Current Topics in Genome Analysis
October 21,2003

Computational Techniques in Comparative Genomics

Elliott H. Margulies, Ph.D.
Genome Technology Branch
National Human Genome Research Institute
elliott@nhgri.nih.gov

Outline

- Fundamental concepts of comparative genomics
- Alignment and visualization tools
- Information available through genome browsers
- Gene prediction and identification
- Identifying regulatory sequences
- Insights from Human-Mouse sequence comparisons
- Multi-species sequence analysis

Genome Wide Sequence Data

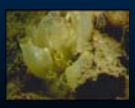
'Finished'

Human

Draft assemblies Available

Pufferfish

Mouse



Sea Squirt

Rat

Draft assemblies on the way

Dog

Chicken

Tetraodon

Chimpanzee

Zebrafish

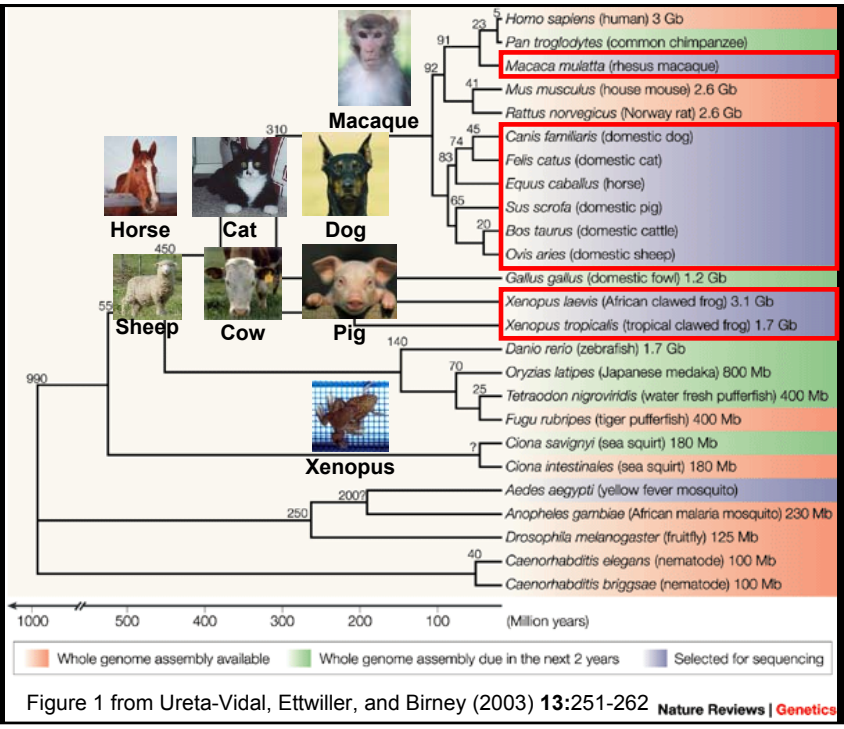


Figure 1 from Ureta-Vidal, Ettwiller, and Birney (2003) **13:**251-262 Nature Reviews | Genetics

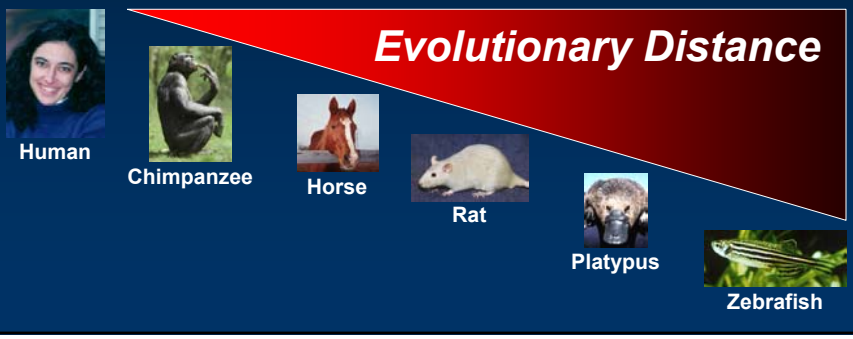
What can be more curious *than that the hand of a man, formed for grasping, that of a mole for digging, the leg of the horse, the paddle of the porpoise, and the wing of the bat, should all be constructed on the same pattern, and should include similar bones, in the same relative positions?*

Charles Darwin, The Origin of Species (1859)

Similarity in Morphology ⟸ **Similarity in Genetics**

Comparative Genomics

- **Find sequences that have diverged less than we expect**
  These sequences are likely to have a functional role
- **Our expectation is related to the time since the last common ancestor**



Comparative Genomics

Similarity in Identity

Similarity in Function

My name is Elliott. I am presenting a seminar about comparative genomics.

Mon nom est Elliott. Je vous présente un séminaire à propos de génomique comparative.

Meine name ist Elliott. Ich präsentiere ein Seminar über Comparativ Genomics.

Outline

- Fundamental concepts of comparative genomics
- Alignment and visualization tools
- Information available through genome browsers
- Gene prediction and identification
- Identifying regulatory sequences
- Insights from Human-Mouse sequence comparisons
- Multi-species sequence analysis

Sequence Alignments

100% Identical

```
Species 1 CATGGGCAAATTGGCCCATTGGCCATGGGGGCCCACCGTA
          |||||||||||||||||||||||||||||||||||||||
Species 2 CATGGGCAAATTGGCCCATTGGCCATGGGGGCCCACCGTA
```

80% Identical

```
Species 1 CATGGGCAAATTGGCCCATTGGCCATGGGGGCCCACCGTA
          ||  ||||  |||   |||  ||||||  ||||||  ||||  ||||
Species 2 CACGGGCTAATCCGCCAATTGGCTATGGGG-CCCAGCGTA
```

30% Identical

```
Species 1 CATGGGCAAATTGGCCCATTGGCCATGGGGGCCCACCGTA
            |  | |     |  ||     |   |     |  | |
Species 2 CACGAACTAATCCGCCAATAGCCTATAGCG-CACAGCGAA
```

Tools for Aligning Genomic Sequences

Resource

Genome Research (2000) **10**:577-586

PipMaker—A Web Server for Aligning Two Genomic DNA Sequences

Scott Schwartz,[1] Zheng Zhang,[1] Kelly A. Frazer,[2] Arian Smit,[3] Cathy Riemer,[1] John Bouck,[4] Richard Gibbs,[4] Ross Hardison,[5] and Webb Miller[1,6]

Departments of [1]Computer Science and Engineering and [5]Biochemistry and Molecular Biology and Center for Gene Regulation, The Pennsylvania State University, University Park, Pennsylvania USA 16802; [2]Genome Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, California USA 94720; [3]Axys Pharmaceuticals, La Jolla, California USA 92037; [4]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas USA 77030

