Comparative genomics tools for biological discovery

Inna Dubchak, Ph.D.

Staff scientist

Lawrence Berkeley National Laboratory

ildubchak@lbl.gov

Outline

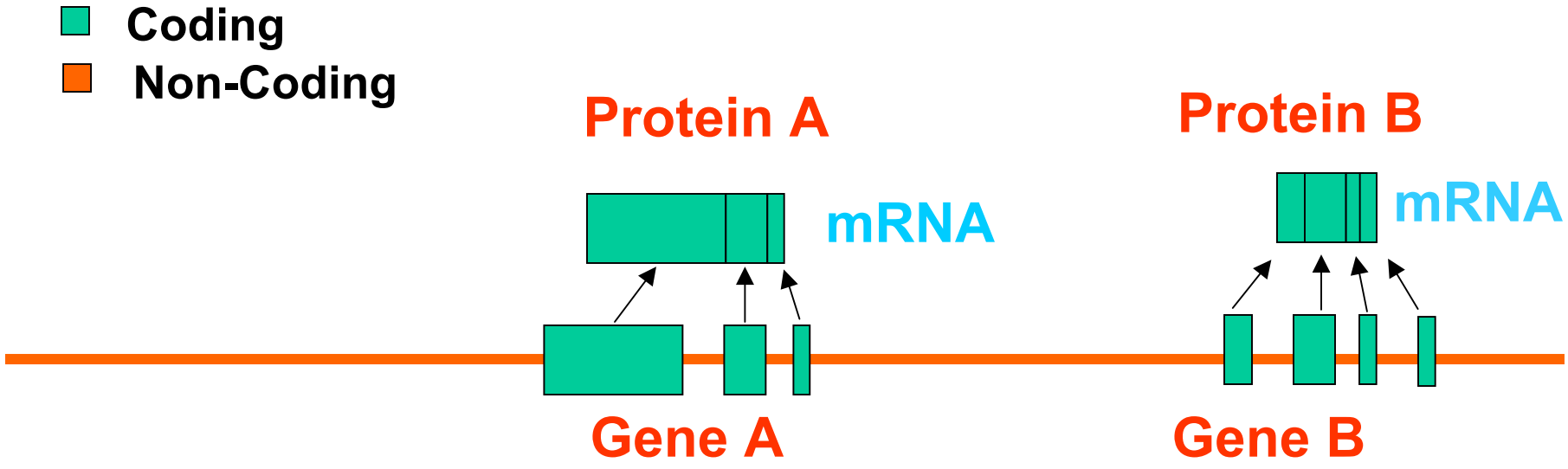
What is comparative genomics?

VISTA tools developed for comparative genomics.

Large scale VISTA applications including aligning whole genome assemblies

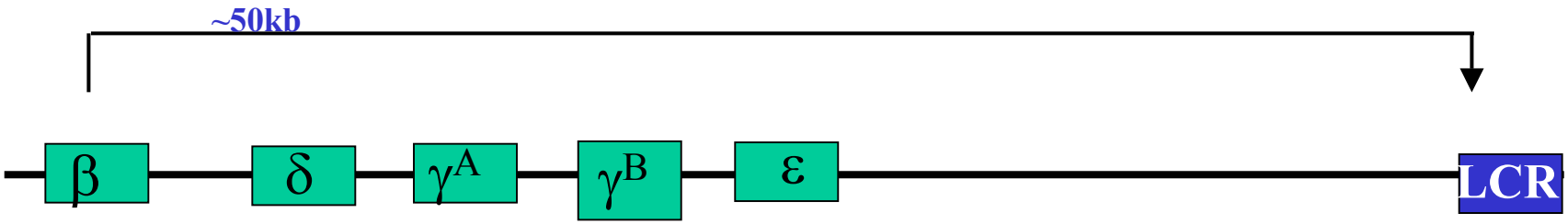
Related biological stories

1-2% Coding



Distant Non-Coding Sequences Causing Disease

β–Thalassemia

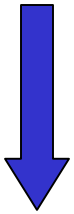


Disease	Gene	Distance
Campomelic displasia	**SOX9**	850kb
Aniridia	**PAX6**	125kb
X-Linked Deafness	**POU3F4**	900kb
Saethre-Chotzen syndrome	**TWIST**	250kb
Rieger syndrome	**PITX2**	90kb
Split hand/split foot malformation	**SHFM1**	450kb

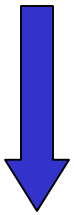
Background

Evolution can help!

In general, functionally important sequences are conserved

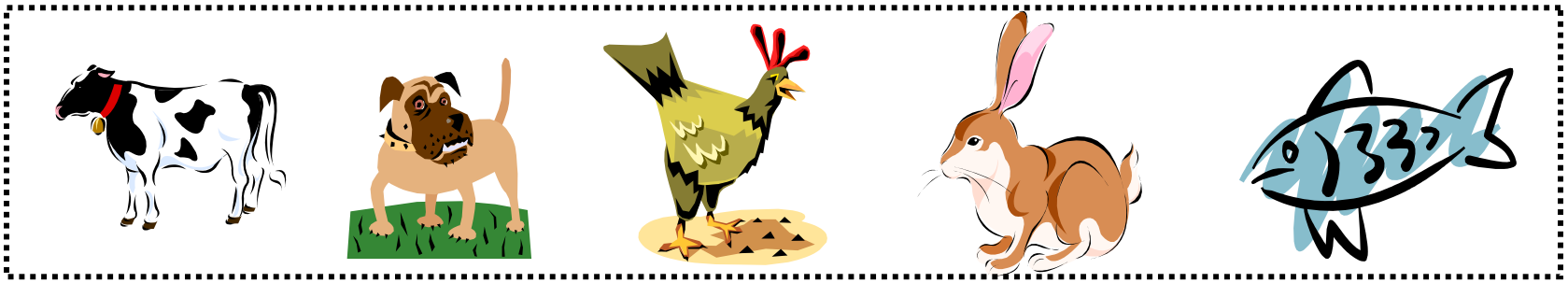


Conserved sequences are functionally important



Raw sequence can help in finding biological function

Comparing sequences of different organisms



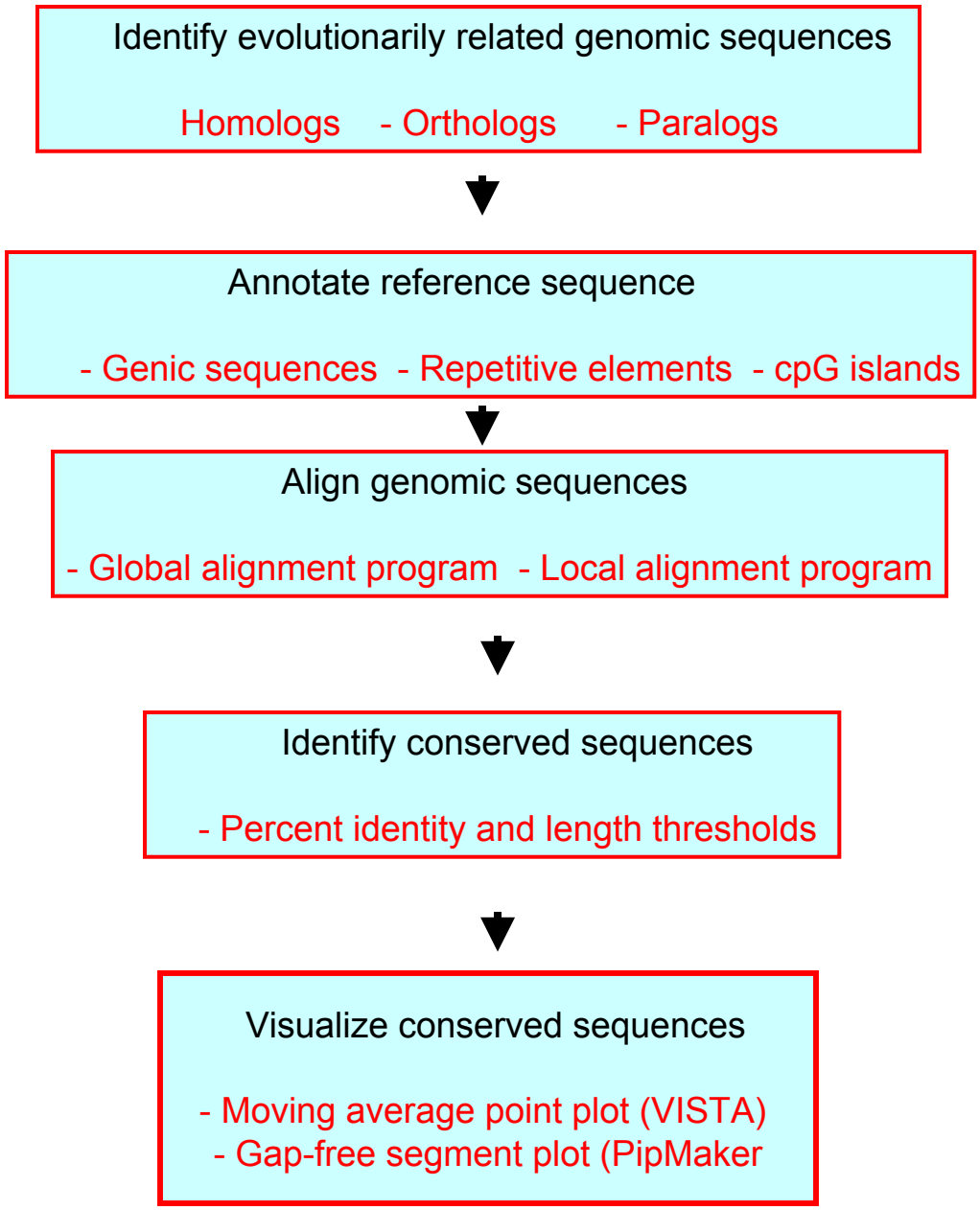
- Helps in gene predictions
- Helps in understanding evolution
- Conserved between species non-coding sequences are reliable guides to regulatory elements
- Differences between evolutionary closely related sequences help to discover gene functions

Challenges

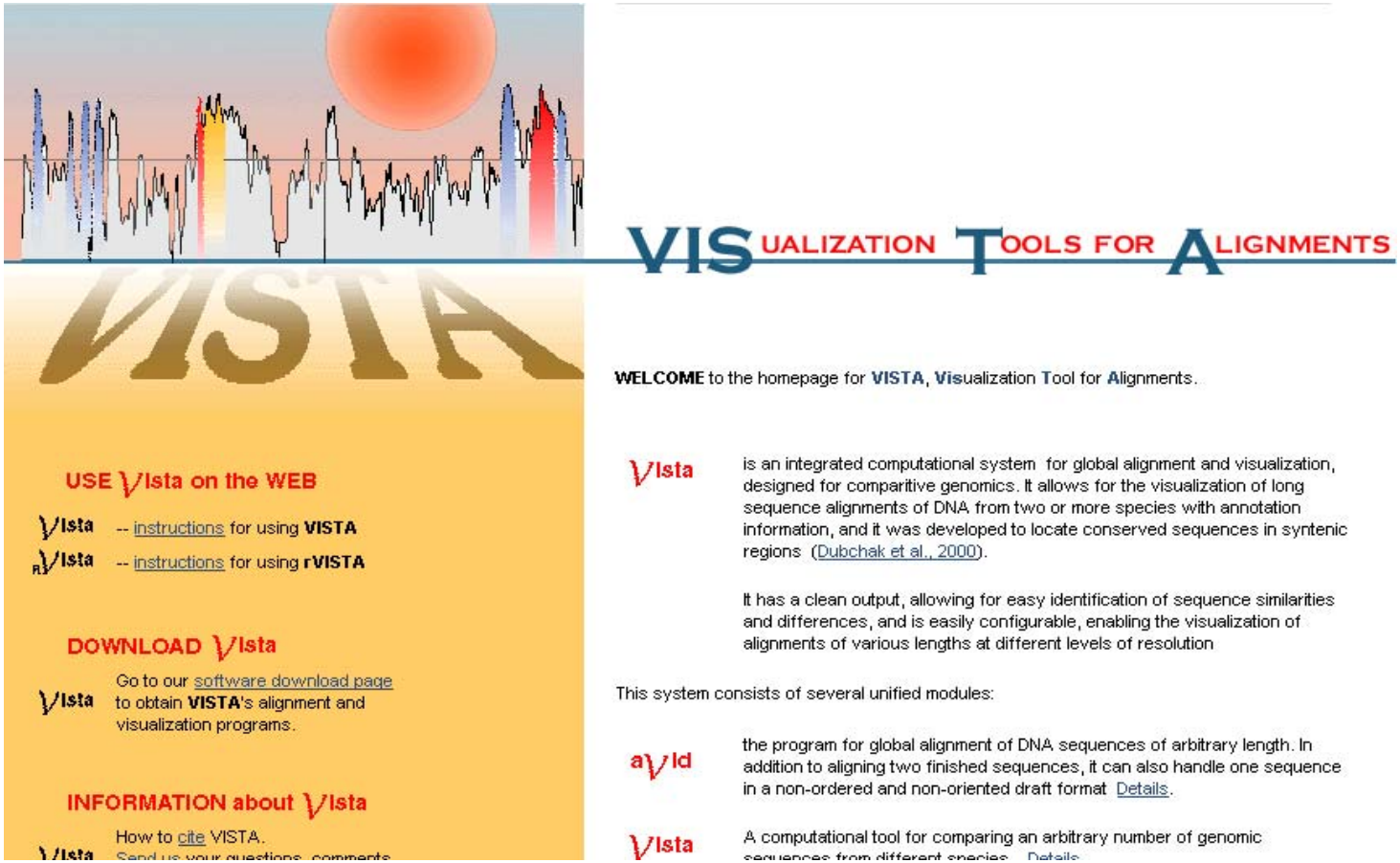
- Sequence at different stages of completion, difficult to compare

