

Comparative genomics tools for biological discovery

Inna Dubchak, Ph.D.

Staff scientist

Lawrence Berkeley National Laboratory

ildubchak@lbl.gov

Outline

What is comparative genomics?

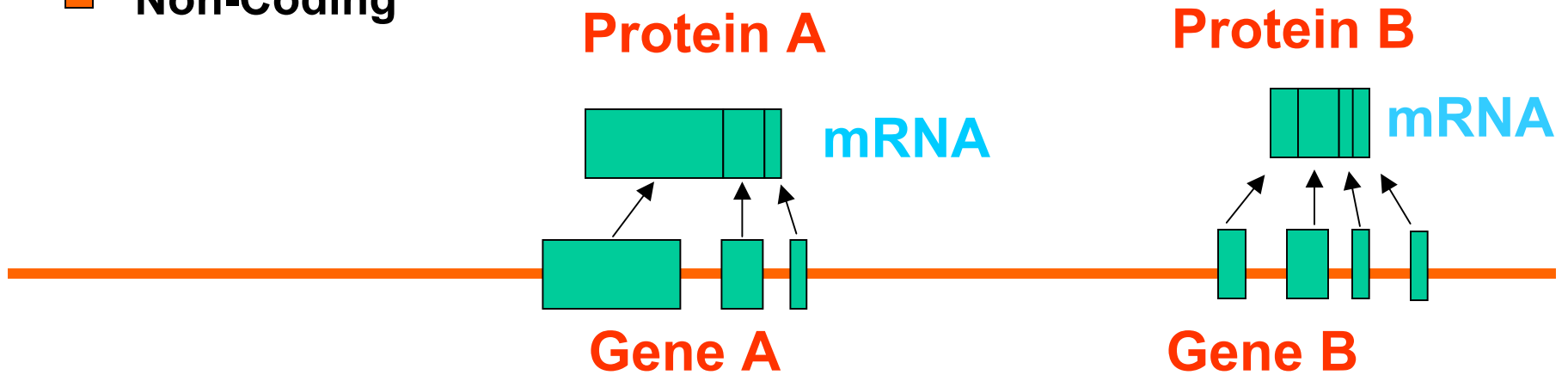
VISTA tools developed for comparative genomics.

Large scale VISTA applications including aligning whole genome assemblies

Related biological stories

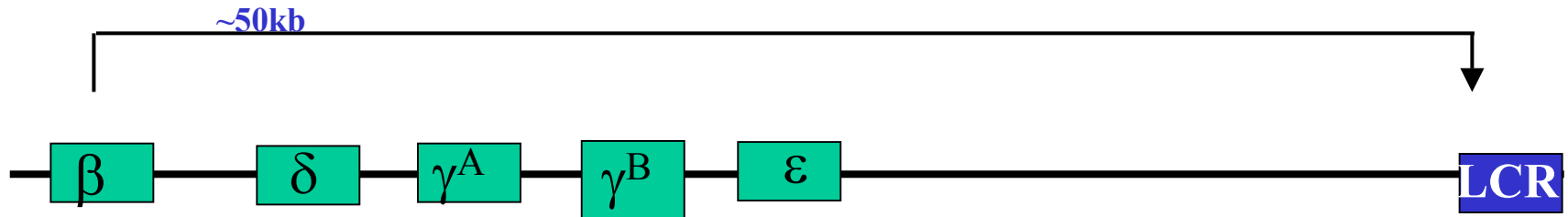
1-2% Coding

- Coding
- Non-Coding



Distant Non-Coding Sequences Causing Disease

β -Thalassemia

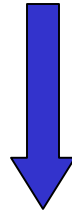


Disease	Gene	Distance
Campomelic displasia	SOX9	850kb
Aniridia	PAX6	125kb
X-Linked Deafness	POU3F4	900kb
Saethre-Chotzen syndrome	TWIST	250kb
Rieger syndrome	PITX2	90kb
Split hand/split foot malformation	SHFM1	450kb

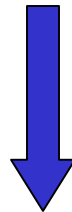
Background

Evolution can help!

In general, functionally important sequences are conserved

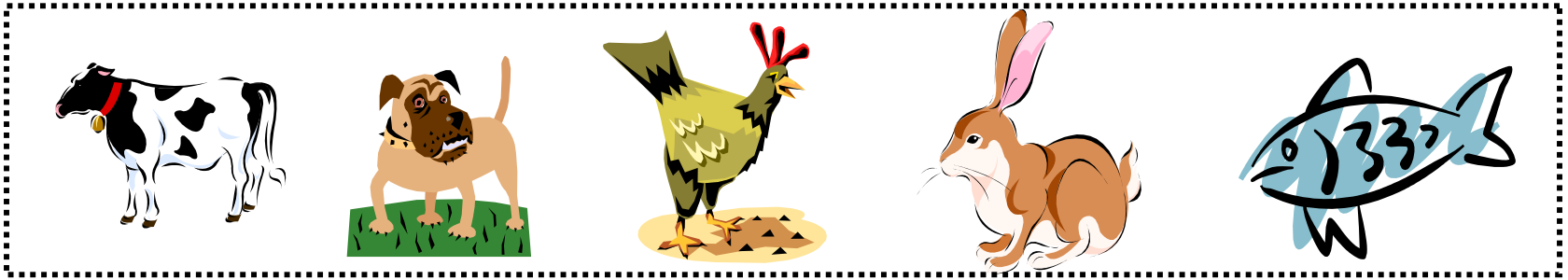


Conserved sequences are functionally important



Raw sequence can help in finding biological function

Comparing sequences of different organisms



- Helps in gene predictions
- Helps in understanding evolution
- Conserved between species non-coding sequences are reliable guides to regulatory elements
- Differences between evolutionary closely related sequences help to discover gene functions

Challenges

- Sequence at different stages of completion, difficult to compare

Whole genome shotgun → Partial Assemblies
Finished BACs

- Fast and accurate analysis
- Scaling up to the size of whole genomes

Identify evolutionarily related genomic sequences

Homologs - Orthologs - Paralogs



Annotate reference sequence

- Genic sequences - Repetitive elements - cpG islands



Align genomic sequences

- Global alignment program - Local alignment program



Identify conserved sequences

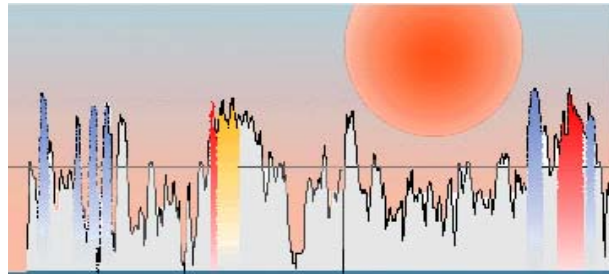
- Percent identity and length thresholds



Visualize conserved sequences

- Moving average point plot (VISTA)
- Gap-free segment plot (PipMaker)

<http://www-gsd.lbl.gov/vista>



VISUALIZATION TOOLS FOR ALIGNMENTS

VISTA

WELCOME to the homepage for VISTA, Visualization Tool for Alignments.

USE VISTA on the WEB

Vista -- [instructions](#) for using **VISTA**
aVista -- [instructions](#) for using **rVISTA**

DOWNLOAD VISTA

Vista Go to our [software download page](#) to obtain **VISTA**'s alignment and visualization programs.

INFORMATION about VISTA

Vista How to [cite](#) VISTA.
Send us your questions, comments

Vista

is an integrated computational system for global alignment and visualization, designed for comparative genomics. It allows for the visualization of long sequence alignments of DNA from two or more species with annotation information, and it was developed to locate conserved sequences in syntenic regions (Dubchak et al., 2000).

It has a clean output, allowing for easy identification of sequence similarities and differences, and is easily configurable, enabling the visualization of alignments of various lengths at different levels of resolution.

This system consists of several unified modules:

aVid

the program for global alignment of DNA sequences of arbitrary length. In addition to aligning two finished sequences, it can also handle one sequence in a non-ordered and non-oriented draft format [Details](#).

Vista

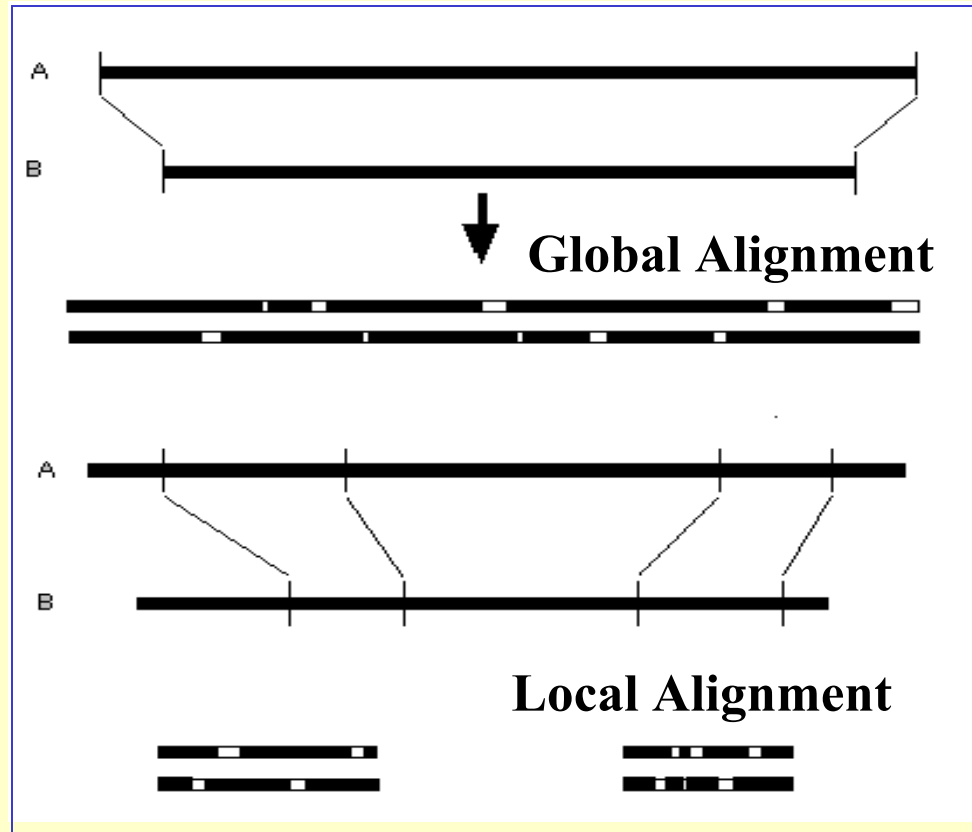
A computational tool for comparing an arbitrary number of genomic sequences from different species. [Details](#).

Processed ~ 16000 queries on-line, distributed > 700 copies of the program in 35 countries

Modules of VISTA:

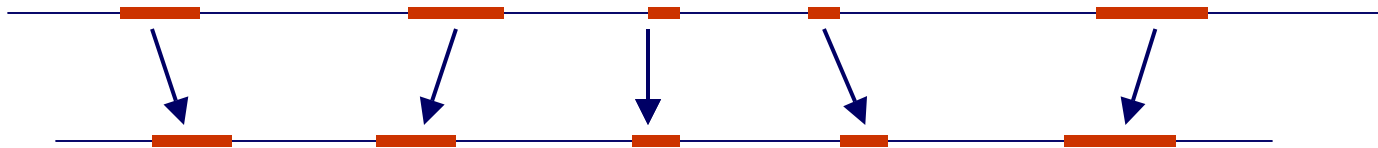
- Program for global alignment of DNA fragments of any length
- Visualization of alignment and various sequence features for any number of species
- Evaluation and retrieval of all regions with predefined levels of conservation

Local vs global alignment



AVID- the alignment engine behind VISTA

- **Very fast** global alignment of megabases of sequence.
- **Provides details** about ordered and oriented contigs, and accurate placement in the finished sequence.
- **Full integration** with repeat masking.



