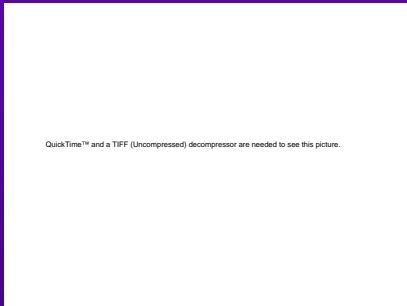


# Sifting Genomic Sequence through Comparative Analysis



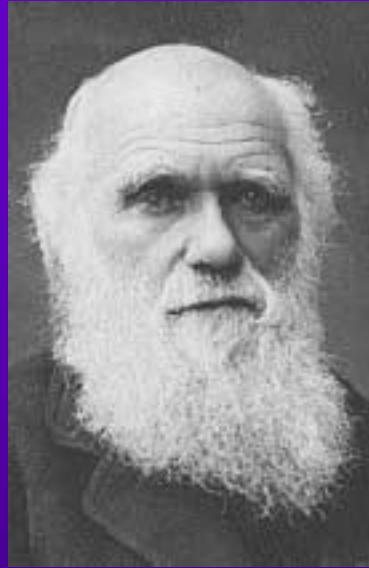
**Len Pennacchio, Ph.D**  
**Staff Scientist**  
**Genome Sciences Department &**  
**Joint Genome Institute (JGI)**  
**Lawrence Berkeley National Laboratory**

# Subset of Genomes with Significant Sequence Available

---



**Mouse**



**Human**



**Fugu**



**Rat**



**Ciona**



**Tetraodon**

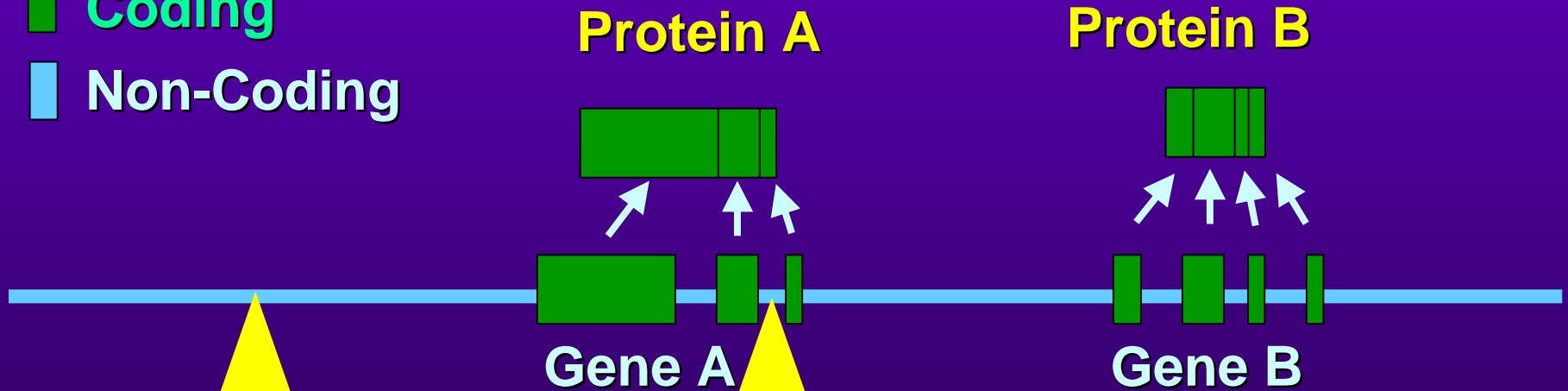
GAAAGACCTGTTGGAGGCTATGAATGCAATCAAGGTGACAGACAACCTGGTGCAATGATGGTAGTGAAATGGAGGAGAGGGGATTGATTC AAGATGCATT  
TAGGACCAAGAAATCGGGAGCTTGTGAACGTGTGTATGAGTACTGTAGACGGAGTGGGTGTGTATCAGAGAAGATCTGAGCATTGGGCTTGCTCTCCTC  
AGAGGCCCTGCGAGTGGAGTTCAGCTTTTCCCTCATGGGGCAAATCTCACTTTGCTCCAGTTCCTGGGGCTCAGAGTCCCTGGCCCAGATGCCTCTTGCC  
ATCTCATCTTCACCCTGCCTGGCTTCCCTTGCTTGTTCAGGATTGTTTCATAAAGAGGGATGTGGTTGGTCTTTAACCCATATGAATGCTGGCTGAGGAT  
GCCTGCGGAACCTGTAGTGAAGCTTTAGGGGCTGCTCGGGTTCTGGCTGGTAGGTGAACACTGTCCATCTTGCCGGCTGGGACACAGTGACTCTGGGTA  
GTTGTGTAAGAGAGGGGCCCTTGGCAGACAAACAGGTTCTTCTCTGTTGGTGGGCCAGCCAGCAGGTGAGTGGGAAGGTTAAAGGTCATGGGGTTTGGGA  
GAAACTGGGTGAGGAGTTCAGCCCCATCCCCGTAAAGCTCCTGGGAAGCACTTCTCTACTGGGGCAGCCCCTGATACCAGGGCACTCATTAACCCCTCTG  
GGTGCCAGGGAAAGGGCAGGAGTGTAGTGTGGGAGGCAGCTGAGGTCAACTTCTTTTGAACCTCCACGTGGTATTTACTCAGAGCAATTGGTGCCAGAG  
GCTCAGGGCCCTGGAGTATAAAGCAGAATGTCTGCTCTCTGTGCCAGACGTGAGCAGGTGAGCAGCTGGGGCGAAAGACCTGTTGGAGGCTATGAATGC  
AATCAAGGTGACAGACAACCTGGTGCAATGATGGTAGTGAAATGGAGGAGAGGGGATTGATTC AAGATGCATTTAGGACCAAGAATCGGGAGCTTGTGAA  
CGTGTGTATGAGTACTGTAGACGGAGTGGGTGTGTATCAGAGAAGATCTGAGCATTGGGCTTGCTCTCCTCAGAGGCCCTGCGAGTGGAGTTCAGCTT  
TTCTCATGGGGCAAATCTCACTTTGCTCCAGTTCCTGGGGCTCAGAGTCCCTGGCCCAGATGCCTCTTGCCATCTCATCTTCACCCTGCCTGGCTTCC  
CTTGCTTGTTCAGGATTGTTTCATAAAGAGGGATGTGGTTGGTCTTTAACCCATATGAATGCTGGCTGAGGATGCCTGCGGAACCTGTAGTGAAGCTTTC  
AGGGGCTGCTCGGGTTCTGGCTGGTAGGTGAACACTGTCCATCTTGCCGGCTGGGACACAGTGACTCTGGGTAGTTGTGTAAGAGAGGGGGCCCTTGGCAG  
ACAAACAGGTTCTTCTCTGTTGGTGGGCCAGCCAGCAGGTGAGTGGGAAGGTTAAAGGTCATGGGGTTTGGGAGAACTGGGTGAGGAGTTCAGCCCCATC  
CCCCGTAAAGCTCCTGGGAAGCACTTCTCTACTGGGGCAGCCCCTGATACCAGGGCACTCATTAACCCCTCTGGGTGCCAGGGAAAGGGCAGGAGGTGAGT  
GCTGGGAGGCAGCTGAGGTCAACTTCTTTTGAACCTCCACGTGGTATTTACTCAGAGCAATTGGTGCCAGAGGCTCAGGGCCCTGGAGTATAAAGCAGAA  
TGTCTGCTCTCTGTGCCAGACGTGAGCAGGTGAGCAGCTGGGGCTGTCTGCTCTCTGTGCCAGACGTGAGCAGGTGAGCAGCTGGGGCTGTCTGCTCT  
CTGTGCCAGACGTGAGCAGGTGAGCAGCTGGGGCTGTCTGCTCTCTGTGCCAGACGTGAGCAGGTGAGCAGCTGGGGCTGTCTGCTCTCTGTGCCAG  
GAAAGACCTGTTGGAGGCTATGAATGCAATCAAGGTGACAGACAACCTGGTGCAATGATGGTAGTGAAATGGAGGAGAGGGGATTGATTC AAGATGCATT  
TAGGACCAAGAAATCGGGAGCTTGTGAACGTGTGTATGAGTACTGTAGACGGAGTGGGTGTGTATCAGAGAAGATCTGAGCATTGGGCTTGCTCTCCTC  
AGAGGCCCTGCGAGTGGAGTTCAGCTTTTCCCTCATGGGGCAAATCTCACTTTGCTCCAGTTCCTGGGGCTCAGAGTCCCTGGCCCAGATGCCTCTTGCC  
ATCTCATCTTCACCCTGCCTGGCTTCCCTTGCTTGTTCAGGATTGTTTCATAAAGAGGGATGTGGTTGGTCTTTAACCCATATGAATGCTGGCTGAGGAT  
GCCTGCGGAACCTGTAGTGAAGCTTTAGGGGCTGCTCGGGTTCTGGCTGGTAGGTGAACACTGTCCATCTTGCCGGCTGGGACACAGTGACTCTGGGTA  
GTTGTGTAAGAGAGGGGCCCTTGGCAGACAAACAGGTTCTTCTCTGTTGGTGGGCCAGCCAGCAGGTGAGTGGGAAGGTTAAAGGTCATGGGGTTTGGGA  
GAAACTGGGTGAGGAGTTCAGCCCCATCCCCGTAAAGCTCCTGGGAAGCACTTCTCTACTGGGGCAGCCCCTGATACCAGGGCACTCATTAACCCCTCTG  
GGTGCCAGGGAAAGGGCAGGAGTGTAGTGTGGGAGGCAGCTGAGGTCAACTTCTTTTGAACCTCCACGTGGTATTTACTCAGAGCAATTGGTGCCAGAG  
GCTCAGGGCCCTGGAGTATAAAGCAGAATGTCTGCTCTCTGTGCCAGACGTGAGCAGGTGAGCAGCTGGGGCGAAAGACCTGTTGGAGGCTATGAATGC  
AATCAAGGTGACAGACAACCTGGTGCAATGATGGTAGTGAAATGGAGGAGAGGGGATTGATTC AAGATGCATTTAGGACCAAGAATCGGGAGCTTGTGAA  
CGTGTGTATGAGTACTGTAGACGGAGTGGGTGTGTATCAGAGAAGATCTGAGCATTGGGCTTGCTCTCCTCAGAGGCCCTGCGAGTGGAGTTCAGCTT  
TTCTCATGGGGCAAATCTCACTTTGCTCCAGTTCCTGGGGCTCAGAGTCCCTGGCCCAGATGCCTCTTGCCATCTCATCTTCACCCTGCCTGGCTTCC  
CTTGCTTGTTCAGGATTGTTTCATAAAGAGGGATGTGGTTGGTCTTTAACCCATATGAATGCTGGCTGAGGATGCCTGCGGAACCTGTAGTGAAGCTTTC  
AGGGGCTGCTCGGGTTCTGGCTGGTAGGTGAACACTGTCCATCTTGCCGGCTGGGACACAGTGACTCTGGGTAGTTGTGTAAGAGAGGGGGCCCTTGGCAG  
ACAAACAGGTTCTTCTCTGTTGGTGGGCCAGCCAGCAGGTGAGTGGGAAGGTTAAAGGTCATGGGGTTTGGGAGAACTGGGTGAGGAGTTCAGCCCCATC  
CCCCGTAAAGCTCCTGGGAAGCACTTCTCTACTGGGGCAGCCCCTGATACCAGGGCACTCATTAACCCCTCTGGGTGCCAGGGAAAGGGCAGGAGGTGAGT  
GCTGGGAGGCAGCTGAGGTCAACTTCTTTTGAACCTCCACGTGGTATTTACTCAGAGCAATTGGTGCCAGAGGCTCAGGGCCCTGGAGTATAAAGCAGAA  
TGTCTGCTCTCTGTGCCAGACGTGAGCAGGTGAGCAGCTGGGGCTGTCTGCTCTCTGTGCCAGACGTGAGCAGGTGAGCAGCTGGGGCTGTCTGCTCT  
CTGTGCCAGACGTGAGCAGGTGAGCAGCTGGGGCTGTCTGCTCTCTGTGCCAGACGTGAGCAGGTGAGCAGCTGGGGCTGTCTGCTCTCTGTGCCAG