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

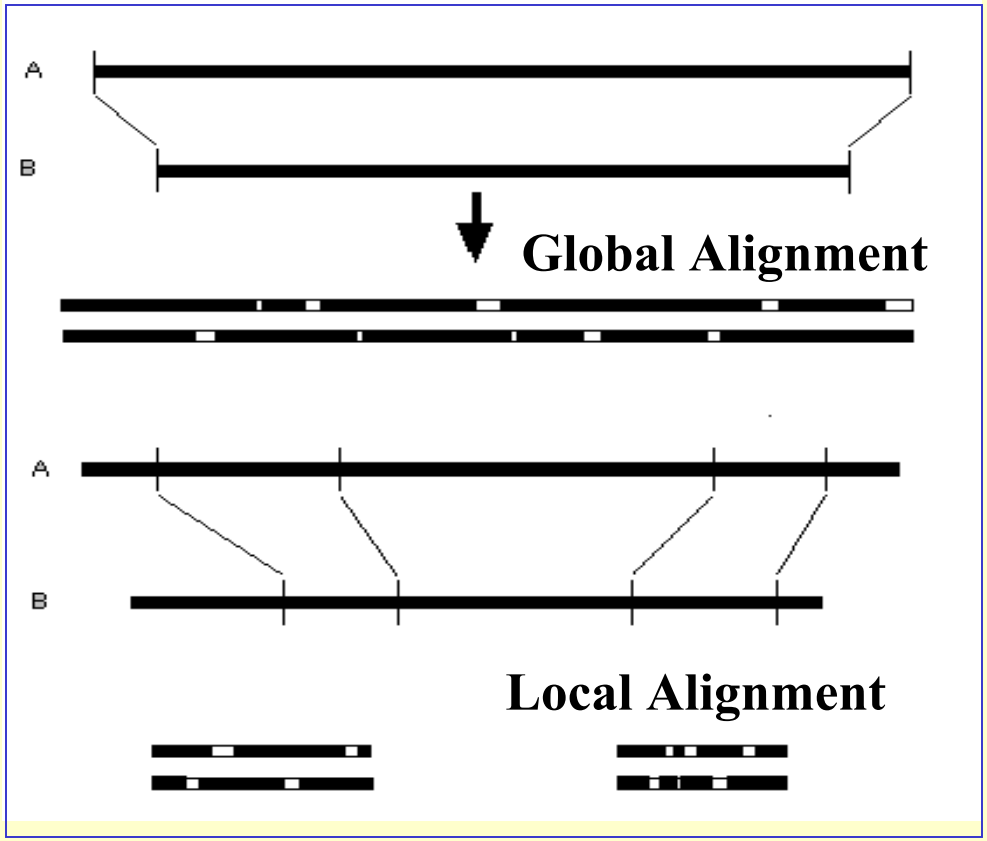


Processed ~ 16000 queries on-line, distributed > 700 copies of the program in 35 countries

Modules of VISTA:

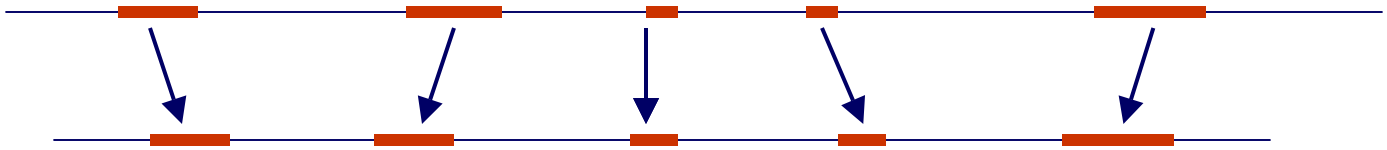
- Program for global alignment of DNA fragments of any length
- Visualization of alignment and various sequence features for any number of species
- Evaluation and retrieval of all regions with predefined levels of conservation

Local vs global alignment



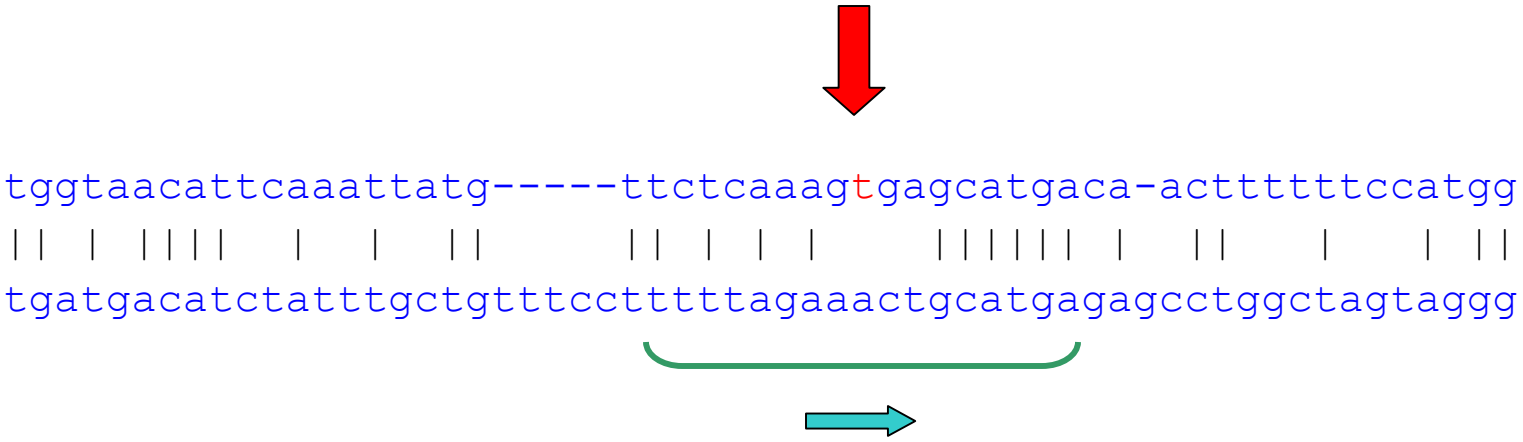
AVID- the alignment engine behind VISTA

- 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



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------TTTTTGCAACAGTA
    |||| |||||||| | |||||||||||||    | ||| |      |||||  |||||
GACACTGGAAAAGCAGAGGAAGTGGTTTATTGACCTGCCCCCCCCTTTTTTATAACAGTG
```

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
```

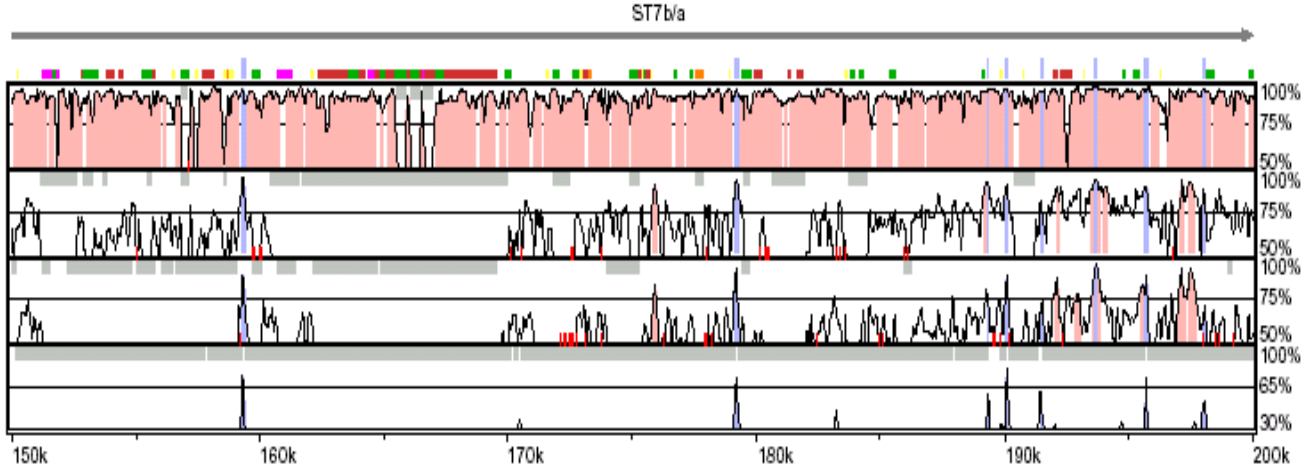
VISTA file

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

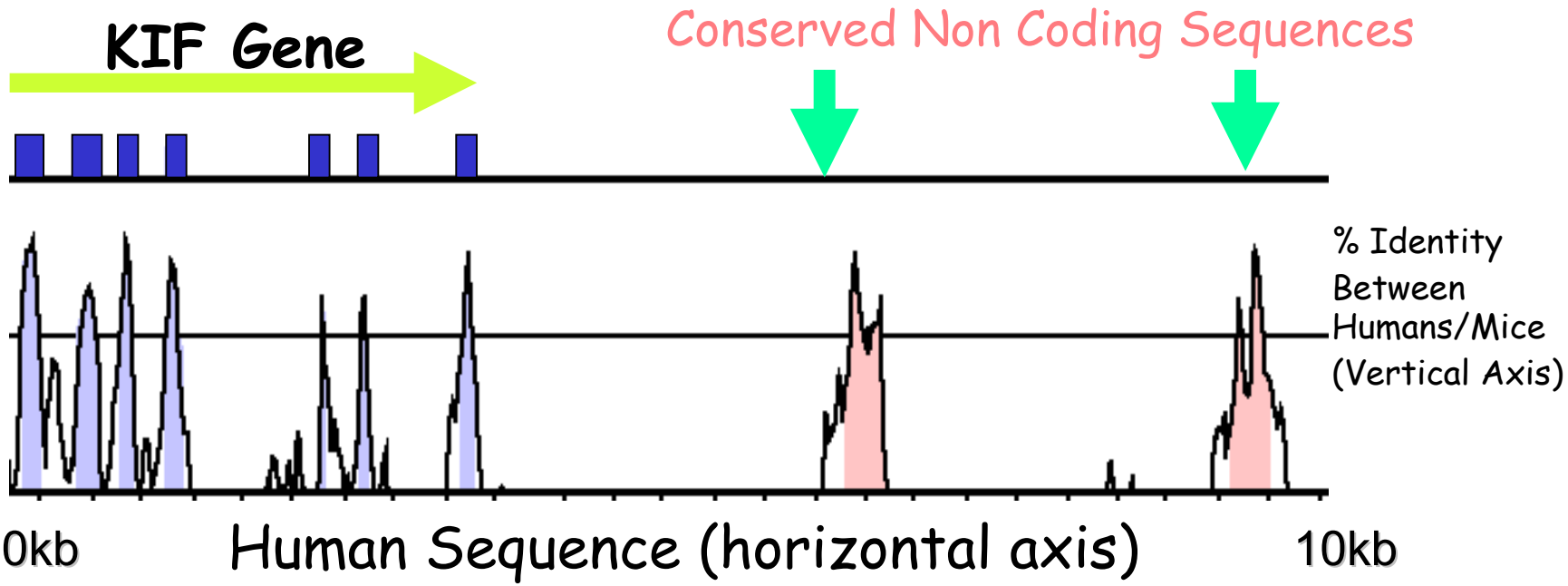
Alignment 2
Seqs: human/cow
Criteria: 93%, 100 bp
Regions: 49

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

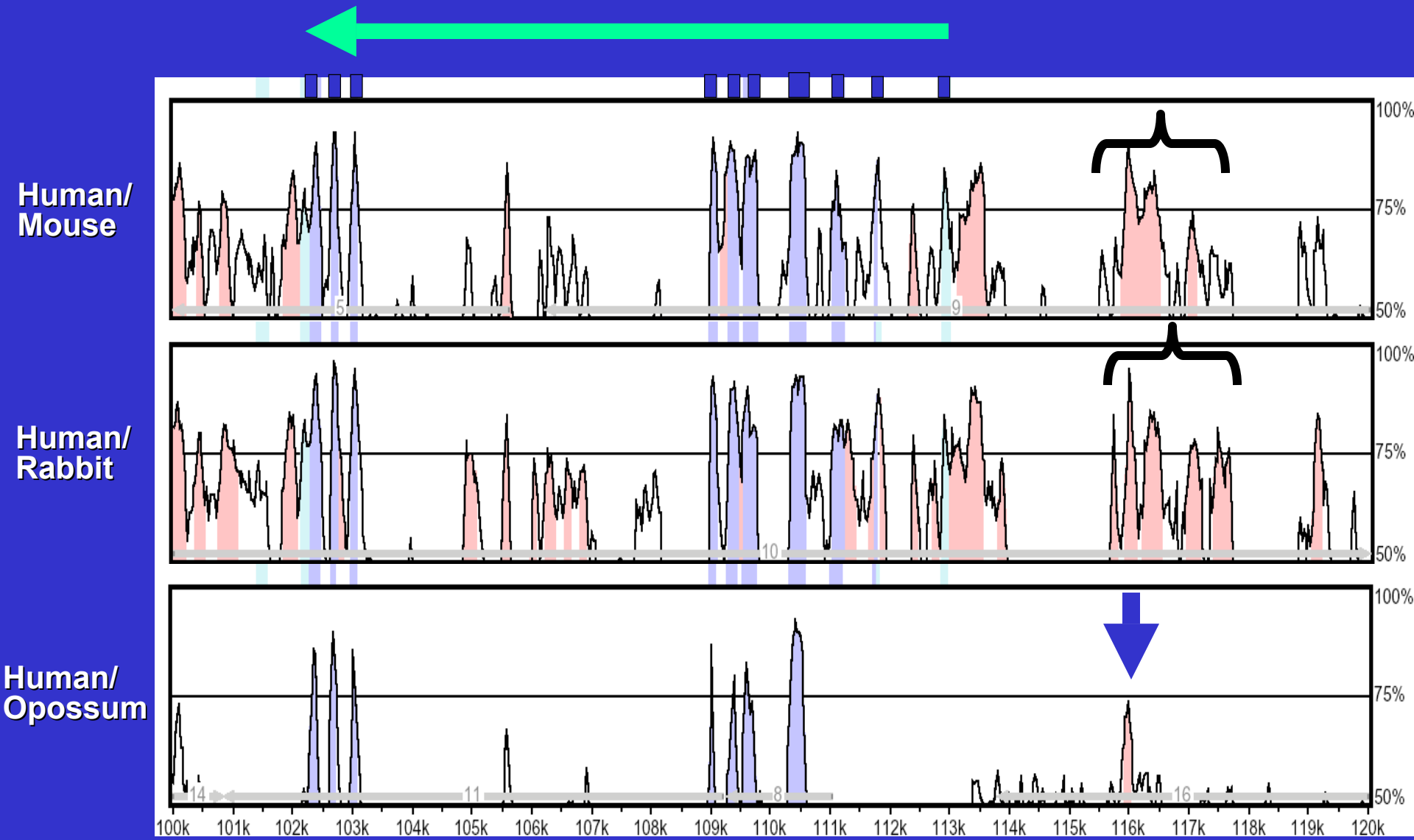
Alignment 4
Seqs: human/fugu
Criteria: 55%, 100 bp
Regions: 15



