Current Topics in Genome Analysis
March 1, 2005

# Computational Techniques in Comparative Genomics

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

## Outline

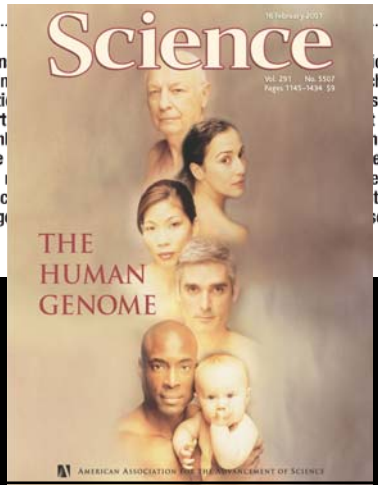
- Fundamental concepts of comparative genomics
- Alignment and visualization tools
  - Pair-wise and multi-species methods
  - Combining with transcription factor binding site data
- Motif Identification
- Comparative genomics resources available at UC Santa Cruz -- `http://genome.ucsc.edu`
  - Genome-wide sequence availability
  - Gene prediction and identification Finding orthologous sequences in other species
  - Identifying conserved sequences
- Insights from vertebrate genome sequence comparisons
- Multi-species sequence analysis

# Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium*

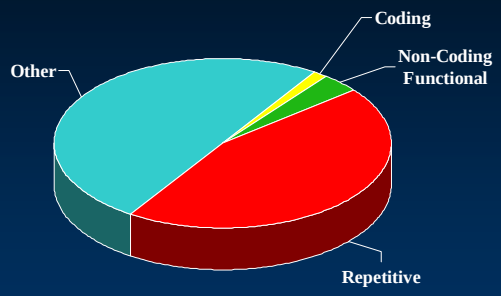
* *A list of authors and their affiliations appears in the* Supplementary Information

The sequ nstruction on about
human ev Sequencin hromatic
portion of collaborati sequence
with high we report genome
sequence pted by onl nome and
is accura any of the egmental
duplicatio ods. The e, greatly
improves enome inc tably, the
human ge -coding g serve as a
firm foun ead.


# Why Compare Genomic Sequences from Different Species?

- Explore evolutionary relationships




- Enhanced gene prediction algorithms
- Identify *functionally constrained* sequence

## Charles Darwin

- Served as *naturalist* on a British science expedition around the world (1831 -- 1836)
- *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*
- *The Origin of Species* (1859)
  - All species evolved from a single life form
  - "Variation" within a species occurs randomly
  - Natural selection
  - Evolutionary change is gradual

## Other Intellectual Foundations

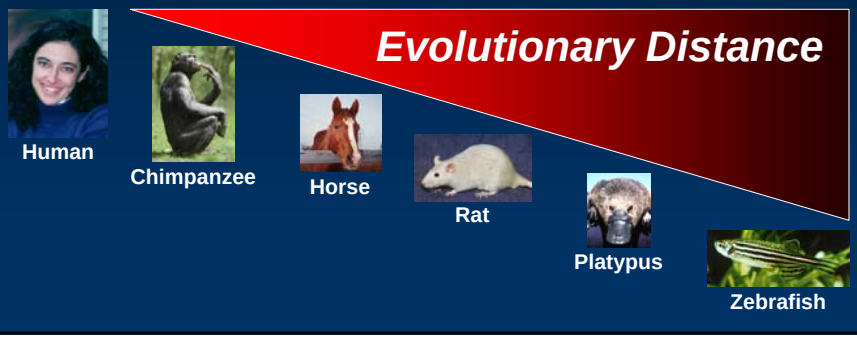
- Darwin (1859)
  - Theories of Evolution
- Mendel (1866) *(rediscovered in 1900)*
  - Genes are units of heredity
- Avery, McCarty & MacLeod (1944)
  - DNA as the "transforming principle"
- Watson & Crick (1953)
  - Structure of DNA
- Sanger (1977)
  - Methods of sequencing DNA

## Rationale

- DNA represents a “blueprint” for structure and physiology of all living things
- All species use DNA
- Mutations in *functional* DNA are less likely to be tolerated

## 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



## What's in a Name?

- Highly conserved sequences
- Sequences under purifying selection
- Functionally constrained sequences
- ECOR – Evolutionary COnserved Region
  - Variant: ECR
- CNS – Conserved Non-coding Sequence
- CNGs – Conserved Non-Genic sequence
- MCS – Multi-species Conserved Sequence

## Outline

- Fundamental concepts of comparative genomics
- Alignment and visualization tools
  - Pair-wise and multi-species methods
  - Combining with transcription factor binding site data
- Motif Identification
- Comparative genomics resources available at UC Santa Cruz -- `http://genome.ucsc.edu`
  - Genome-wide sequence availability
  - Gene prediction and identification
  - Finding orthologous sequences in other species
  - Identifying conserved sequences
- Insights from vertebrate genome 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 [2]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

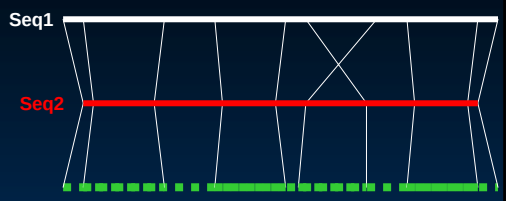
VISTA: *visualizing global DNA sequence alignments of arbitrary length*

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

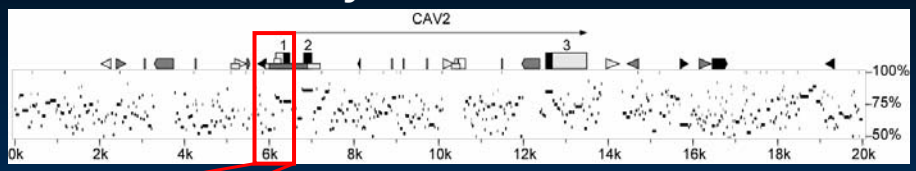
- Visualization
- Alignment Strategy
  - VISTA: `avid`
  - PipMaker: `blastz`
- East Coast – West Coast

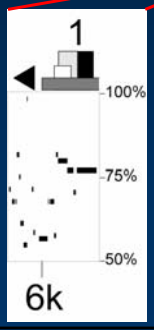




# PipMaker
# http://bio.cse.psu.edu/pipmaker/

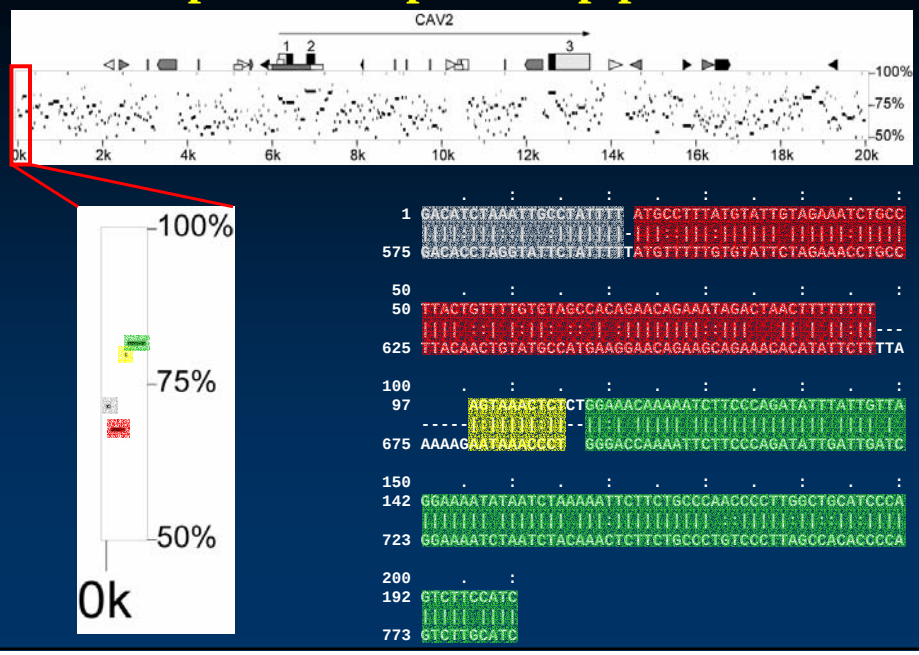
- **P**ercent **I**dentity **P**lot



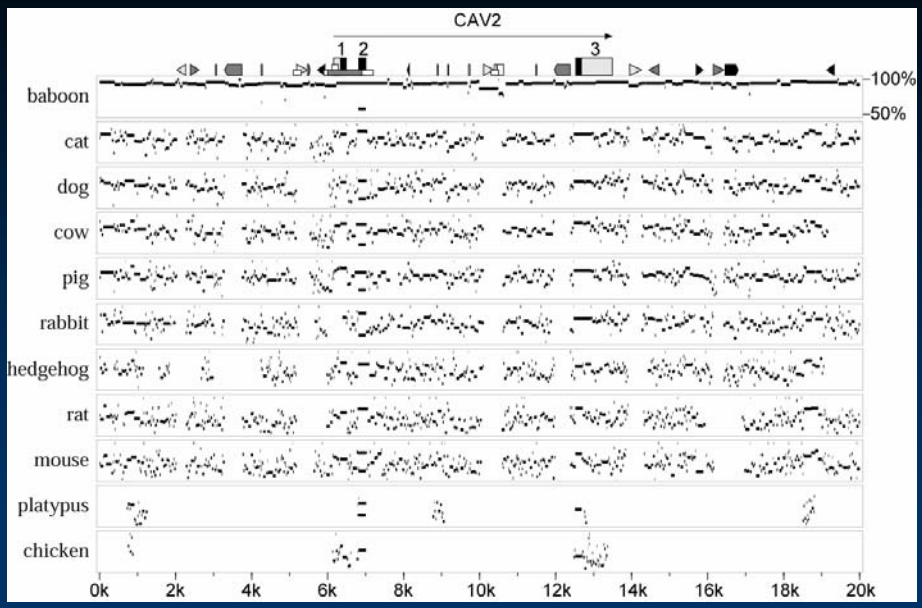




- X-axis is the reference sequence
- Horizontal lines represent gap-free alignments