- **ORDER and ORIENT**
- **FIND** all common k-long words (k-mers)
- **ALIGN** k-mers scoring by local homology
- **FIX** k-mers with good local homology
- **RECURSE** with smaller k (shorter words)

Visualization



```
tggtaacattcaaattatg-----ttctcaaagtgagcatgaca-acttttttccatgg  
|| | |||| | | || | | | | | | | | | | | | | |  
tgatgacatctatgtgctgtttccttttagaaactgcatgagagcctggctagtaggg
```



Window of length L is centered at a particular nucleotide in the base sequence

Percent of identical nucleotides in L positions of the alignment is calculated and plotted

Move to the next nucleotide

Exons file

```
> 12877 289557 ST7b/a
+ 13076 282515
12877 13226
159297 159379
179096 179255
189328 189382
190026 190141
191420 191495
193659 193727 b only
195616 195770
197970 198067
230397 230511
248856 248928
250369 250471
269322 269472
278619 278711
281458 281597
282396 283253 3'b
289297 289557 3'a
```

Sequences in FastA format

```
> Human ST7 gene
CTGAATGGCTCGTAGAAA
TATTGCATTAACCTGCTG
GACATGCTGAATAGCAAT
CGACTACAGT. . . .
```

```
> Cow ST7 gene
CTGAATGGCTCGTAGAAA
TAATGCATTCCCCTGCTG
GACATGCTGAATAGCAAT
CGACTACAGT. . . .
```

Repeat Masker

VISTA

Alignment files

```
185140 185150 185160 185170 185180
GACATTGGAAAAGTAAAGGAAGTGGTTTAT---CTTGCTC-----TTTTTGC AACAGTA
||||| ||||||| | ||||||| ||| | ||||| |||
GACACTGGAAAAGCAGAGGAAGTGGTTTATTGACCTGCCCCCCCTTTTTTATAACAGTG
```

Conservation files

```
80078 (149626) to 80171 (149724) = 99bp at 63.6% noncoding
159297 (158141) to 159379 (158223) = 83bp at 80.7% exon
179096 (159067) to 179253 (159224) = 158bp at 75.9% exon
189328 (159566) to 189382 (159620) = 55bp at 81.8% exon
190026 (159996) to 190139 (160109) = 114bp at 80.7% exon
191420 (160192) to 191495 (160267) = 76bp at 73.7% exon
```

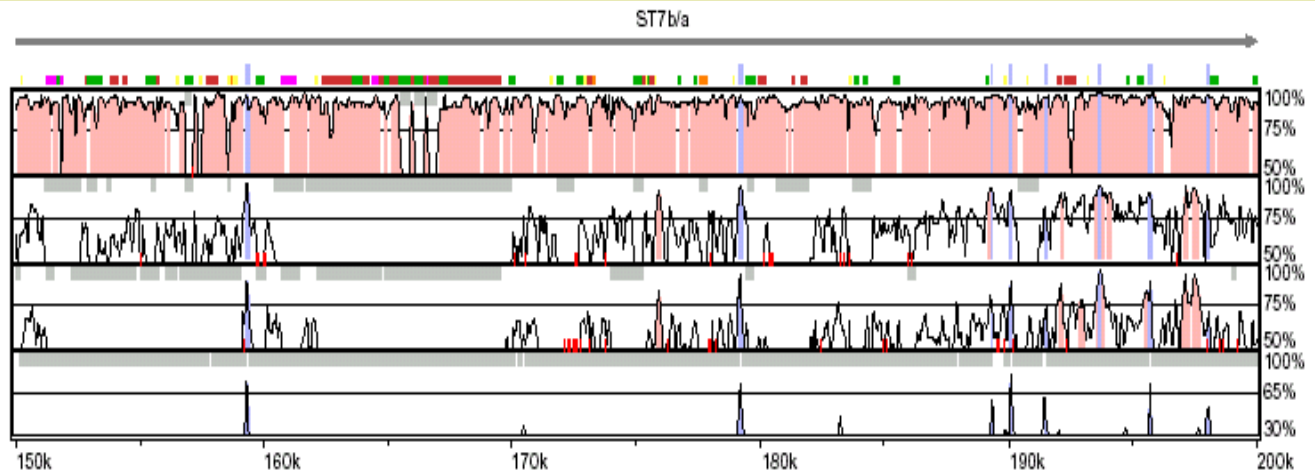
VISTA file

```
Alignment 1
Seqs: human/baboon
Criteria: 95%, 100 bp
Regions: 817

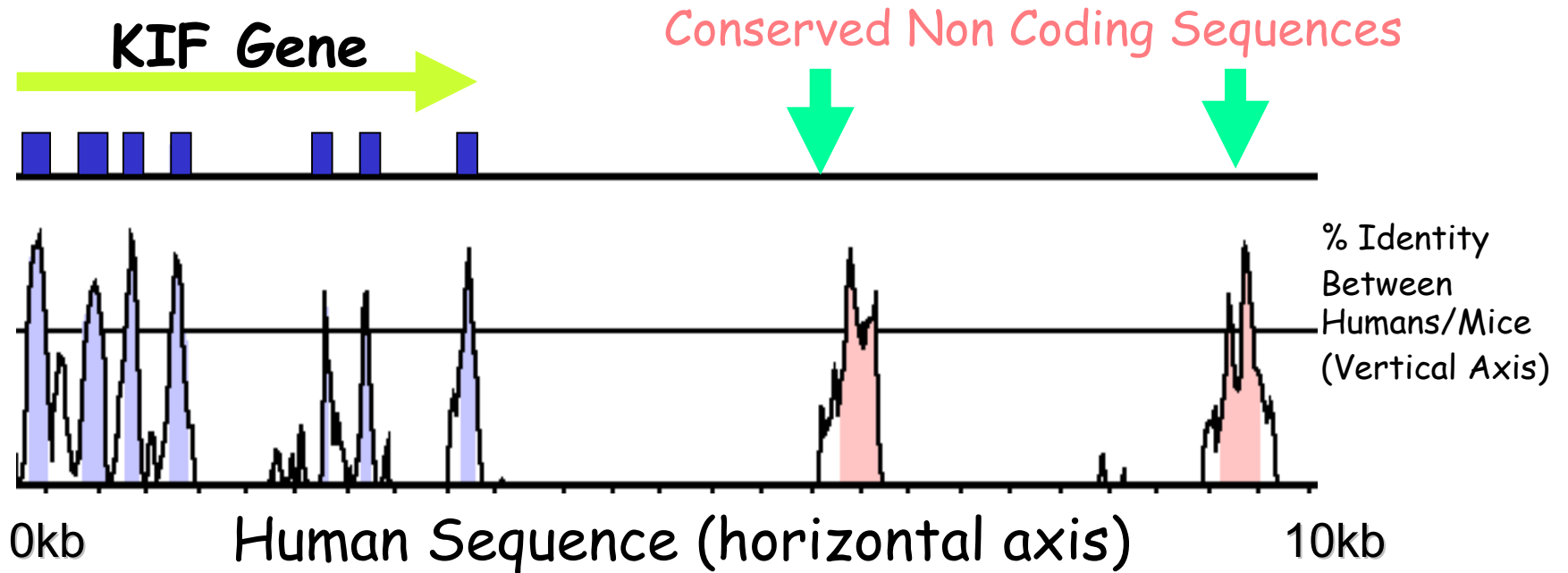
Alignment 2
Seqs: human/cow
Criteria: 93%, 100 bp
Regions: 49

Alignment 3
Seqs: human/mouse
Criteria: 86%, 100 bp
Regions: 51

Alignment 4
Seqs: human/fugu
Criteria: 55%, 100 bp
Regions: 15
```