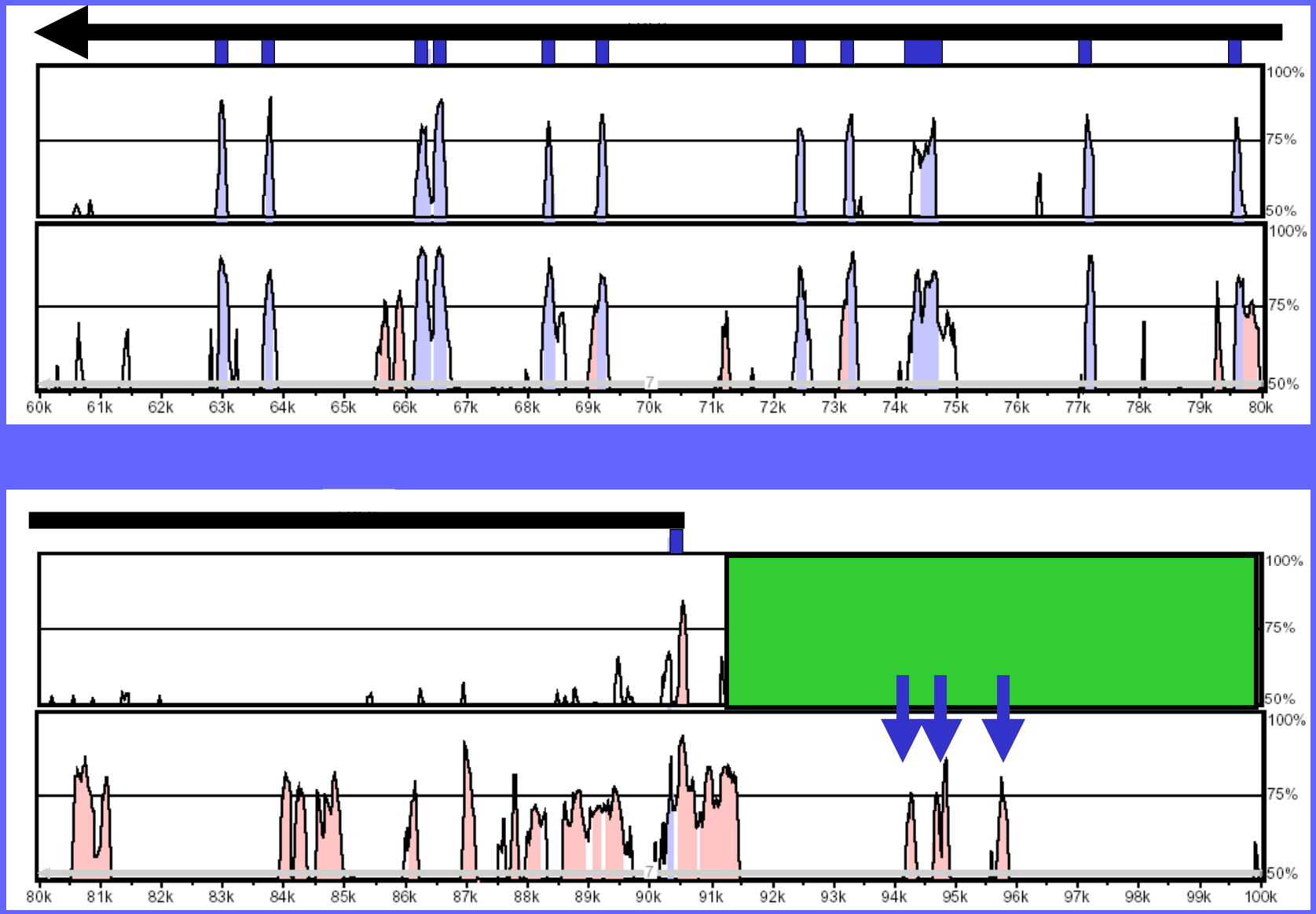
VISTA plot


KIF Gene
Conserved Non Coding Sequences
% Identity Between Humans/Mice (Vertical Axis)
0kb
Human Sequence (horizontal axis)
10kb

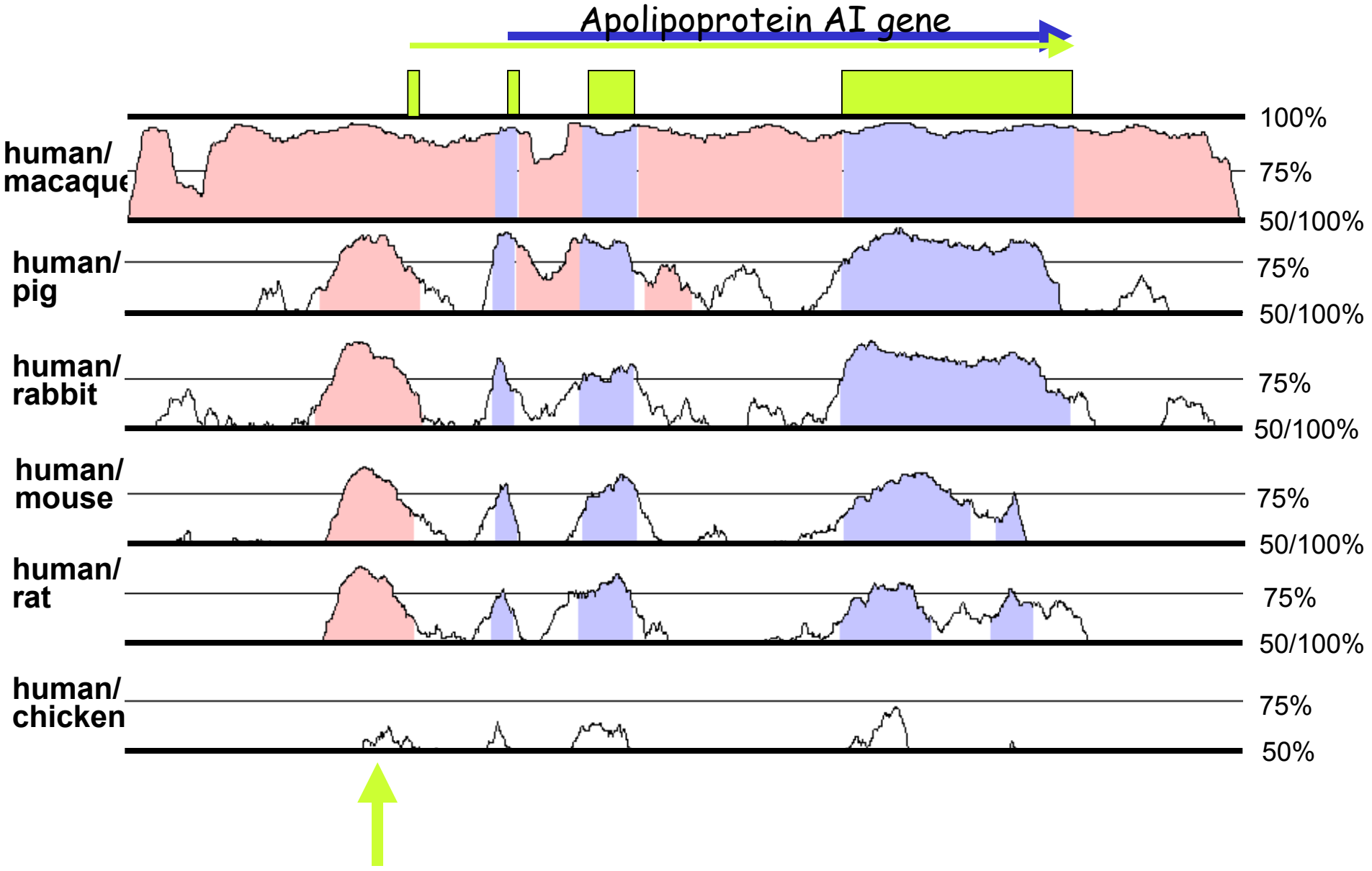
Nuclear Hormone Receptor:LXR-Alpha



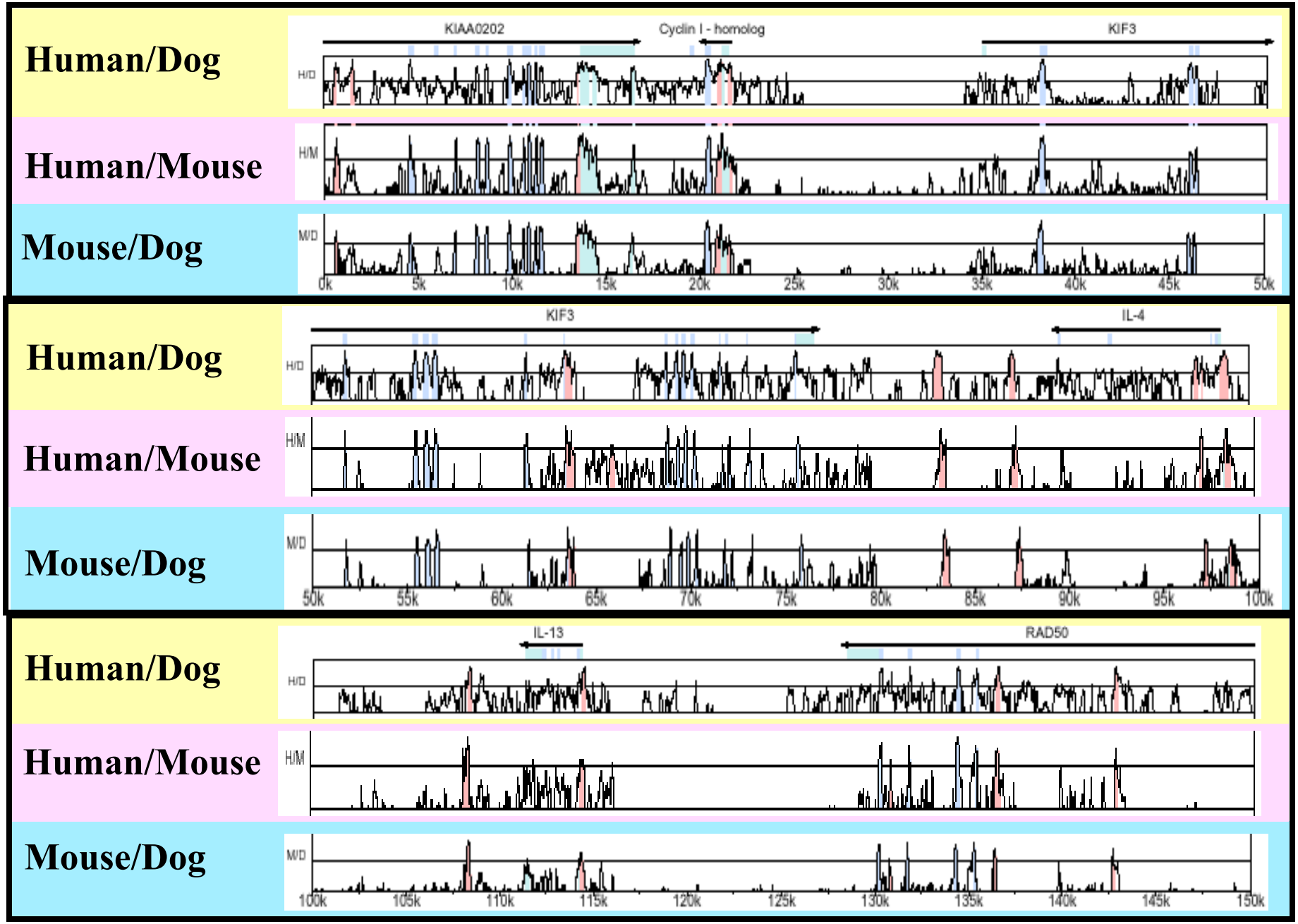
Low-Density Lipoprotein Receptor (LDLR)



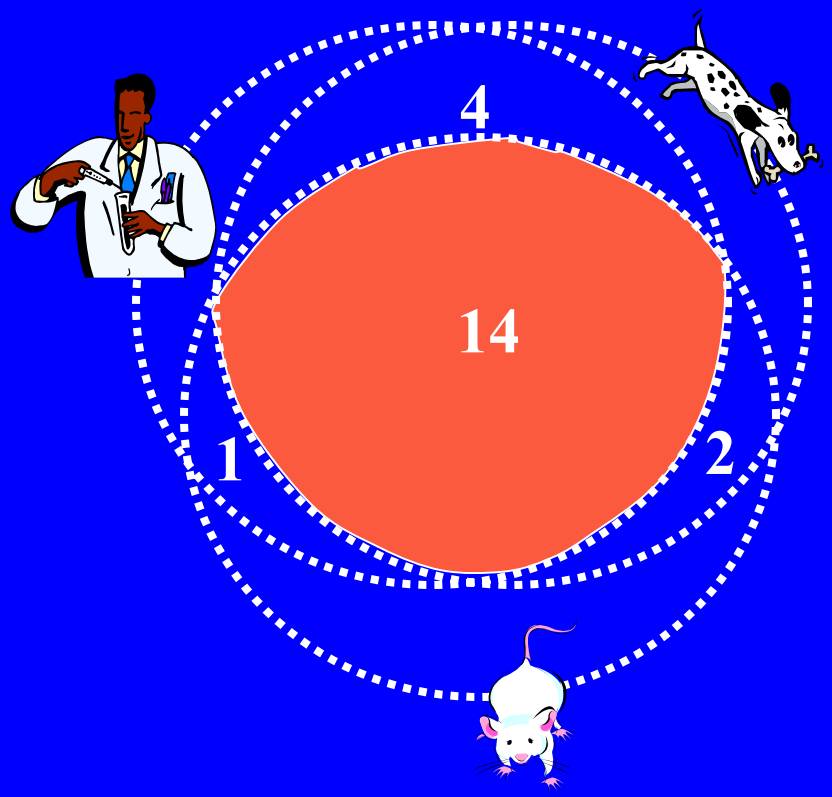
Multi-Species Comparative Analysis (VISTA)