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




# MultiPipMaker



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

VISUALIZATION TOOLS FOR ALIGNMENTS

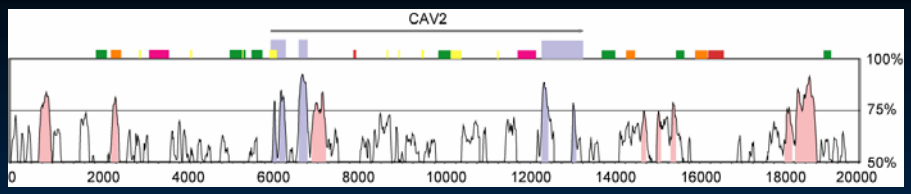
VISTA

- 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

```
GACATCTAAATTGCCTATTTT ATGCCTTTATGTATTGTAGAAATCTGCCTTACTGTTTTGTGTAGCCACAGAACAGAAATAGACTAACTTTTTTTT
||||:|||::|  :||||||||-|||::||||:|||||| |||||||:||||||||| ::| |:||: :: | :|||||||||::|||   || | ||:||
GACACCTAGGTATTCTATTTTTATGTTTTTGTGTATTCTAGAAACCTGCCTTACAACTGTATGCCATGAAGGAACAGAAGCAGAAACACATATTCTT
```

72% 68% 80% 88% 76% 52% 56% 64%

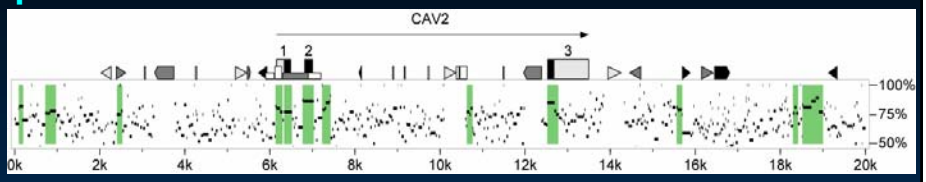
# VISTA



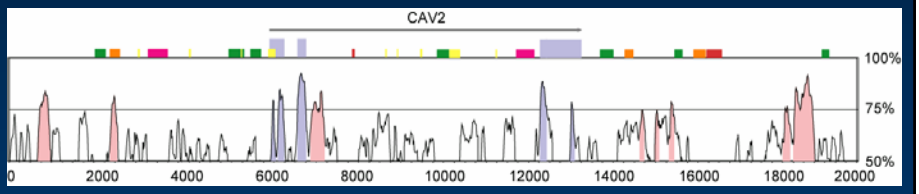

- Percent Identity is plotted from:
  - 100 base windows
  - Moved every 15 bases
- Colored regions meet certain alignment criteria
  - >100 bp >75% Identity

# What's Your Preference?

**PipMaker**




**VISTA**




# East & West Coast Unite

`http://zpicture.dcode.org/`

Resource

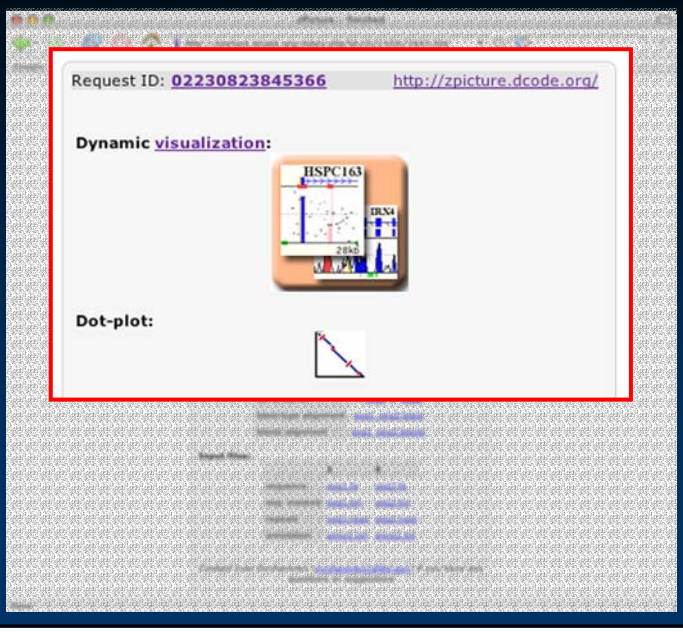
## zPicture: Dynamic Alignment and Visualization Tool for Analyzing Conservation Profiles

Ivan Ovcharenko,[1,2] Gabriela G. Loots,[2] Ross C. Hardison,[3] Webb Miller,[4,5] and Lisa Stubbs[2,6]

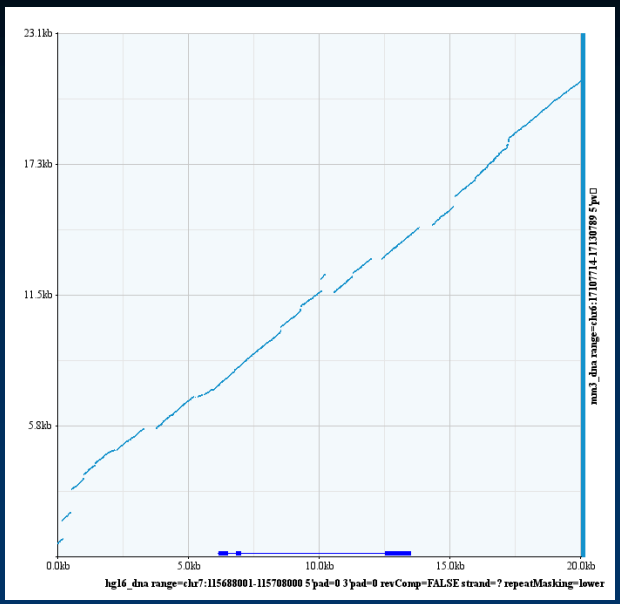
[1]*Energy, Environment, Biology and Institutional Computing, Lawrence Livermore National Laboratory, Livermore, California 94550, USA;* [2]*Genome Biology Division, Lawrence Livermore National Laboratory, Livermore, California 94550, USA;* [3]*Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA;* [4]*Department of Computer Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802, USA;* [5]*Department of Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA*