BIOINFORMATICS **APPLICATIONS NOTE** Vol. 16 no. 11 2000 Pages 1046–1047

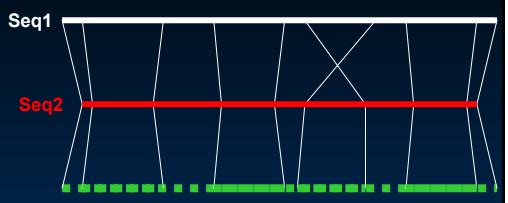
VISTA: *visualizing global DNA sequence alignments of arbitrary length*

Chris Mayor[1], Michael Brudno[1], Jody R. Schwartz[2], Alexander Poliakov[2], Edward M. Rubin[2], Kelly A. Frazer[2], Lior S. Pachter[3,*] and Inna Dubchak[1,*]

[1]National Energy Research Scientific Computing Center, [2]Genome Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA and [3]Department of Mathematics University of California at Berkeley, Berkeley, CA 94720, USA

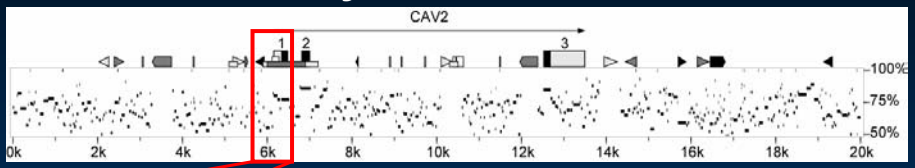
PipMaker *vs.* VISTA

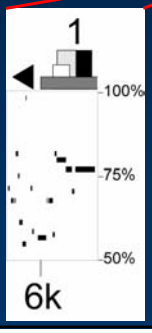
- Visualization
- Alignment Strategy
  - VISTA: `avid`
  - PipMaker: `blastz`
- East Coast – West Coast



PipMaker
http://bio.cse.psu.edu/pipmaker/

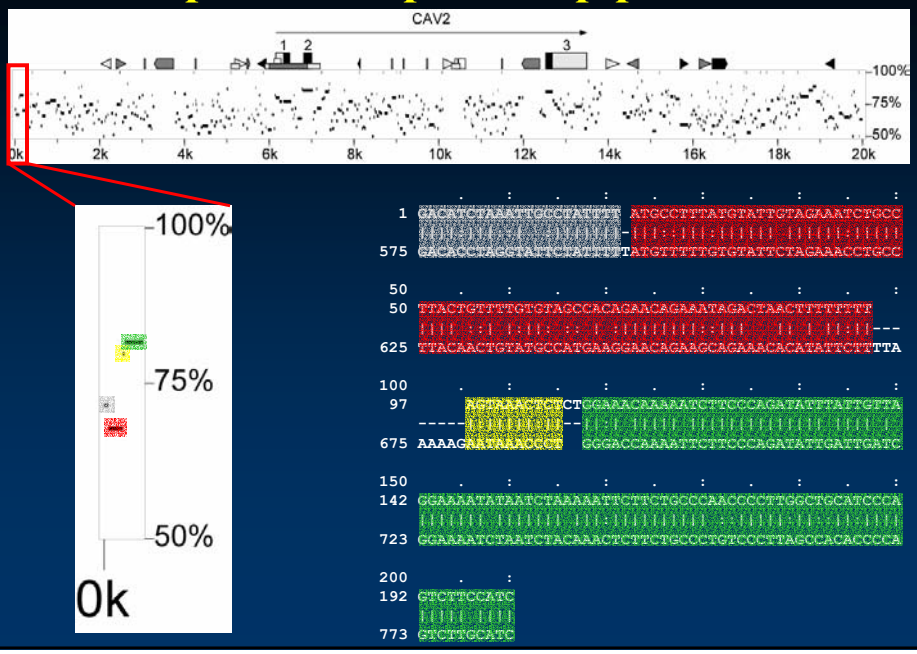
- Percent Identity Plot





- X-axis is the reference sequence
- Horizontal lines represent gap-free alignments

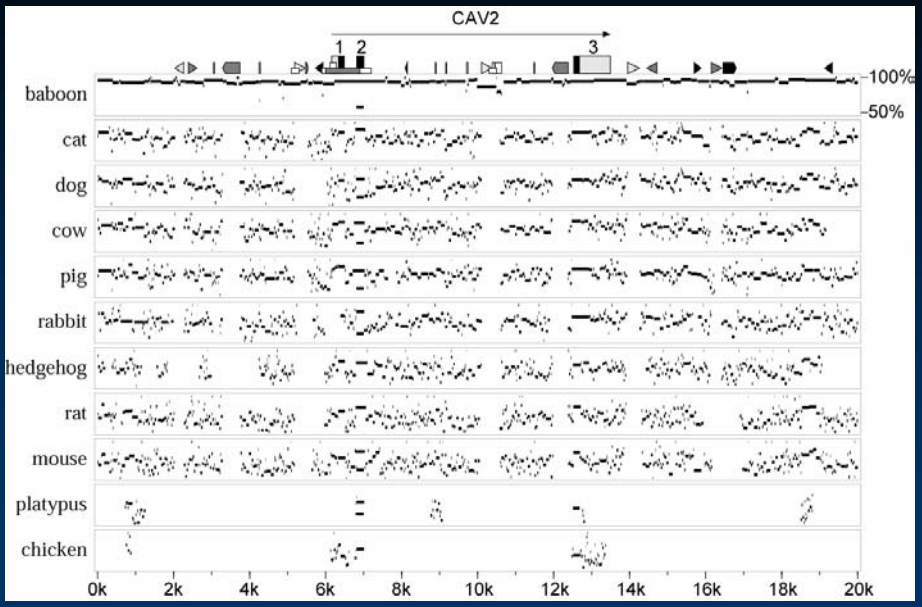
http://bio.cse.psu.edu/pipmaker/



PipMaker
http://bio.cse.psu.edu/pipmaker/

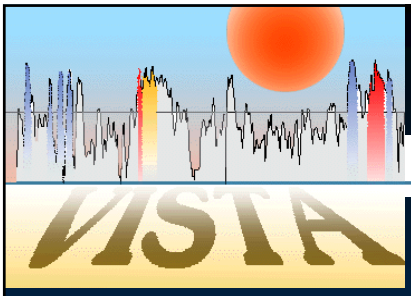
- **P**ercent **I**dentity **P**lot
- **Available in 3 Flavors:**
  - **Regular**
    - No Additional Options
  - **Advanced** go to submission page
    - Different Alignment Strategies
    - Additional Output Summaries
  - **MultiPipMaker**
    - Multi-species display