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%

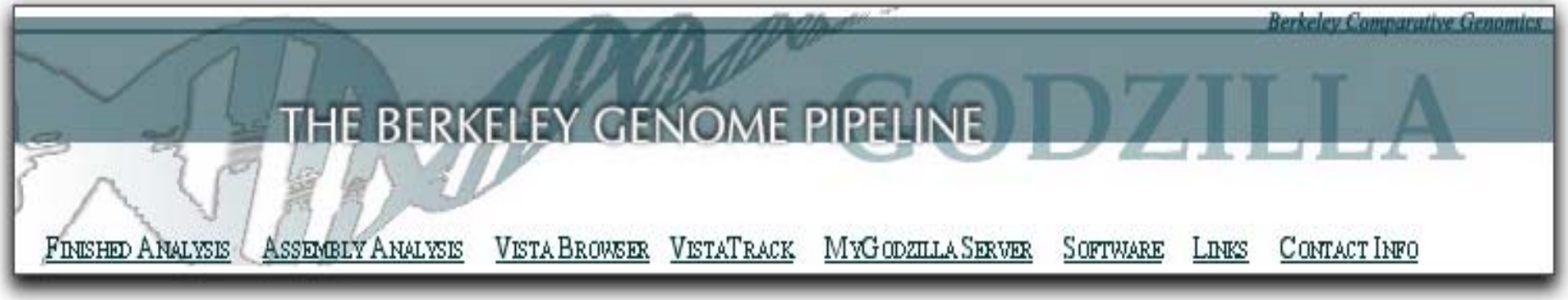
D/M > 77%

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



Godzilla - automatic computational system for comparative analysis of genomes

http://pipeline.lbl.gov http://www-gsd.lbl.gov/vista

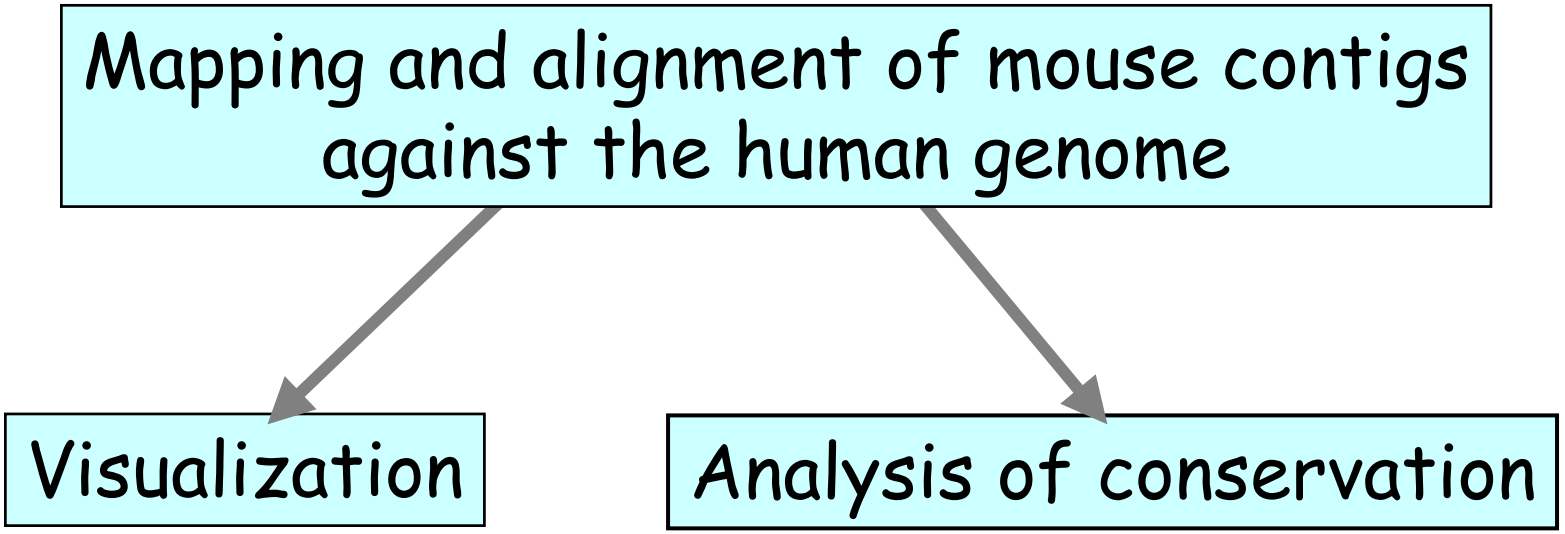
DATA

Base Human Genome – Golden Path Assembly

Mouse assemblies:

Arachne	October 2001
Phusion	November 2001
MGSC v3	April 2002

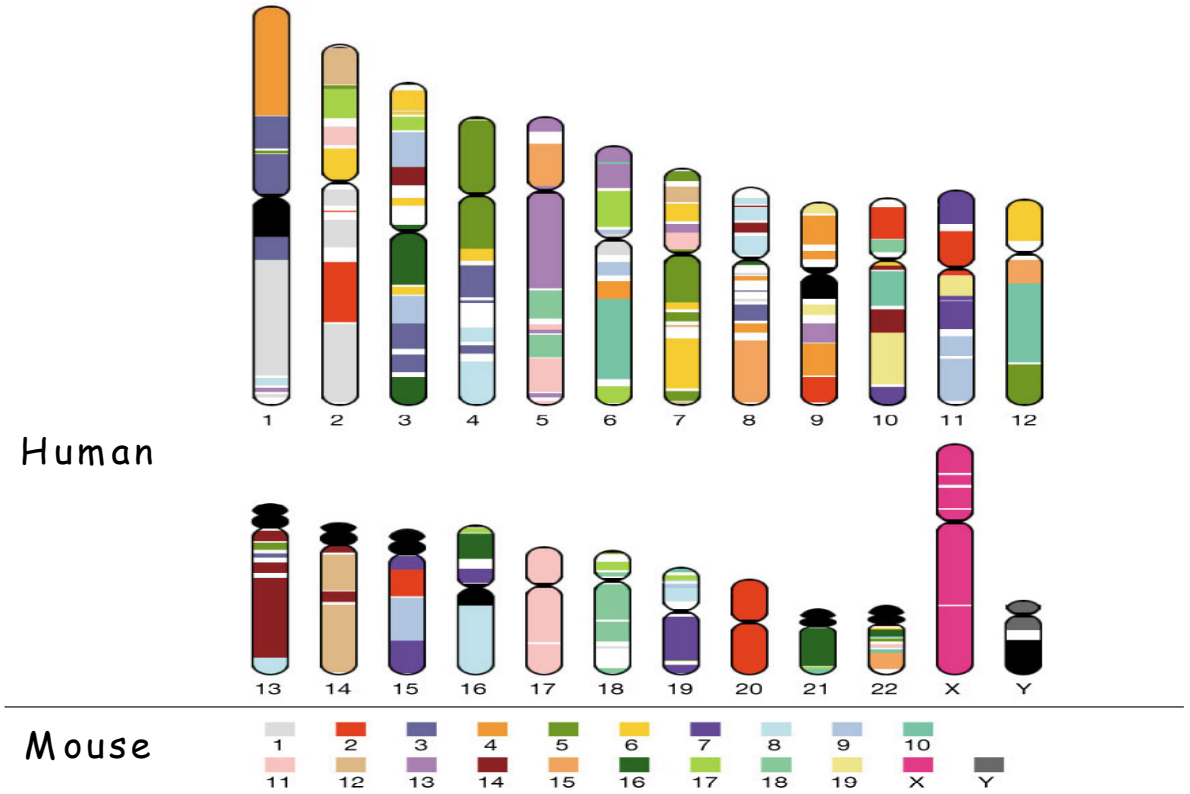
Main modules of the system



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

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

Chromosome Comparison

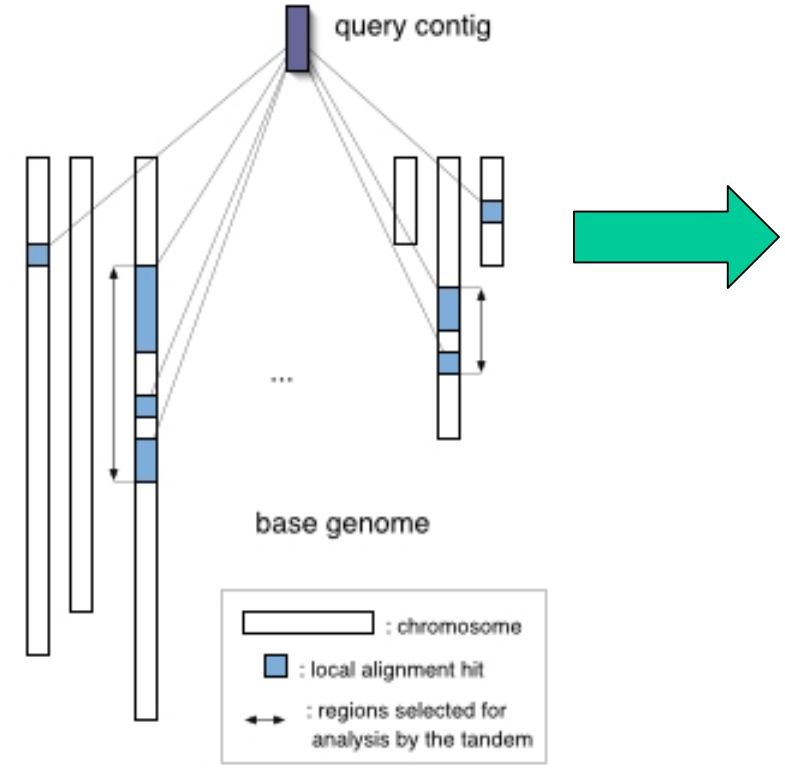


Base pair alignment

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

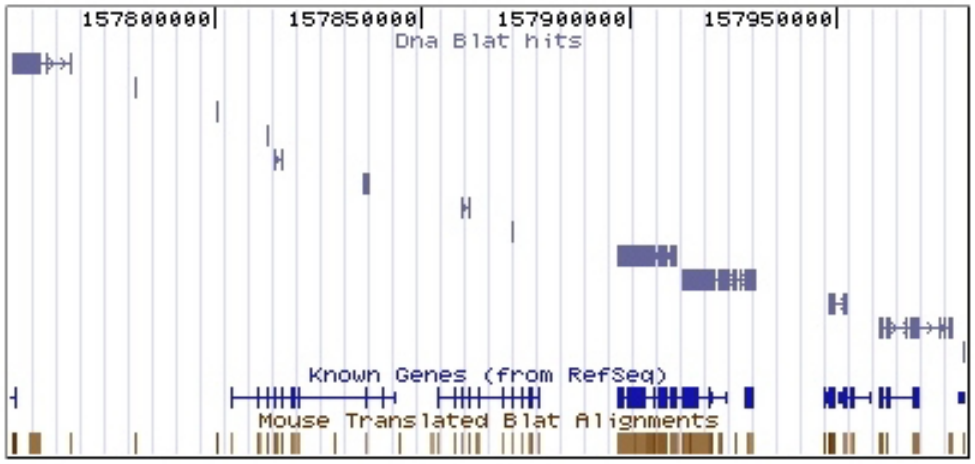
Tandem Local/Global Alignment Approach

Sequence fragment anchoring (DNA and/or translated BLAT)
Multi-step verification of potential regions using global alignment (AVID)

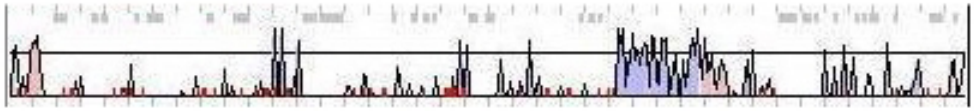


ANCHORS FROM LOCAL ALIGNMENT HITS

LOCAL ALIGNMENT HITS



GLOBAL ALIGNMENT



Advantage of the tandem approach:

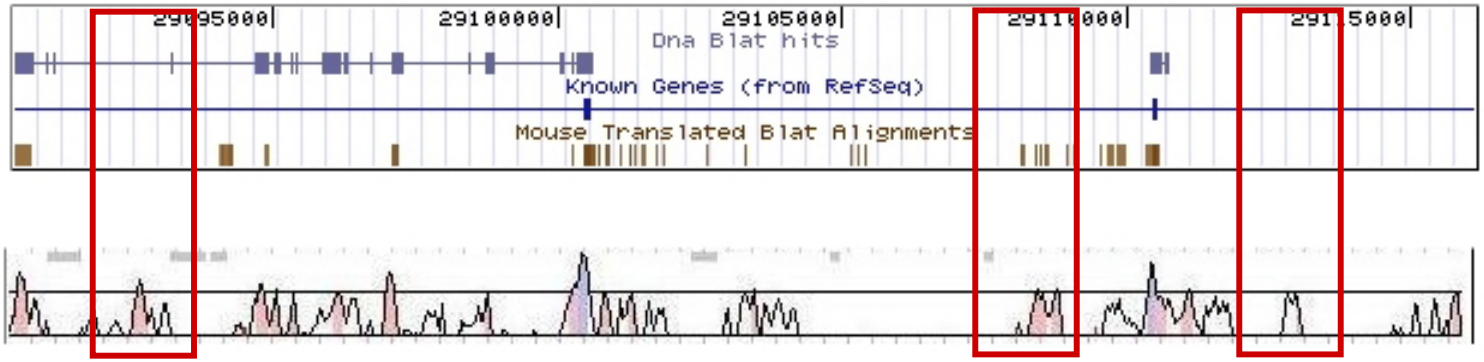
better sensitivity/specificity trade-off