# Inventory of Mammalian DNA

Coding ~2%

Non-coding ~98%

■ Coding  
■ Non-Coding



**Intergenic**  
(~70%)  
Junk?

**Intronic**  
(~30%)  
Junk?

**FUNCTION?**  
-Study which  
portion first?

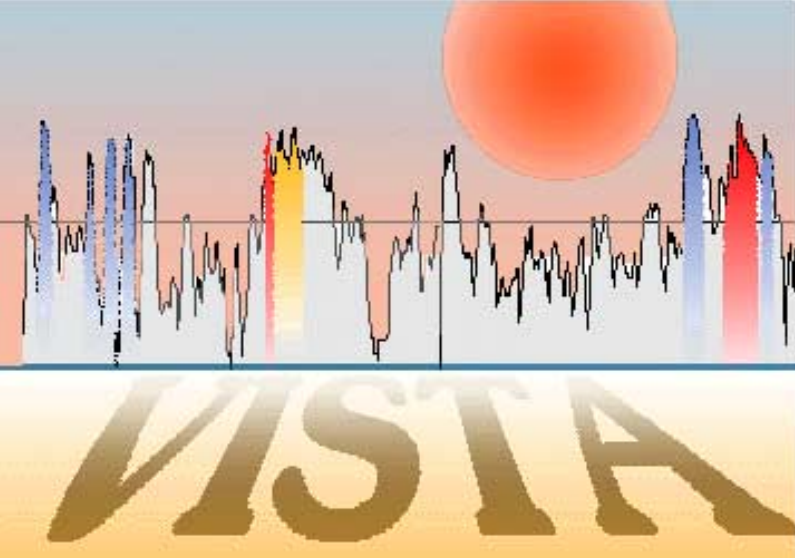


# **Sifting Comparative Sequence: Human Genes and Regulation**

---

- **Comparative Genomic Tools and Databases**
  - Prioritizing Genomic Sequence for Biological Study
- **Examples of Insights into Gene Regulation**
  - Mouse Transgenic Studies
  - Power of Non-Mammalian Sequence (Chicken/Fugu)
- **Identification of a Novel Gene (ApoAV)**
  - Functional Characterization
  - Human Genetic Association Studies





**VIS**UALIZATION **T**OOLS FOR **A**LIGNMENTS

VISTA is an integrated system for **global sequence alignment** and **visualization**, designed for comparative genomic analysis

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

# 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.

## Visualization

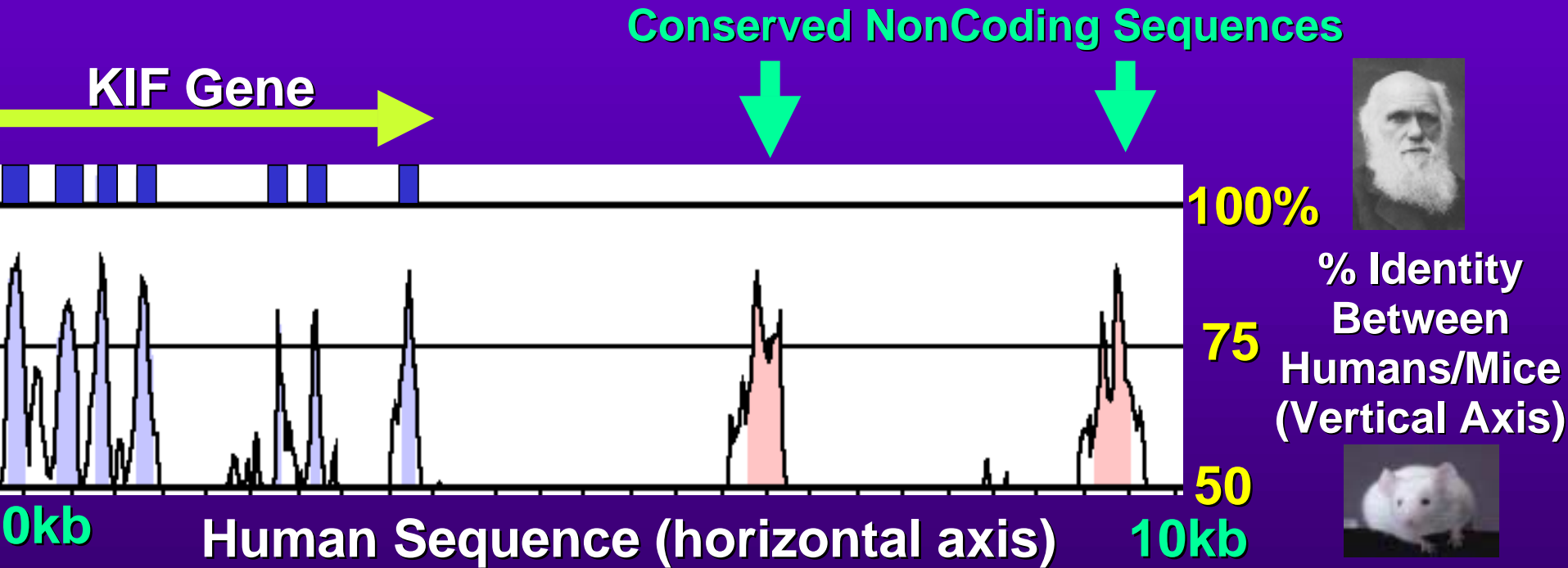


tggtaacattcaaattatg — tctcaaagttagcatgaca-actttttccatgg  
||| ||| | | || ||| | ||| | | |||  
tgatgacatctatttgctgtttccttttagaaactgcatgagagcctggctagtaggg



# VISTA Plot

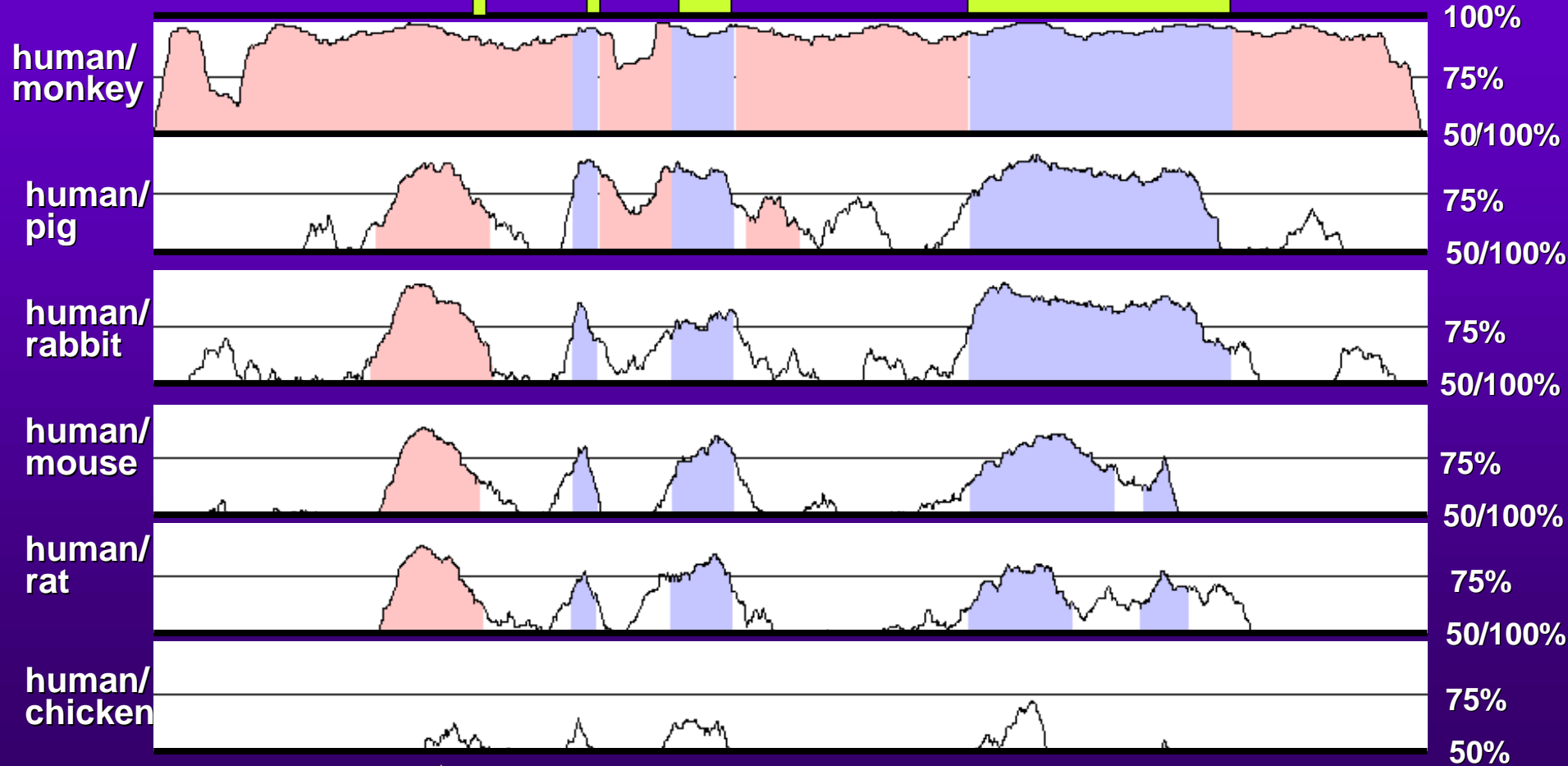
(VISual Tool for Alignment)