***Genome Research*, 2004, 14(3):472 7**

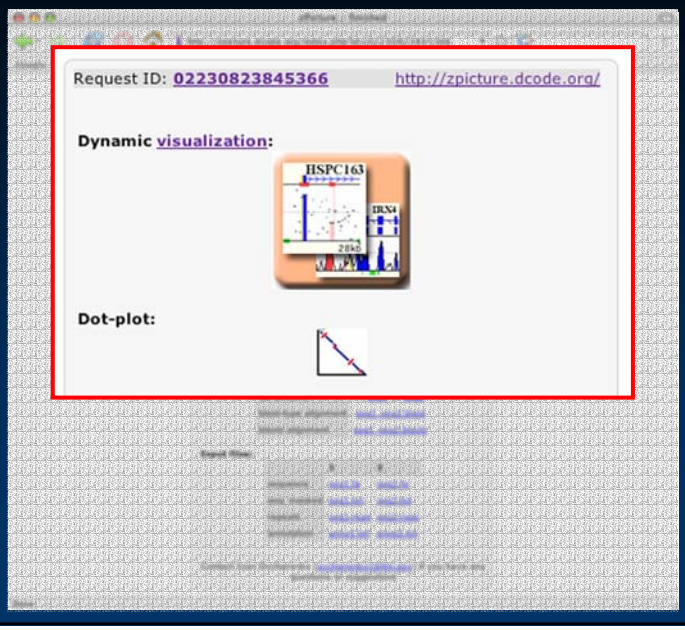
# zPicture Output Summary Page




# zPicture: Dot Plot View



# zPicture Output Summary Page

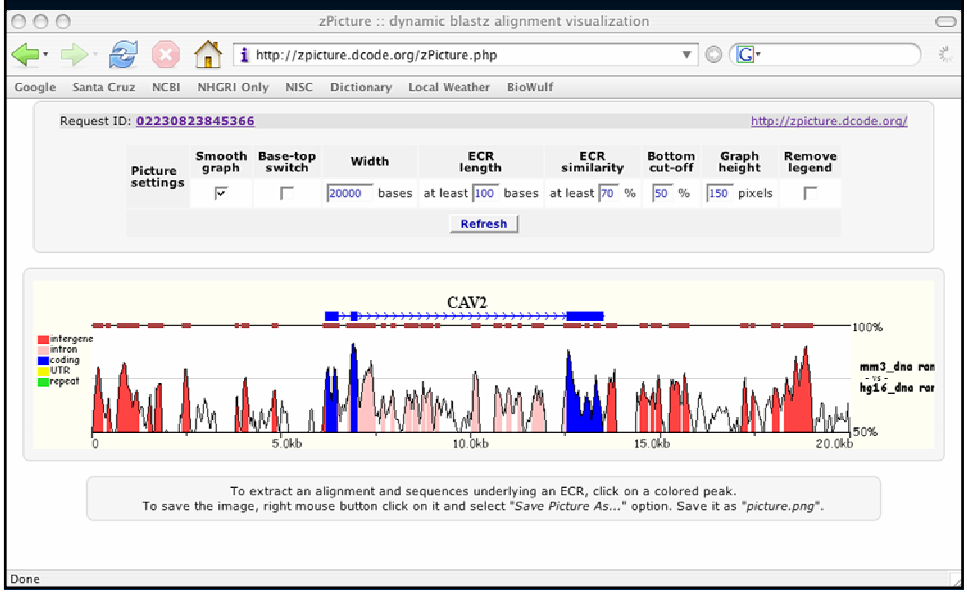


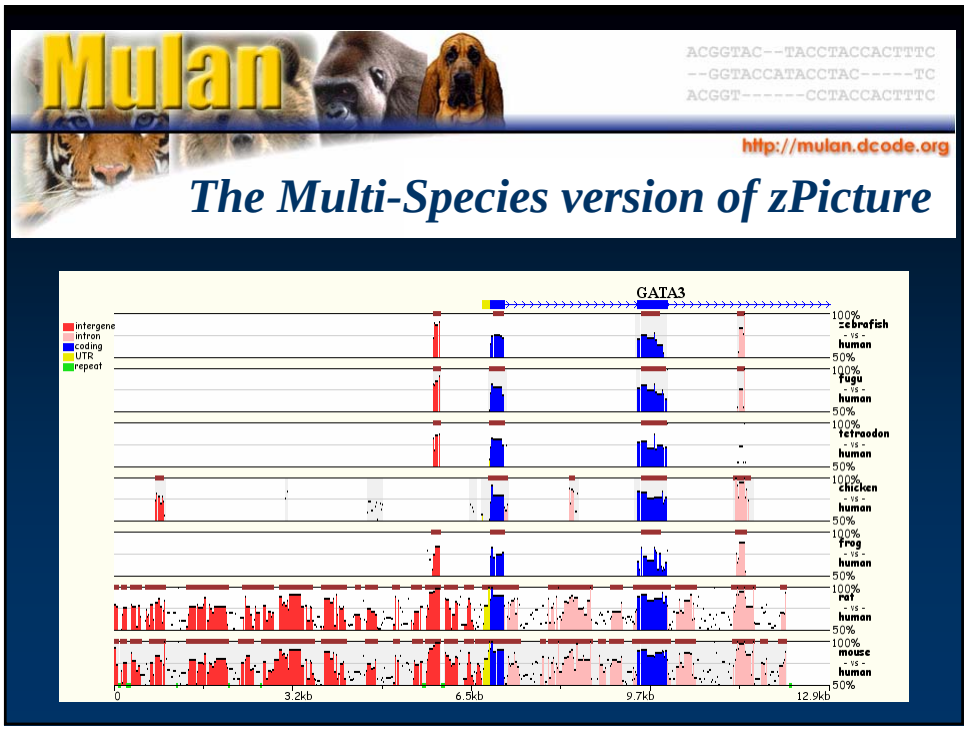

# Dynamic Visualization Options

## PipMaker-style

# Dynamic Visualization

## VISTA-style

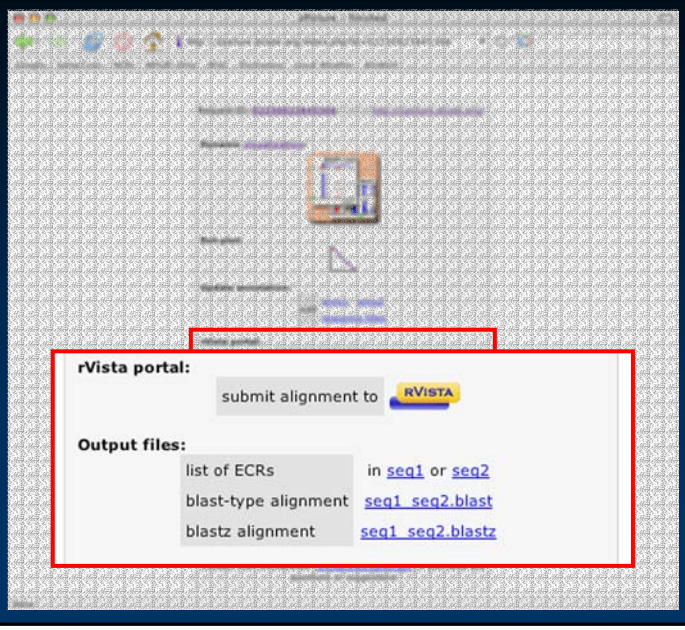


# Outline

- Fundamental concepts of comparative genomics
- Alignment and visualization tools
  - Pair-wise and multi-species methods
  - Combining with transcription factor binding site data
- Motif Identification
- Comparative genomics resources available at UC Santa Cruz -- `http://genome.ucsc.edu`
  - Genome-wide sequence availability
  - Gene prediction and identification
  - Finding orthologous sequences in other species
  - Identifying conserved sequences
- Insights from vertebrate genome sequence comparisons
- Multi-species sequence analysis

# zPicture Output Summary Page



# Are there any transcription factor binding sites in my alignment?


## *Pre-Computed Sequence Alignments*


*TRANSFAC*

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

**http://www.gene-regulation.com/**

# TRANSFAC

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

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

- A database of:
  - Eukaryotic transcription factors
  - Their genomic binding sites
  - And DNA binding profiles
- Data are collected from published studies
  - Non curated
  - Redundant data

# JASPAR: An Alternative to TRANSFAC

*Nucleic Acids Research, 2004, Vol. 32, Database issue* ***D91–D94***
*DOI: 10.1093/nar/gkh012*

**JASPAR: an open-access database for eukaryotic transcription factor binding profiles**

**Albin Sandelin, Wynand Alkema, Pär Engström, Wyeth W. Wasserman[1] and Boris Lenhard***

Center for Genomics and Bioinformatics, Karolinska Institutet, Berzelius väg 35, S-17177 Stockholm, Sweden and [1]Centre for Molecular Medicine and Therapeutics, Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada

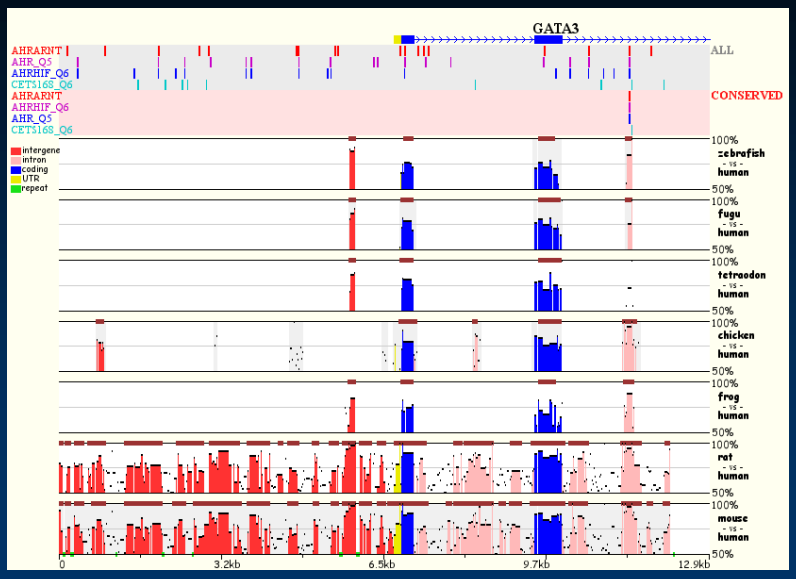
- **Differences from TRANSFAC:**
  - **Manually curated for "high quality" experiments**
  - **Non-redundant collection**

**http://jaspar.cgb.ki.se/**

# TRANSFAC Data are inherently "noisy"

- **Binding sites are very short**
  **6-10 bases in length**
- **Low complexity**
  **Only 4 "letters" in the DNA alphabet**
- **Frequently observe binding site by chance**
- ***Conservation can help reduce the noise***
- **rVISTA 2.0 for pair-wise alignments**
- **multiTF for multi-species alignments**

# Example of multiTF Output




# Summary of Alignment Tools

- **PipMaker (`blastz`)**
- **VISTA (`avid`)**
- **zPicture and MULAN**
- **Lagan and mLagan (glocal alignments)**
  - `http://lagan.stanford.edu/`
- **rVISTA 2.0**
- **Box 1 from:**

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

- **Table 1 from:**

Miller, Makova, Nekrutenko, and Hardison (2004) Comparative Genomics *Annual Reviews in Human Genetics* **5**:15-56

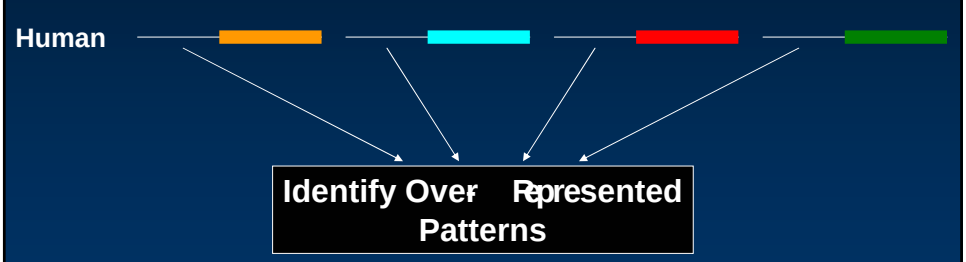
# Outline

- Fundamental concepts of comparative genomics
- Alignment and visualization tools
  - Pair-wise and multi-species methods
  - Combining with transcription factor binding site data
- **Motif Identification**
- Comparative genomics resources available at UC Santa Cruz -- `http://genome.ucsc.edu`
  - Genome-wide sequence availability
  - Gene prediction and identification Finding orthologous sequences in other species
  - Identifying conserved sequences
- Insights from vertebrate genome sequence comparisons
- Multi-species sequence analysis