- fill-in effect
- scoring longer alignments

NT_002606 at Chr.17:2909457-29116113

BLAT
Local
alignment

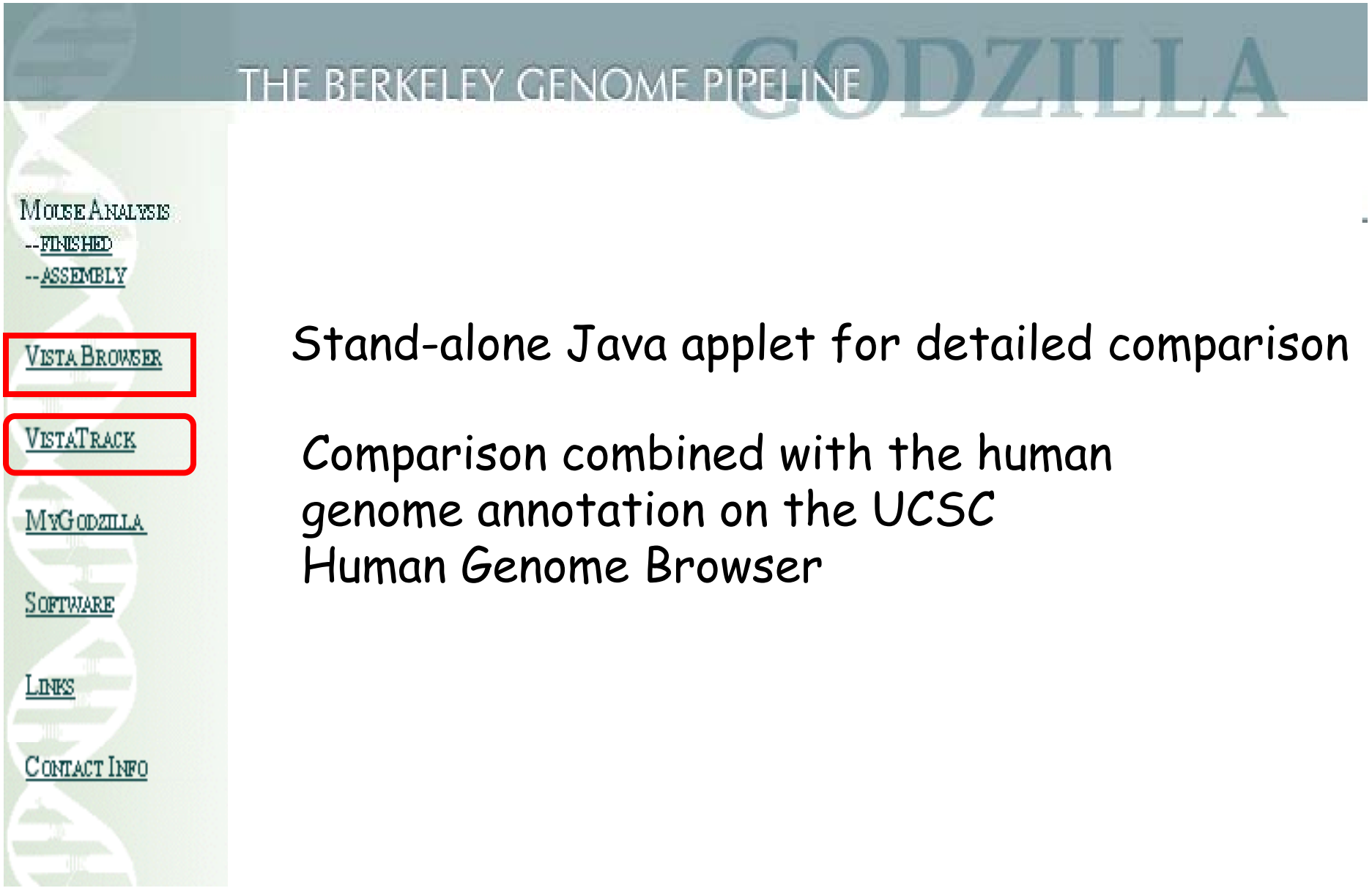
AVID
Lobal
alignment



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

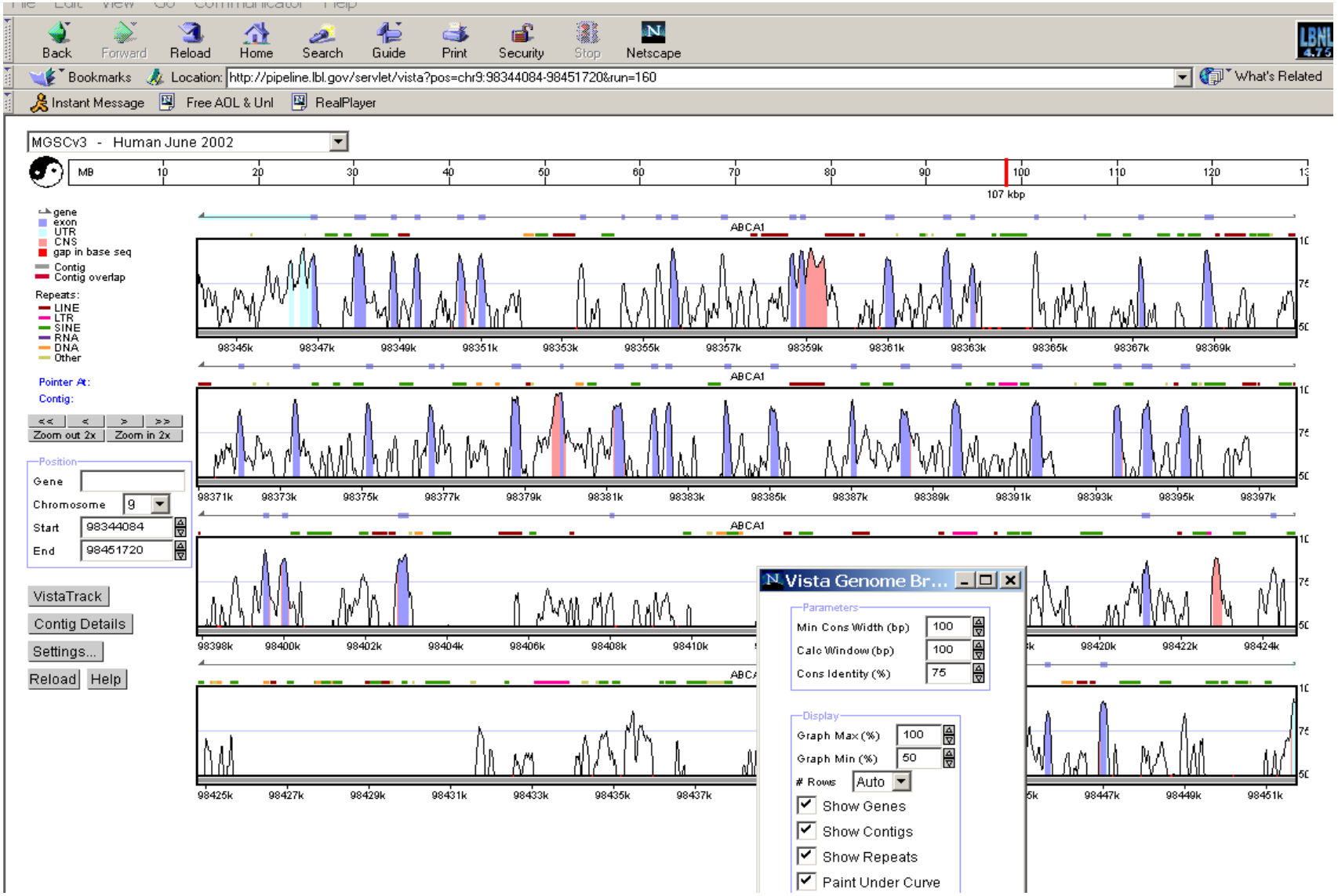
Visualization – VistaBrowser & VistaTrack



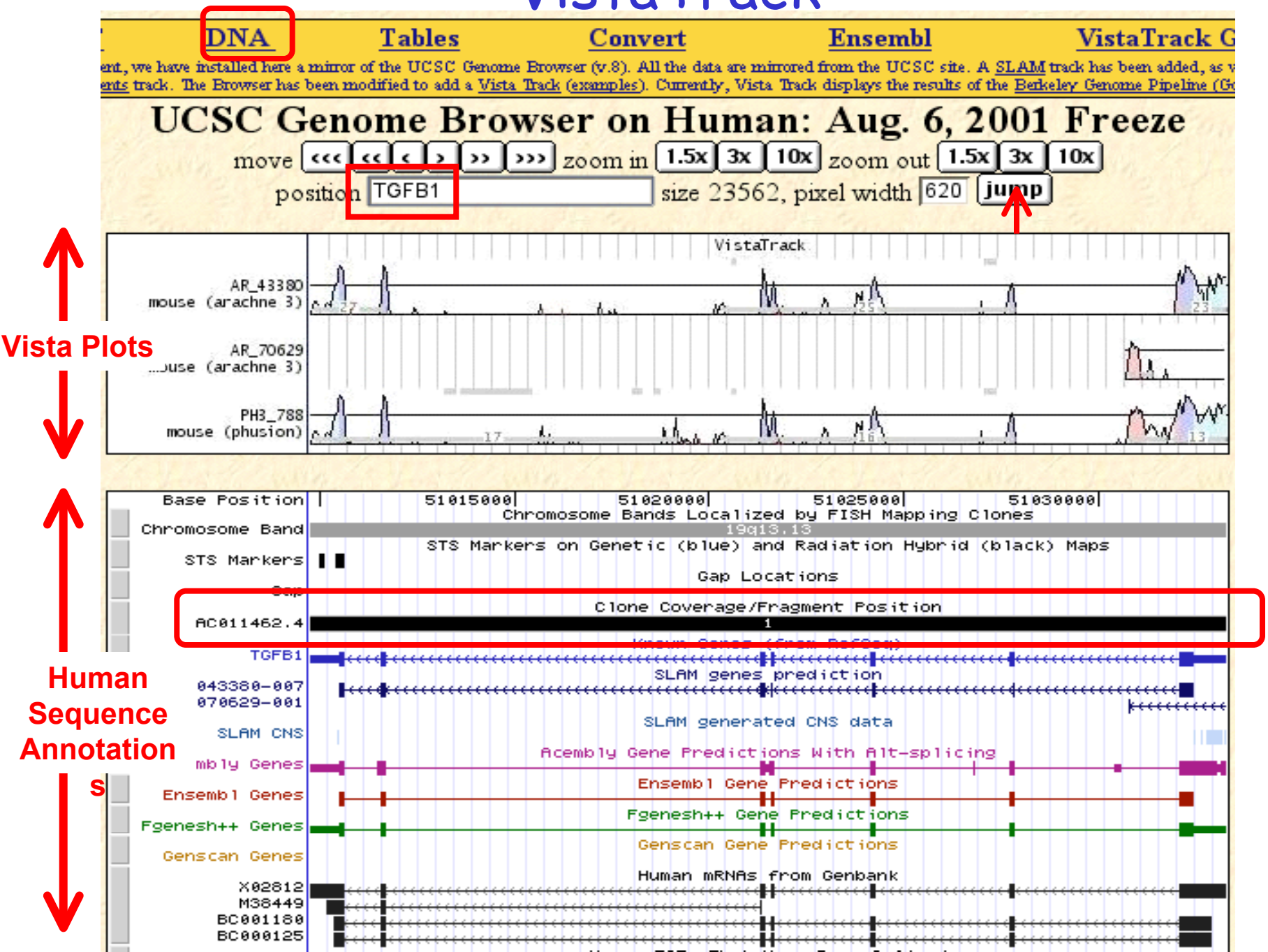
Stand-alone Java applet for detailed comparison

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

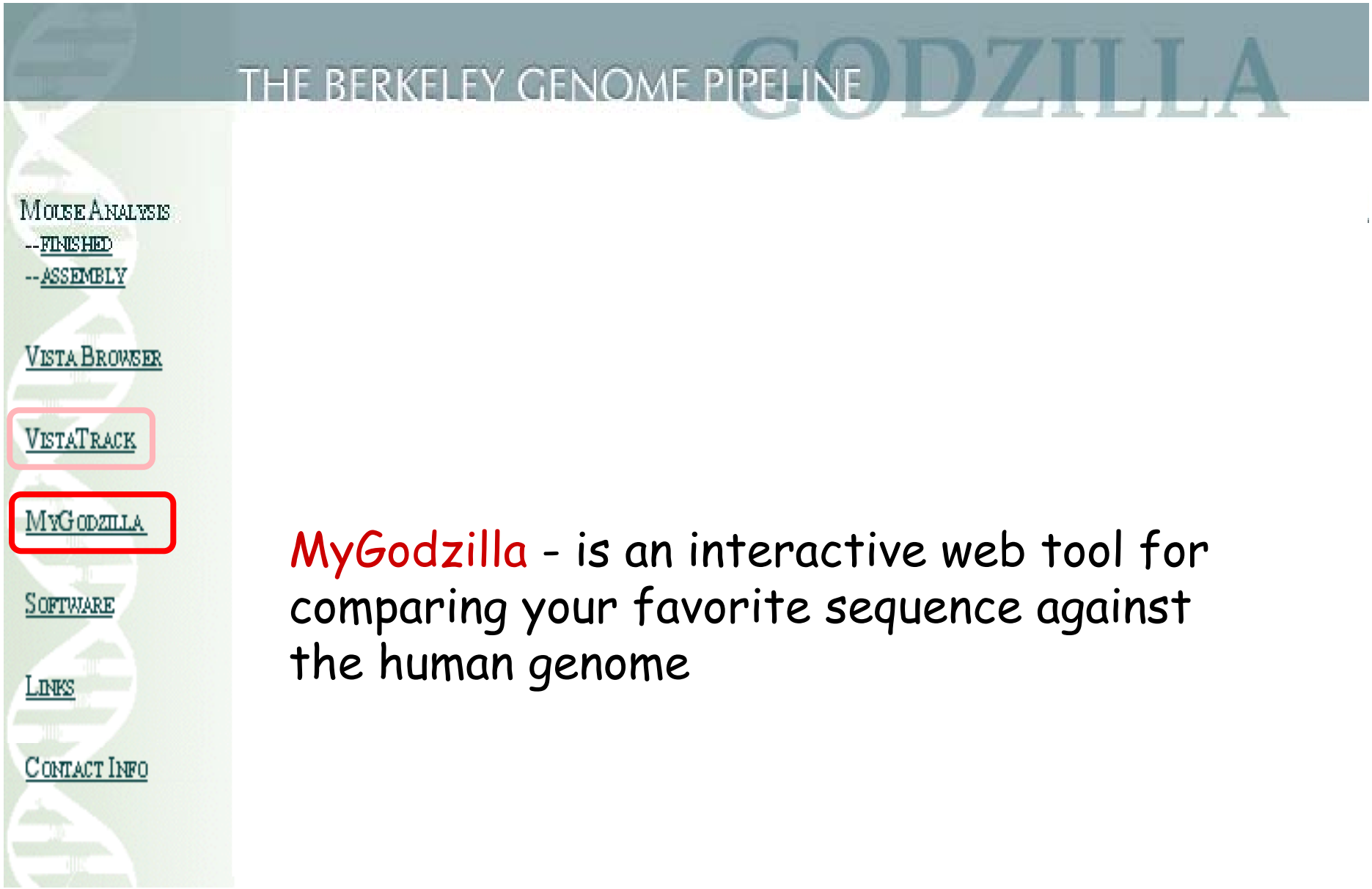
VistaBrowser



VistaTrack


DNA
Tables
Convert
Ensembl
VistaTrack 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
VistaTrack
AR_43380
mouse (arachne 3)
AR_70629
mouse (arachne 3)
PH3_788
mouse (phusion)
Vista Plots
Base Position
51015000
51020000
51025000
51030000
Chromosome Band
Chromosome Bands Localized by FISH Mapping Clones
19q13.13
STS Markers
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
Gap Locations
Clone Coverage/Fragment Position
AC011462.4
TGFB1
SLAM genes prediction
043380-007
070629-001
SLAM CNS
SLAM generated CNS data
Acembly Gene Predictions With Alt-splicing
Ensembl Genes
Ensembl Gene Predictions
Fgenesh++ Genes
Fgenesh++ Gene Predictions
Genscan Genes
Genscan Gene Predictions
Human mRNAs from Genbank
X02812
M38449
BC001180
BC000125
Human Sequence Annotations