MultiPipMaker



PipTools

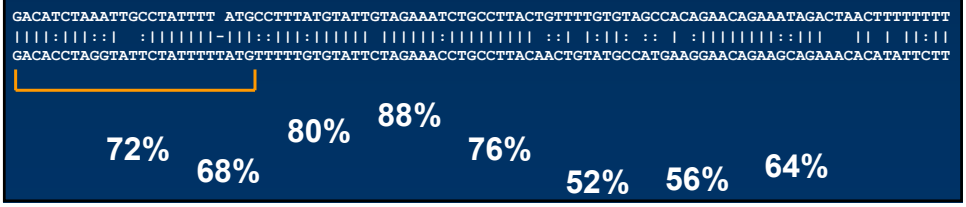
- "Show me all alignments that are at least X% identity and Y% length"
  - `strong-hits`
- Coordinate Conversion
  - `transform-pos`
  - `shift-pos`
  - `where-hit`
- Coordinate Extraction from GenBank Files
  - `genbank2exons`
  - `genbank2repeats`
- Laj – Interactive Alignment Viewer – Open Laj



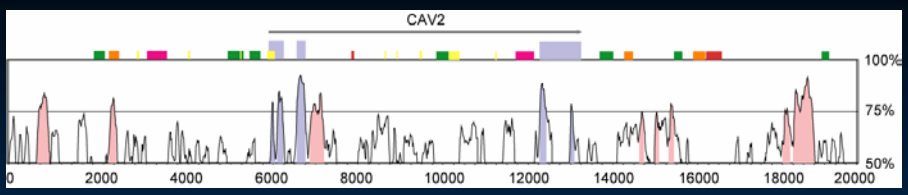
http://www-gsd.lbl.gov/vista/

VISUALIZATION TOOLS FOR ALIGNMENTS

- **Global Alignment (`avid`)**
  - Bray et al. (2003) *Genome Res* **13**:97-102
- **Sliding Window Approach to Visualization**
  - **Plot Percent Identity within a Fixed Window-Size, at Regular Intervals**



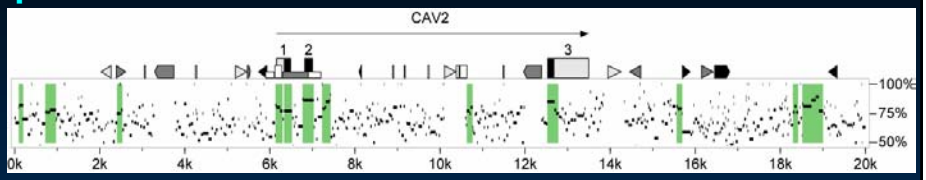
VISTA



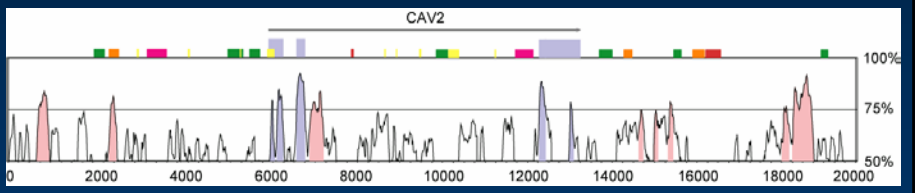
- **Percent Identity is plotted from:**
  - **100 base windows**
  - **Moved every 15 bases**
- **Colored regions meet certain alignment criteria**
  - **>100 bp >75% Identity**

What's Your Preference?

PipMaker



VISTA



Summary of Alignment Tools

- **PipMaker (`blastz`)**
- **VISTA (`avid`)**
- **Lagan and mLagan (glocal alignments)**
  - **http://lagan.stanford.edu/**
- **Box 1 from:**

Ureta-Vidal, Ettwiller, and Birney (2003) Comparative Genomics: Genome-Wide Analysis in Metazoan Eukaryotes *Nature Reviews Genetics* **4:** 251-262

Outline

- Fundamental concepts of comparative genomics
- Alignment and visualization tools
- **Information available through genome browsers**
- Gene prediction and identification
- Identifying regulatory sequences
- Insights from Human-Mouse sequence comparisons
- Multi-species sequence analysis

Genome Browsers

UCSC Genome Bioinformatics

http://genome.ucsc.edu

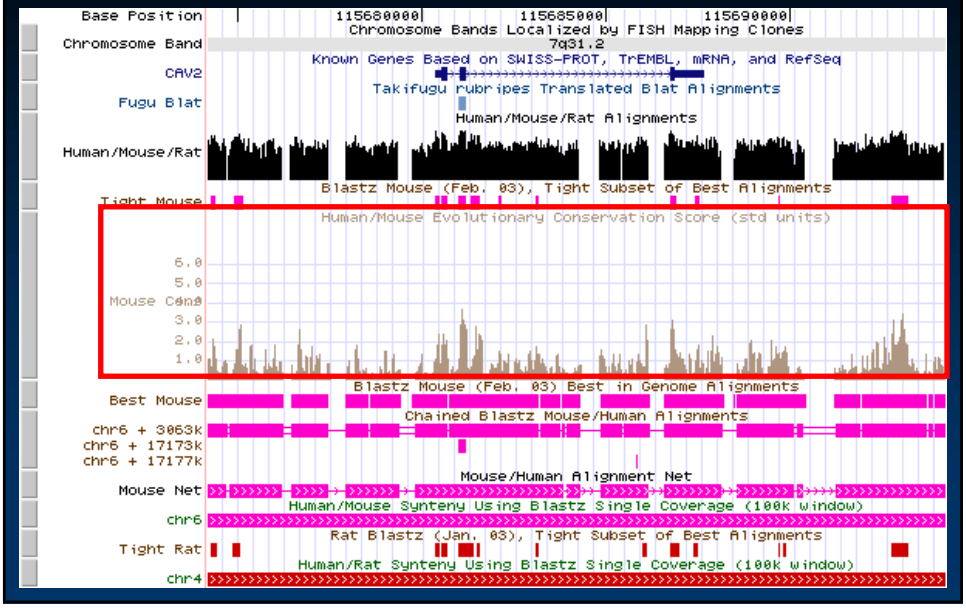
e! project Ensembl

http://www.ensembl.org

NCBI Map Viewer

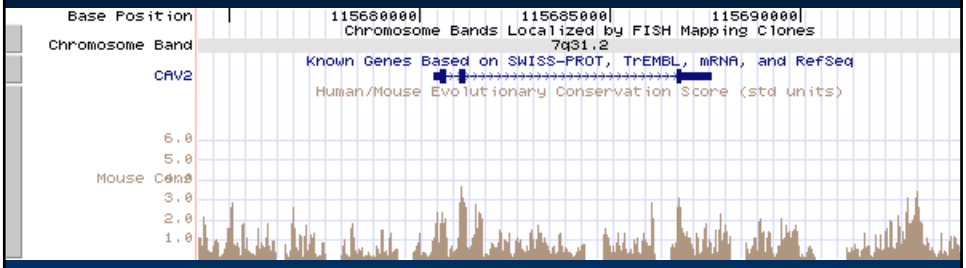
http://www.ncbi.nlm.nih.gov/mapview/

Comparative Sequence Tracks



Mouse Conservation Score at UCSC

Probability that the observed conservation would occur by chance under *Neutral Evolution*