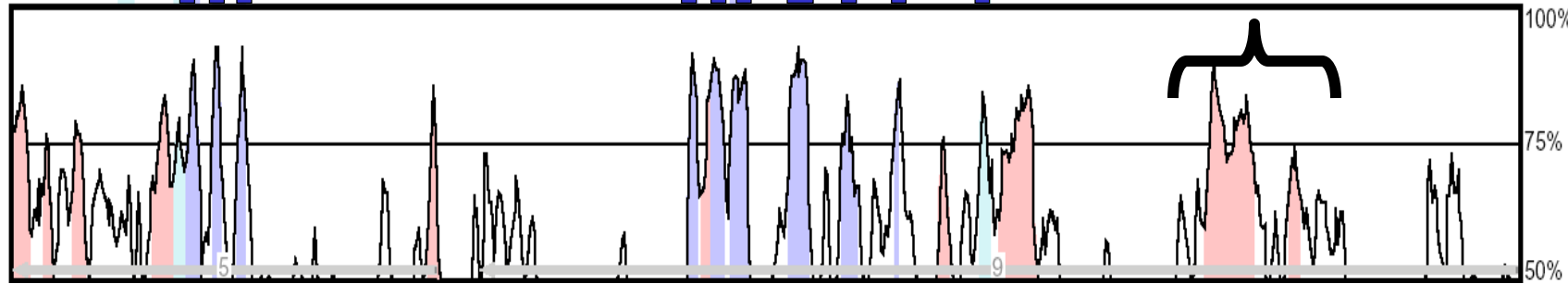
VISTA plot



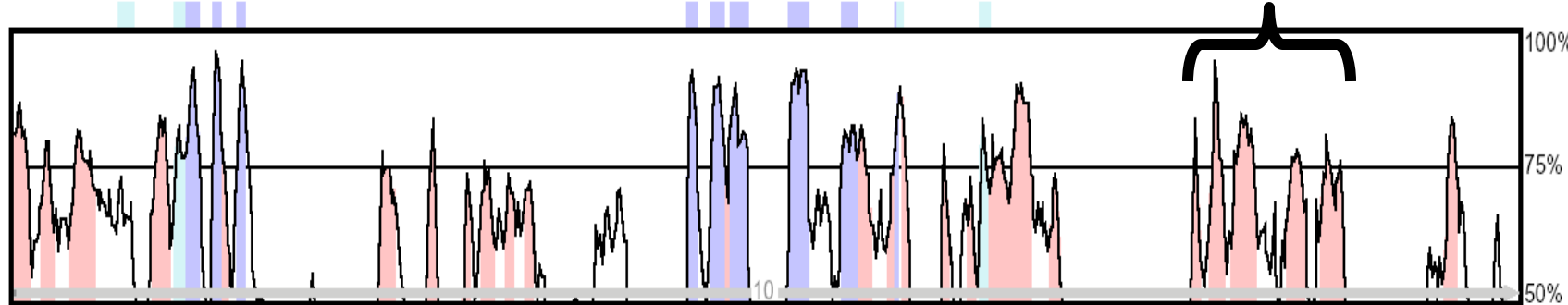
Nuclear Hormone Receptor:LXR-Alpha



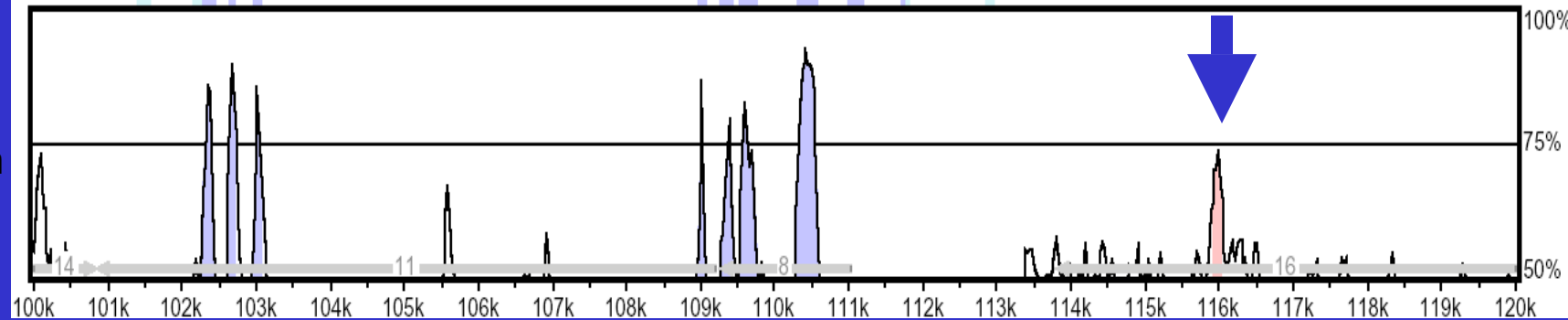
Human/
Mouse



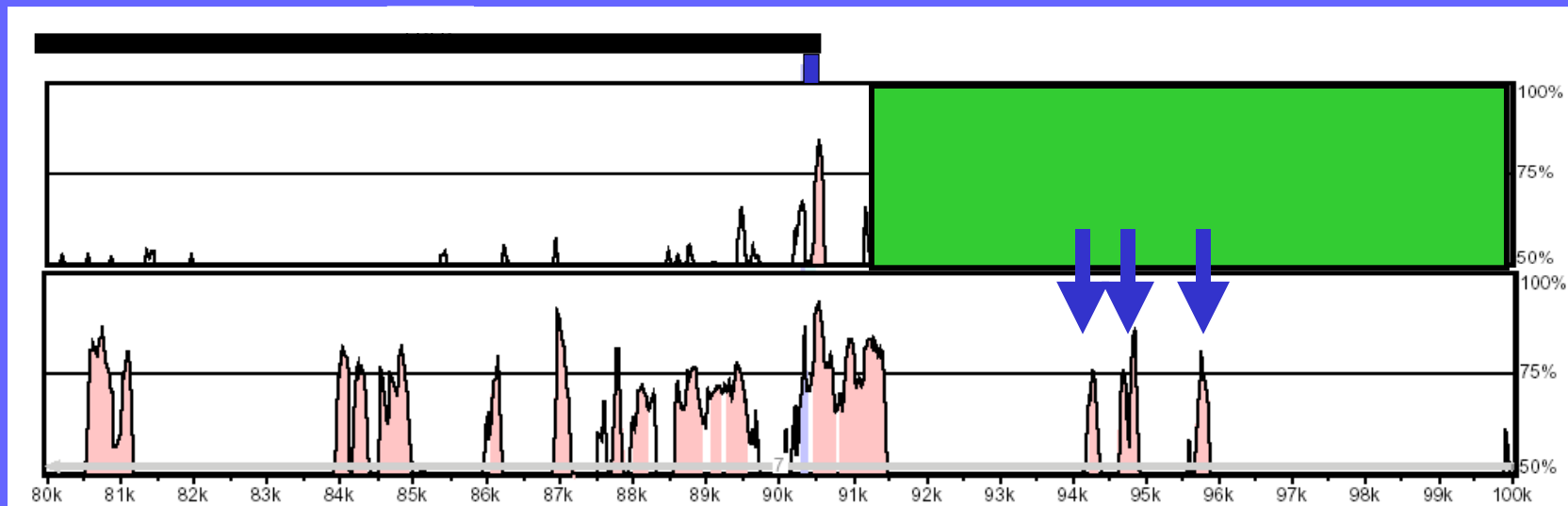
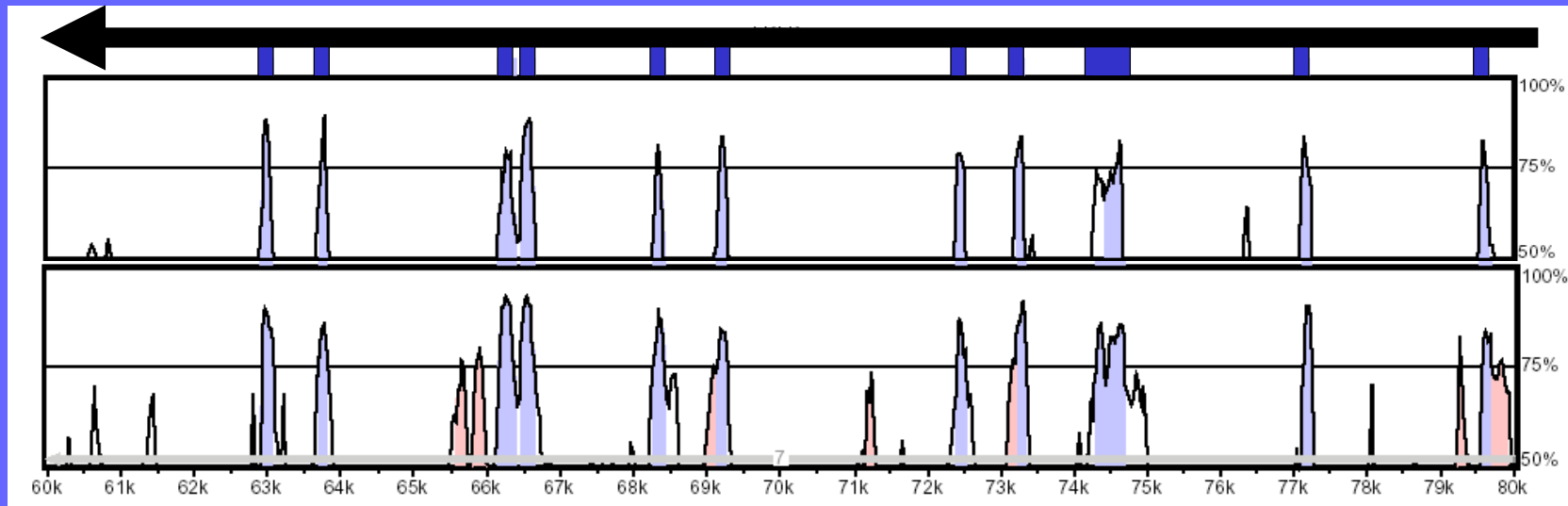
Human/
Rabbit



Human/
Opossum

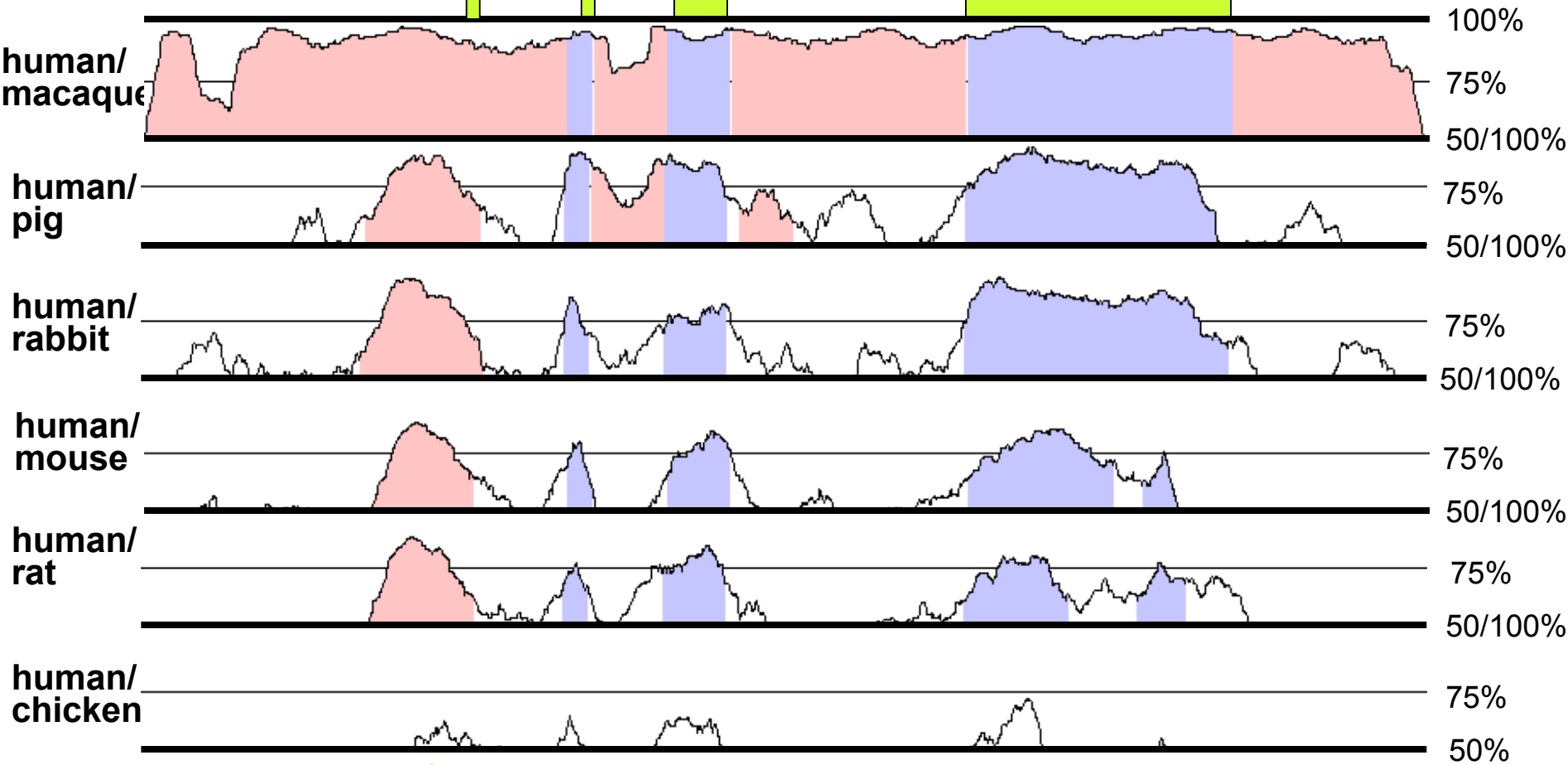
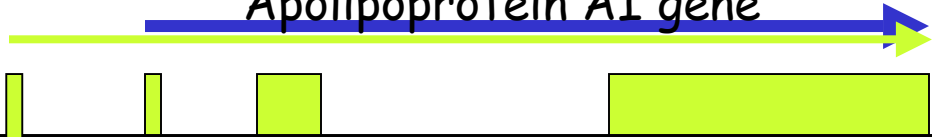


Low-Density Lipoprotein Receptor (LDLR)