# Motif Finding

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

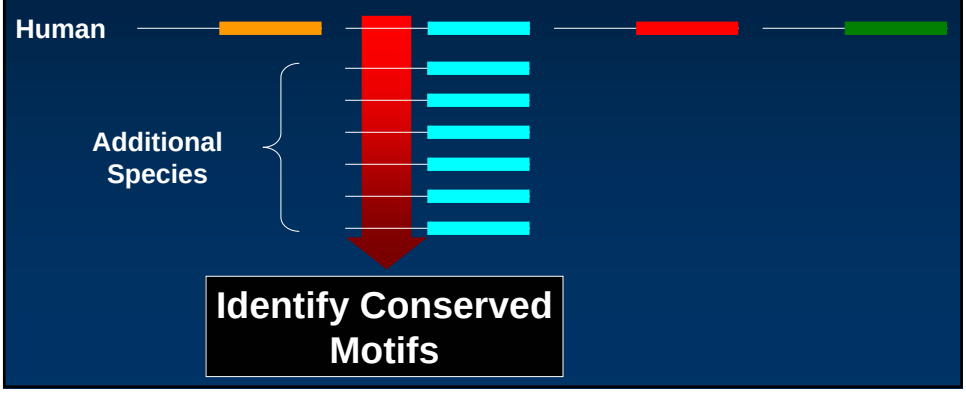
***Coordinately Regulated Genes***



## Phylogenetic Footprinting

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

***Orthologous Genes***



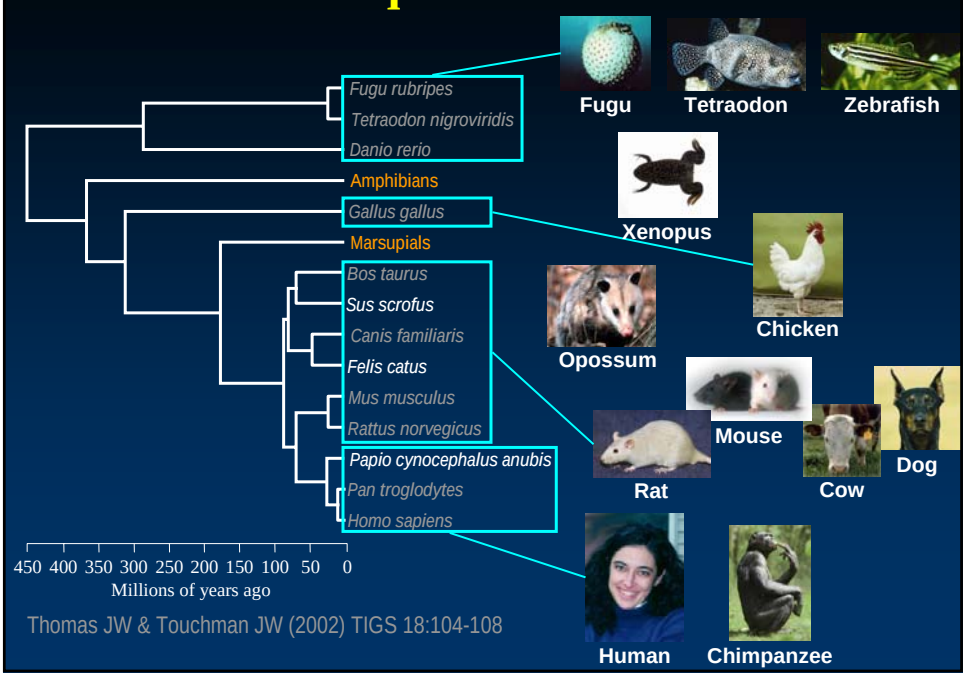

## Summary of Phylogenetic Footprinting Tools

- **FootPrinter –** http://bio.cs.washington.edu/software.html
  - Blanchette and Tompa (2003) *Nucleic Acids Research* **31:**3840–3842
- **phyloCon –** http://oldural.wustl.edu/~twang/PhyloCon/
  - Wang and Stormo (2003) *Bioinformatics* **19:**2369-80
- **phyME**
  - Sinha, Blanchette, and Tompa (2004) *BMC Bioinformatics* **28:**170
- **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

# Outline

- Fundamental concepts of comparative genomics
- Alignment and visualization tools
  - Pair-wise and multi-species methods
  - Combining with transcription factor binding site data
- Motif Identification
- Comparative genomics resources available at UC Santa Cruz -- http://genome.ucsc.edu
  - Genome-wide sequence availability
  - Gene prediction and identification Finding orthologous sequences in other species
  - Identifying conserved sequences
- Insights from vertebrate genome sequence comparisons
- Multi-species sequence analysis

# Genome-Wide Sequences




Thomas JW & Touchman JW (2002) TIGS 18:104-108

# Genome Browsers

UCSC Genome Bioinformatics

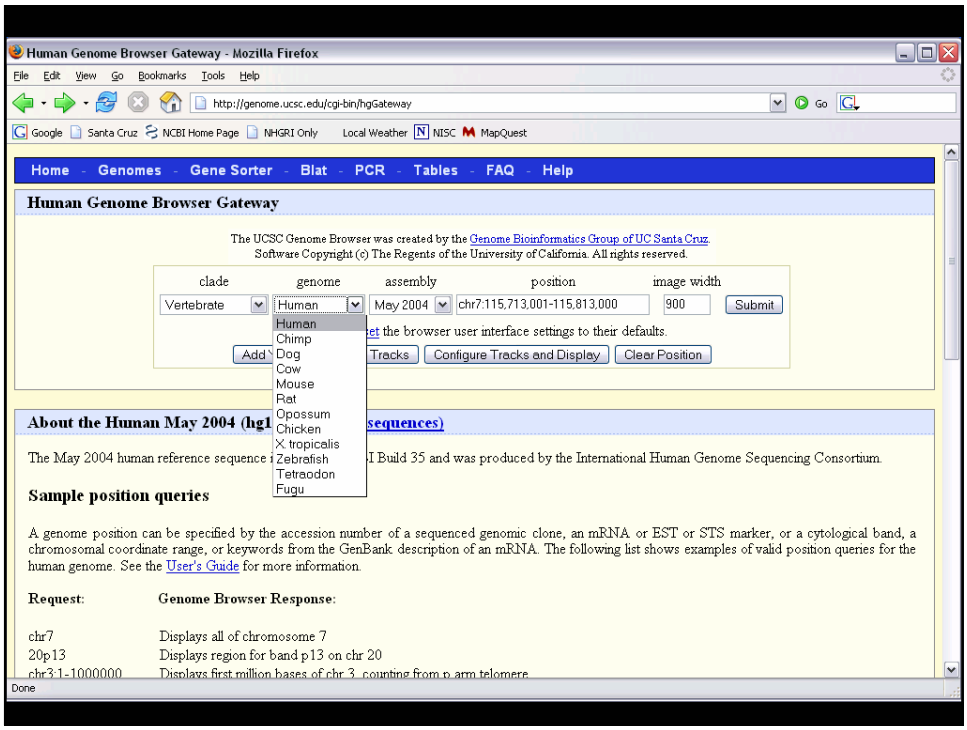
http://genome.ucsc.edu

project Ensembl

http://www.ensembl.org

NCBI Map Viewer

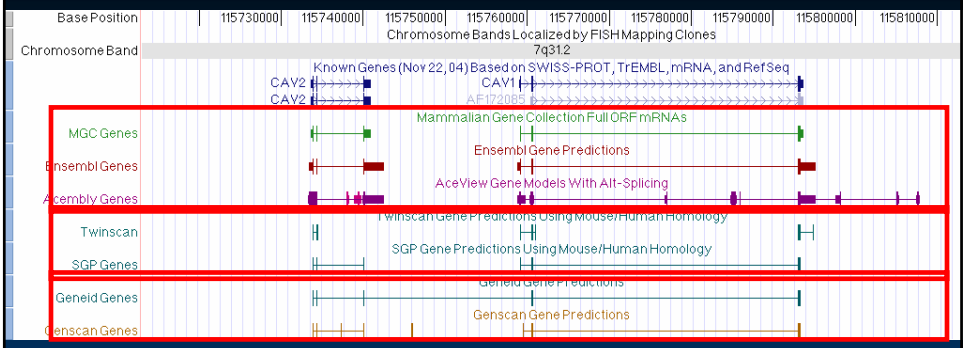
http://www.ncbi.nlm.nih.gov/mapview/



# Outline

- Fundamental concepts of comparative genomics
- Alignment and visualization tools
  - Pair-wise and multi-species methods
  - Combining with transcription factor binding site data
- Motif Identification
- Comparative genomics resources available at UC Santa Cruz -- `http://genome.ucsc.edu`
  - Genome-wide sequence availability
  - Gene prediction and identification
  - Finding orthologous sequences in other species
  - Identifying conserved sequences
- Insights from vertebrate genome sequence comparisons
- Multi-species sequence analysis

# Approaches to Gene Prediction




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

## Additional Gene Prediction Resources

- **Fugu BLAT Track at UCSC**
- **SLAM –** http://baboon.math.berkeley.edu/~syntenic/slam.html
  - Cawley et al. (2003) *Nucleic Acids Research* **31:**3507-3509
- **Exoniphy**

  Siepel and Haussler. Computational identification of evolutionarily conserved exons. *Proc. 8th Annual Int'l Conf. on Research in Computational Biology*, pp. 177-186, 2004.

  `http://www.soe.ucsc.edu/~acs/recomb2004.pdf`
  - **Also see genome "test" browser for data**
- **Box 1 from:**
  - Ureta-Vidal et al. (2003) *Nature Reviews Genetics* **4:**251-262

## Outline

- Fundamental concepts of comparative genomics
- Alignment and visualization tools
  - Pair-wise and multi-species methods
  - Combining with transcription factor binding site data
- Motif Identification
- **Comparative genomics resources available at UC Santa Cruz -- `http://genome.ucsc.edu`**
  - Genome-wide sequence availability
  - Gene prediction and identification
  - **Finding orthologous sequences in other species**
  - **Identifying conserved sequences**
- Insights from vertebrate genome sequence comparisons
- Multi-species sequence analysis

# Chaining Alignments

- Chaining bridges the gulf between large syntenic blocks and base by base alignments.