Neutral Evolution

- No selective pressure/advantage to keep or change the DNA sequence
- Rate of variation should correlate with:
  - Mutation rate
  - Amount of time since the last common ancestor
- The neutral rate can vary across the genome

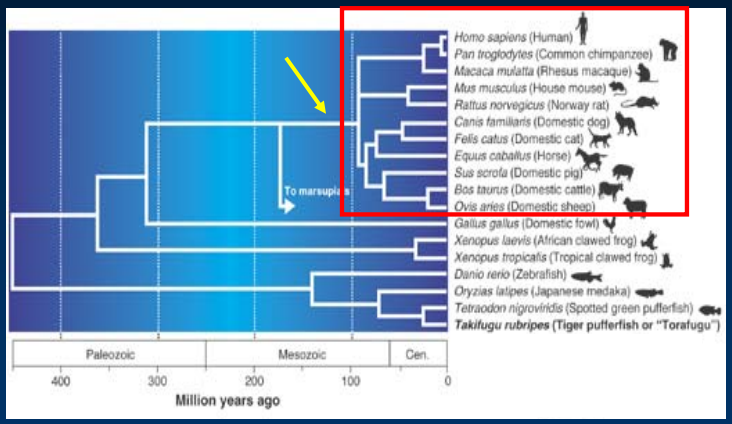
Types of Neutrally Evolving DNA

- 4-Fold Degenerate Sites
  - Third position of codons which can be any base and code for the same amino acid

	Second				
First	U	C	A	G	Last
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Types of Neutrally Evolving DNA

- **Ancestral Repeats**
  - **Ancient Relics of Transposons Inserted Prior to the Eutherian Radiation**

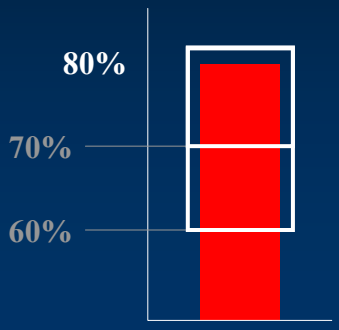


Adapted from Hedges & Kumar, *Science* **297**:1283-5

Mouse Conservation Score at UCSC

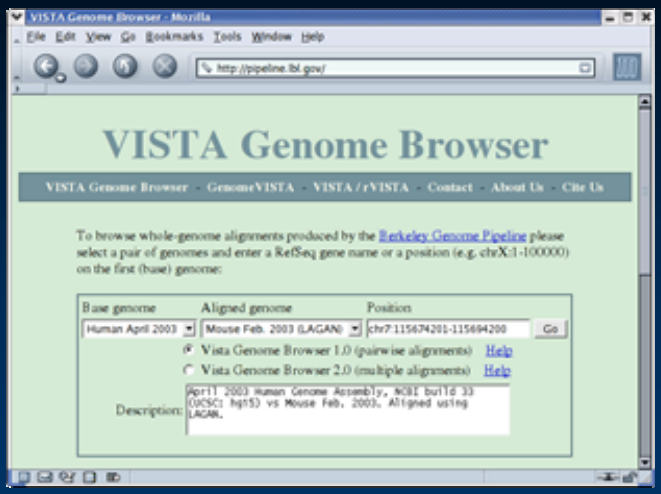
Probability that the observed conservation would occur by chance under *Neutral Evolution*

- **Calculated from 50-base windows**
- **Score is weighted on the surrounding neutral rate**



L-Score	Frequentist Probability	Bayesian Probability
1	0.1	0.32
2	0.01	0.75
3	0.001	0.94
4	0.0001	0.97
5	0.00001	0.98
6	0.000001	0.99

Berkley Genome Pipeline
http://pipeline.lbl.gov/



Link to Static View

Analysis of Comparative Sequence Data

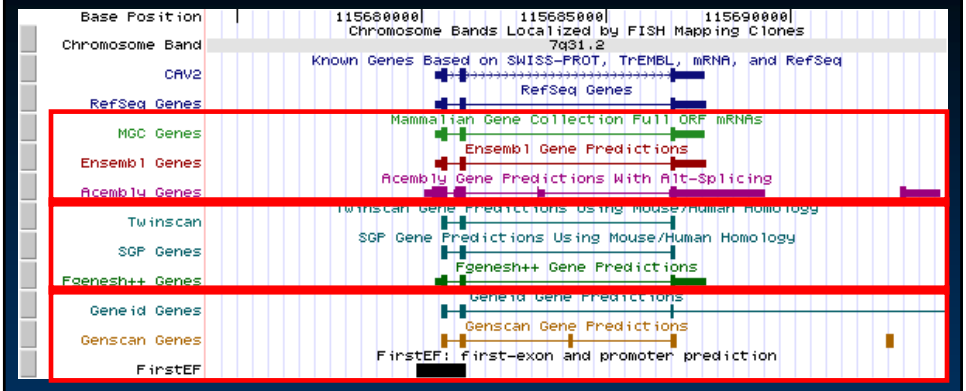
Sequence Conservation → **Identification of Functional Elements**

- **Coding Sequences (i.e. Genes)**
  - Relatively EASY to identify
  - Basic understanding of the 'language'
  - Complementary datasets available (ESTs, cDNAs)
- **Non-Coding Functional Sequences**
  - HARD to identify
  - Very little idea of what to look for
  - Virtually no complementary datasets

Outline

- Fundamental concepts of comparative genomics
- Alignment and visualization tools
- Information available through genome browsers
- **Gene prediction and identification**
- Identifying regulatory sequences
- Insights from Human-Mouse sequence comparisons
- Multi-species sequence analysis

Approaches to Gene Prediction



- **Evidence-Based**
  - MGC
  - Acembly
  - Ensembl
- ***Ab Initio***
  - Genscan
  - Geneid
  - FirstEF
- **Dual-Genome**
  - Twinscan
  - SGP
  - Fgenesh++