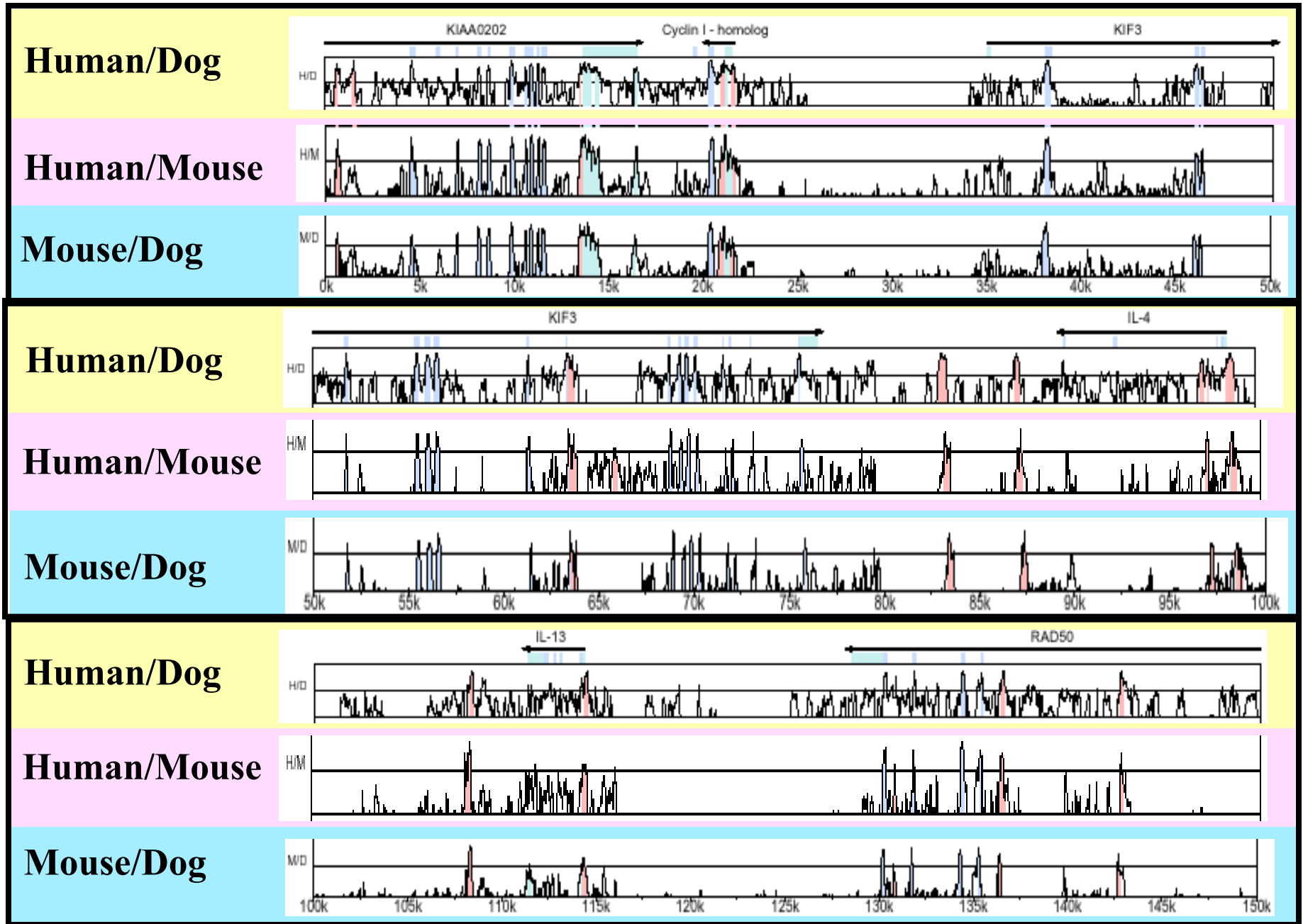
Multi-Species Comparative Analysis (VISTA)

Apolipoprotein AI gene

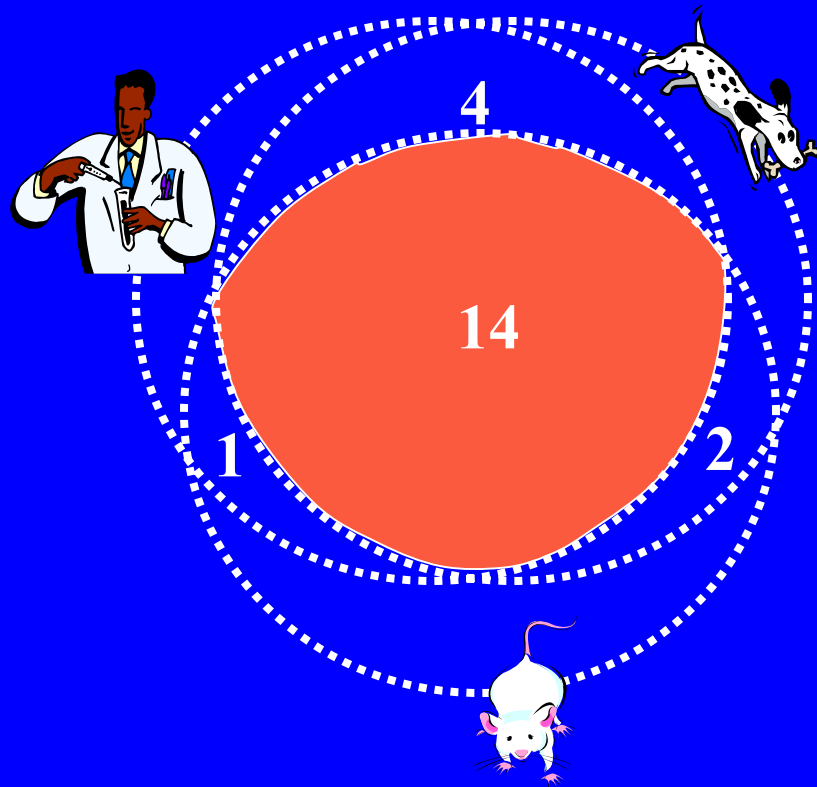


Liver enhancer

Example: Dubchak et al., 2000, *Genome Research*, 10: 1304-1306.



Active conservation of noncoding sequences - present in more than two mammals



% Cutoff

sum of three pair wise
Intersection/Union
values is maximal

Over 120 basepairs:

$H/D > 92\%$

$H/M > 80\%$

$D/M > 77\%$

VISTA flavors

- **VISTA** - comparing DNA of multiple organisms
- **for 3 species** - analyzing cutoffs to define actively conserved non-coding sequences
- **cVISTA** - comparing two closely related species
- **rVISTA** - regulatory VISTA

Main features of VISTA

- Clear , configurable output
- Ability to visualize several global alignments on the same scale
- Alignments up to several megabases
- Working with finished and draft sequences
- Available source code and WEB site

THE BERKELEY GENOME PIPELINE

[FINISHED ANALYSIS](#) [ASSEMBLY ANALYSIS](#) [VISTA BROWSER](#) [VISTA TRACK](#) [MYGODZILLA SERVER](#) [SOFTWARE](#) [LINKS](#) [CONTACT INFO](#)

Godzilla - automatic computational system for comparative analysis of genomes

<http://pipeline.lbl.gov>

<http://www-gsd.lbl.gov/vista>

DATA

Base Human Genome - Golden Path Assembly

Mouse assemblies:

Arachne	October 2001
Phusion	November 2001
MGSC v3	April 2002

Main modules of the system

Mapping and alignment of mouse contigs against the human genome

```
graph TD; A[Mapping and alignment of mouse contigs against the human genome] --> B[Visualization]; A --> C[Analysis of conservation];
```

Visualization

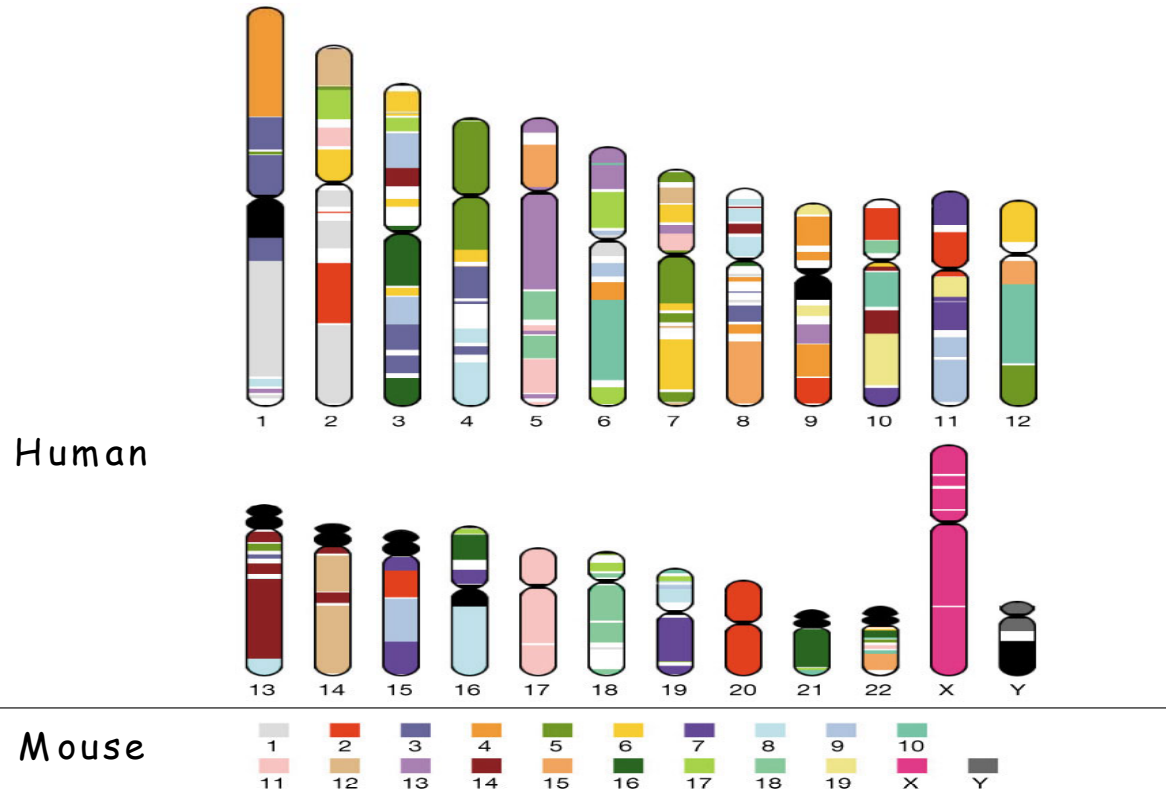
Analysis of conservation



Linux cluster with
15 1.2GHz PC,
750Mb of RAM

Three days to align
the entire mouse
genome against the
human genome

Chromosome Comparison



Base pair alignment

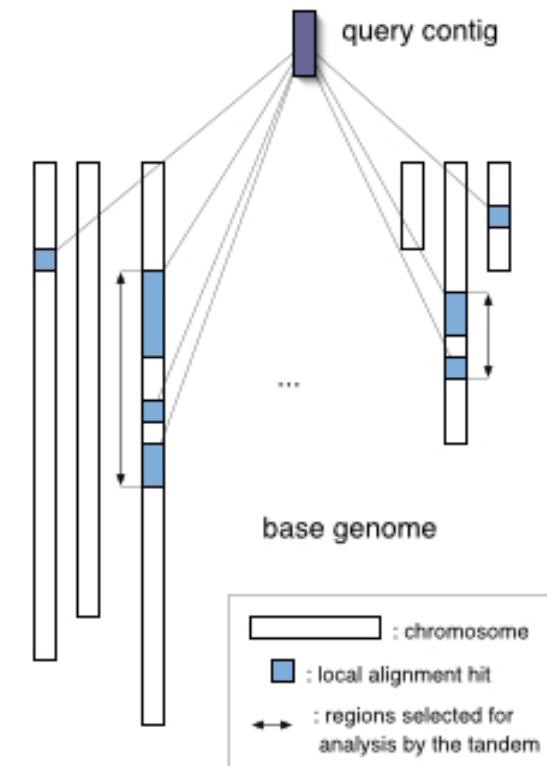
```

247 GGTGAGGTCGAGGACCCTGCA  CGGAGCTGTATGGAGGGCA  AGAGC
    |:  ||  ||||:  ||||  --:||  |||  |::|  |||---|||
368 GAGTCGGGGGAGGGGGCTGCTGTTGGCTCTGGACAGCTTGCATTGAGAGG
  
```

Tandem Local/Global Alignment Approach

Sequence fragment **anchoring** (DNA and/or translated BLAT)

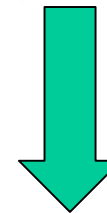
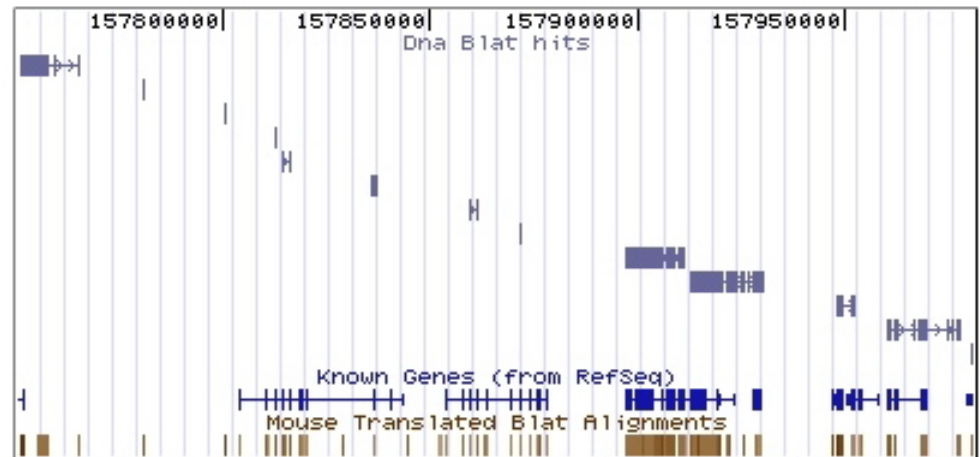
Multi-step verification of potential regions using global alignment (AVID)



