFBS Conserved records that have at most 80
These combinations will discard the gene/alignment st
simple list of position ranges.
Base-pair-wise intersection (AND) of TFBS Conserv
Base-pair-wise union (OR) of TFBS Conserved and
Check the following boxes to complement one or both t
in the intersection if it is not included in the table.
Complement TFBS Conserved before intersection/u
Complement Promoter/Stanford before intersection/u

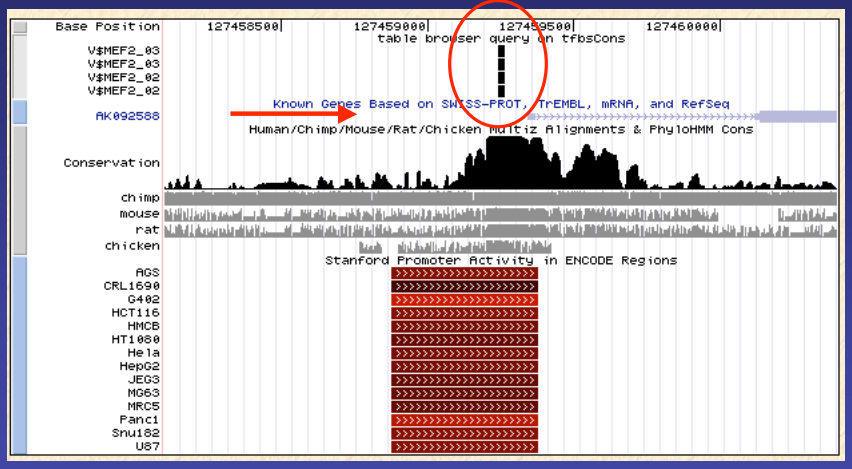

Base Position
127458500
127459000
127459500
127460000
table browser query on tfbsCons
V$MEF2_03
V$MEF2_03
V$MEF2_02
V$MEF2_02
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
AK092588
Human/Chimp/Mouse/Rat/Chicken Multiz Alignments & PhyloHMM Cons
Conservation
chimp
mouse
rat
chicken
Stanford Promoter Activity in ENCODE Regions
AGS
CRL1690
G402
HCT116
HMCB
HT1080
Hela
HepG2
JEG3
MG63
MRC5
Panc1
Snu182
U87

Table Browser - GALA Query

Request:

Identify promoters that are regulated by muscle-specific factors

- Of all functional ENCODE promoters (Stanford Promoters Track)
- *How many correspond to conserved regions in mammals?*
- *How many have a conserved MEF-2 site?*
- *Can we identify clusters of binding sites in active promoters?*

GALA Query

Rationale: Identify clusters of MEF-2 and MYOD binding sites in promoters analyzed by the Stanford Group

Expression and Regulation
Transcription factor binding sites
The transcription factor binding sites were produced with tffind , tfloc, and TRANSFAC (free registration requir

Query ☐ all TF binding sites (interval is required)
☐ only binding sites conserved in hg16Gg2 cutoff used was 0.85
☑ only binding sites conserved in hg16Mm3Rn3 cutoff used was 0.85
☐ only binding sites conserved in hg16Pt1Mm3Rn3Gg2 cutoff used was 0.75

Previous user queries

These queries will stay in the history for 14 days from last use.
☐ 1: table browser query on ChIP/~~LI Pol2 HeLa~~ status: ready Found 24339 range(s)
☐ 2: table browser query on Promoter/Stanford status: ready Found 642 range(s)
☐ 3: (only binding sites conserved in hg16Mm3Rn3) AND (factor ID = V$MEF2_01, V$MEF2_02, V$MEF2_..) status: ready Found 2284 range(s)
☐ 4: (only binding sites conserved in hg16Mm3Rn3) AND (factor ID = V$MYOD_01, V$MYOD_Q6) statu
Found 158646 range(s)

☑ V$MEF2_01
☑ V$MEF2_02
☑ V$MEF2_03
☑ V$MEF2_04

☑ V$MYOD_01
☑ V$MYOD_Q6