Additional Gene Prediction Resources

- **Fugu BLAT Track at UCSC**
- **EXOFISH –** http://www.genoscope.cns.fr/proxy/cgi-bin/exofish.cgi
- **SLAM –** http://baboon.math.berkeley.edu/~syntenic/slam.html
  - Cawley et al. (2003) *Nucleic Acids Research* **31:**3507-3509
- **Box 1 from:**
  - Ureta-Vidal et al. (2003) *Nature Reviews Genetics* **4:**251-262

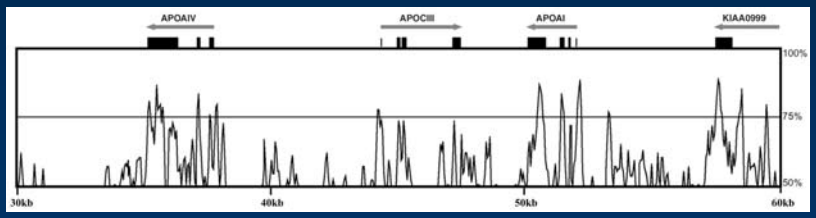
Gene Finding based on Sequence Conservation

Science **294:**169-173

An Apolipoprotein Influencing Triglycerides in Humans and Mice Revealed by Comparative Sequencing

Len A. Pennacchio,[1] Michael Olivier,[2*] Jaroslav A. Hubacek,[3] Jonathan C. Cohen,[3] David R. Cox,[2] Jean-Charles Fruchart,[4] Ronald M. Krauss,[1] Edward M. Rubin[1†]

APOAI/CIII/AIV Gene Cluster



Adapted from Figure 1, Pennacchio et al. *Science* **294:**169-173

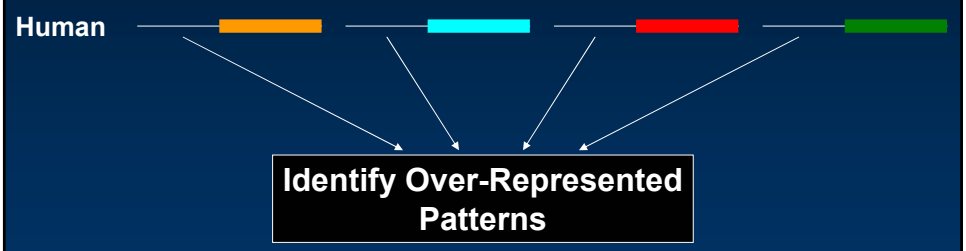
Outline

- Fundamental concepts of comparative genomics
- Alignment and visualization tools
- Information available through genome browsers
- Gene prediction and identification
- **Identifying regulatory sequences**
- Insights from Human-Mouse sequence comparisons
- Multi-species sequence analysis

Motif Finding

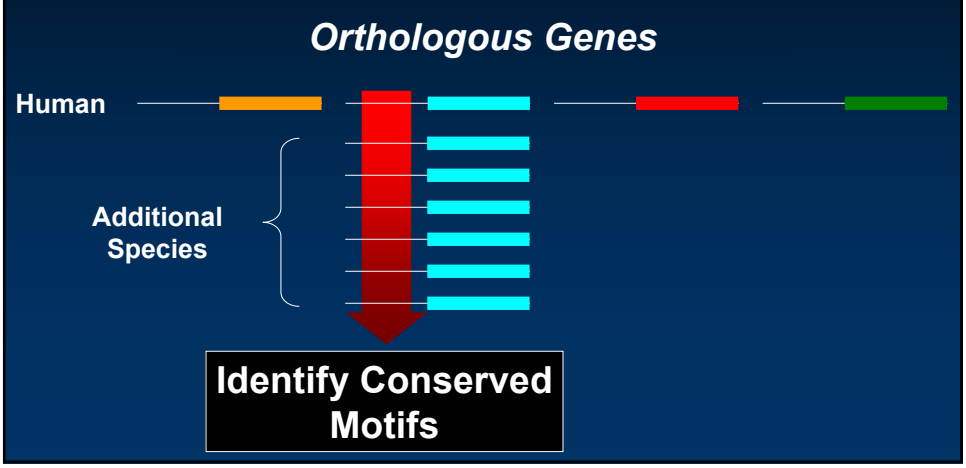
- **Identify Transcription Factor Binding Sites**
- **What sequences should be searched?**

Coordinately Regulated Genes



Phylogenetic Footprinting

- **FootPrinter** – http://bio.cs.washington.edu/software.html
- **Takes the phylogeny into account**



Summary of Phylogenetic Footprinting Tools

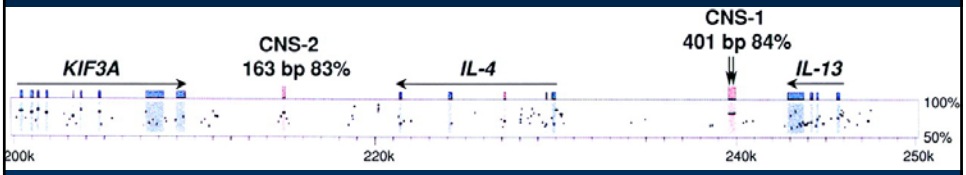
- **FootPrinter –** http://bio.cs.washington.edu/software.html
  - Blanchette and Tompa (2003) *Nucleic Acids Research* **31:**3840–3842
- **rVISTA –** http://www-gsd.lbl.gov/vista/rVistaInput.html
  - Loots et al. (2002) *Genome Research* **12:** 832–839
- **List of motif-finding algorithms:**
  - Box 1 of Ureta-Vidal et al. (2003) *Nature Reviews Genetics* **4:**251-262
- **Bayesian Approaches (and home of the Gibbs sampler)**
  - http://www.wadsworth.org/resnres/bioinfo/
- **Example of motif-finding limited by mouse conservation:**
  - Wasserman et al. (2000) *Nature Genetics* **26:**225-228

Prioritize Non-Coding Sequence Conservation

Science (2000) **288:**136-140

Identification of a Coordinate Regulator of Interleukins 4, 13, and 5 by Cross-Species Sequence Comparisons

G. G. Loots,[1,2] R. M. Locksley,[3] C. M. Blankespoor,[1] Z. E. Wang,[3] W. Miller,[4] E. M. Rubin,[1*] K. A. Frazer[1*]