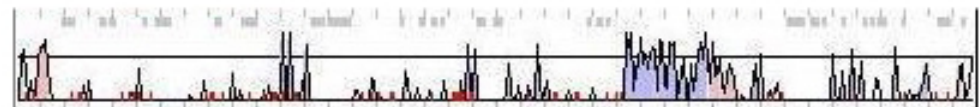
ANCHORS FROM LOCAL ALIGNMENT HITS



LOCAL ALIGNMENT HITS



GLOBAL ALIGNMENT



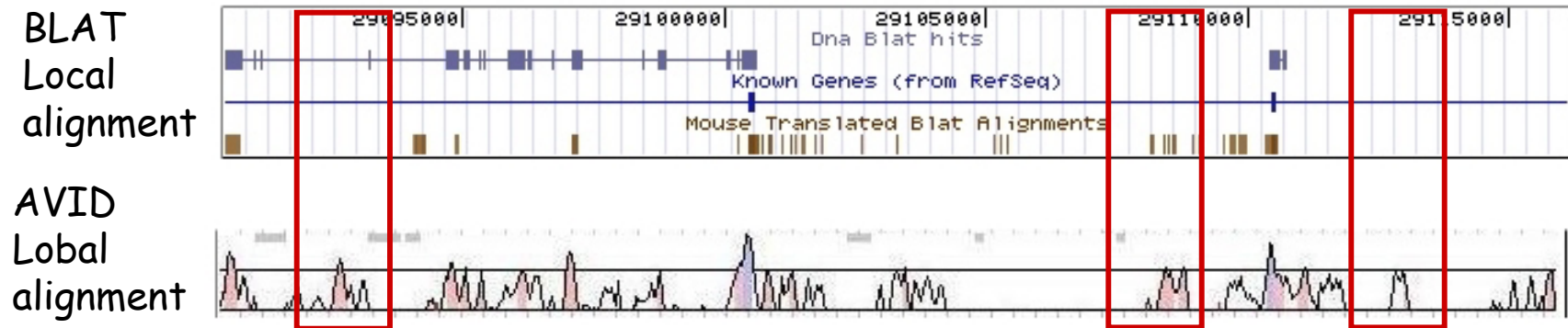
Advantage of the tandem approach:

better sensitivity/specificity trade-off

fill-in effect

scoring longer alignments

NT_002606 at Chr.17:2909457-29116113



Alignment strategies for different types of assemblies.

Method	Scheme of alignment	Examples
Contigs	Individual contigs	Finished BACs
Scaffold	contigs can be reoriented and reordered	Arachne October 2001 Phusion November 2001
Chopped pieces	mouse chromosomes are chopped in 250 kb and aligned to the Human Genome	Celera chromosome 16 MGSC v3

Visualization - VistaBrowser & VistaTrack

THE BERKELEY GENOME PIPELINE

GODZILLA

[MOUSE ANALYSIS](#)

[--FINISHED](#)

[--ASSEMBLY](#)

[VISTA BROWSER](#)

[VISTA TRACK](#)

[MYGODZILLA](#)

[SOFTWARE](#)

[LINKS](#)

[CONTACT INFO](#)

Stand-alone Java applet for detailed comparison

Comparison combined with the human genome annotation on the UCSC Human Genome Browser

VistaBrowser

MGSCv3 - Human June 2002

MB 10 20 30 40 50 60 70 80 90 100 110 120 130

107 kbp

gene
exon
UTR
CNS
gap in base seq
Contig
Contig overlap

Repeats:
LINE
LTR
SINE
RNA
DNA
Other

Pointer At:
Contig:

<< < > >>
Zoom out 2x Zoom in 2x

Position
Gene
Chromosome 9
Start 98344084
End 98451720

VistaTrack
Contig Details
Settings...
Reload Help

Vista Genome Br...
Parameters
Min Cons Width (bp) 100
Calc Window (bp) 100
Cons Identity (%) 75
Display
Graph Max (%) 100
Graph Min (%) 50
Rows Auto
 Show Genes
 Show Contigs
 Show Repeats
 Paint Under Curve

VistaTrack

DNA

Tables

Convert

Ensembl

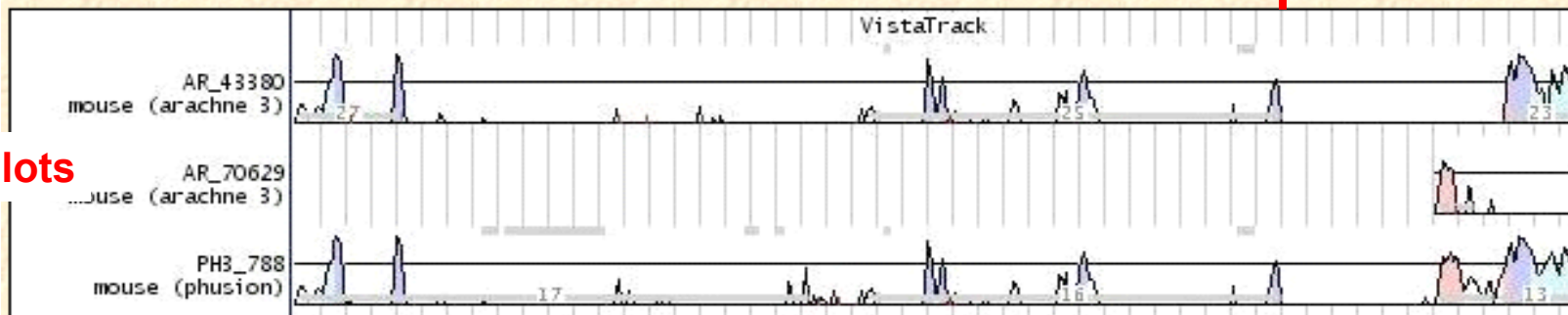
VistaTrack G

ent, we have installed here a mirror of the UCSC Genome Browser (v.8). All the data are mirrored from the UCSC site. A SLAM track has been added, as v
ents track. The Browser has been modified to add a Vista Track (examples). Currently, Vista Track displays the results of the Berkeley Genome Pipeline (G

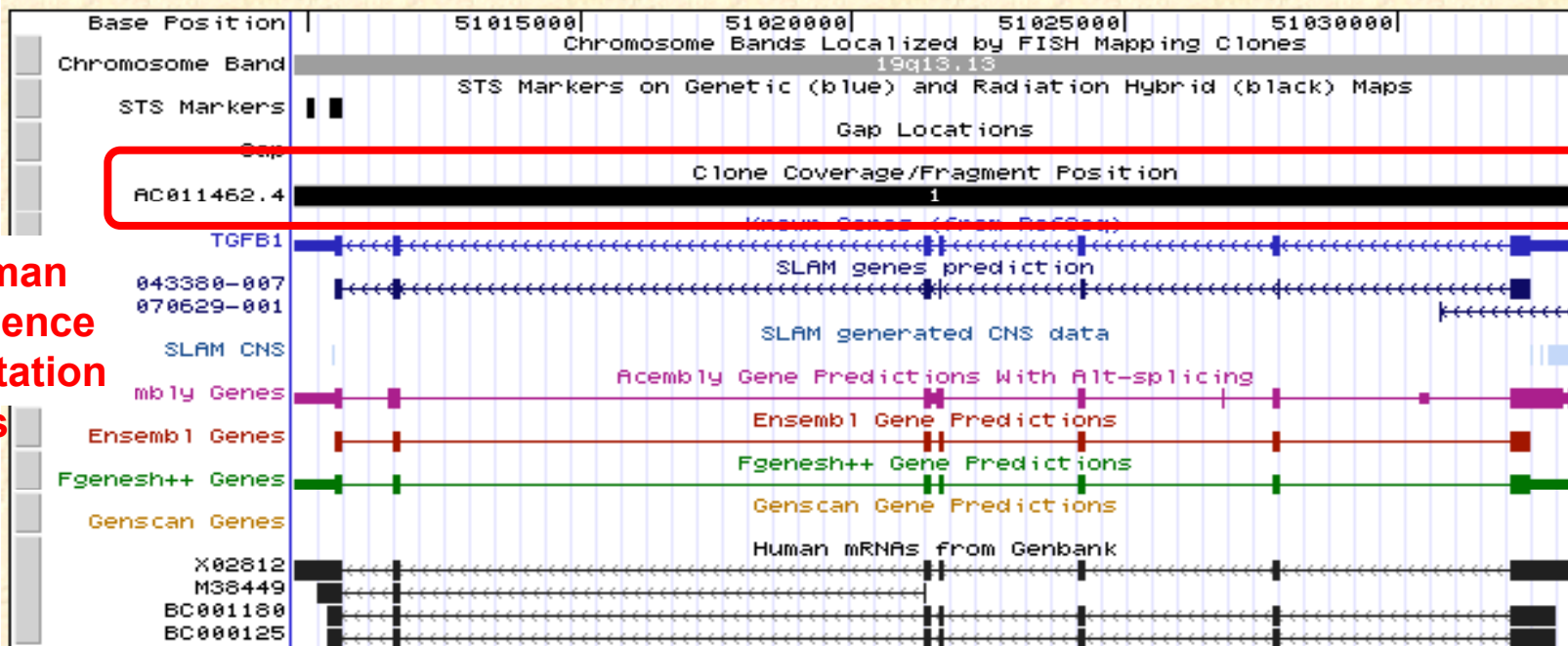
UCSC Genome Browser on Human: Aug. 6, 2001 Freeze

move <<< << < > >> >>> zoom in 1.5x 3x 10x zoom out 1.5x 3x 10x
position **TGFB1** size 23562, pixel width 620 **jump**

Vista Plots



Human
Sequence
Annotation



<http://pipeline.lbl.gov/>

THE BERKELEY GENOME PIPELINE

GODZILLA

[MOUSE ANALYSIS](#)

[--FINISHED](#)

[--ASSEMBLY](#)

[VISTA BROWSER](#)

[VISTA TRACK](#)

[MYGODZILLA](#)

[SOFTWARE](#)

[LINKS](#)

[CONTACT INFO](#)

MyGodzilla - is an interactive web tool for comparing your favorite sequence against the human genome

MyGodzilla Tool



MyGodzilla

Submit a Request

Sequence

(choose one of the three options)

Paste a Query Sequence (FASTA format finished sequences only, 300K max)

Draft sequences can be all entered at once, each contig starting with > and the sequence name

Alternatively, you can also select a file or enter a GenBank identification number:

FASTA

Text files only. Word documents are **not accepted.**
Sequences should be in FASTA format