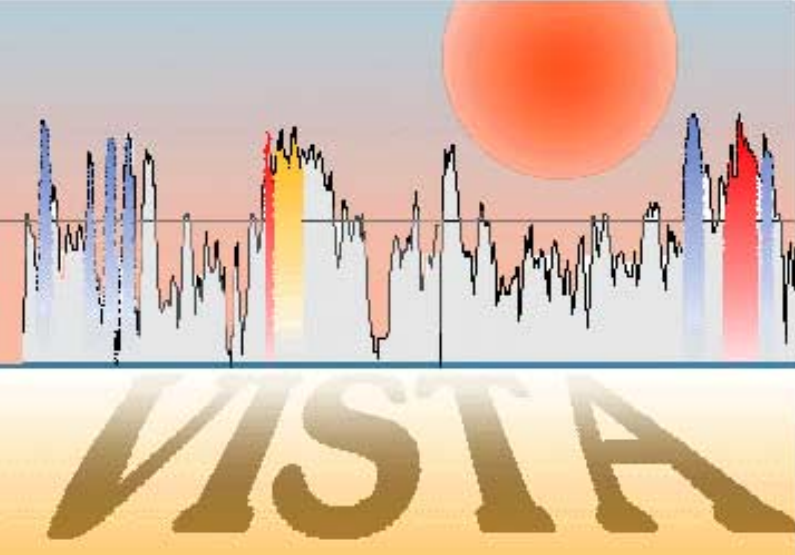
<http://www-gsd.lbl.gov/vista>

# Multi-Species Comparative Analysis (VISTA)

Apolipoprotein A1 gene



Liver enhancer



**VIS**UALIZATION **T**OOLS FOR **A**LIGNMENTS

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

**mVISTA:** main VISTA

-standard comparative sequence plots

**rVISTA:** regulatory VISTA

-conserved transcription factor binding sites

**Availability:**

Web-based

Stand Alone Package

# VISTA Browser

## Pre-processed Whole Human/Mouse Genome Comparison

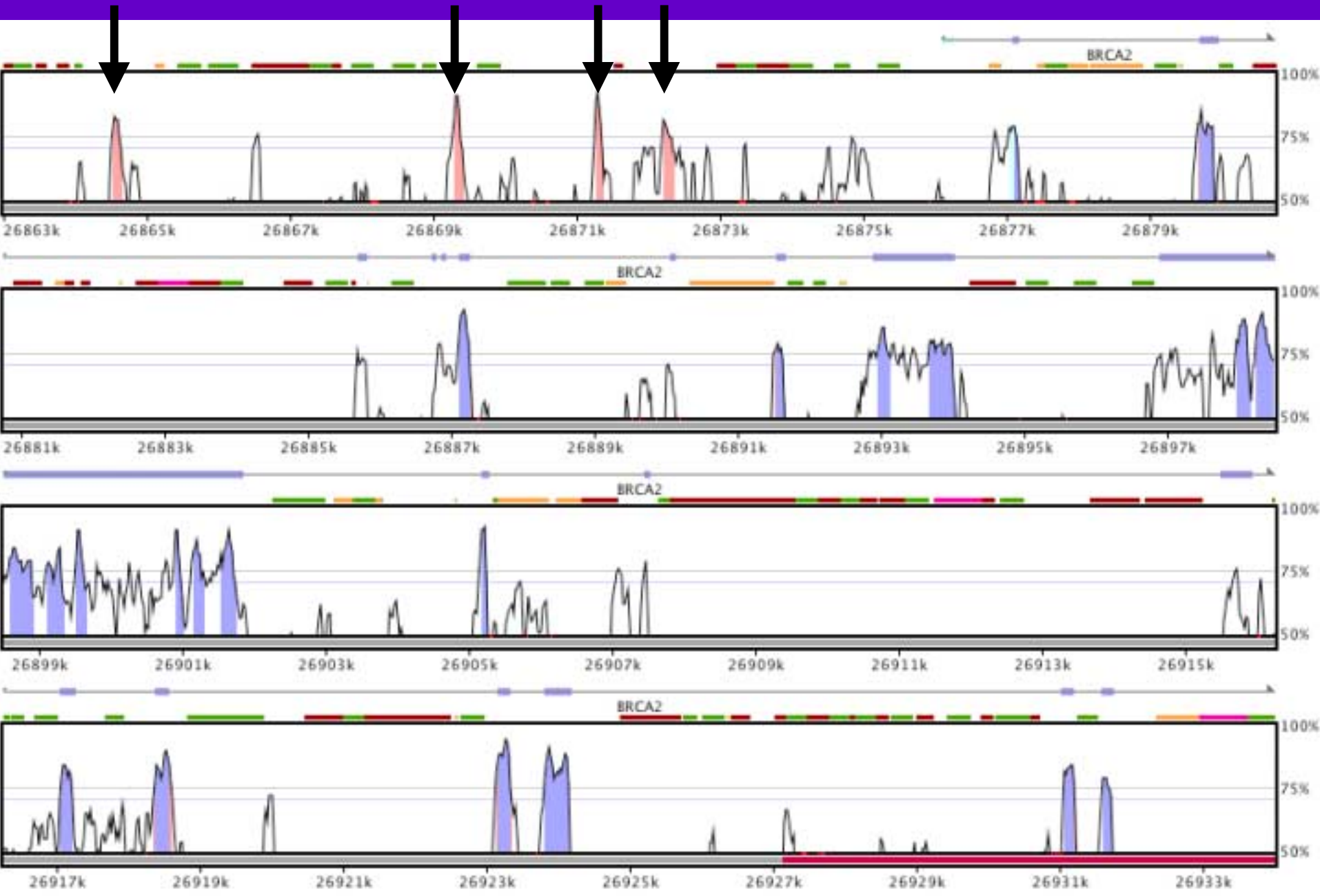
THE BERKELEY GENOME PIPELINE **GODZILLA**

Compare the Human and Mouse Genomes

Please enter a position in the Human Genome (June 2002 draft assembly), select the browser to display comparative analysis results, and press the submit button:

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

# VISTA Browser (Human/Mouse BRCA2 Comparison)



# GenomeVISTA

Self-Input Sequence Comparison to either Human, Mouse or Rat Reference Genomes

## 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**

Browse...