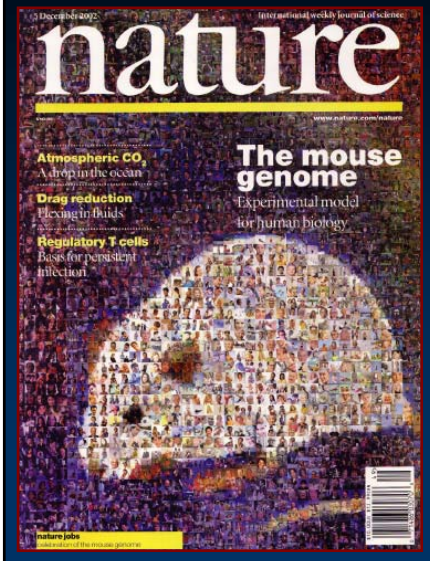


Good review on Comparative Genomics and Regulatory Sequences:
Pennacchio and Rubin, Genomic Strategies to Identify Mammalian Regulatory Sequences, *Nature Reviews Genetics* **2:**100-109

Outline

- Fundamental concepts of comparative genomics
- Alignment and visualization tools
- Information available through genome browsers
- Gene prediction and identification
- Identifying regulatory sequences
- **Insights from Human-Mouse sequence comparisons**
- Multi-species sequence analysis

Insights from Human-Mouse Sequence Comparisons

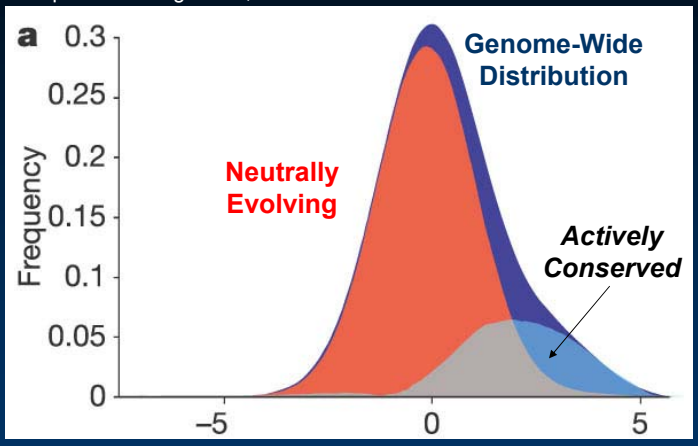


Nature 420:520, 2002

- Similar gene content and linear organization
  - ~340 syntenic blocks
- Difference in genome size
  - Mouse genome is 14% smaller
- Sequence Conservation
  - ~40% in Alignments
  - ~5% Under Selection
    - ~1.5% Protein Coding
    - ~3.5% Non-Coding
- Also see January 2003 issue of *Genome Research*

Actively Conserved Sequence

Adapted From Figure 28, *Nature* **420**:553



Conservation Score

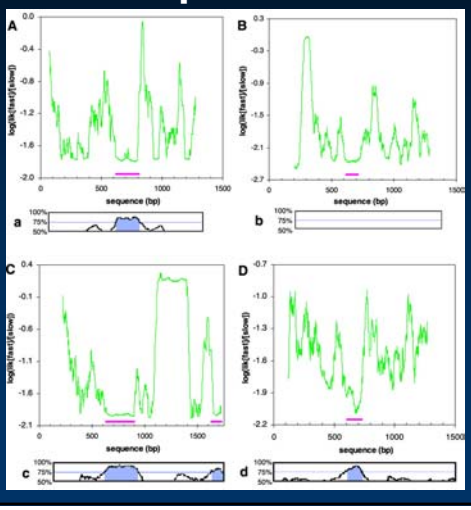
Outline

- Fundamental concepts of comparative genomics
- Alignment and visualization tools
- Information available through genome browsers
- Gene prediction and identification
- Identifying regulatory sequences
- Insights from Human-Mouse sequence comparisons
- **Multi-species sequence analysis**

Phylogenetic Shadowing

Boffelli et al. (2003) *Science* **299:**1391-1394.