Or

GenBank

GenBank Locus: Accession or
GI Number

Treat lower-case letters as repeats

Base Genome

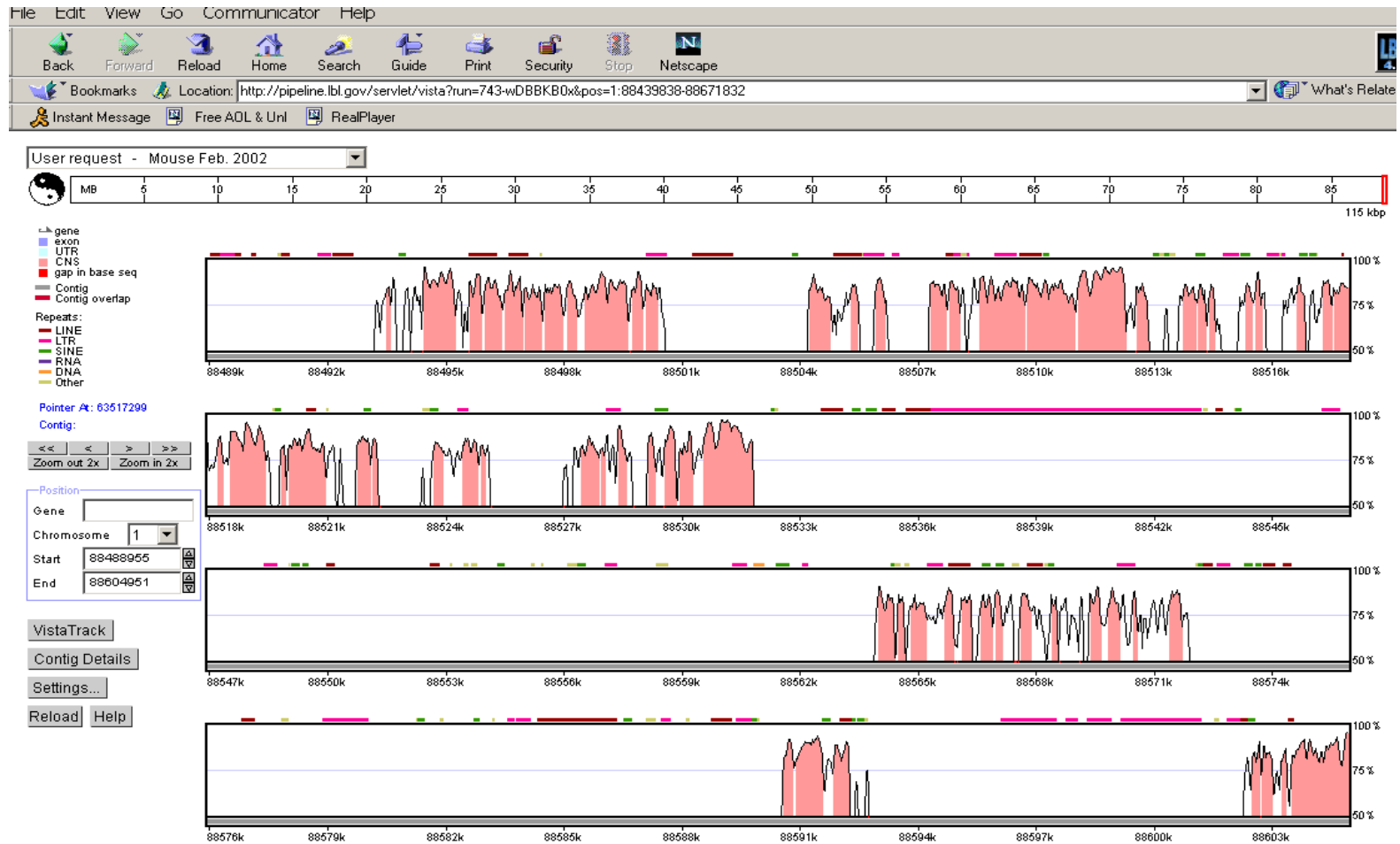
Advanced Options

Submit a DNA sequence of ANY organism...

Query against the human genome assembly- June 2002



Query against the mouse genome assembly - Feb. 2002



Examples of Results

- Understanding the structure of conservation
- Identification of putative functional sites
- Discovery of new genes
- Detection of contamination and misassemblies

Biological stories

O N E

Wayward Discovery of a New Apolipoprotein Gene

T W O

Interleukin Expression Switch

Identification of a New Apo Gene on Human 11q23

Godzilla

UCSC Genome Browser on Human: Aug. 6, 2001 Freeze

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

position APOA1 size 52979, pixel width 620 jump

AR_46908
mouse (arachne 3)

Base Position

Chromosome Band

STS Markers

Gap

Coverage

ZNF259
APOA4
APOC3
APOA1

132720000 132730000 132740000 132750000 132760000

Chromosome Bands Localized by FISH Mapping Clones

11q23.3

STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps

Gap Locations

Clone Coverage/Fragment Position

Known Genes (from RefSeq)

Highly Conserved Region

Zoom In

ApoA4

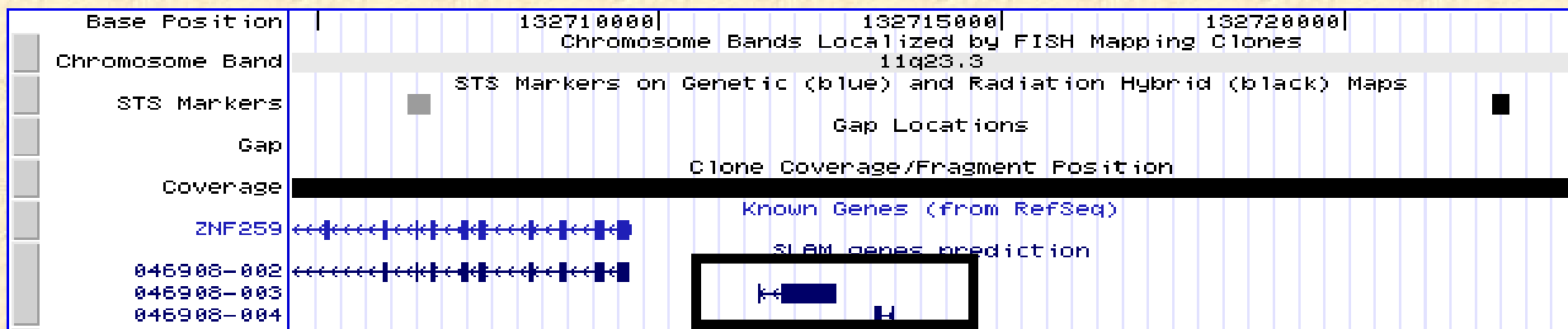
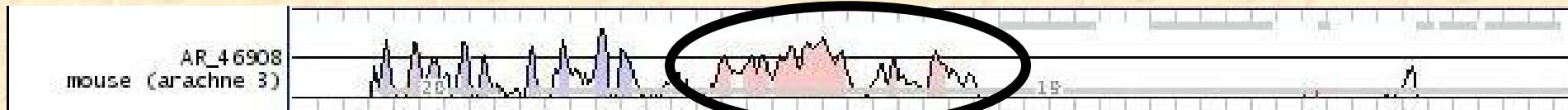
ApoC3

ApoA1

Identification of a New Apo Gene on Human 11q23 Godzilla

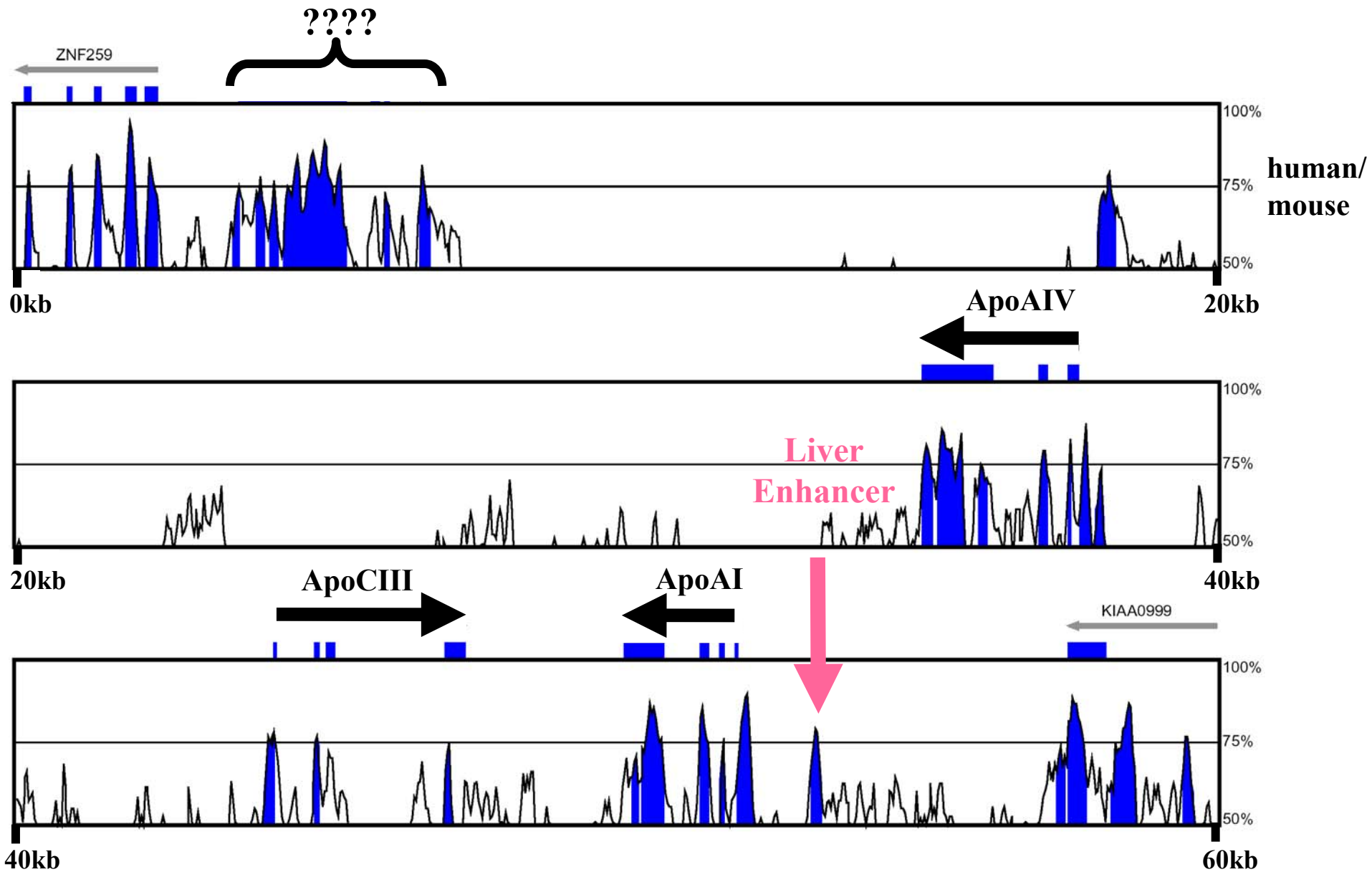
UCSC Genome Browser on Human: Aug. 6, 2001 Freeze

move <<< << < > >> >>> zoom in 1.5x 3x 10x zoom out 1.5x 3x 10x
position APOA1 size 52979, pixel width 620 jump



↑
New Gene (ApoA5)
Pennacchio LA et al.
Science. 2001, 294:169-73.