**The Challenge:**

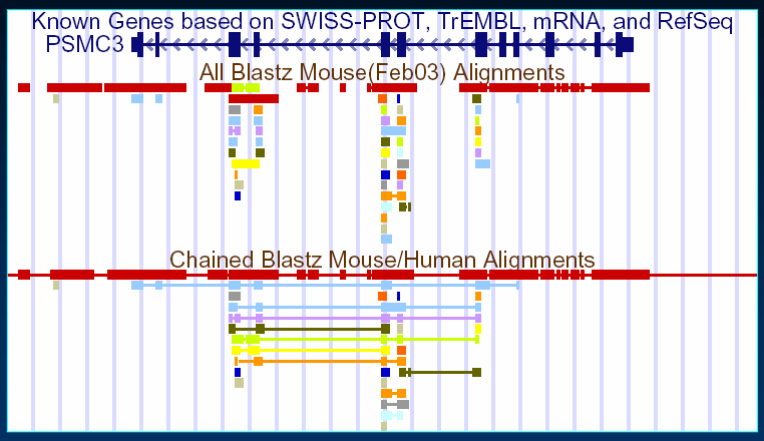
- Local alignments tend to break at transposon insertions, inversions, duplications, etc.
- Global alignments tend to force non-homologous bases to align.

**The Solution:**

- Chaining is a rigorous way of joining together local alignments into larger structures.

*Slide (though modified) Courtesy of Jim Kent*

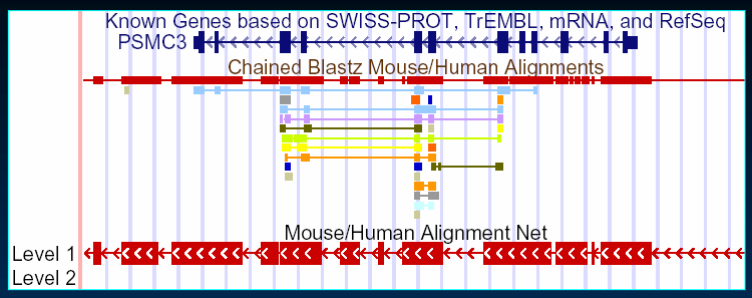
# Chains join together related local alignments




**Protease Regulatory Subunit 3**

*Slide Courtesy of Jim Kent*

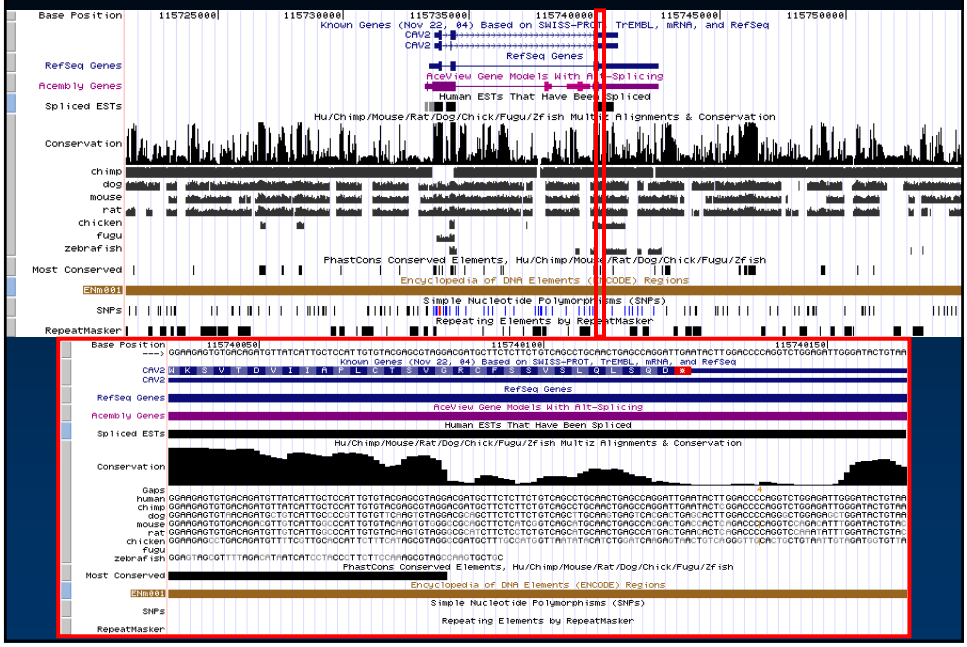
# Net Alignments: Focus on Orthology




- Frequently, there are numerous mouse alignments for any given human region, particularly for coding regions.
- Net finds best mouse match for each human region.

*Slide (though modified) Courtesy of Jim Kent*

# Genome-wide Multiple Sequence Alignments



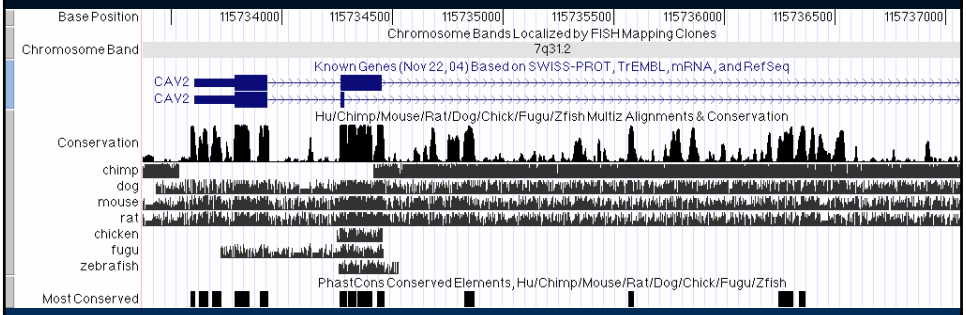
## Conservation Score at UCSC

- Displays evolutionary conservation based on a phylogenetic hidden Markov model

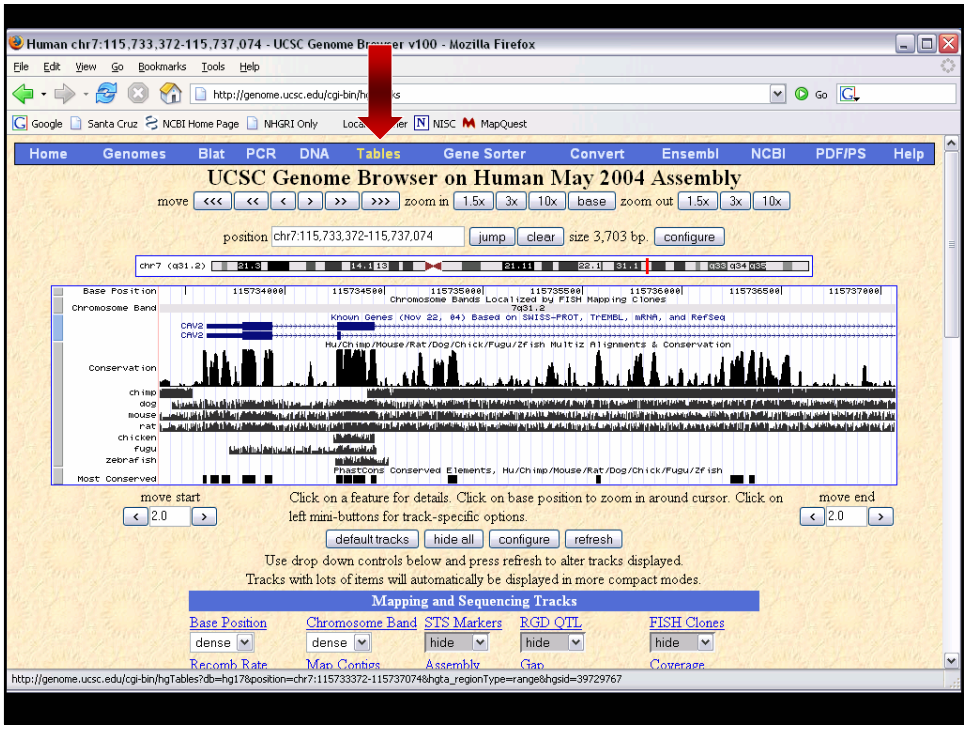
**Probability that the given alignment column was generated by the "conserved" state**

- "Most Conserved" track represents highly conserved regions
  - Tuned to cover ~4% of the genome