- **Identifying sequence *differences* between multiple primate species**

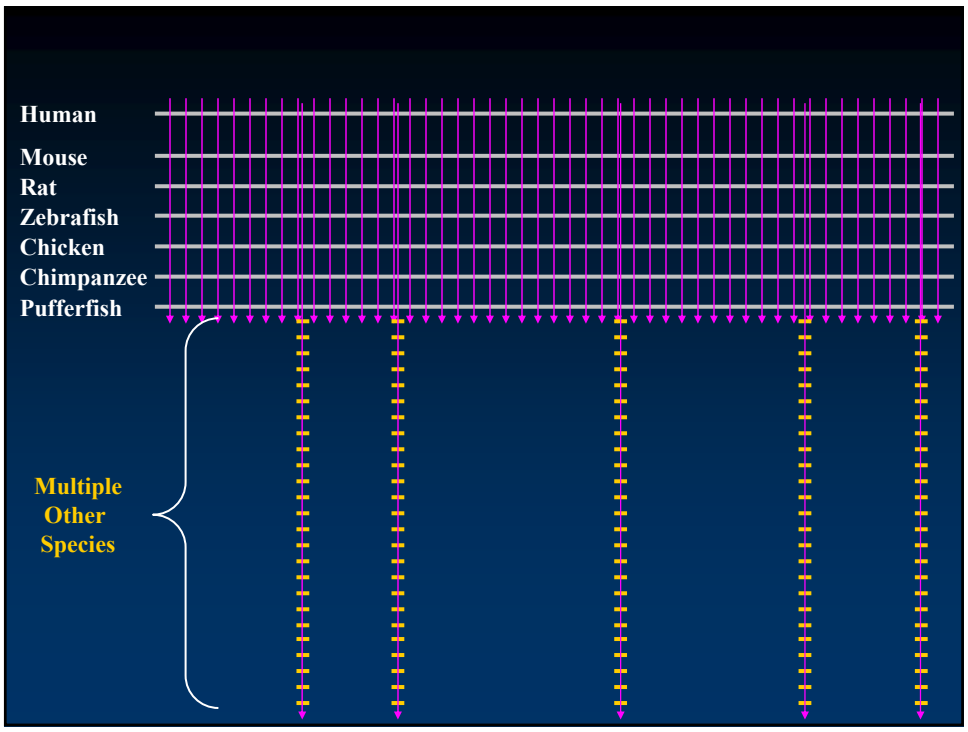


Comparative analyses of multi-species sequences from targeted genomic regions

J. W. Thomas[1*], J. W. Touchman[1,2*], R. W. Blakesley[1,2], G. G. Bouffard[1,2], S. M. Beckstrom-Sternberg[1,2], E. H. Margulies[1], M. Blanchette[3], A. C. Siepel[3], P. J. Thomas[2], J. C. McDowell[2], B. Maskeri[2], N. F. Hansen[2], M. S. Schwartz[3], R. J. Weber[3], W. J. Kent[3], D. Karolchik[3], T. C. Bruen[3], R. Bevan[3], D. J. Cutler[4], S. Schwartz[5], L. Elnitski[5], J. R. Idol[1], A. B. Prasad[1], S.-Q. Lee-Lin[1], V. V. B. Maduro[1], T. J. Summers[1], M. E. Portnoy[1], N. L. Dietrich[2], N. Akhter[2], K. Ayele[2], B. Benjamin[2], K. Cariaga[2], C. P. Brinkley[2], S. Y. Brooks[2], S. Granite[2], X. Guan[2], J. Gupta[2], P. Haghighi[2], S.-L. Ho[2], M. C. Huang[2], E. Karlins[2], P. L. Laric[2], R. Legaspi[2], M. J. Lim[2], Q. L. Maduro[2], C. A. Masiello[2], S. D. Mastrian[2], J. C. McCloskey[2], R. Pearson[2], S. Stantripop[2], E. E. Tiongson[2], J. T. Tran[2], C. Tsurgeon[2], J. L. Vogt[2], M. A. Walker[2], K. D. Wetherby[2], L. S. Wiggins[2], A. C. Young[2], L.-H. Zhang[2], K. Osoegawa[6], B. Zhu[6], B. Zhao[6], C. L. Shu[6], P. J. De Jong[6], C. E. Lawrence[7], A. F. Smit[8], A. Chakravarti[4], D. Haussler[3,9], P. Green[10], W. Miller[5] & E. D. Green[1,2]

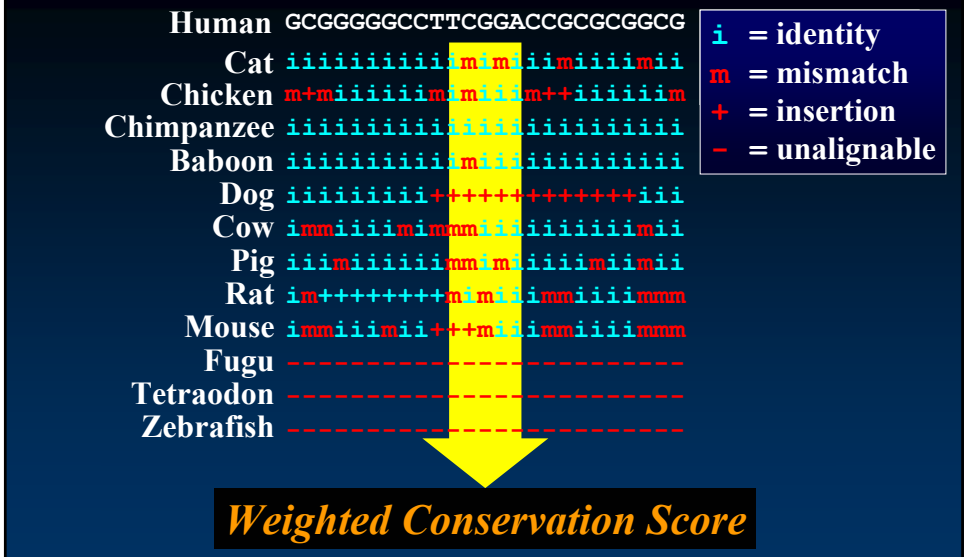
Nature **424**:788-793

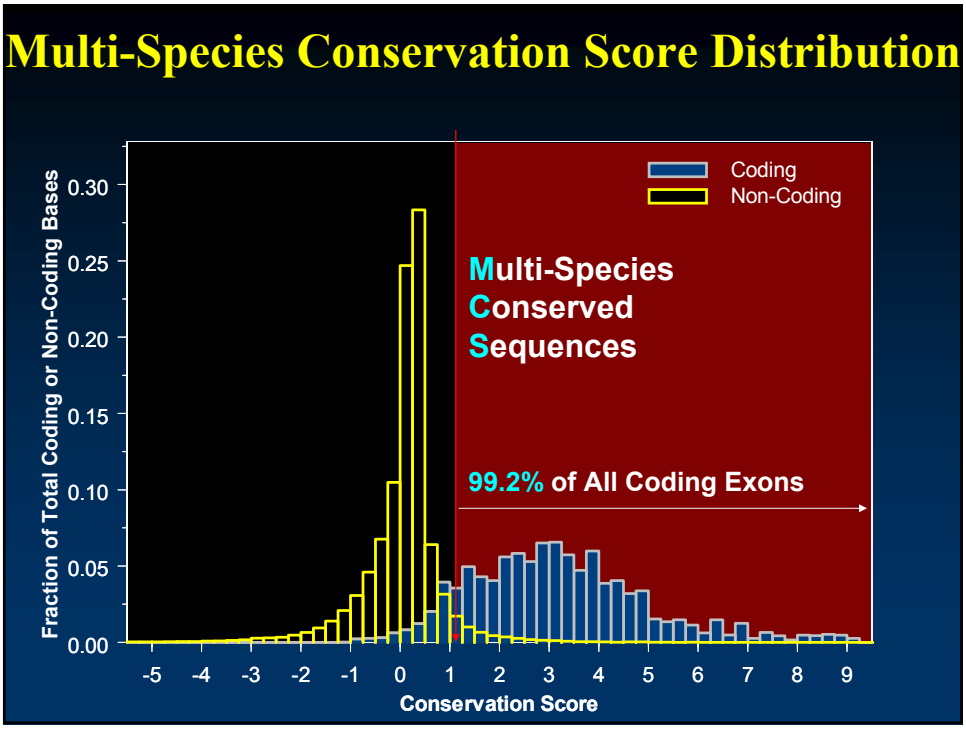
http://genome.ucsc.edu/cgi-bin/hgGateway?org=Zoo