http://pipeline.lbl.gov/



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

MyGodzilla Tool

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

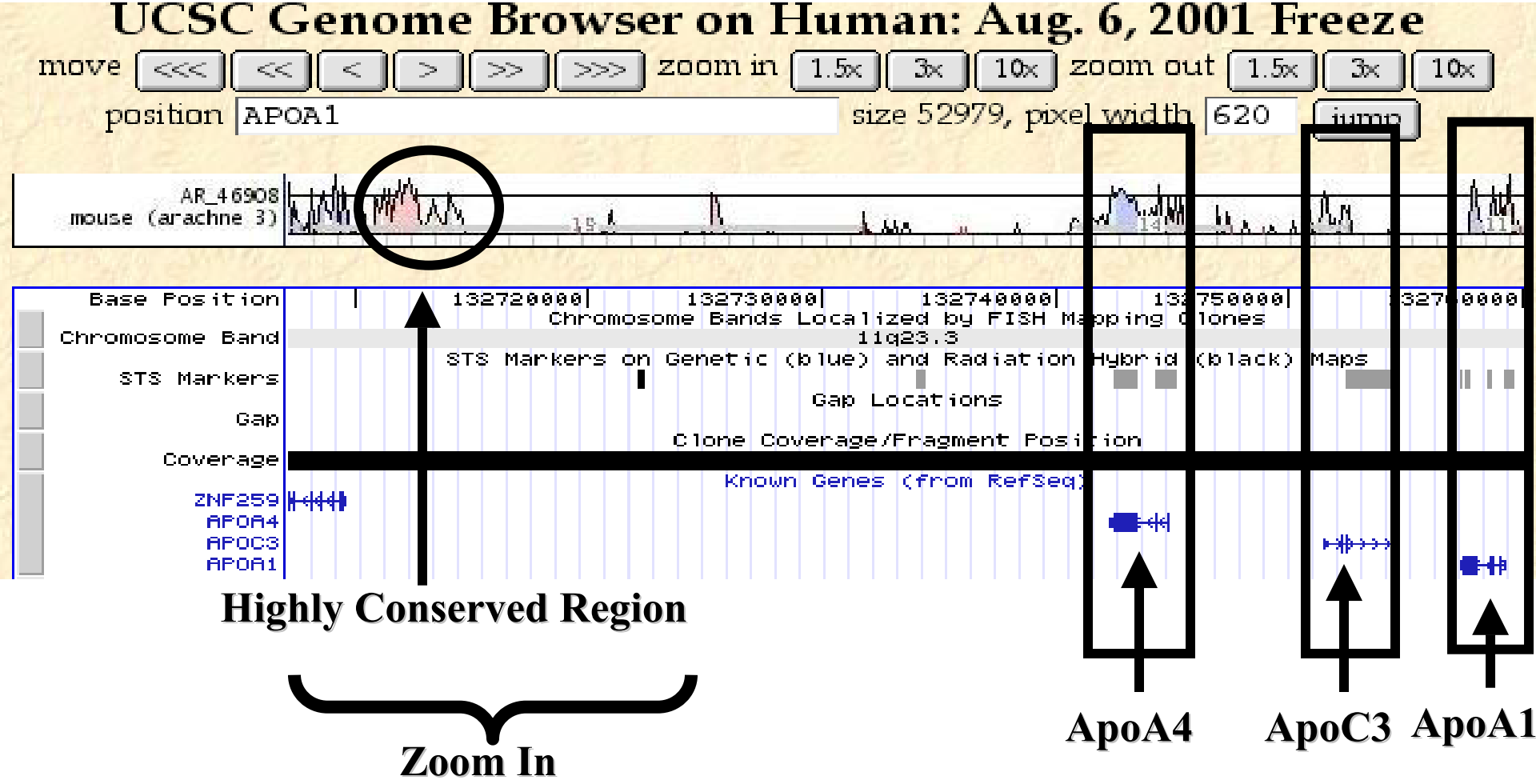
Wayward Discovery of a New Apolipoprotein Gene

T W O

Interleukin Expression Switch

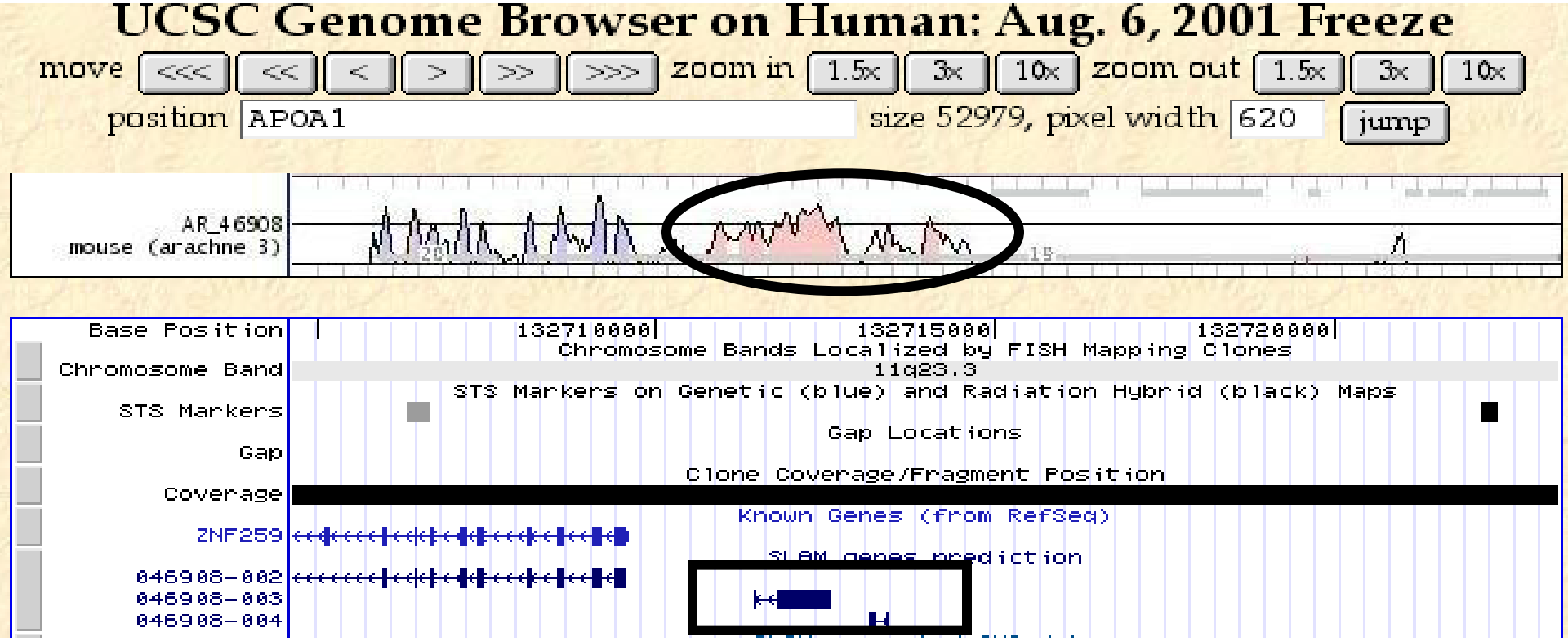
Identification of a New Apo Gene on Human 11q23

Godzilla



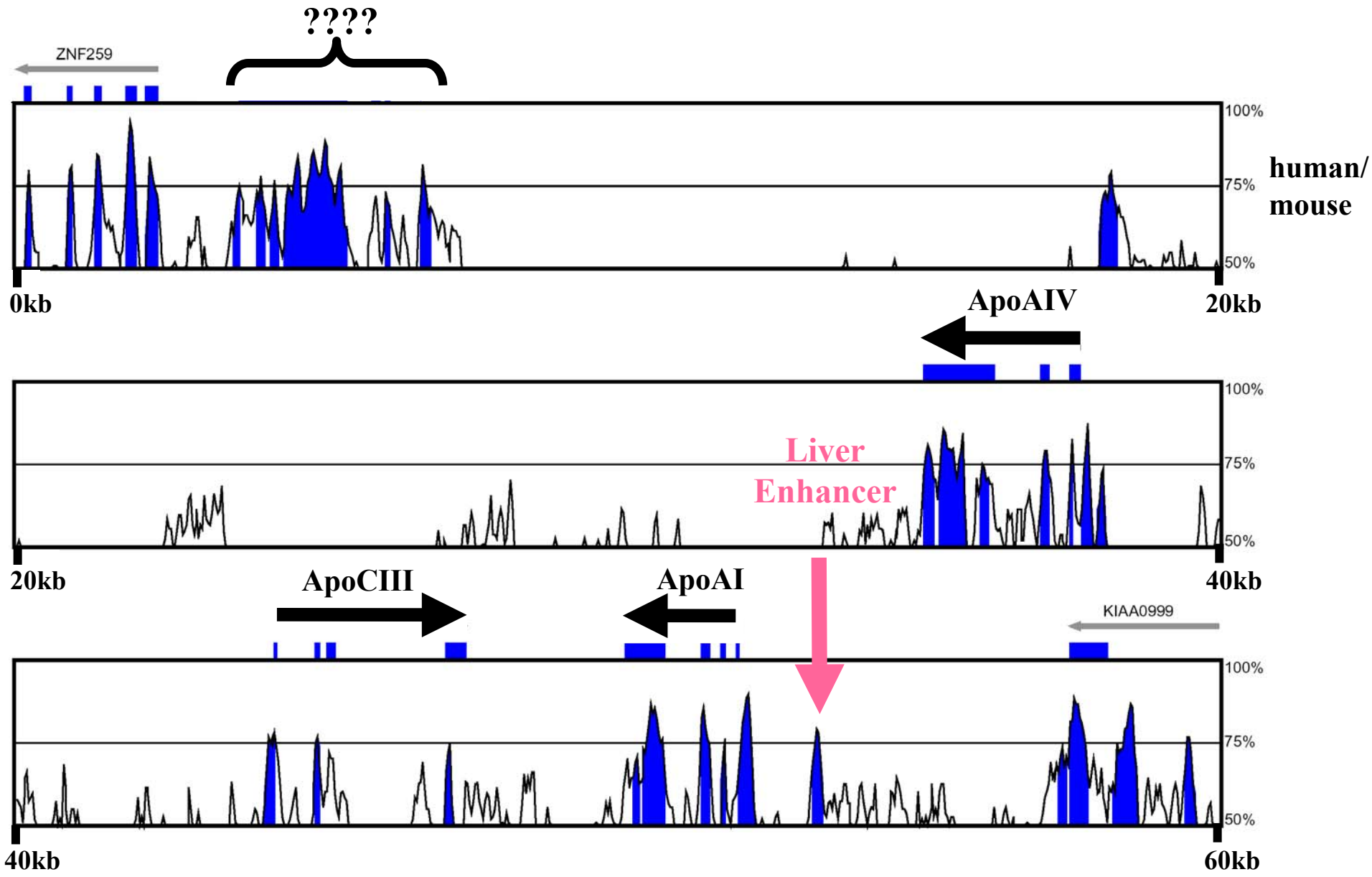
Identification of a New Apo Gene on Human 11q23