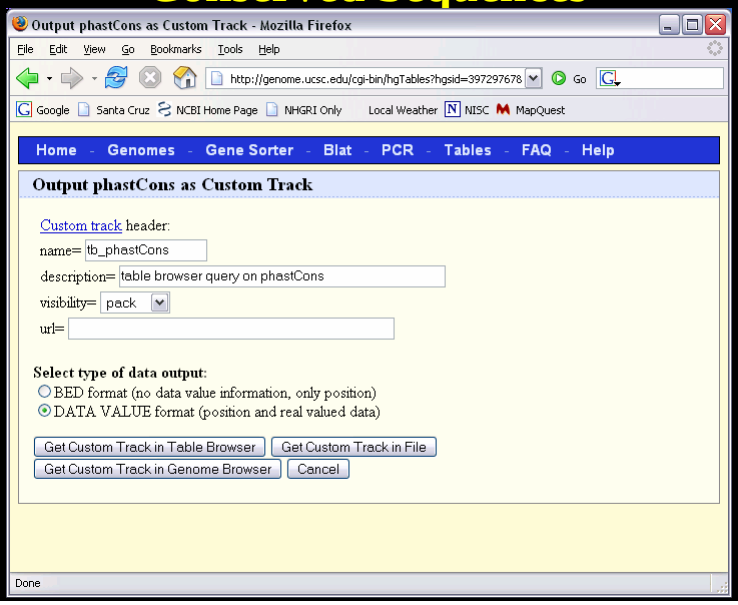
## "Most Conserved" Track at UCSC



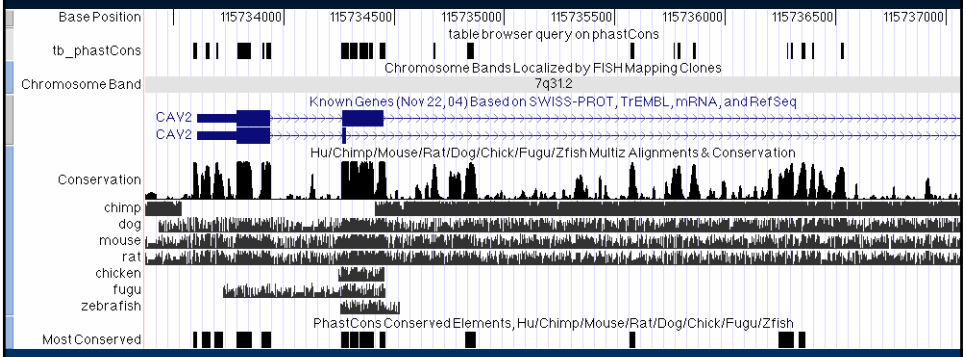

***What it you want a different stringency than was used for the Most Conserved Track?***




# Using the Table Browser to get Highly Conserved Sequences



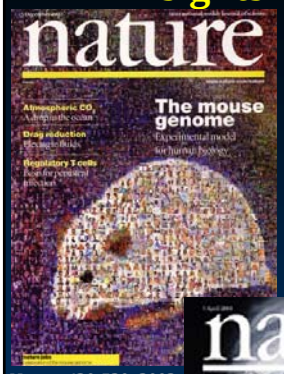
# Using the Table Browser to get Highly Conserved Sequences




# Outline

- Fundamental concepts of comparative genomics
- Alignment and visualization tools
  - Pair-wise and multi-species methods
  - Combining with transcription factor binding site data
- Motif Identification
- Comparative genomics resources available at UC Santa Cruz -- `http://genome.ucsc.edu`
  - Genome-wide sequence availability
  - Gene prediction and identification Finding orthologous sequences in other species
  - Identifying conserved sequences
- Insights from vertebrate genome sequence comparisons
- Multi-species sequence analysis

## Insights from Human-Rodent Sequence Comparisons




*Nature* 420:520, 2002


*Nature* 428:493, 2004

- 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
- See Jan 2003 & April 2004 issues of *Genome Research*

## 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

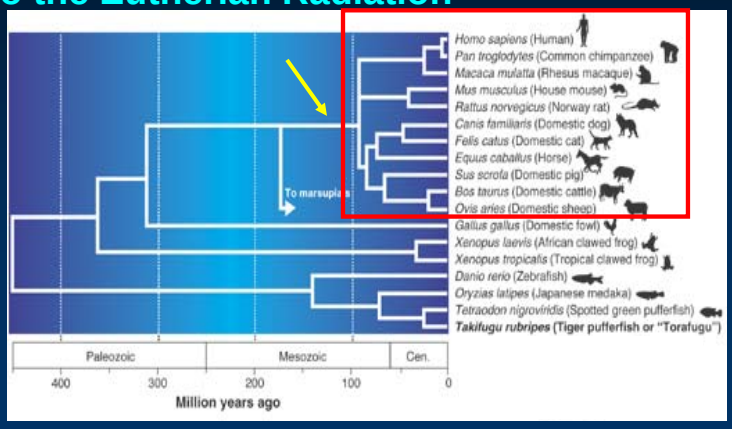
# Types of Neutrally Evolving DNA

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

| First | Second: 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

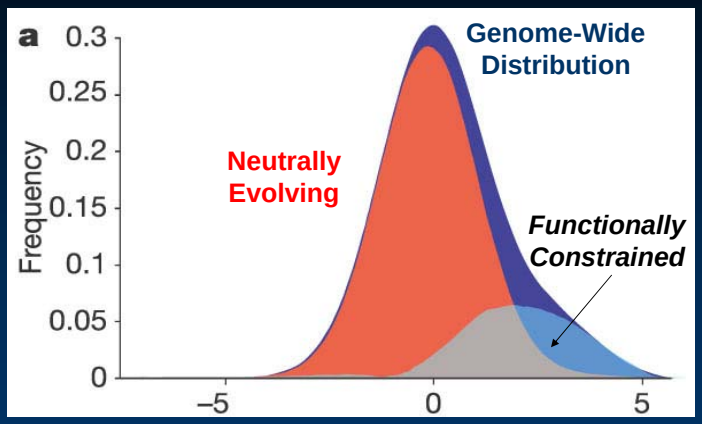
- **Ancestral Repeats**
  - Ancient Relics of Transposons Inserted Prior to the Eutherian Radiation




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

# Determining the Fraction of Sequence Under Purifying Selection

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




**Conservation Score**

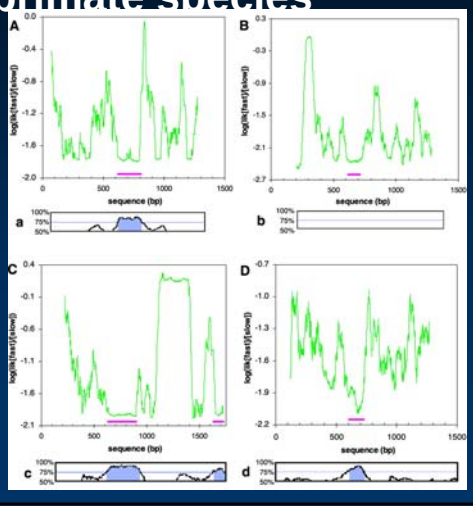
# Outline

- Fundamental concepts of comparative genomics
- Alignment and visualization tools
  - Pair-wise and multi-species methods
  - Combining with transcription factor binding site data
- Motif Identification
- Comparative genomics resources available at UC Santa Cruz -- `http://genome.ucsc.edu`
  - Genome-wide sequence availability
  - Gene prediction and identification Finding orthologous sequences in other species
  - Identifying conserved sequences
- Insights from vertebrate genome sequence comparisons
- Multi-species sequence analysis

# Phylogenetic Shadowing

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

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



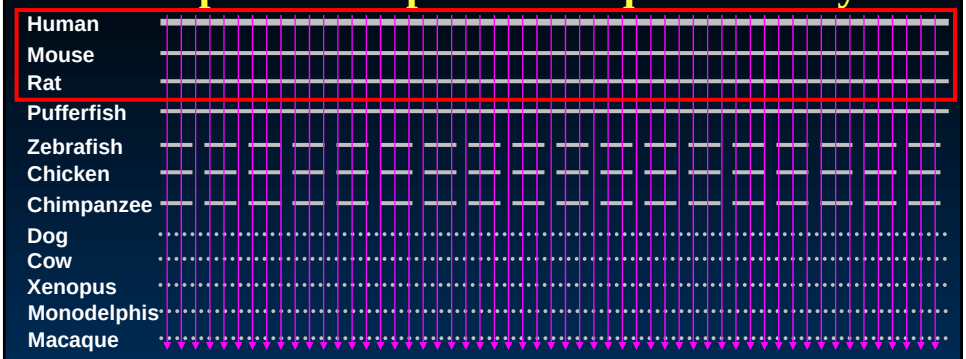
# Multi-Species Comparative Sequence Analysis

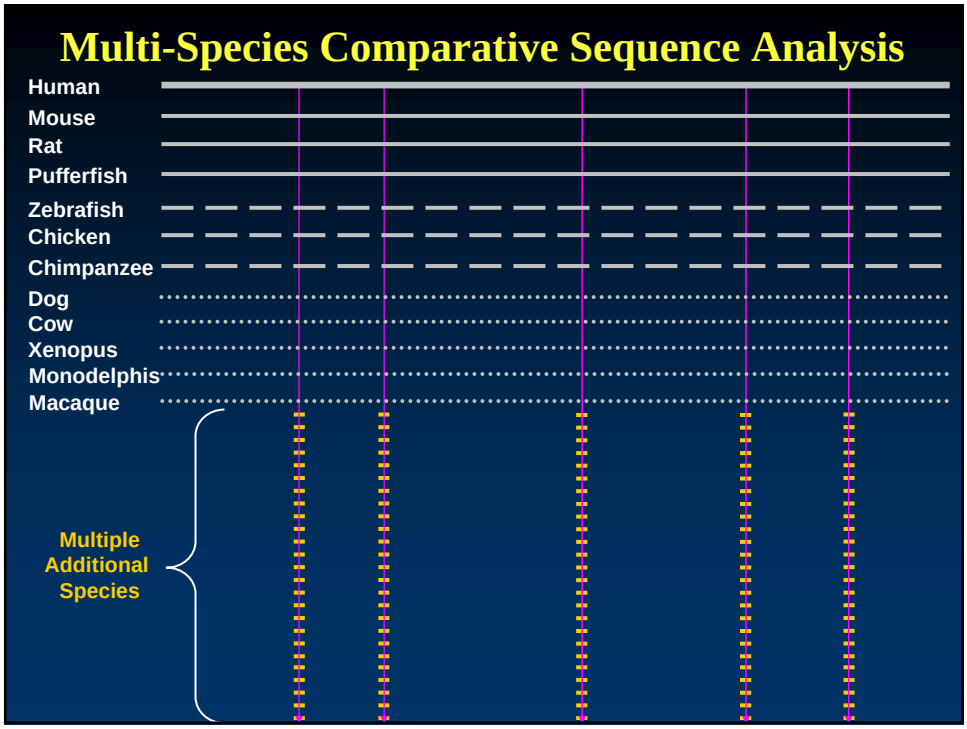
**Comparative analyses of multi-species sequences from targeted genomic regions**