Or

**GenBank**

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

GenBank Locus:  
Accession or GI Number



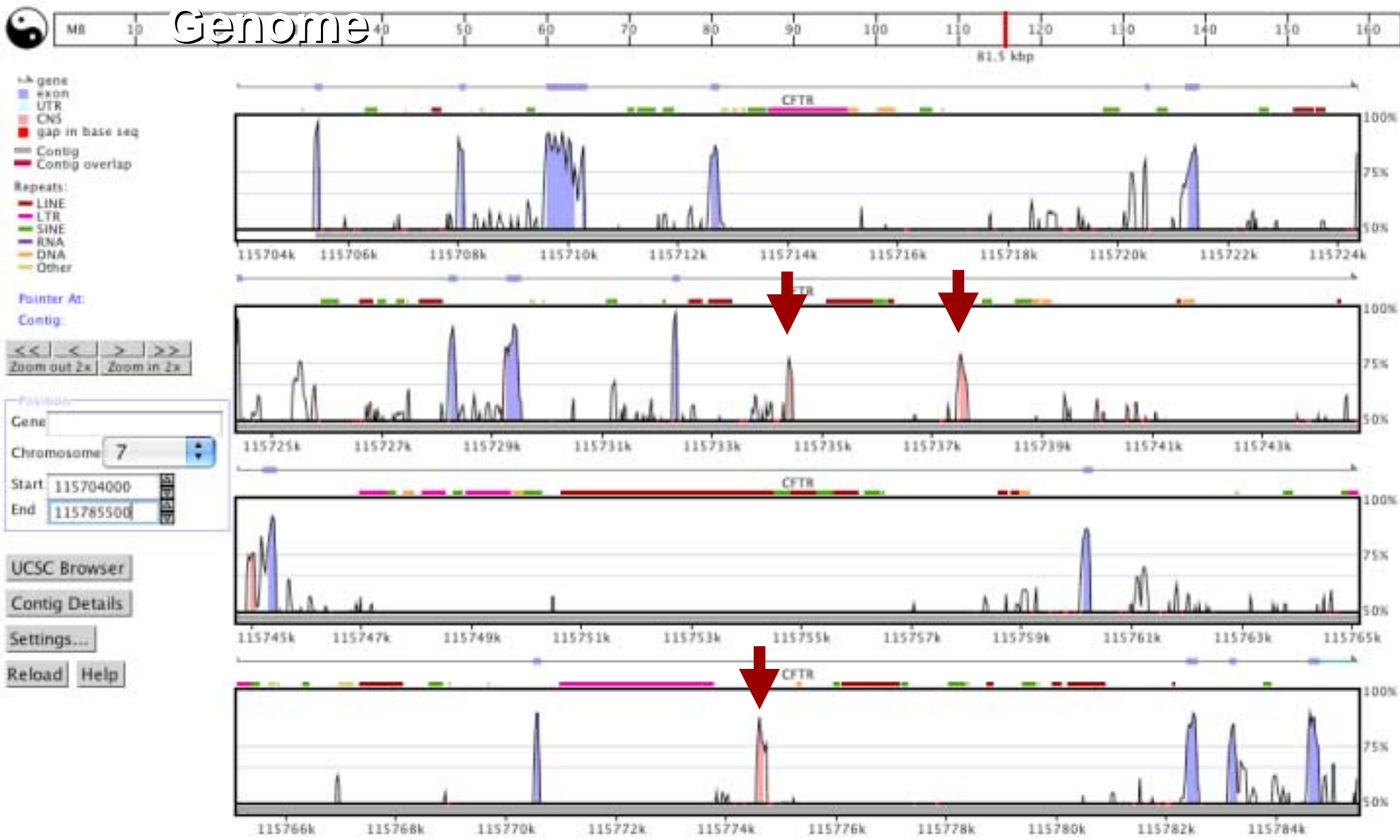
opossum

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

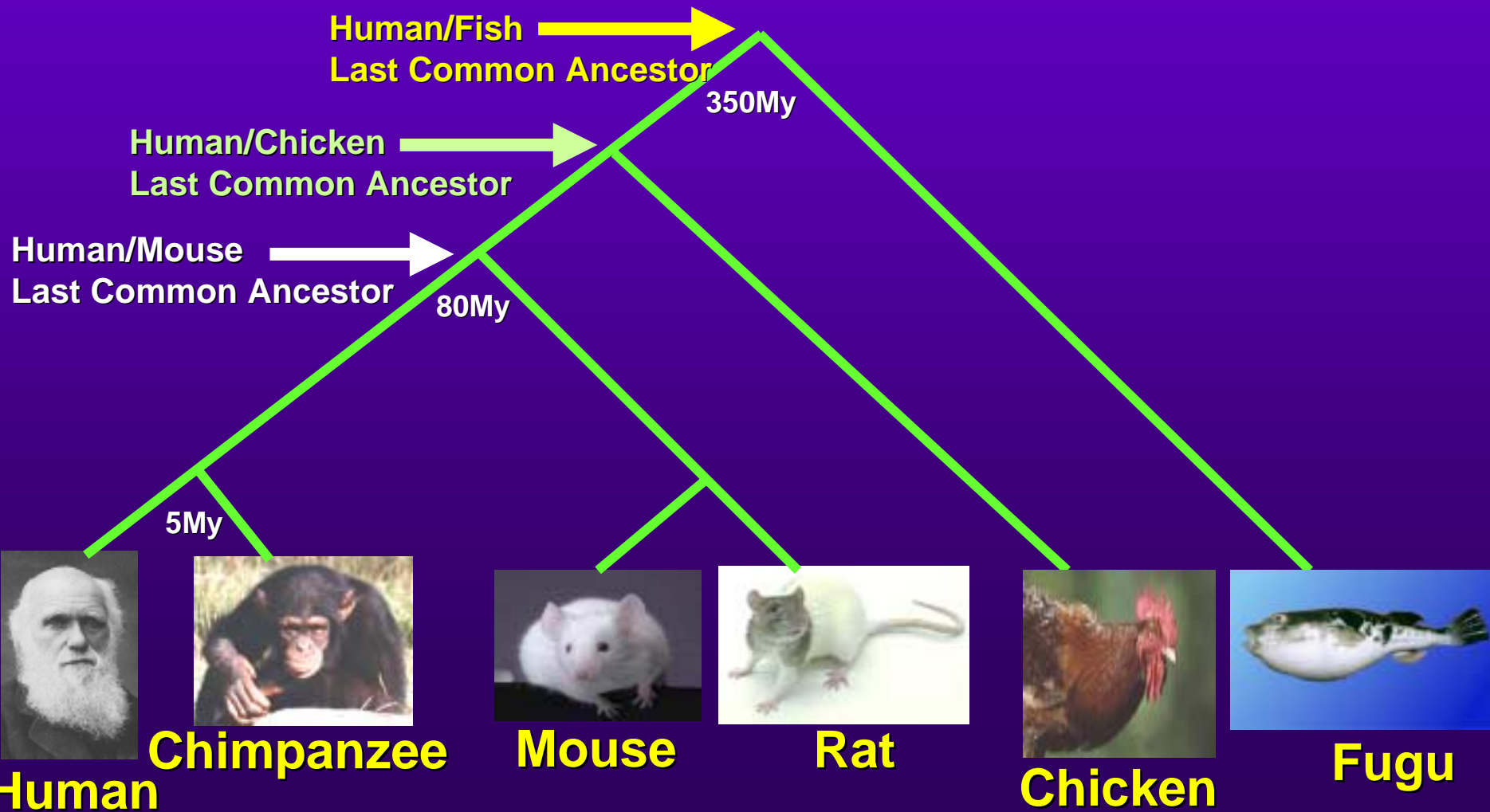


# GenomeVISTA

## Random Opposum BAC versus Human

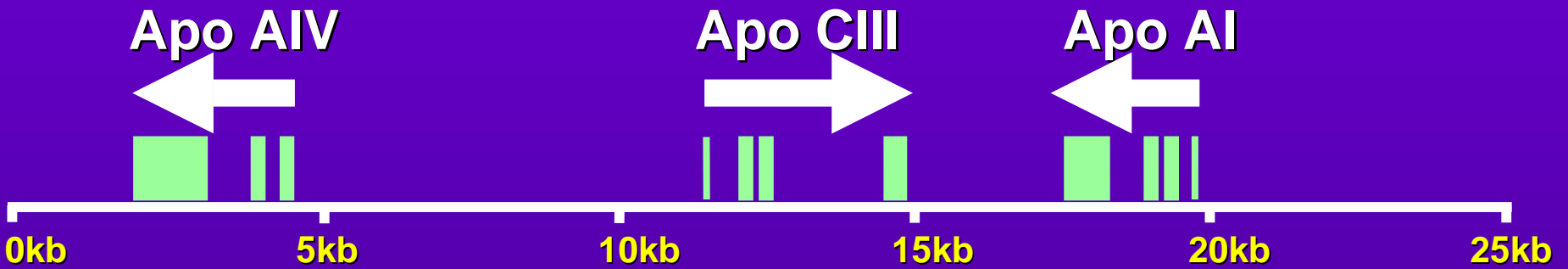


# Biological Insights from Comparative Sequence Analysis

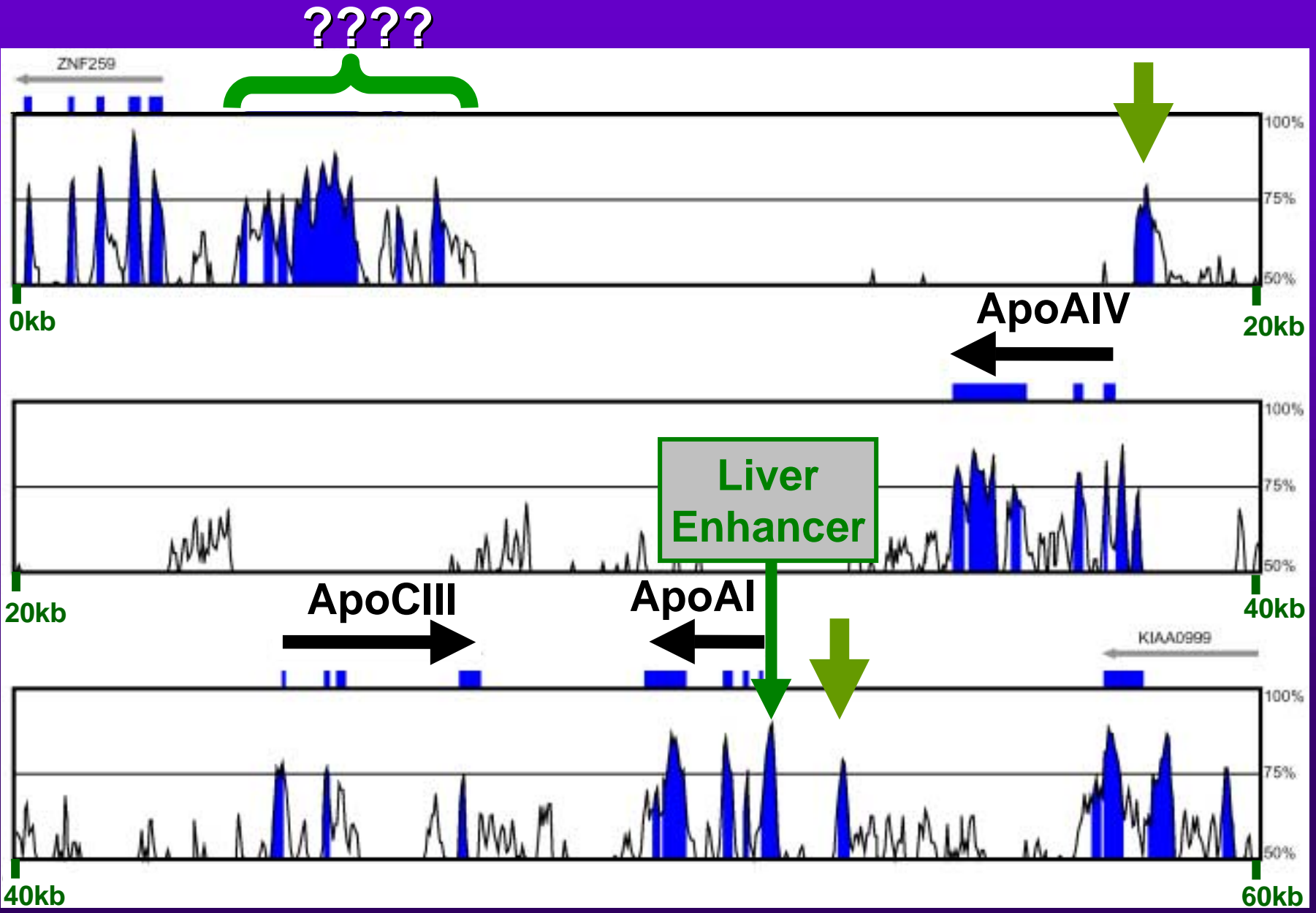


**Identification of a novel gene  
through comparative genomics**

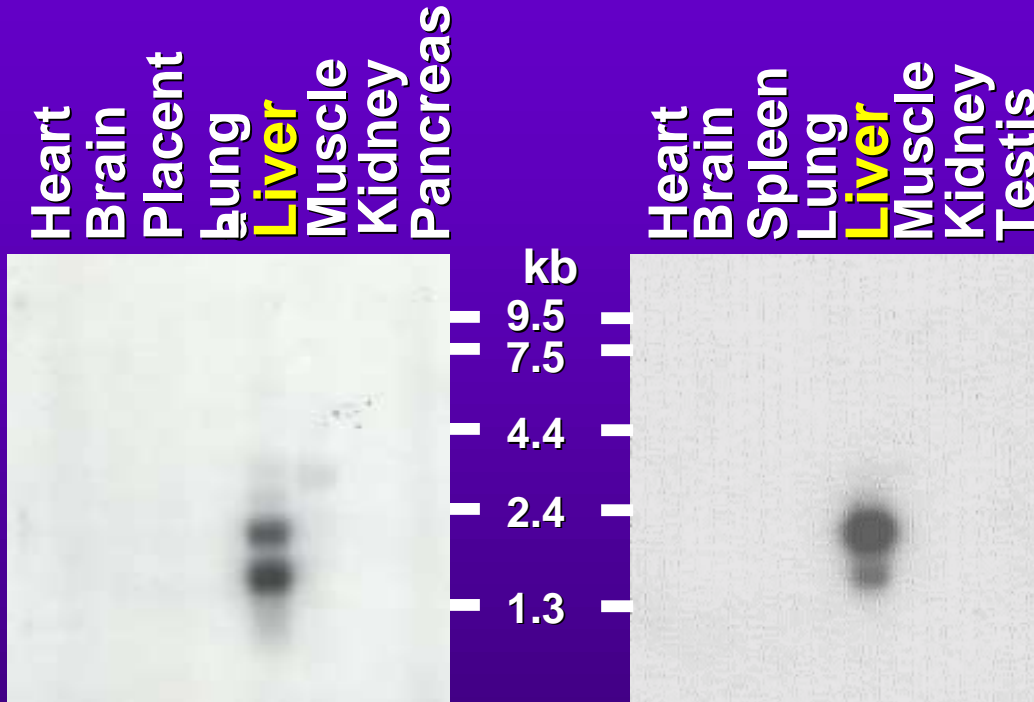
# Human Chromosome 11q23 Apolipoprotein Gene Cluster



# Human/Mouse Apolipoprotein Gene Cluster Sequence Comparison



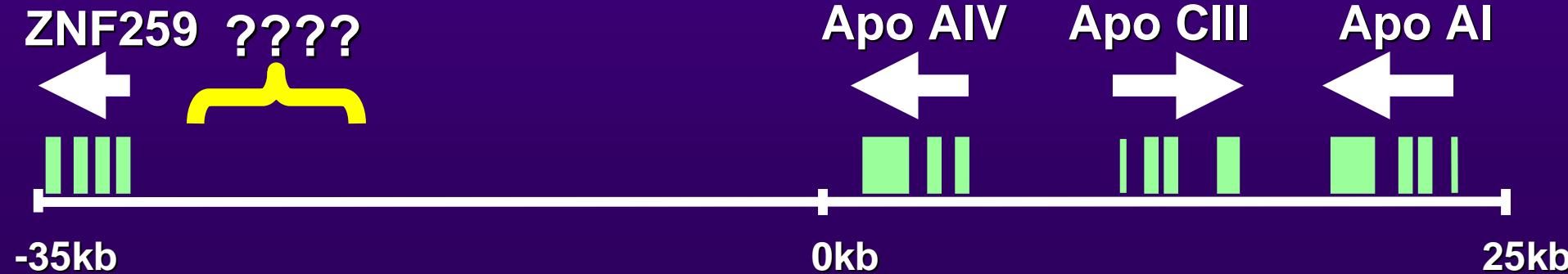
# Northern Blot Analysis of Conserved Sequence



Liver transcripts detected in both human and mouse

Human

Mouse



# Predicted protein has homology to ApoAIV

predicted protein  
human apoAIV

```

---MAAVLTWALALLS----AFSATQARKGFWDYFSQTSG-DKGRVEQIH
MFLKAVVLTLLALVAVAGARA EVSADQVATVMWDYFSQLSNNAKEAVEHLQ

QQKMAREP-ATLKDSL EQDLNMMNKFLEKLRPLSGSEAPRIPQDPVGMRR
KSELTQQLNALFQDKLGEVNTYAGDLQKKLVPFATELHERIAKDSEKLKE

QLQEEL E EVKARLQPYMAEAHEL VGWNLEGLRQQLKPYTMDLMEQVALRV
EIGKELEELRARRLLPHANEVSQKIGDNLRELQORLEPYADQLRTQVNTQA

QELQEQLRVG EDTKAQLLGGVDEAWALLQG----LQSRVVHHTGRFKEL
EQRRQLDPLAQRMERVLRENADSLQASLRPHADELKAKIDQNVEELKGR

FHPYAESLVSGIGR HVQELHRSVAPHAPASPARLSRCVQVLSRKLTLKAK
LTPYADEFKVKIDQTV EELRRSLAPYAQDTQEKLNHQLEGLTFQMKKNAE

ALHARIQQNLDQLREELSRAFAGT-----GTEEGAGPDPQMLSEEVQRRL
ELKARISASA EELRQRLAPLAEDVRGNLKGNT EGLQKSLAELGGHLDQOV

QAFRQDTYLQIAAFTRAIDQETEEVQQQLAPPPPGHSAFAPEFQQQTDGK
EEFRRRVEPYGENFNKALVQOMEQLRQKLGPHAGDVEGHLSFLEKDLRDK

VLSKLQARLDDLWEDITHSLHDQGHSHLGDP-----
VNSFFSTFK EKESQDKT LSLPELEQQQEQQQEQQQEQQVQMLAPLES
    
```

Identity: 26%  
Similarity: 45%

ZNF259    ????