Human/Mouse Apolipoprotein Gene Cluster Sequence Comparison



Predicted Protein Sequence Has Homology to ApoAIV

predicted protein
human apoAIV

```

---MAAVLTWALALIS----AFSATQARKGEWDYFSQTSG-DKGRVEQIH
MFLKAVVLTLLALVAVAGARAEV SADQVATVMWDYFSQLSNNAKEAVEHLQ

QQKMAREP-ATLKDSLEQDLNMMNKFLEKLRPLSGSEAPRLPQDPVGMRR
KSELTQQLNALFQDKLGEVNTYAGDLQKKLVPFATELHERLAKDSEKLKE

QLQEELEEVKARLQPYMAEAHELVGWNLGLRQQLKPYTMDLMEQVALRV
EIGKELEELRARLLPHANEVVSQKIGDNLRELQQRLEPYADQLRTQVNTQA

QELQEQLRVVGEDTKAQLLGGVDEAWALLQG----LQSRVVHHTGRFKEL
EQLRRQLDPLAQRMERVLRENADSLQASLRPHADELKAKIDQNVEELKGR

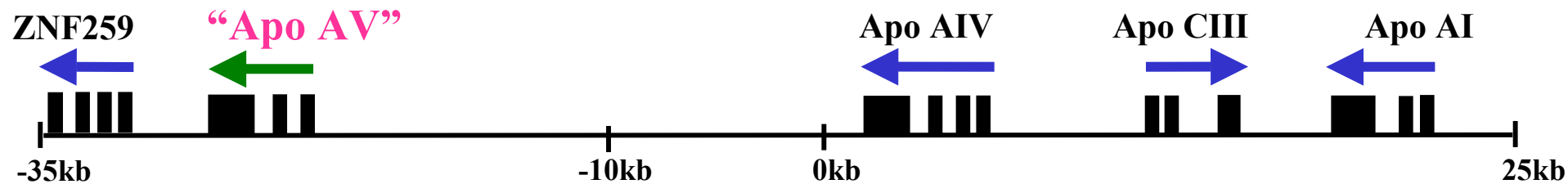
FHPYAESLVSGIGRHHVQELHRSVAPHAPASPARLSRCVQVLSRKITLKAK
LTPYADEFKVKIDQTV EELRRSLAPYAQDTQEKLNHQLEGLTFQMKNNAE

ALHARIQQNLDQLREELSRAFAGT-----GTEEGAGPDPQMLSEEVRQRI
ELKARISASAEELRQRLAPLAEDVRGNLKGNTTEGLQKSLAELGGHLDQQV

QAFRQDTYLQIAAFTRAIDQETEEVQQQLAPPFGHSFAFAPEFQQTDSGK
EEFRRRVEPYGENFNKALVQQMEQLRQKLGPHAGDVEGHLSFLEKDLRDK

VLSKQLQARLDDLWEDITHSLHDQGHSHLGDP-----
VNSFFSTFKKESQDKTLLSLEPELEQQQEQQQEQQQEQQVQMLAPLES
    
```

Identity: 26%
Similarity: 45%

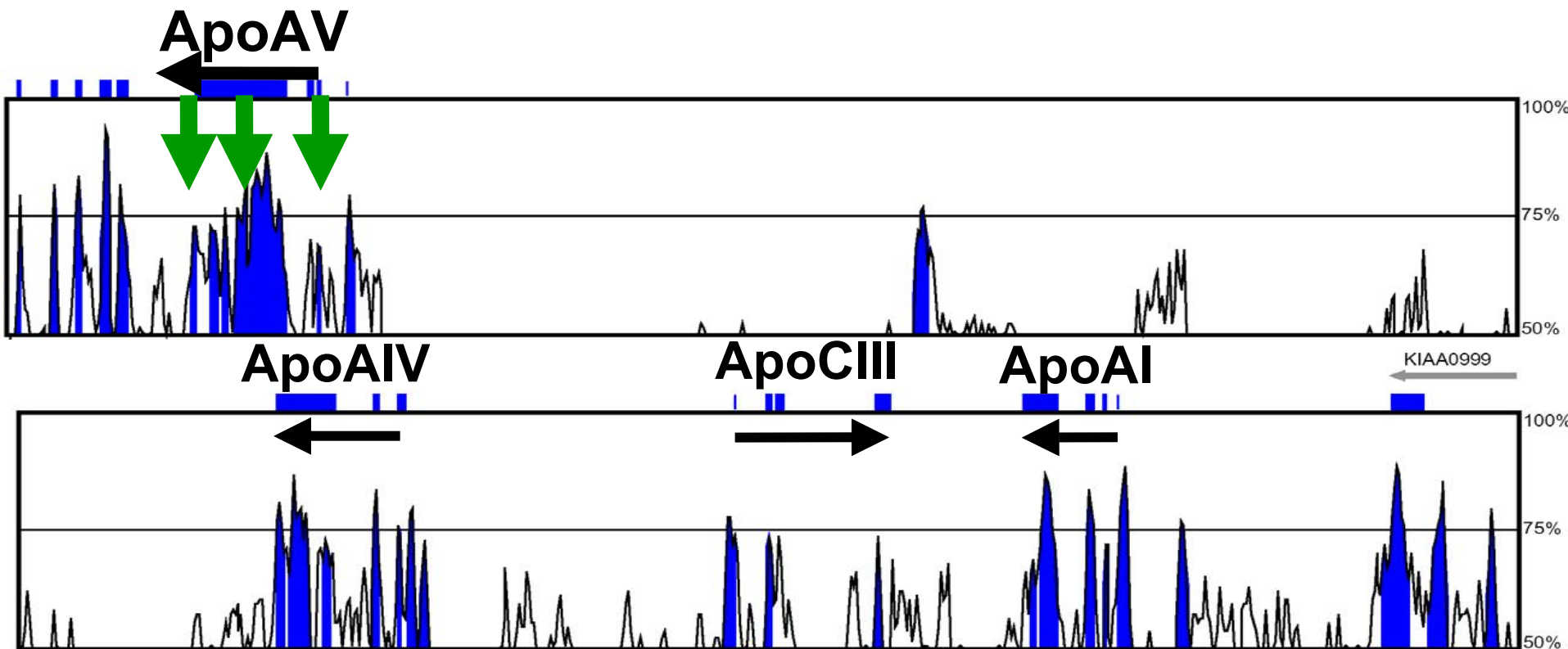


Summary: ApoAV

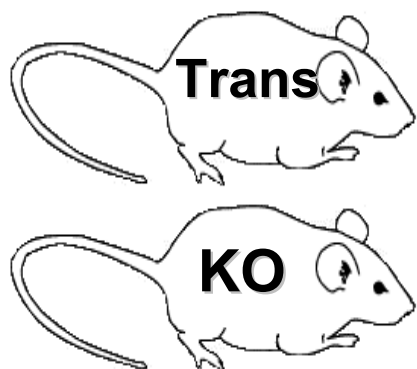
- A new apolipoprotein belonging to the ApoAI/CIII/AIV gene cluster.
- Expressed in the liver & associates with HDL/VLDL.
- An important modulator of triglycerides (TG) in mice.



Is ApoAV involved in human biology/disease?

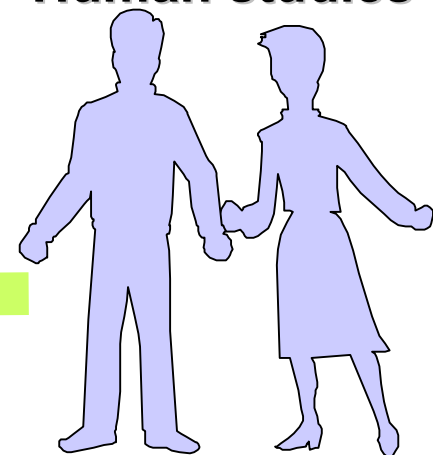


Mouse studies

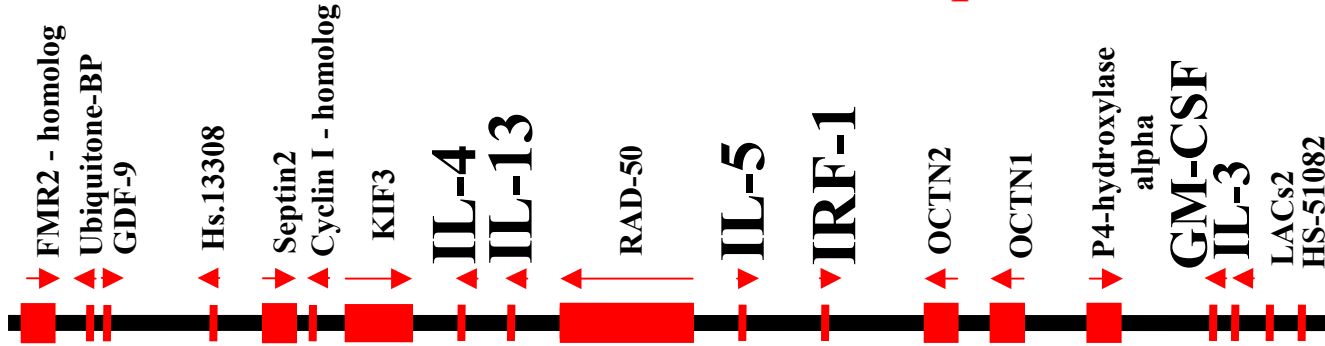


**Importance of ApoAV
on Triglyceride Metabolism**

Human studies



IL Cluster HUM 5q31

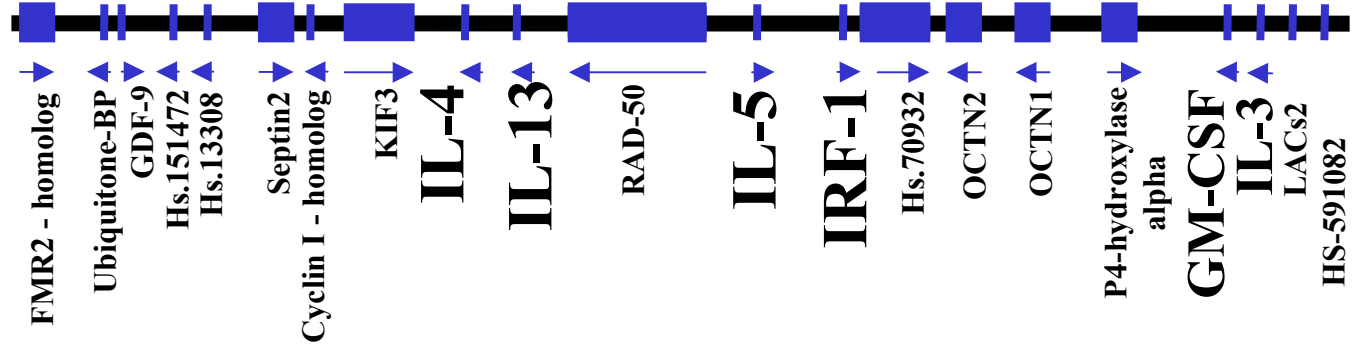


Coding

Exons 3%

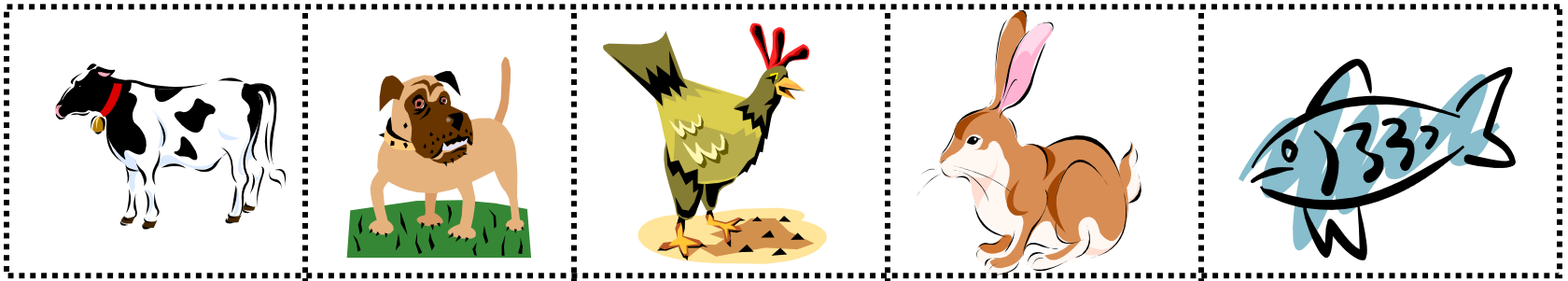
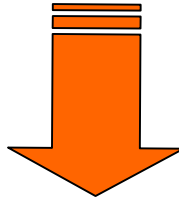
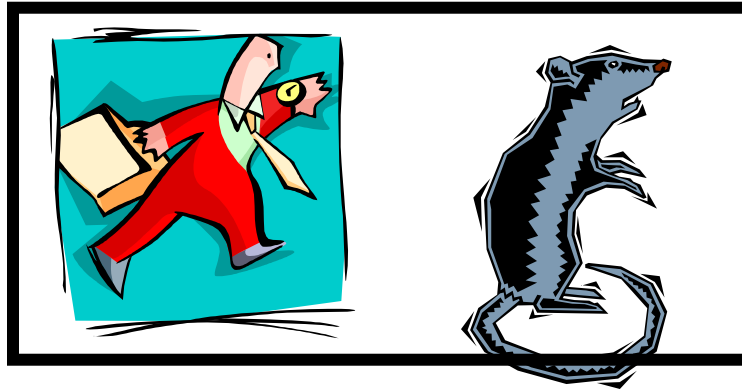
Non-Coding