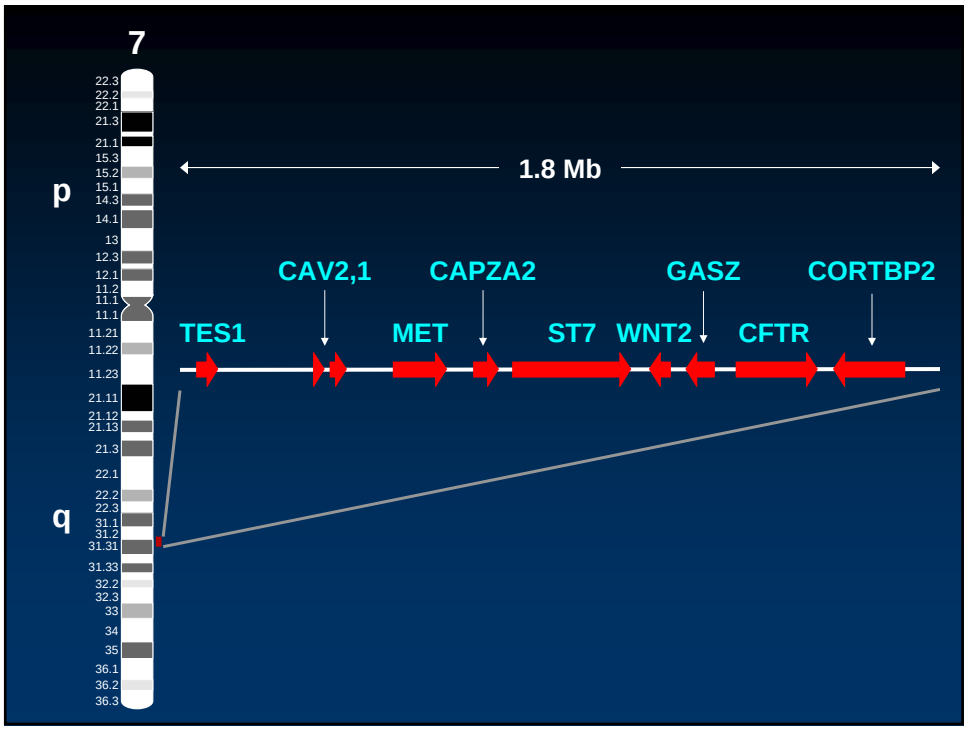
J. W. Thomas[1*], J. W. Touchman[1,2*], R. W. Blakesley[1,2], G. G. Bouffard[1,2], S. M. Beckstrom-Sternberg[1,2], E. H. Margulies[1], M. Blanchette[3], A. C. Siepel[3], P. J. Thomas[2], J. C. McDowell[2], B. Maskeri[2], N. F. Hansen[2], M. S. Schwartz[3], R. J. Weber[3], W. J. Kent[3], D. Karolchik[3], 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]

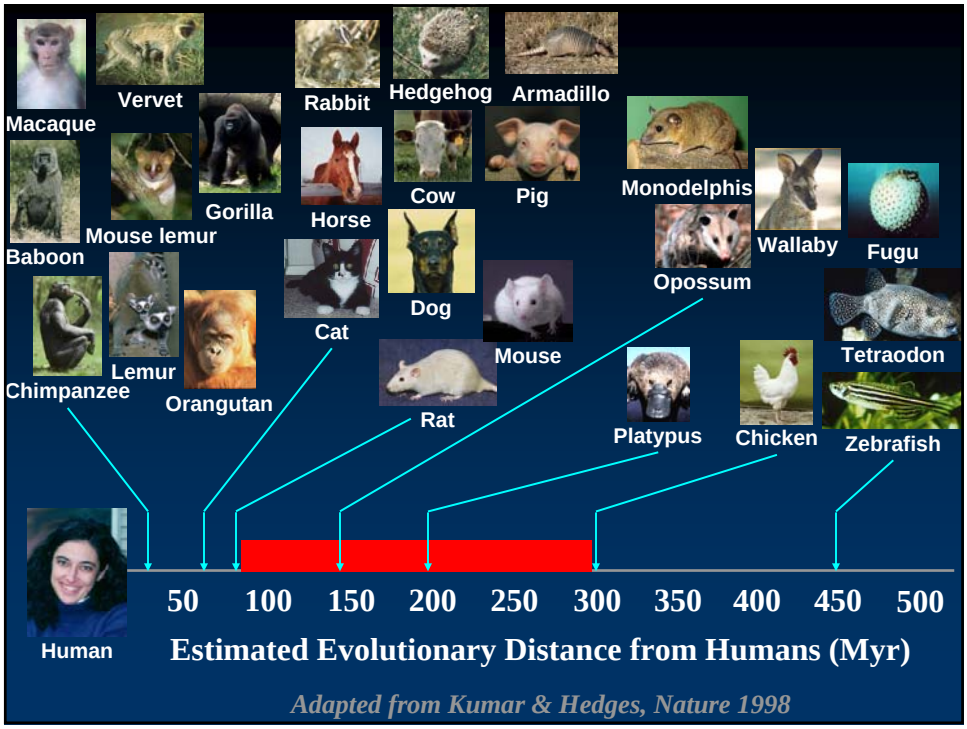
*Nature 424:788, 2003*

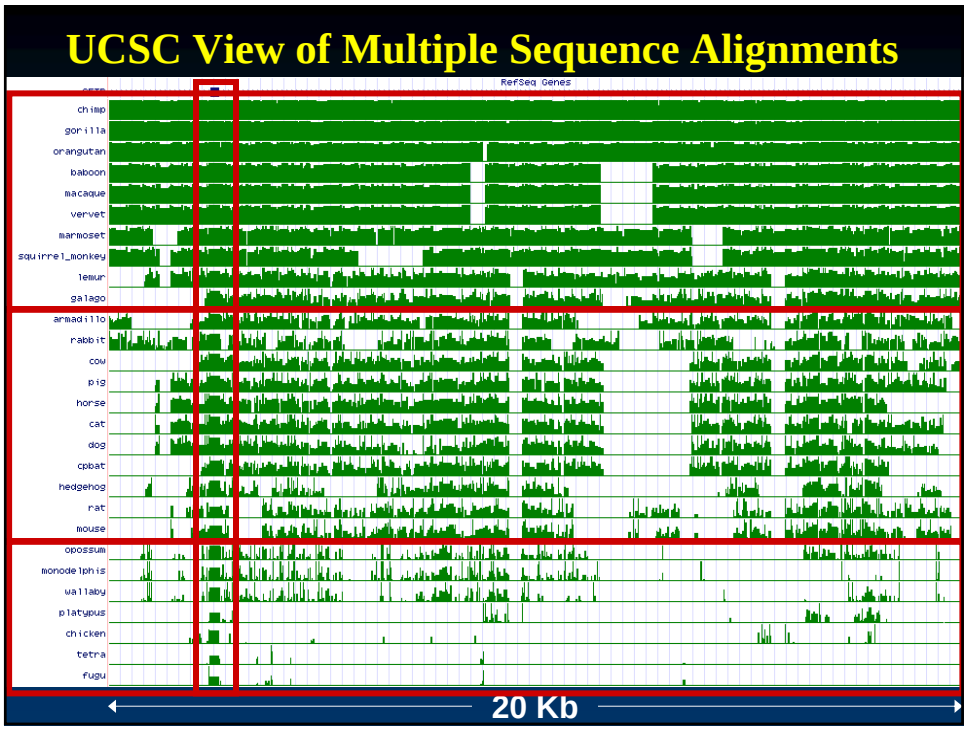
# Multi-Species Comparative Sequence Analysis




Multi-Species Comparative Sequence Analysis
Human
Mouse
Rat
Pufferfish
Zebrafish
Chicken
Chimpanzee
Dog
Cow
Xenopus
Monodelphis
Macaque
Multiple Additional Species


7
p
q
1.8 Mb
CAV2,1
CAPZA2
GASZ
CORTBP2
TES1
MET
ST7
WNT2
CFTR


Macaque
Vervet
Rabbit
Hedgehog
Armadillo
Baboon
Mouse lemur
Gorilla
Horse
Cow
Pig
Monodelphis
Wallaby
Fugu
Opossum
Chimpanzee
Lemur
Orangutan
Cat
Dog
Mouse
Tetraodon
Rat
Platypus
Chicken
Zebrafish
Human
50 100 150 200 250 300 350 400 450 500
Estimated Evolutionary Distance from Humans (Myr)
Adapted from Kumar & Hedges, Nature 1998


UCSC View of Multiple Sequence Alignments
RefSeq Genes
chimp
gorilla
orangutan
baboon
macaque
vervet
marmoset
squirrel_monkey
lemur
galago
armadillo
rabbit
cow
pig
horse
cat
dog
cpbat
hedgehog
rat
mouse
opossum
monodelphis
wallaby
platypus
chicken
tetra
fugu
20 Kb