Apo AIV



Apo CIII



Apo AI



-35kb

0kb

25kb

# Predicted protein has homology to ApoAIV

predicted protein  
human apoAIV

```

---MAAVLTWALALLS----AFSATQARKGFWDYFSQTSG-DKGRVEQIH
MFLKAVVLTLLALVAVAGARA EVSADQVATVMWDYFSQLSNNAKEAVEHLQ

QQKMAREP-ATLKDSL EQDLN MNKFLEKLRPLSGSEAPRIPQDPVGMRR
KSELTQQLNALFQDKLGEVNTYAGDLQKKLVPFATELHERIAKDSEKLKE

QLQEEL E EVKARLQPYMAEAHEL VGWNLEGLRQQLKPYTMDLMEQVALRV
EIGKELEELRARLLPHANEVSQKIGDNLRELQORLEPYADQLRTQVNTQA

QELQEQLRVG EDTKAQLLGGVDEAWALLQG----LQSRVVHHTGRFKEL
EQRRQLDPLAQRMERVLRENADSLQASLRPHADEL KAKIDQNVEELKGR

FHPYAESLVSGIGRHVQELHRSVAPHAPASPARLSRCVQVLSRKLTLKAK
LTPYADEFKVKIDQTV EELRRSLAPYAQDTQEKLNHQLEGLTFQMKKNAE

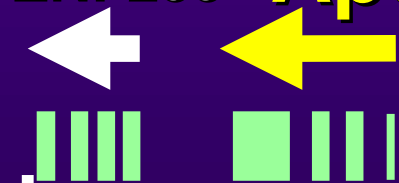
ALHARIQQNLDQLREELSRAFAGT-----GTEEGAGPDPQMLSEEVQRRL
ELKARISASA EELRQRLAPLAEDVRGNLKGNT EQLQKSLAELGGHLDQQV

QAFRQDTYLQIAAFTRAIDQETEEVQQQLAPPPPGHSAFAPEFQQQTDGK
EEFRRRVEPYGENFNKALVQOMEQLRQKLGPHAGDVEGHLSFLEKDLRDK

VLSKLQARLDDLWEDITHSLHDQGHSHLGD P-----
VNSFFSTFK EKESQDKTLSLPELEQQQEQQQEQQQEQQVQMLAPLES
    
```

Identity: 26%  
Similarity: 45%

ZNF259 “Apo AV”



Apo AIV



Apo CIII



Apo AI



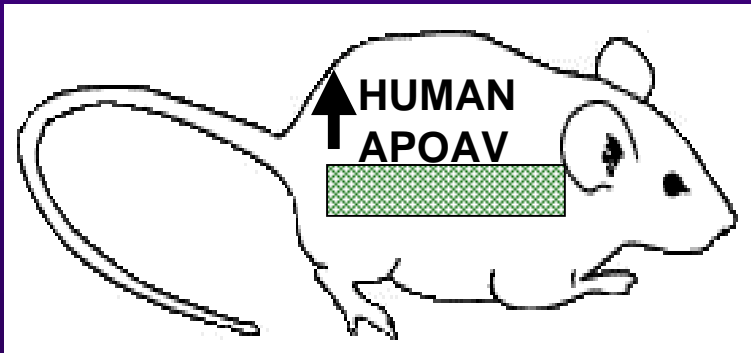
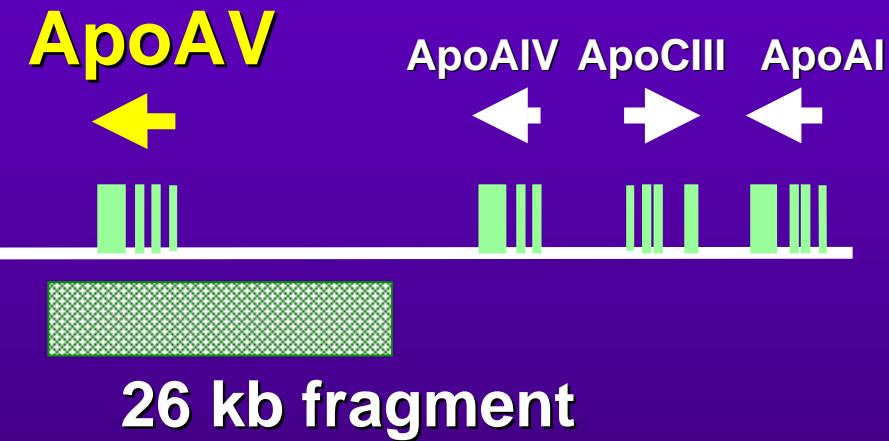
-35kb

0kb

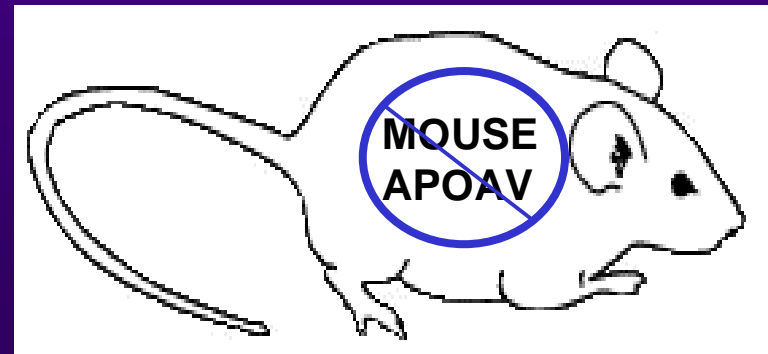
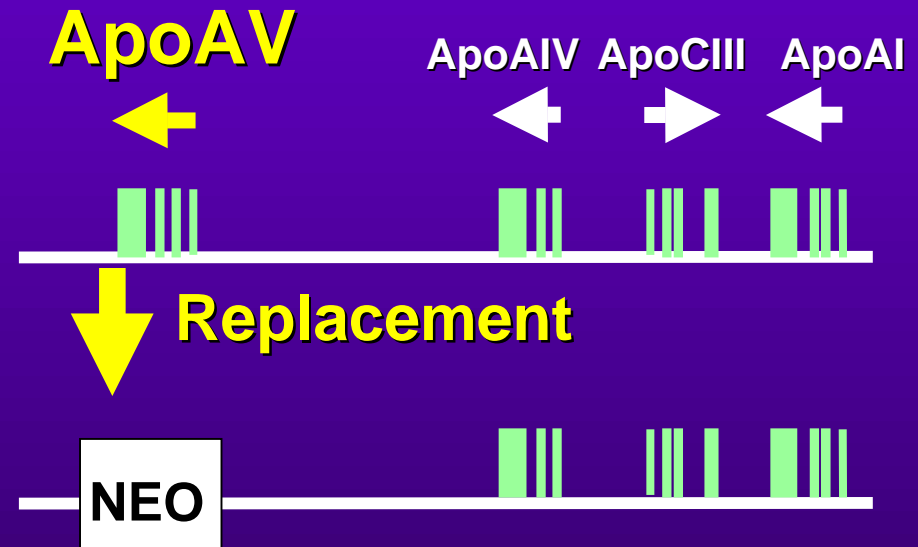
25kb

# ApoAV Mouse Studies

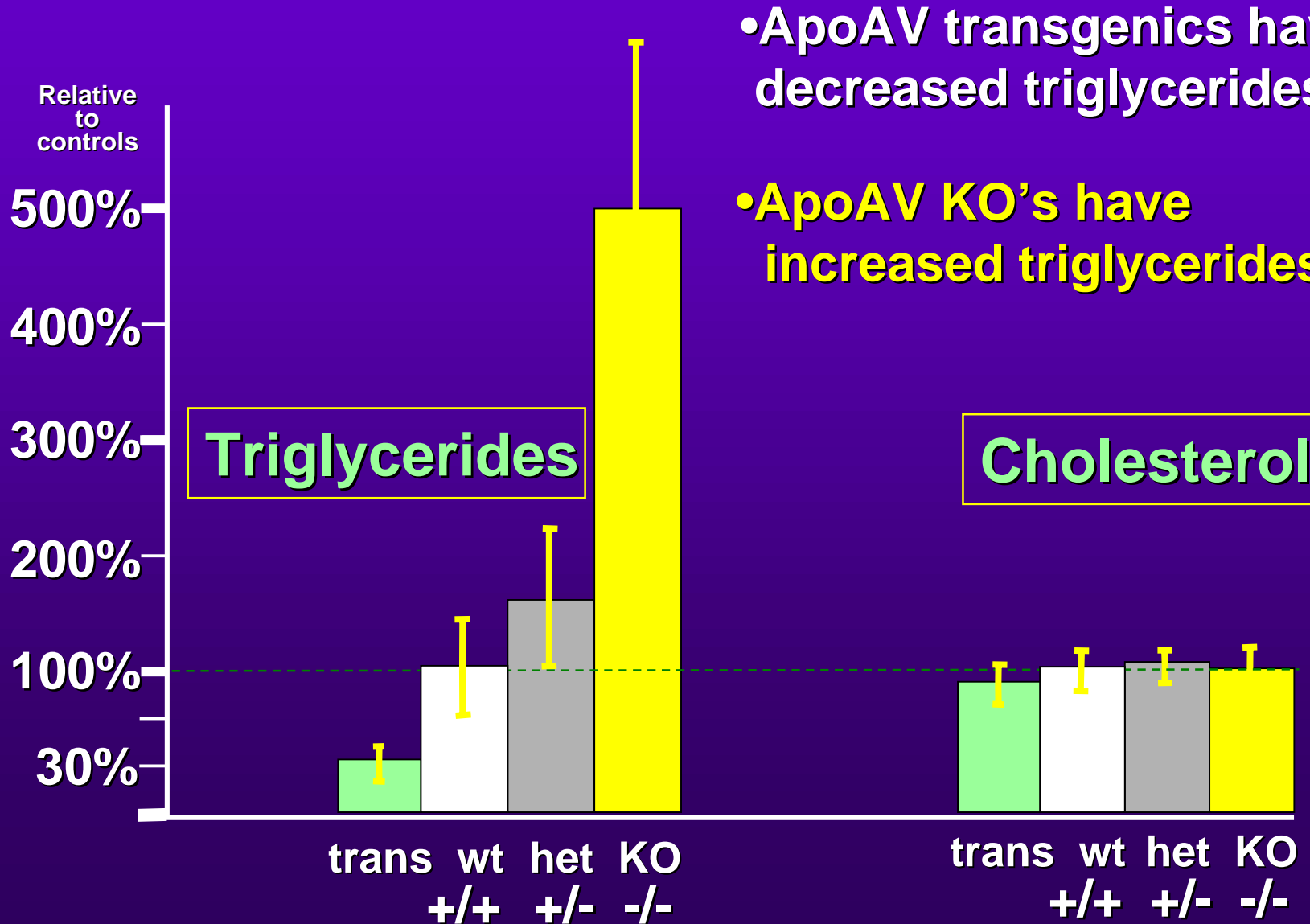
## TRANSGENICS



## KNOCKOUTS



# ApoAV Transgenic and Knockout Plasma Levels



- ApoAV transgenics have decreased triglycerides.