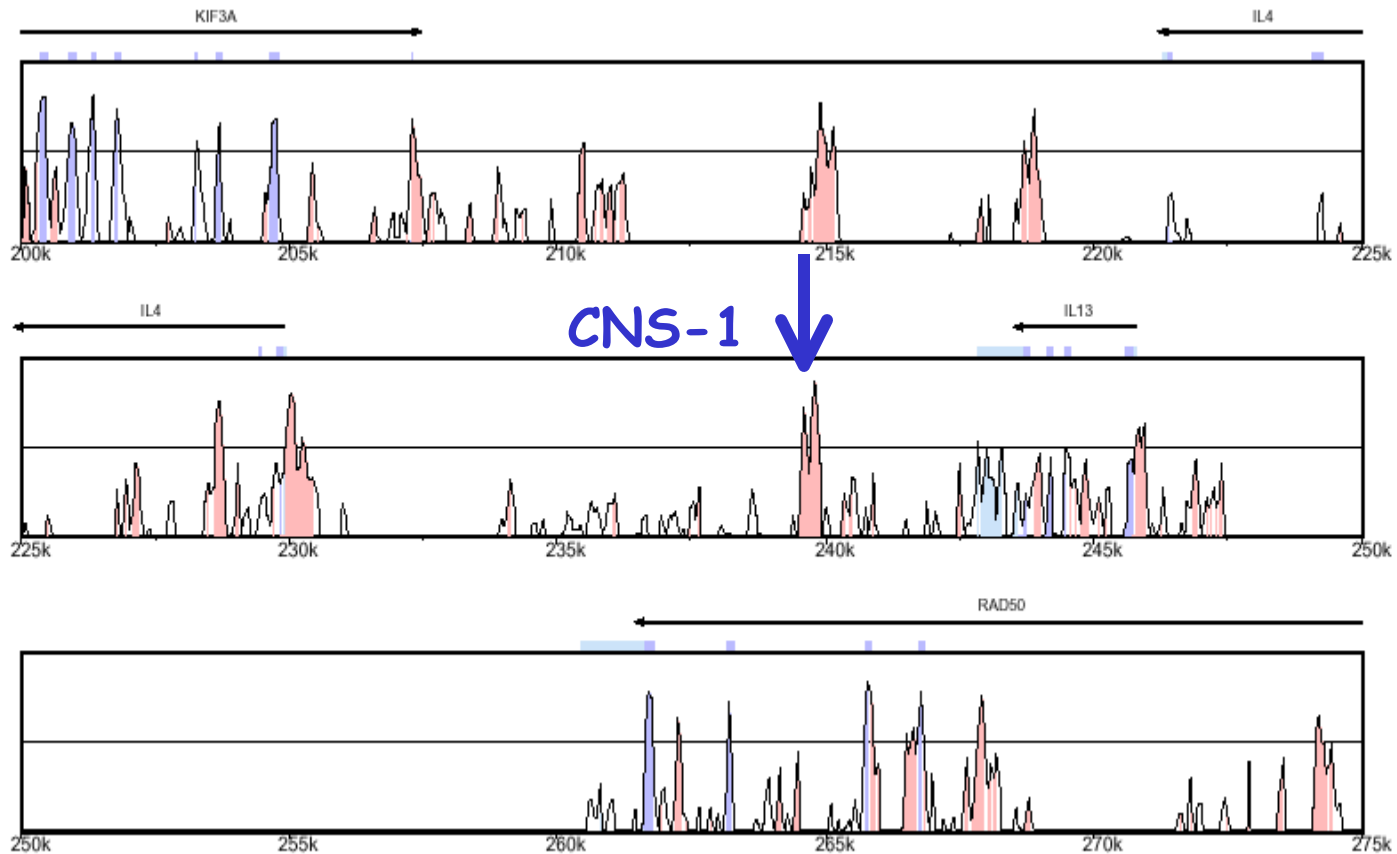
Conserved 2.6%
(>100bp > 75%)



IL Cluster MU Ch 11

A Filtering Strategy





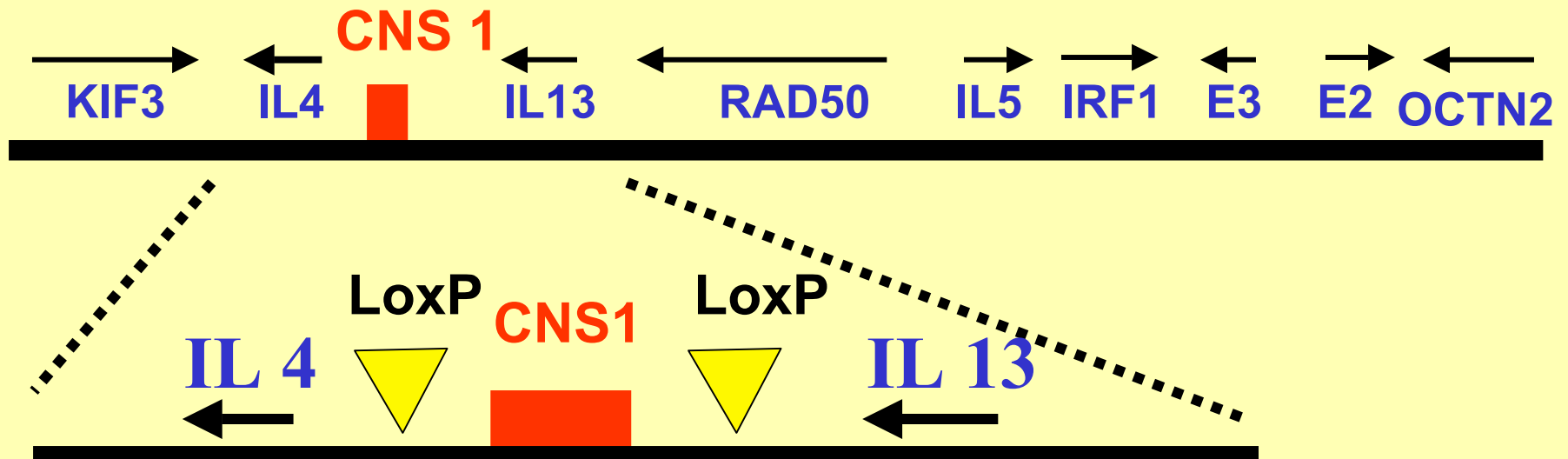
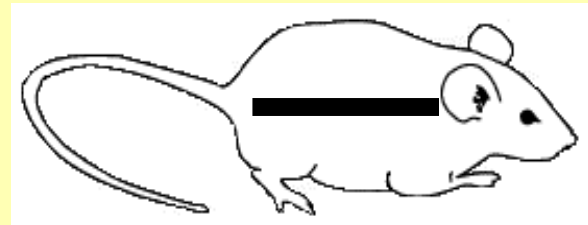
Present in other species: Cow (86%), Dog (81%), Rabbit (73%)

Genomic position conserved in human, mouse, dog, baboon

Single copy in the human genome. Two hypersensitive sites mapped.

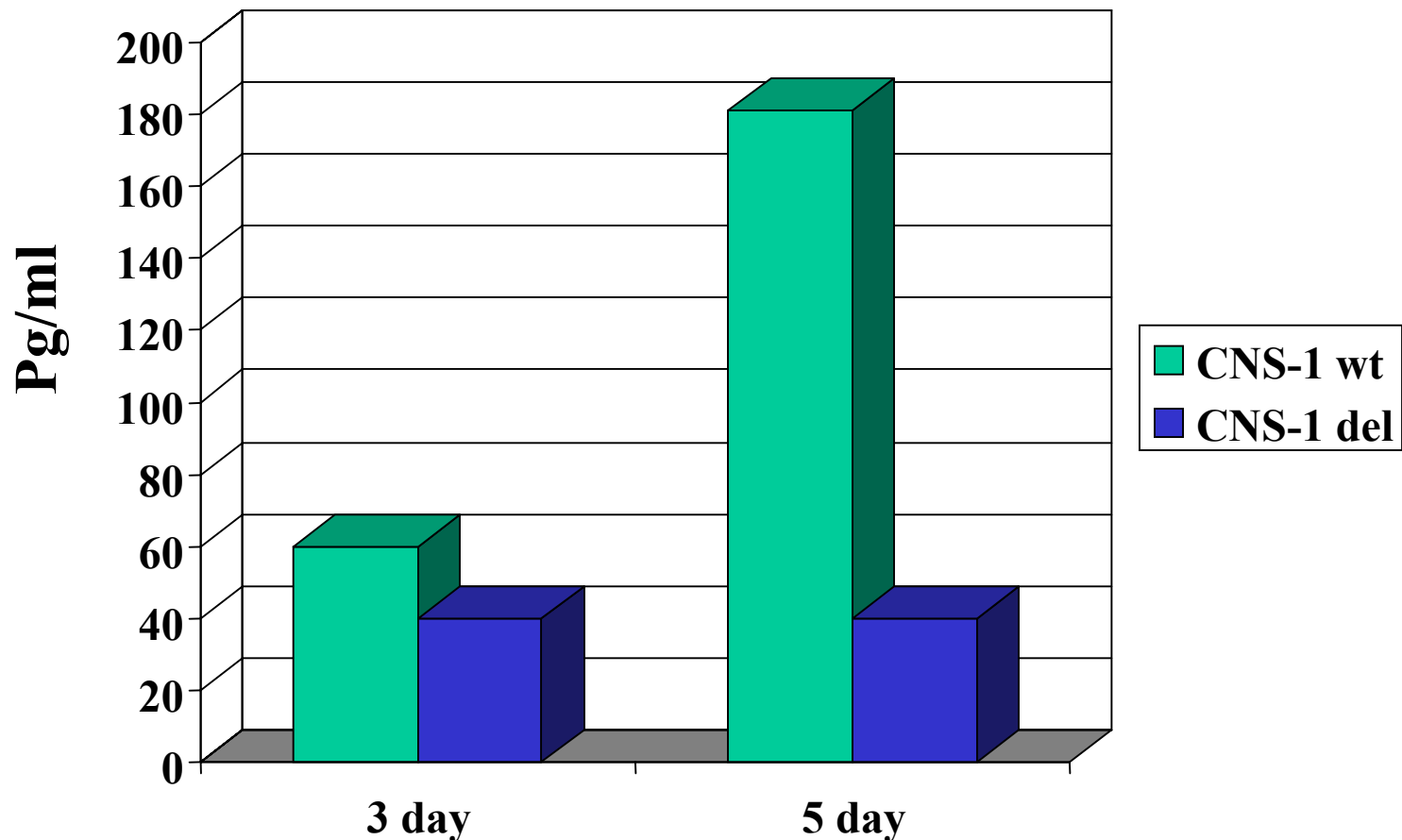
Functional Analysis of CNS1

Generate Human 5q31 YAC Transgenic Mice



Human IL 4 Production in YAC Transgenics Containing and Lacking CNS1

IL-5 & IL13 Expression is also reduced in CNS-1^{del} mice





Science. 2001 Oct 5;294(5540):169-73.

Thanks

Biology

Kelly Frazer

Gaby Loots

Len Pennacchio

Eddy Rubin

Bioinformatics

Michael Brudno

Olivier Couronne

Brian Klock

Chris Mayor

Ivan Ovcharenko

Alexander Poliakov

Jody Schwartz

Lior Pachter (UCB)

Funding - Programs for Genomic Applications (PGA) by NHLBI