Godzilla

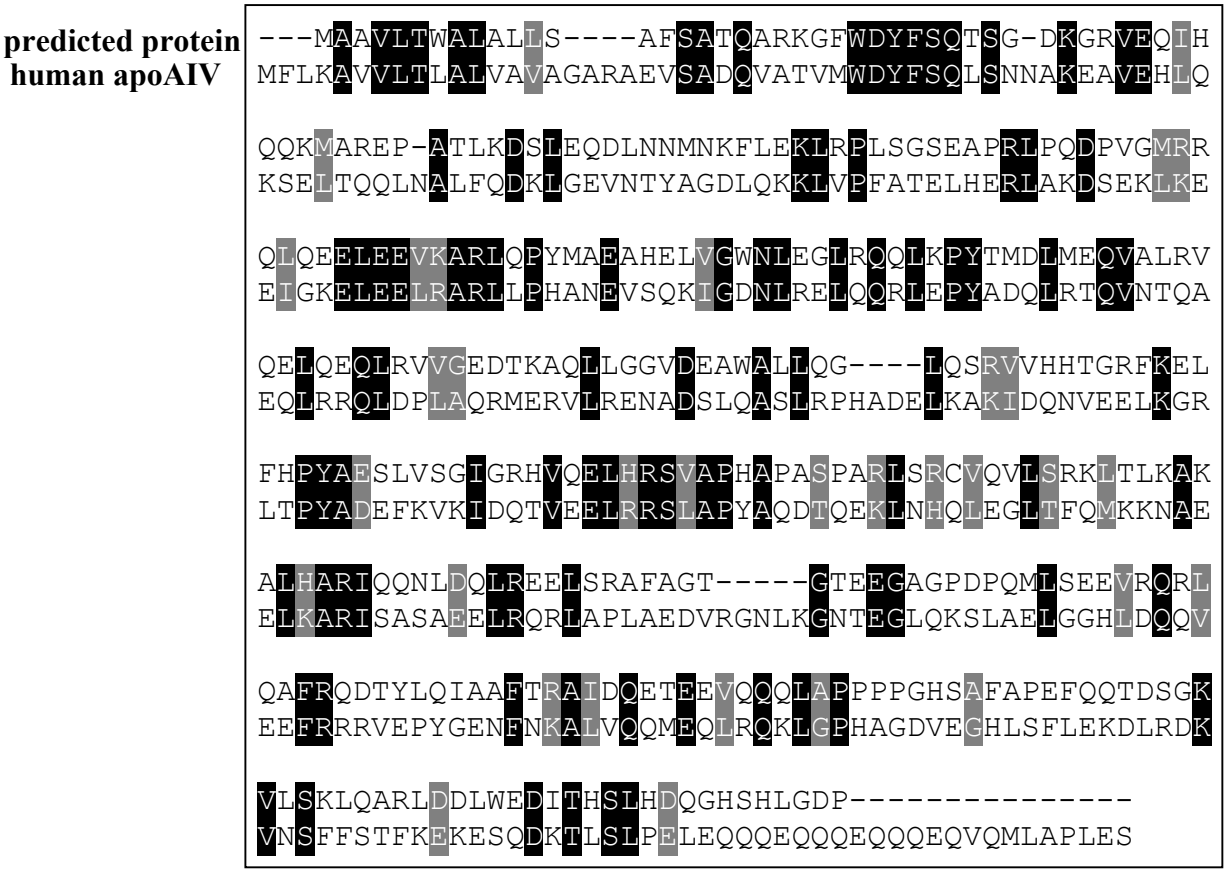


New Gene (ApoA5)
Pennacchio LA et al.
Science. 2001, 294:169-73.

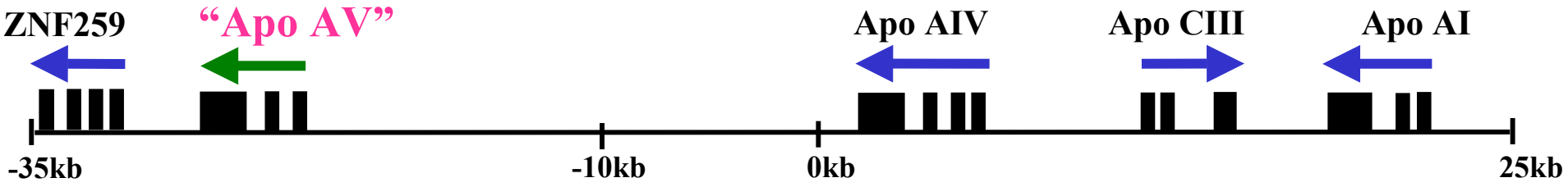
Human/Mouse Apolipoprotein Gene Cluster Sequence Comparison



Predicted Protein Sequence Has Homology to ApoAIV



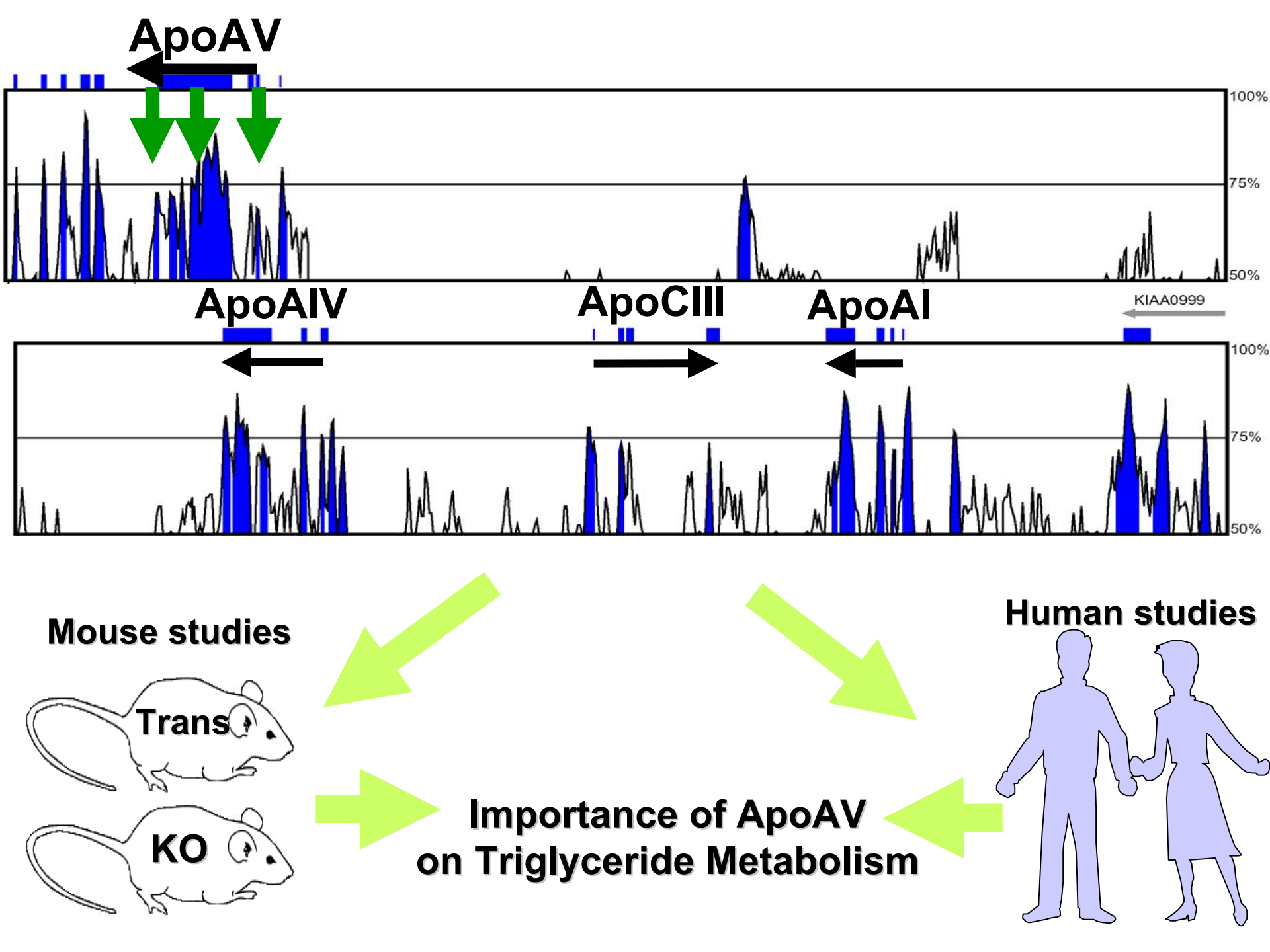
Identity: 26%
Similarity: 45%



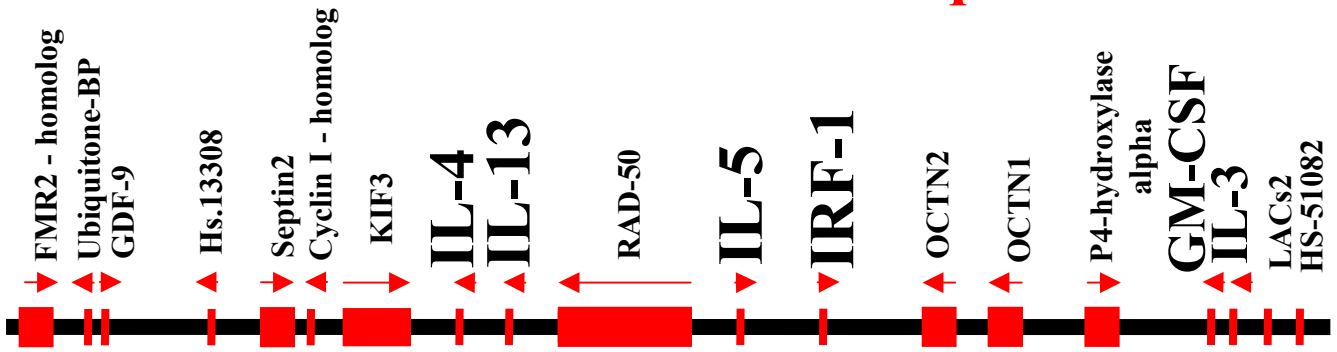
Summary: ApoAV

- 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?


ApoAV
100%
75%
50%
ApoAIV
ApoCIII
ApoAI
KIAA0999
100%
75%
50%
Mouse studies
Trans
KO
Human studies
Importance of ApoAV
on Triglyceride Metabolism

IL Cluster HUM 5q31



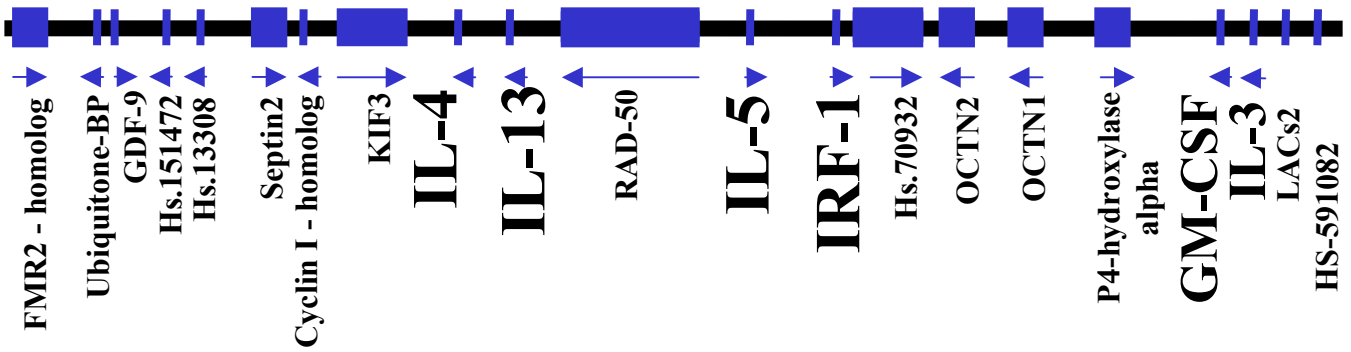
Coding

Exons 3%

Non-Coding

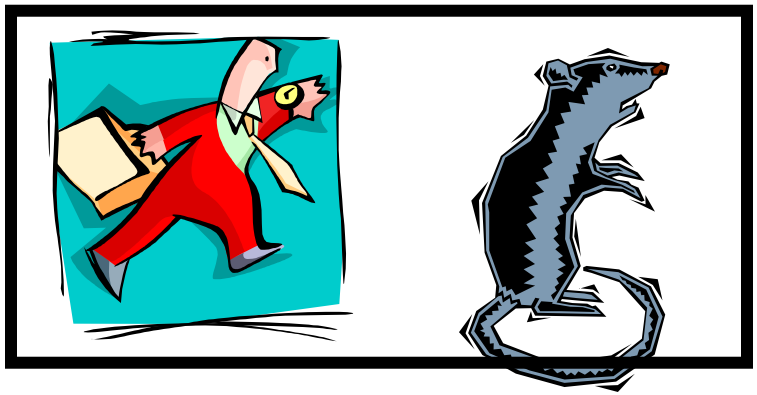
Conserved 2.6%

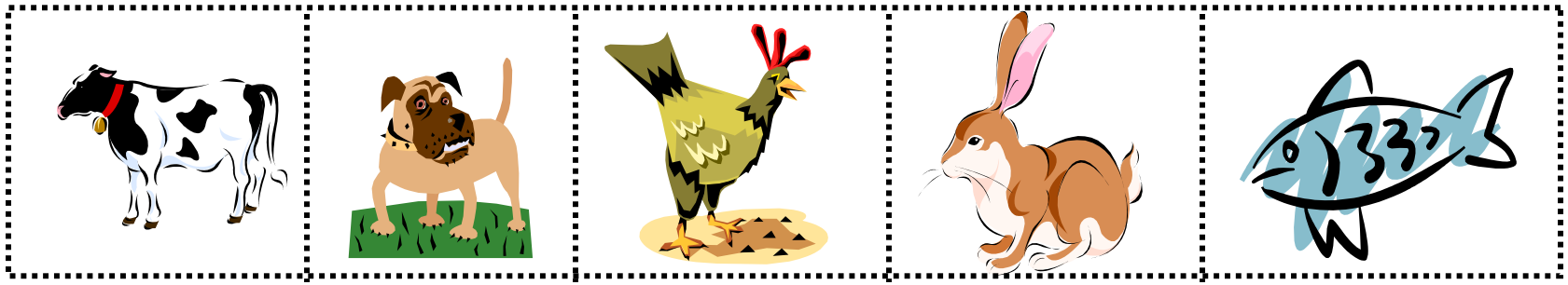
(>100bp > 75%)