- ApoAV KO's have increased triglycerides.

# ApoAV

- Expressed in the liver & associates with HDL/VLDL.
- An important modulator of triglycerides (TG) in mice.



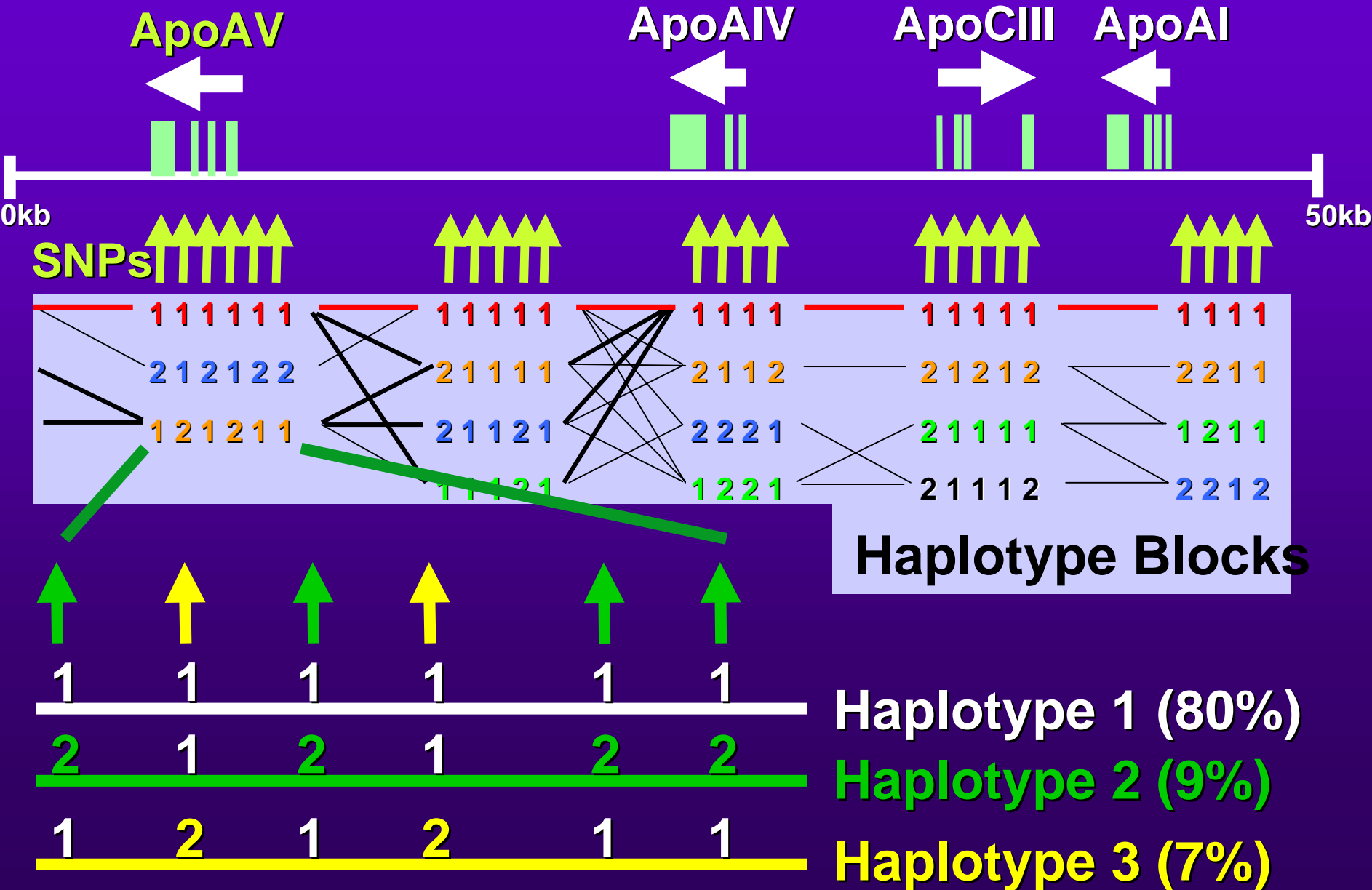
---

**Is ApoAV involved in human biology/disease?**

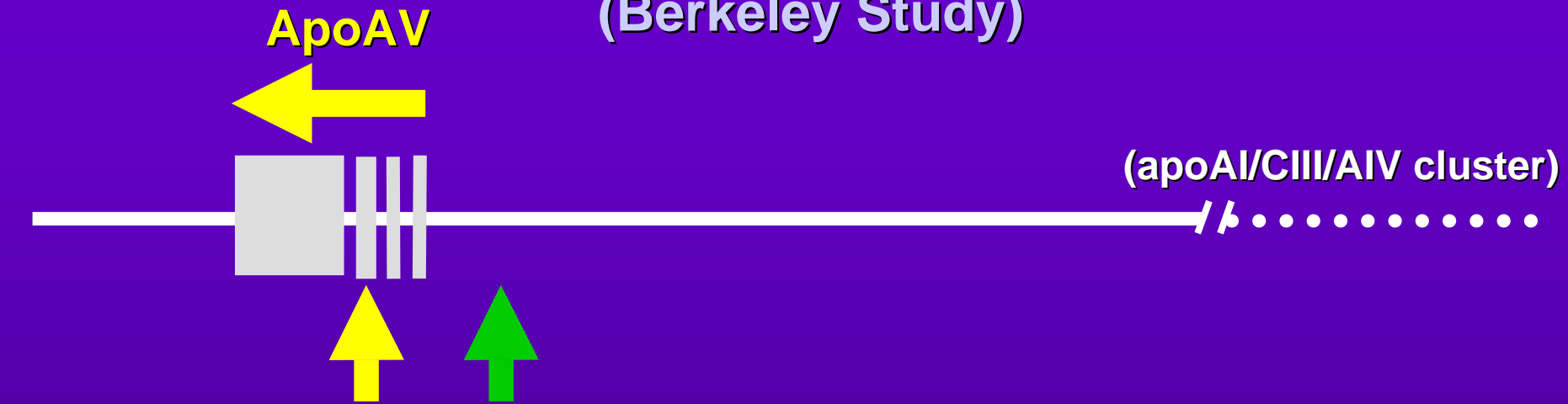
-Hyperlipidemia??

-Common Variation in Plasma Lipid Levels??

# SNP Identification/Haplotype Structure



# Association study I: ApoAIV polymorphisms and plasma parameters (Berkeley Study)



———— Haplotype 1 (80%)

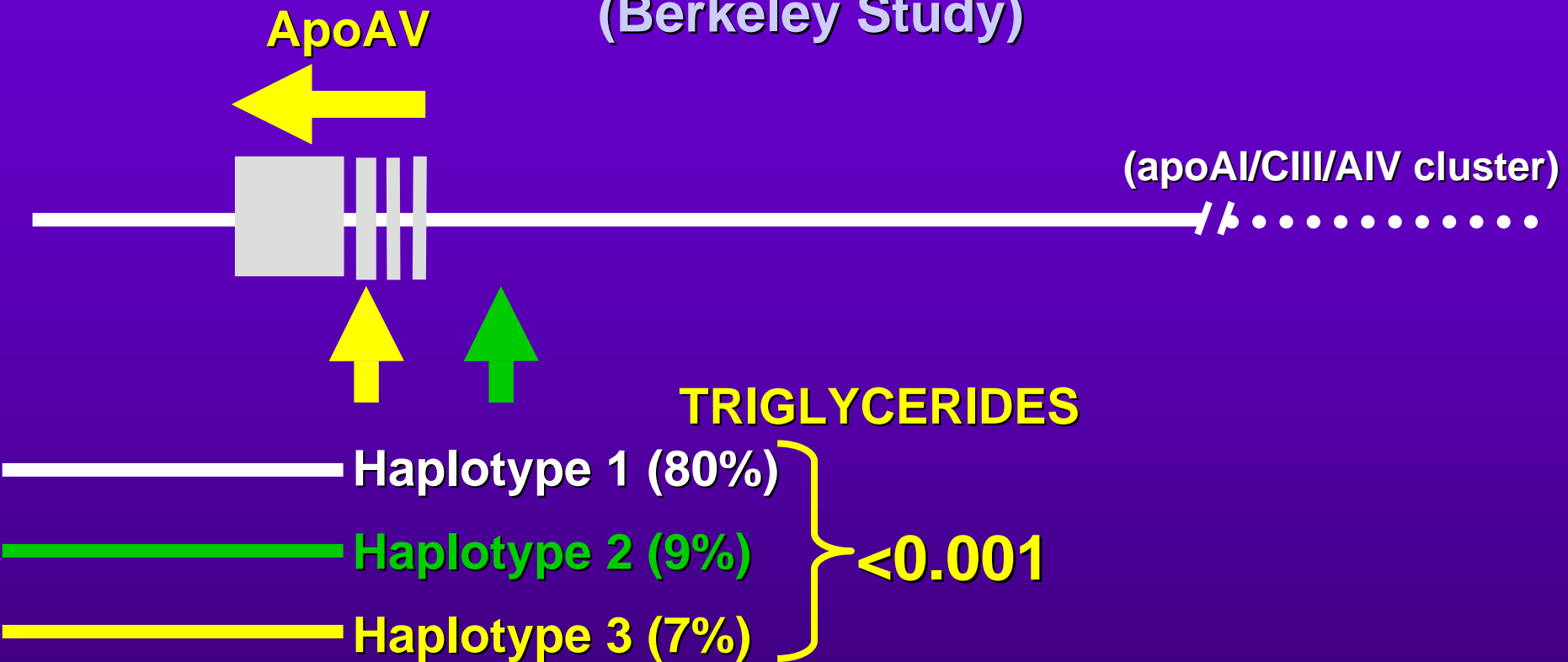
———— Haplotype 2 (9%)

———— Haplotype 3 (7%)

**Genotyped 500 normal individuals phenotyped for plasma:**

- Triglycerides
- IDL, LDL, HDL, VLDL Mass
- HDL, LDL Cholesterol
- ApoAII, ApoB

# Association study I: ApoAV polymorphisms and plasma parameters (Berkeley Study)



Genotyped 500 normal individuals phenotyped for plasma:

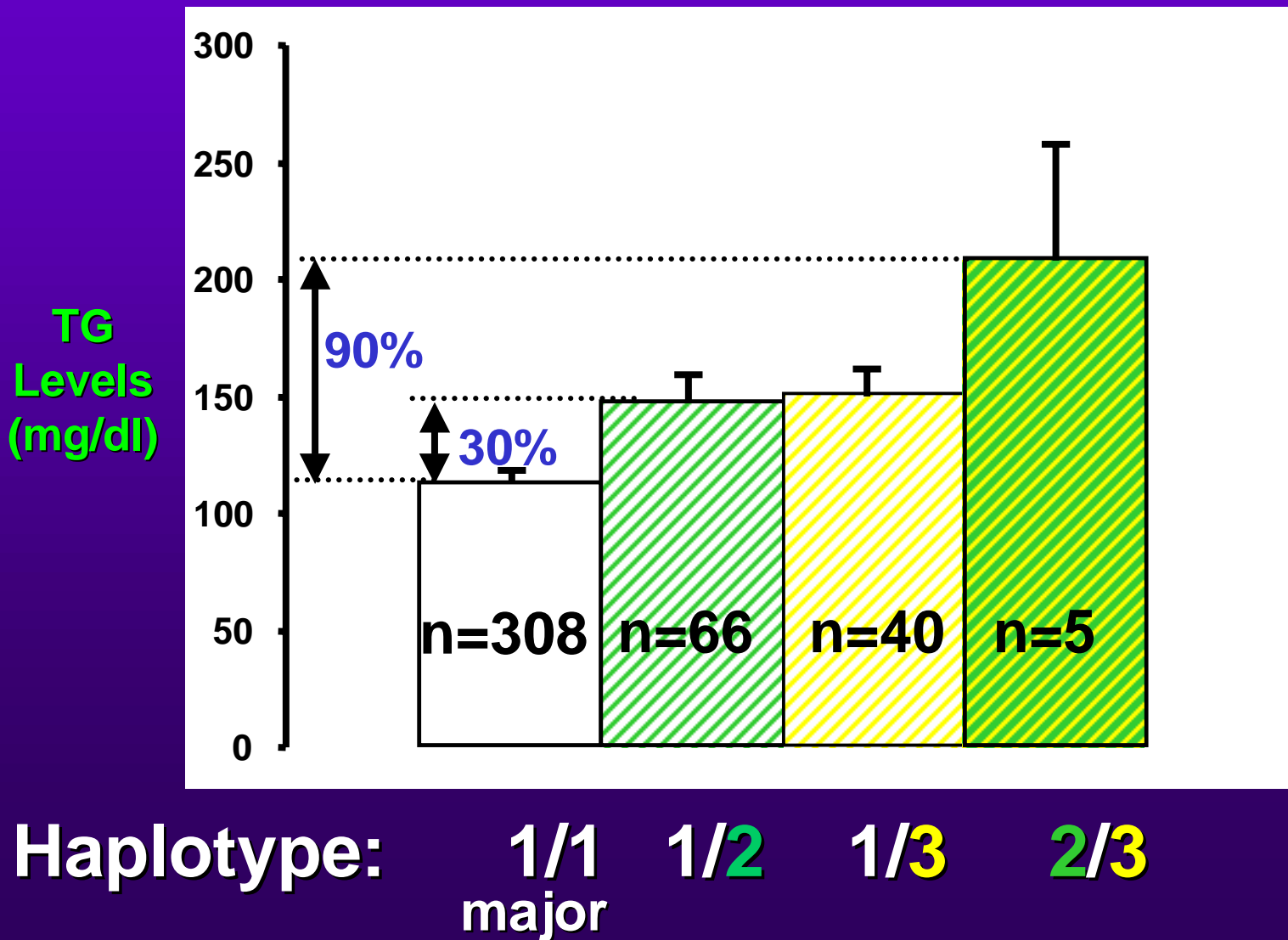
-TRIGLYCERIDES\*

-IDL, LDL, HDL, VLDL Mass

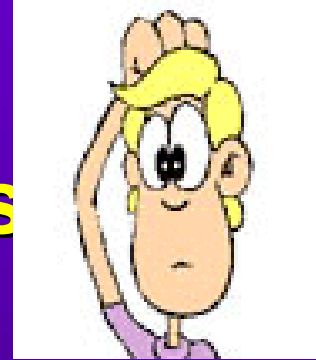
-HDL, LDL Cholesterol

-ApoA1, ApoB\*

# Association between ApoAV and Triglyceride Levels (Berkeley Study)



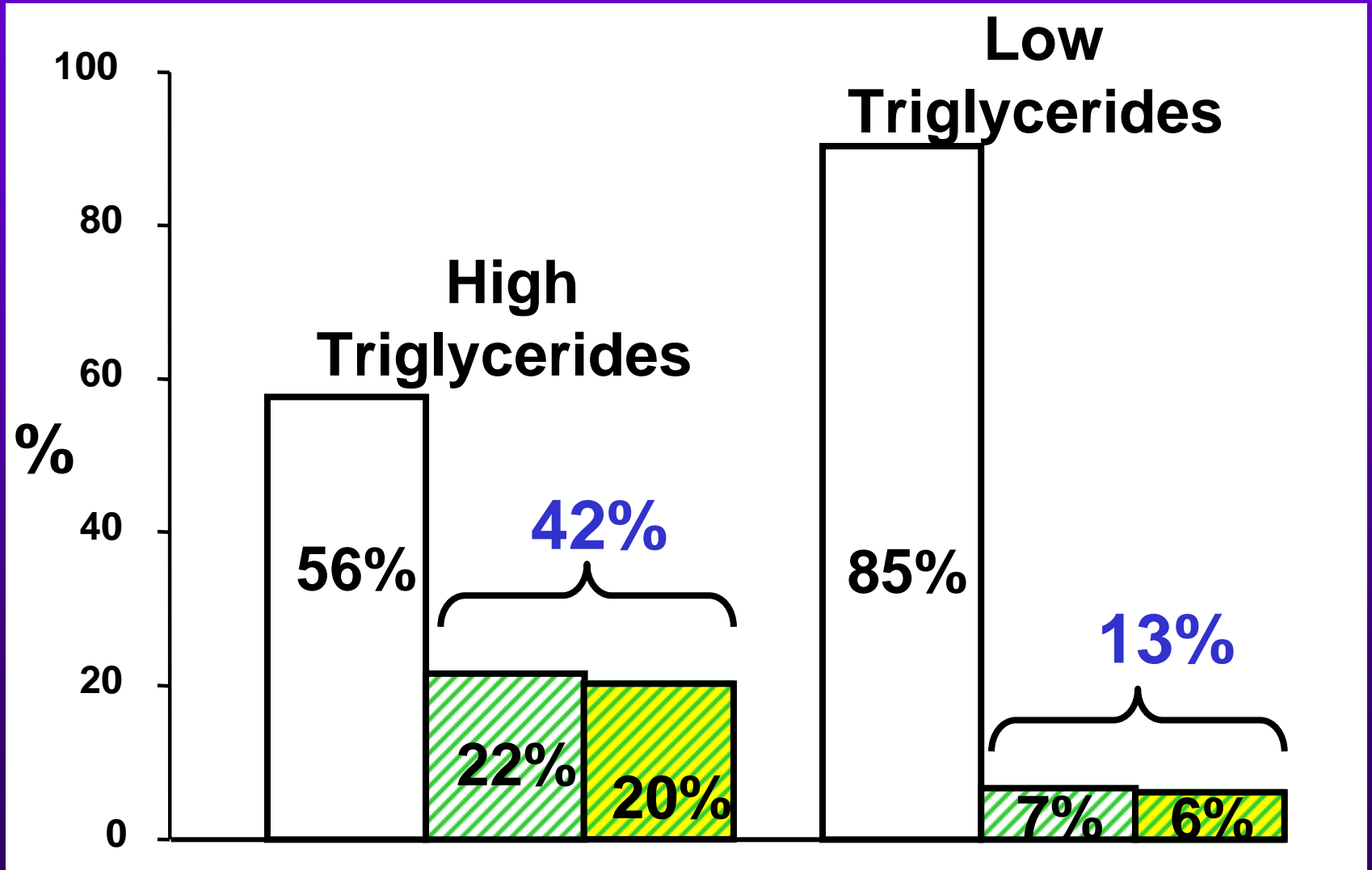
# Association Studies



Is this finding reproducible????

# Association study II: ApoAV polymorphisms and plasma parameters

(Dallas Study: ~500 Individuals)