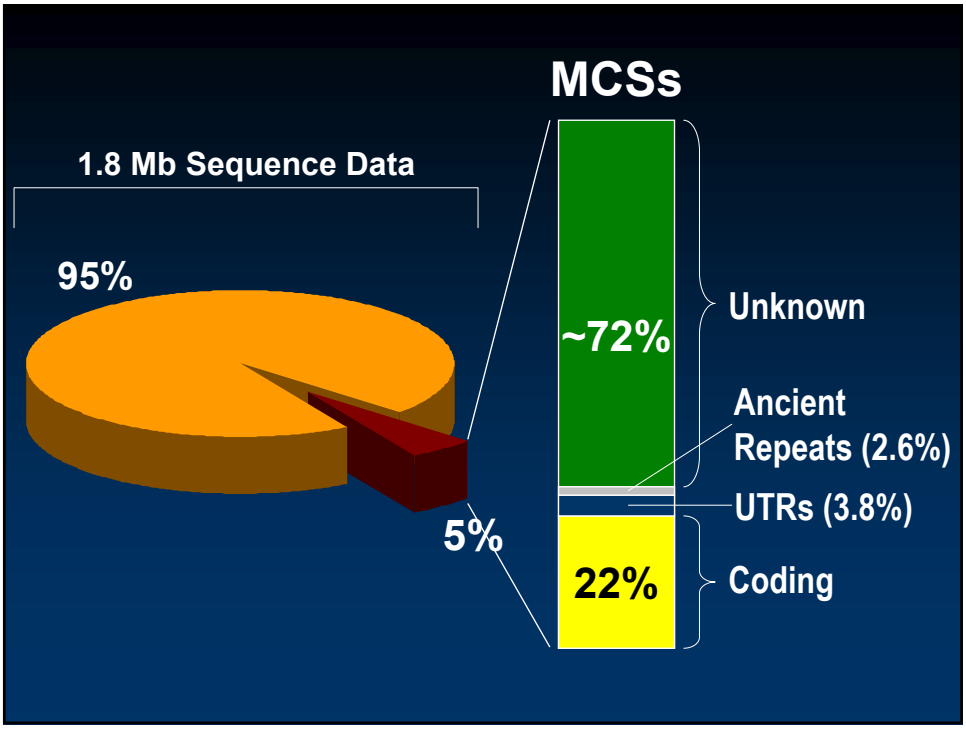


Multi-Species Weighted Conservation Score

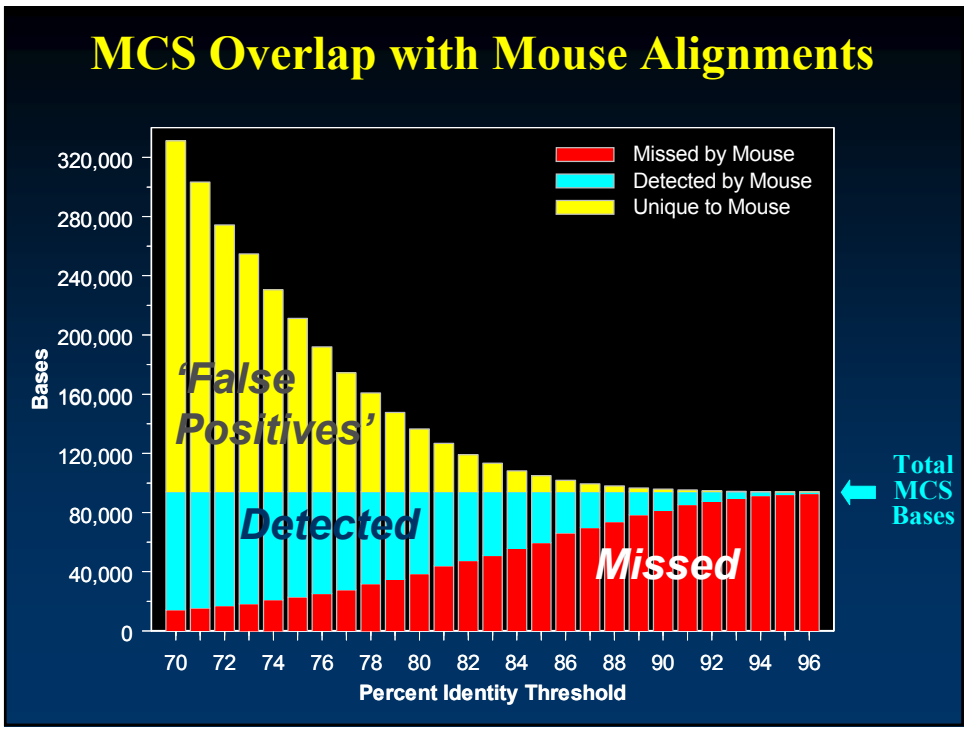
- Takes into Account the Different Divergence Rates of Each Species
  - *"A Chicken Alignment Will Contribute More Than a Baboon Alignment"*
- Based On the Substitution Rates at Bases under Neutral Selection
  - Calculated from 4-Fold Degenerate Positions



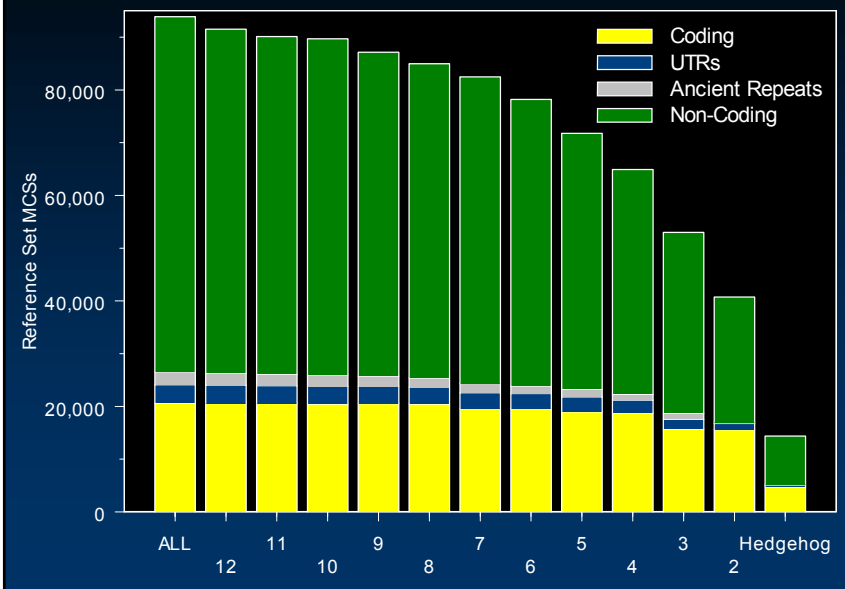

Multi-Species Conservation Score Distribution
Coding
Non-Coding
Multi-Species
Conserved
Sequences
99.2% of All Coding Exons
Fraction of Total Coding or Non-Coding Bases
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9
Conservation Score


MCSs
1.8 Mb Sequence Data
95%
5%
~72%
22%
Unknown
Ancient
Repeats (2.6%)
UTRs (3.8%)
Coding


MCS
CFTR
Human mRNAs from Genbank
DNAase I Hypersensitive Sites
Conservation Score
Binomial_25 Conservation Scores
Cat
Dog
Cow
Pig
Horse
Rabbit
Hedgehog
Rat
Mouse
Platypus
Opossum
Chicken
Fugu


MCS Overlap with Mouse Alignments
Missed by Mouse
Detected by Mouse
Unique to Mouse
Bases
320,000
280,000
240,000
200,000
160,000
120,000
80,000
40,000
0
'False Positives'
Detected
Missed
Total MCS Bases
70
72
74
76
78
80
82
84
86
88
90
92
94
96
Percent Identity Threshold

Subsets of Species Perform Better



More Species is Better