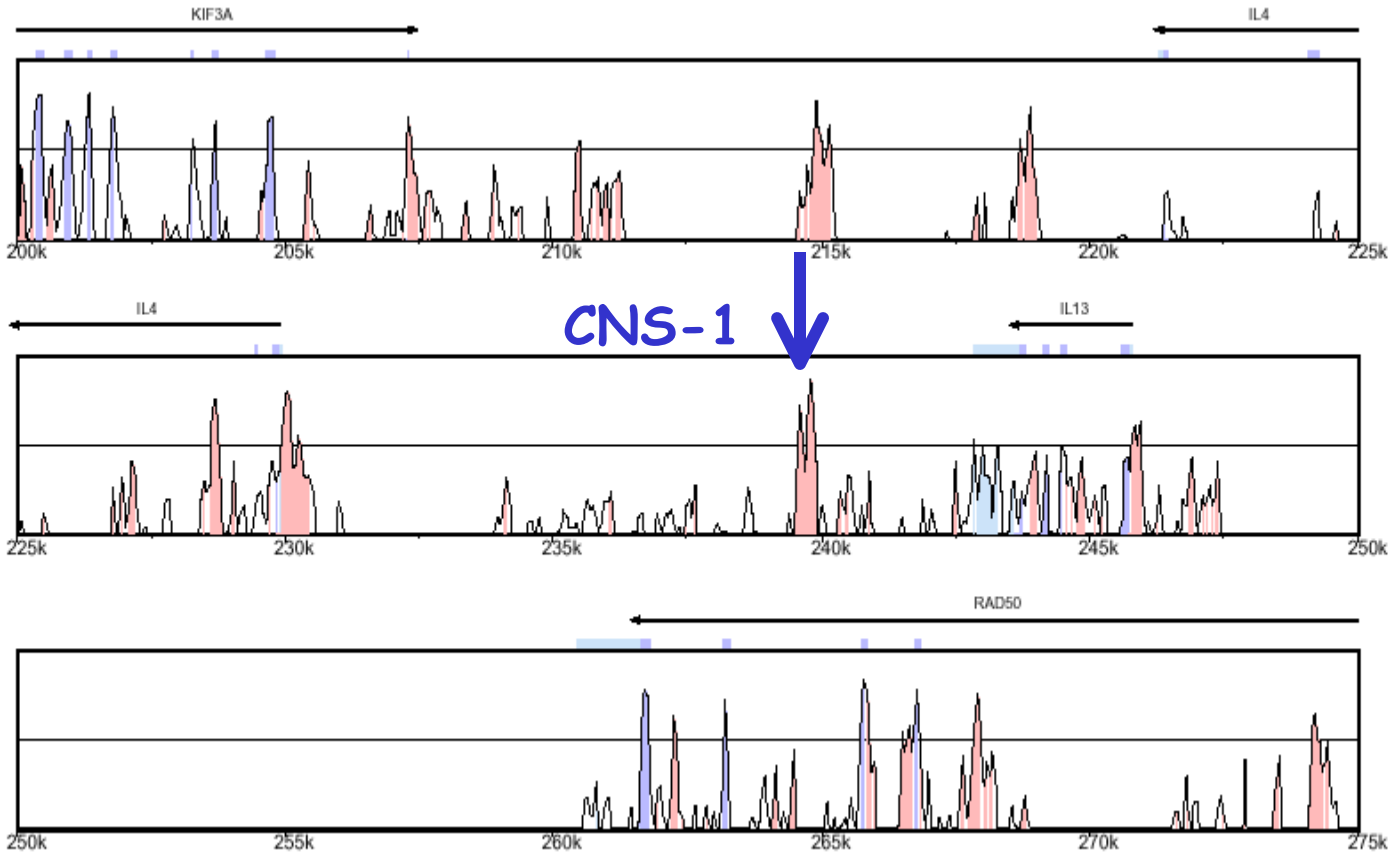


IL Cluster MU Ch 11

A Filtering Strategy







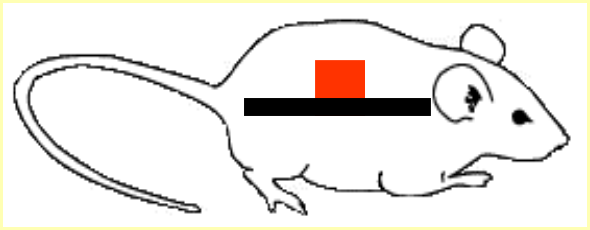
Present in other species: Cow (86%), Dog (81%), Rabbit (73%)

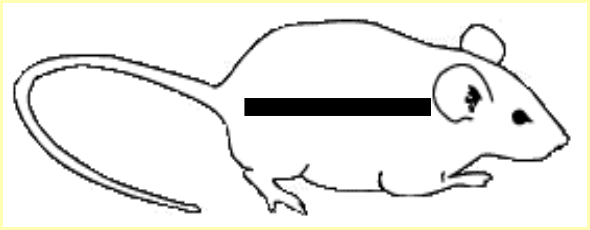
Genomic position conserved in human, mouse, dog, baboon

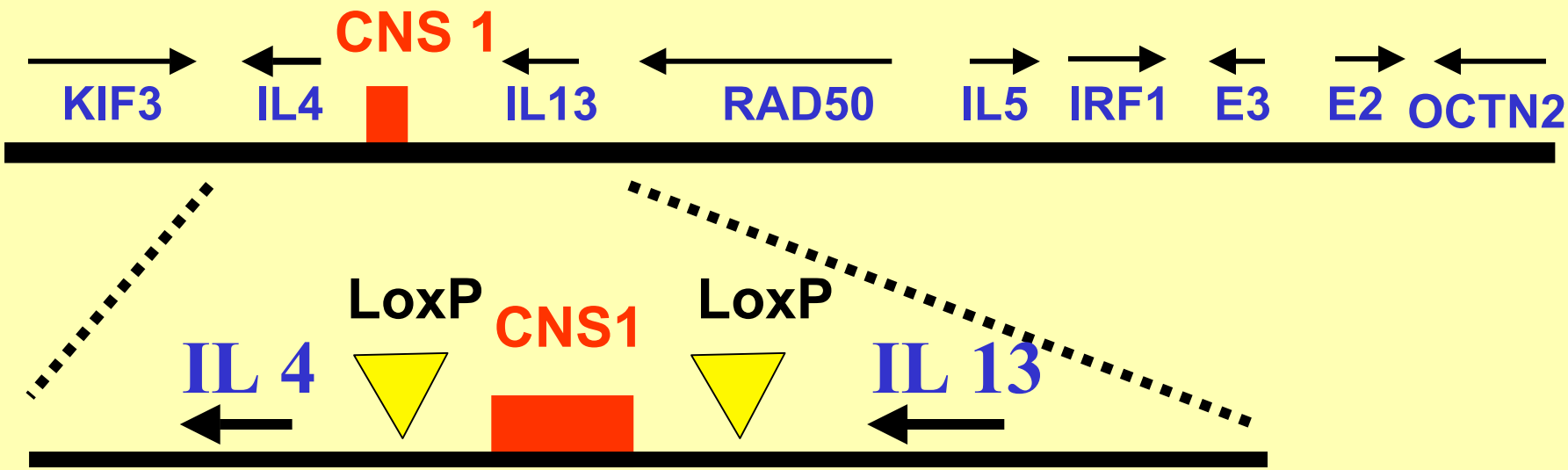
Single copy in the human genome. Two hypersensitive sites mapped.

Functional Analysis of CNS1

Generate Human 5q31 YAC Transgenic Mice

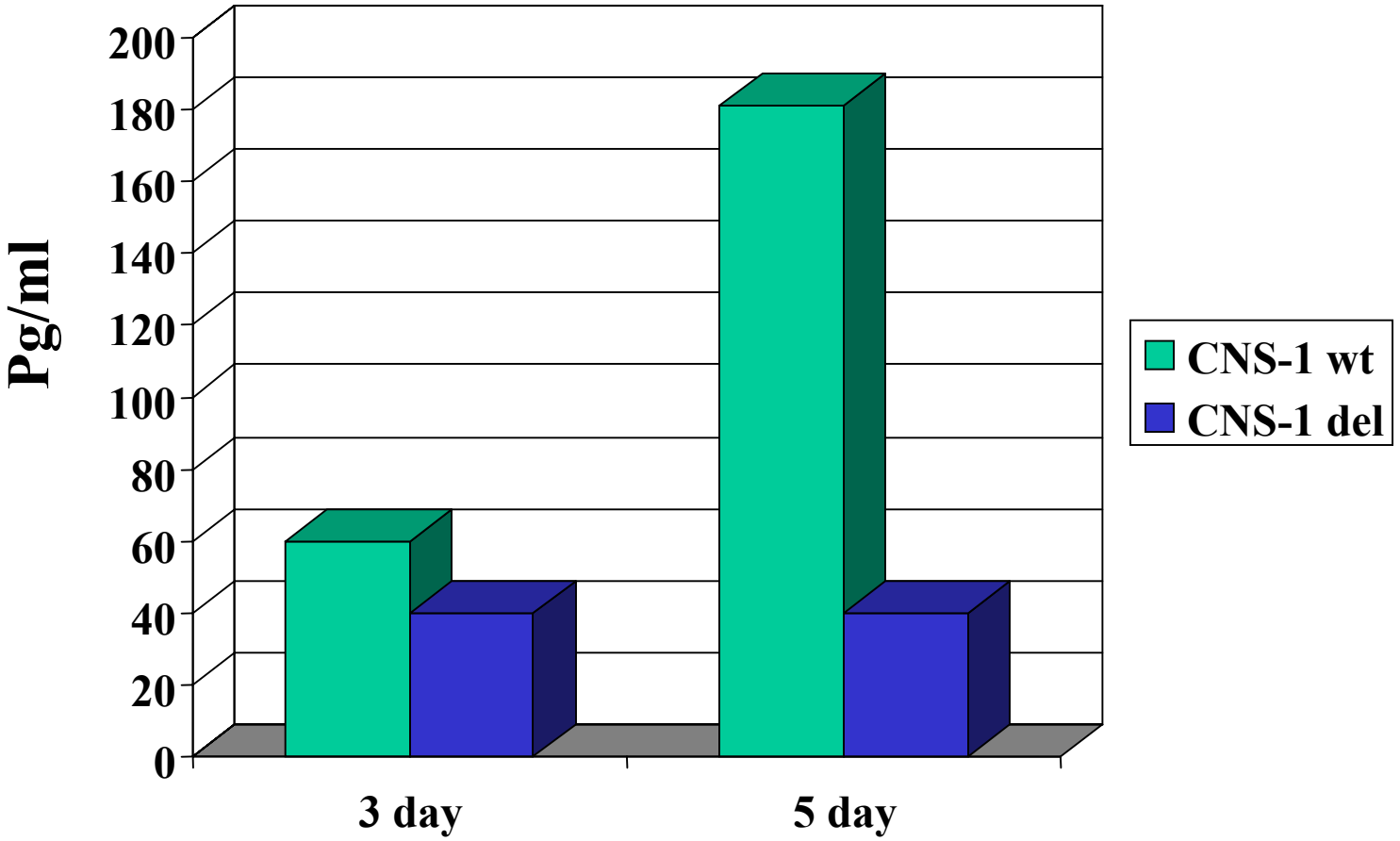


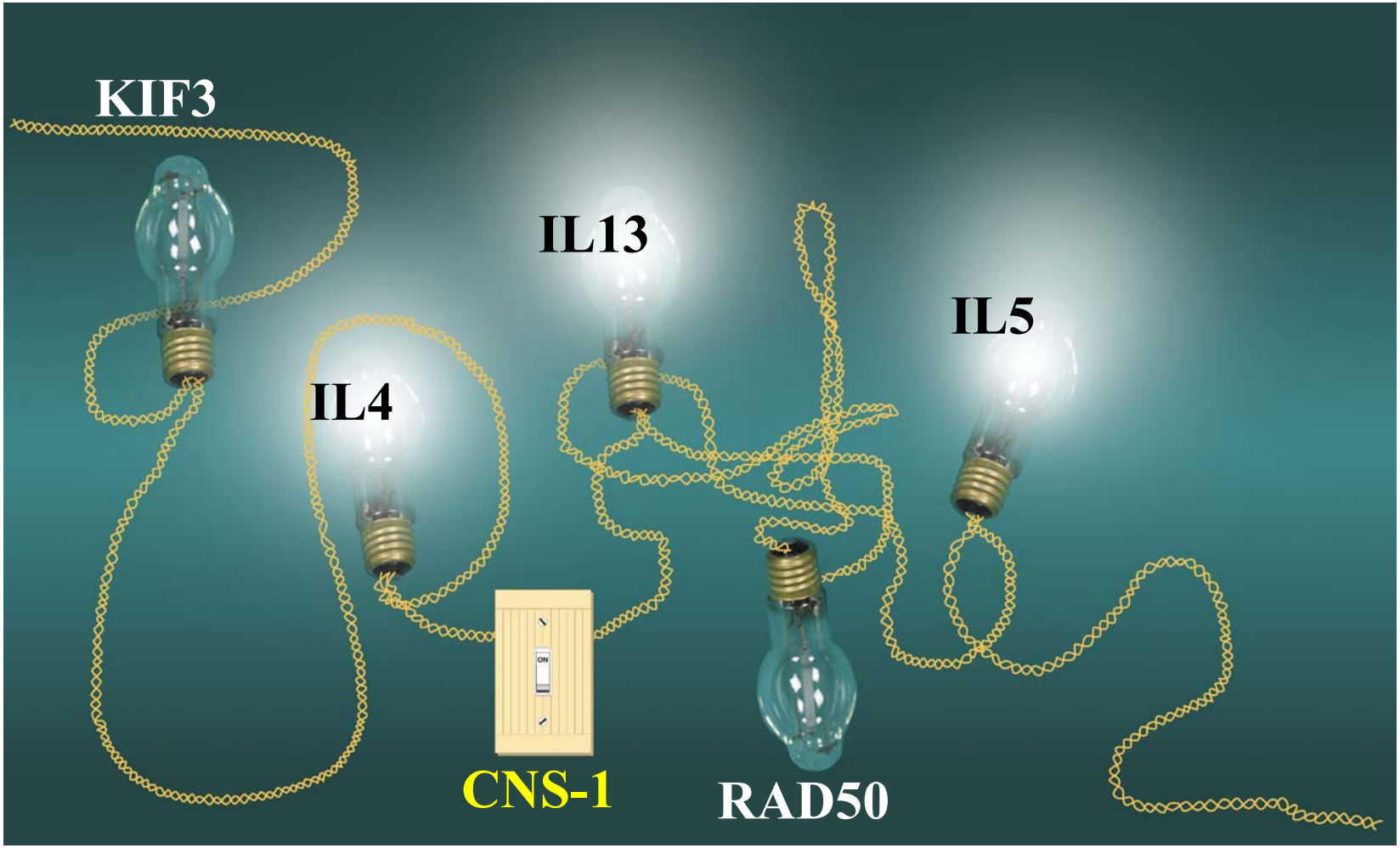




Human IL 4 Production in YAC Transgenics Containing and Lacking CNS1

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