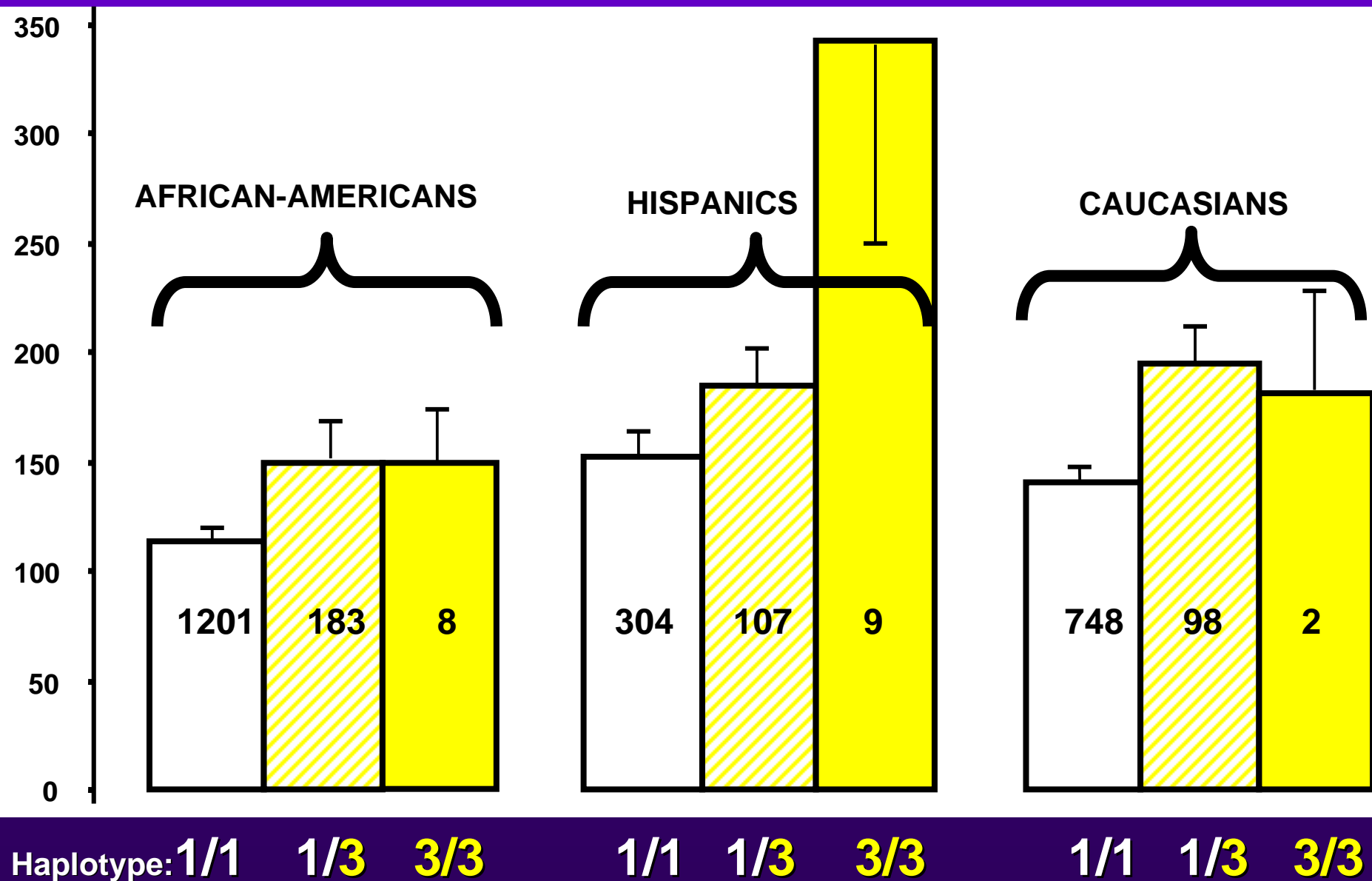


Haplotype: 1/1 1/2 1/3

1/1 1/2

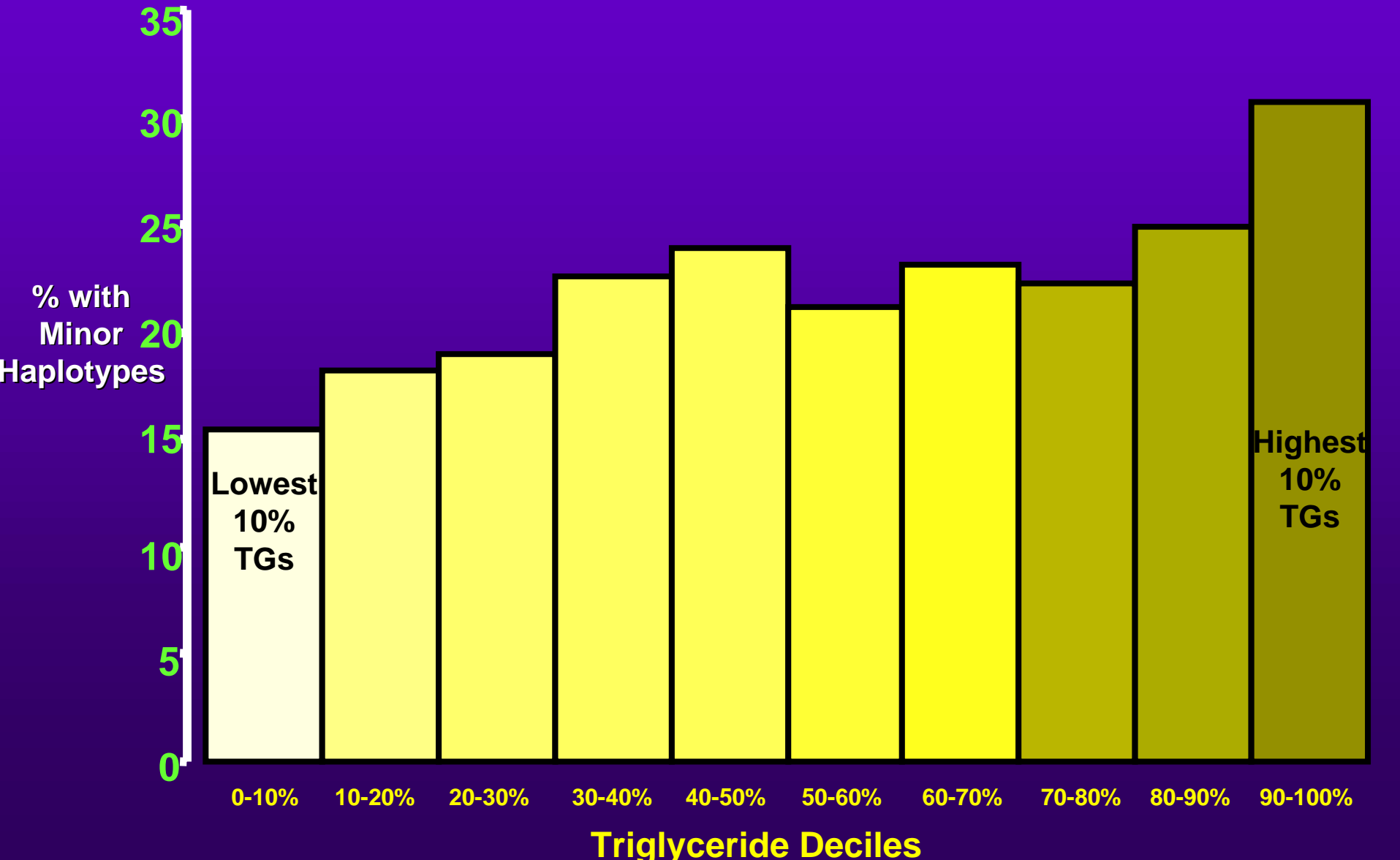
1/3

# Association Study III: Dallas, TX



# Common human variation contributing to a quantitative phenotype

## IV: Analysis of 3000 Caucasian individuals separated by triglyceride



# ApoA5 and Triglyceride Levels

An example of common human variation

contributing to a quantitative phenotype

## Carriers of Minor

### Ethnicity:

Caucasian

African American

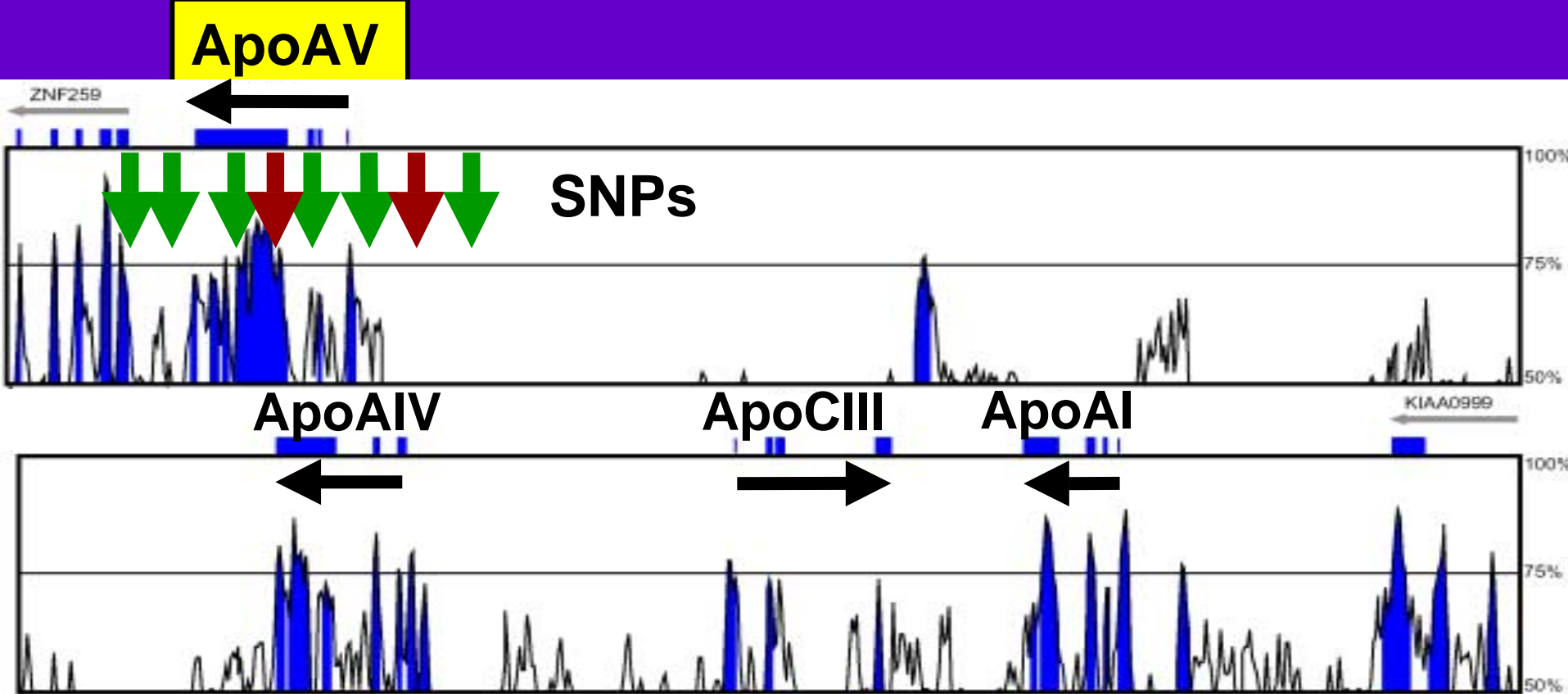
Hispanic

### Haplotype 2 and/or 3

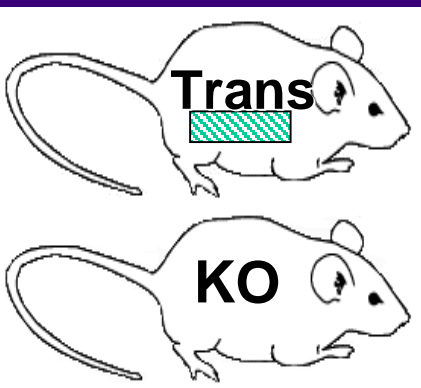
24%

36%

51%



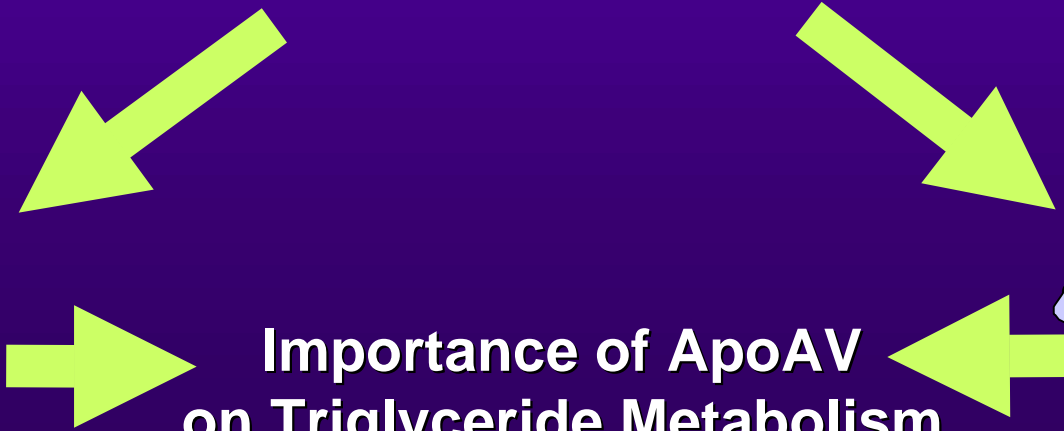
Mouse studies



Human studies



Importance of ApoAV  
on Triglyceride Metabolism



# Acknowledgements

## LBNL

Edward Rubin  
Len Pennacchio  
Marcelo Nobrega  
Nadine Barouhk  
Elaine Gong  
Jennifer Akiyama  
Keith Lewis  
Willow Dean  
Jan-Fang Cheng  
Inna Dubchak  
Lior Pachter  
Ivan Ovcharenko  
Jody Schwartz  
Veena Afzal  
Ronald Krauss  
Patricia Blanche  
Laura Holl

## UT-SW

Jonathan Cohen  
Helen Hobbs  
Jaroslav Hubacek

## Rayne Institute

Philippa Talmud  
Steve Humphries

## Pasteur Institute-Lille

Jamila Fruchart  
Jean-Charles Fruchart

## UCSF

Brian Black

## MCW

Michael Olivier

## NIH/NHLBI

<http://pga.lbl.gov>

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