## 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

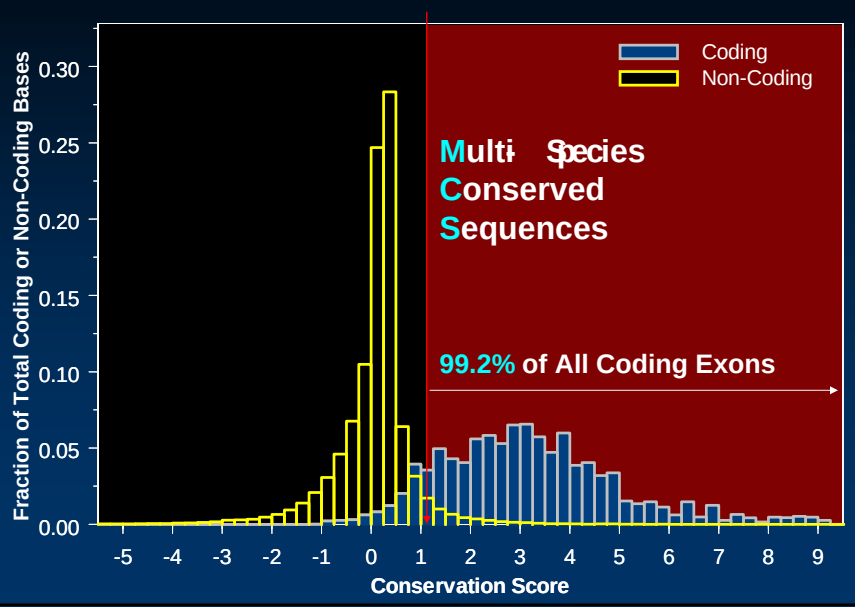
```
Human      GCGGGGGCCTTCGGACCGCGCGGCG
Cat        iiiiiiiiiimimiimiiiimii
Chicken    m+miiiiimimiiim++iiiiim
Chimpanzee iiiiiiiiiiiiiiiiiiiiiii
Baboon     iiiiiiiiimiiiiiiiiiiii
Dog        iiiiiiii+++++++++++++iii
Cow        immiiimimmmiiiiiiiimii
Pig        iiimiiiiimmimiiiimiimii
Rat        im++++++++mimiimmiiimmm
Mouse      immiiimii+++miimmiiimmm
Fugu       ---------------------
Tetraodon  ---------------------
Zebrafish  ---------------------
```

i = identity
m = mismatch
+ = insertion
- = unalignable

*Weighted Conservation Score*

skip

# Multi-Species Conservation Score Distribution




# Multi-Species Conservation Score

Article

Identification and Characterization of Multi-Species Conserved Sequences

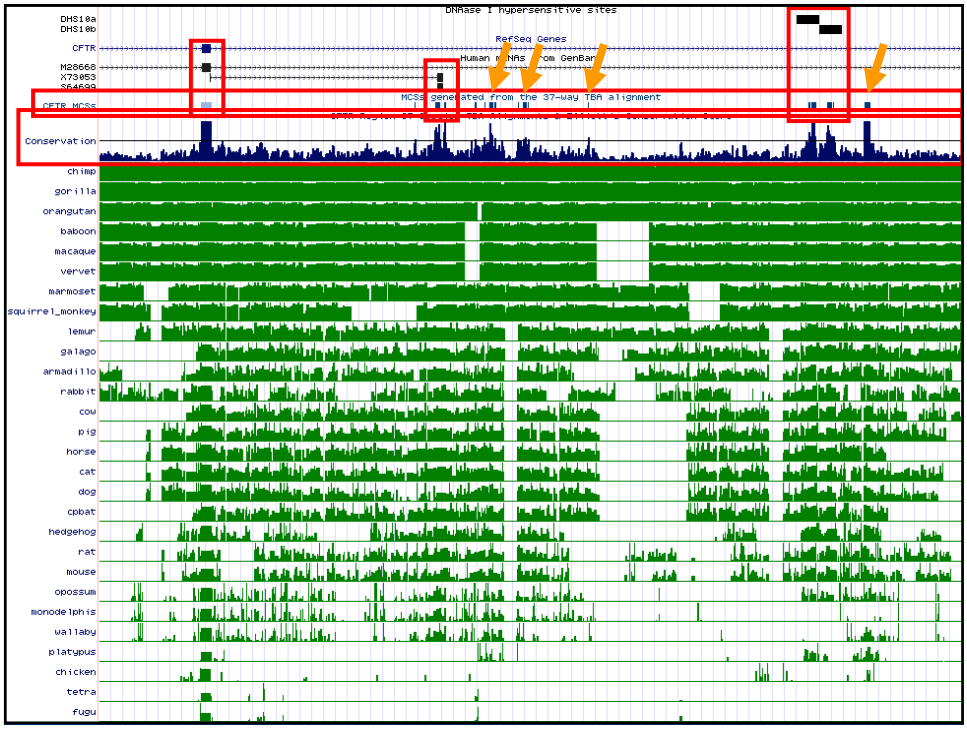
Elliott H. Margulies,[1] Mathieu Blanchette,[3] NISC Comparative Sequencing Program,[1,2] David Haussler,[3,4,5] and Eric D. Green[1,2,5]

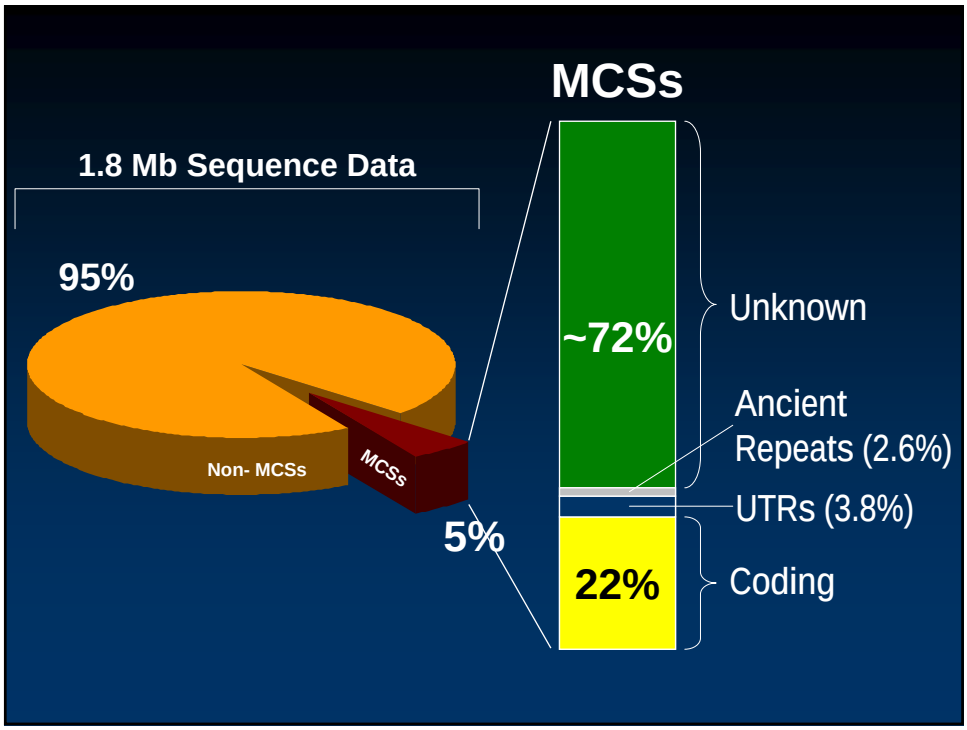
*Genome Research* (2003) 13:2507-2518

**MCS**

**Multispecies Conserved Sequence**

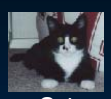
"Noisy" Sequence Alignments → "MCS Generator" → Discrete Regions of Highly Conserved Sequence


DNase I hypersensitive sites
DHS10a
DHS10b
CFTR
RefSeq Genes
M28668
X73053
S64699
CFTR MCSs
Conservation
chimp
gorilla
orangutan
baboon
macaque
vervet
marmoset
squirrel_monkey
lemur
galago
armadillo
rabbit
cow
pig
horse
cat
dog
cpbat
hedgehog
rat
mouse
opossum
monodelphis
wallaby
platypus
chicken
tetra
fugu


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

# Lineage-Specificity of MCSs in Mammals

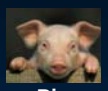
**Carnivores**


Cat


Dog

**Artiodactyls**


Cow


Pig

**Rodents**


Mouse


Rat

**Monotreme**


Platypus

**Marsupials**

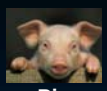

Wallaby


Opossum

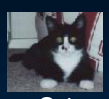
**For each MCS:**
**Catalog All the Species in Which it is Present**

*skip*


Cow


Pig

Artiodactyls


Cat


Dog

Carnivores


Mouse


Rat

Rodents


Platypus

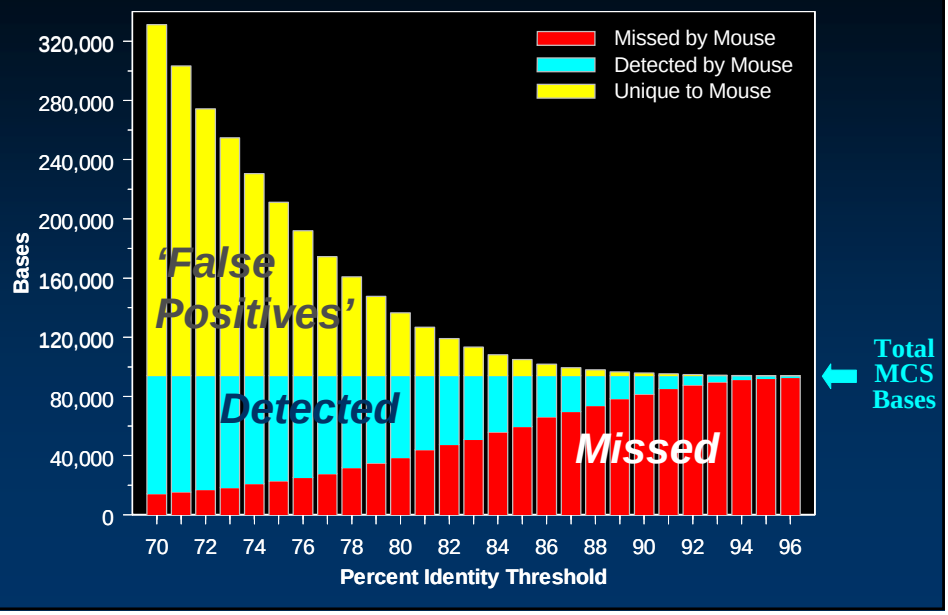
Monotreme


Wallaby


Opossum

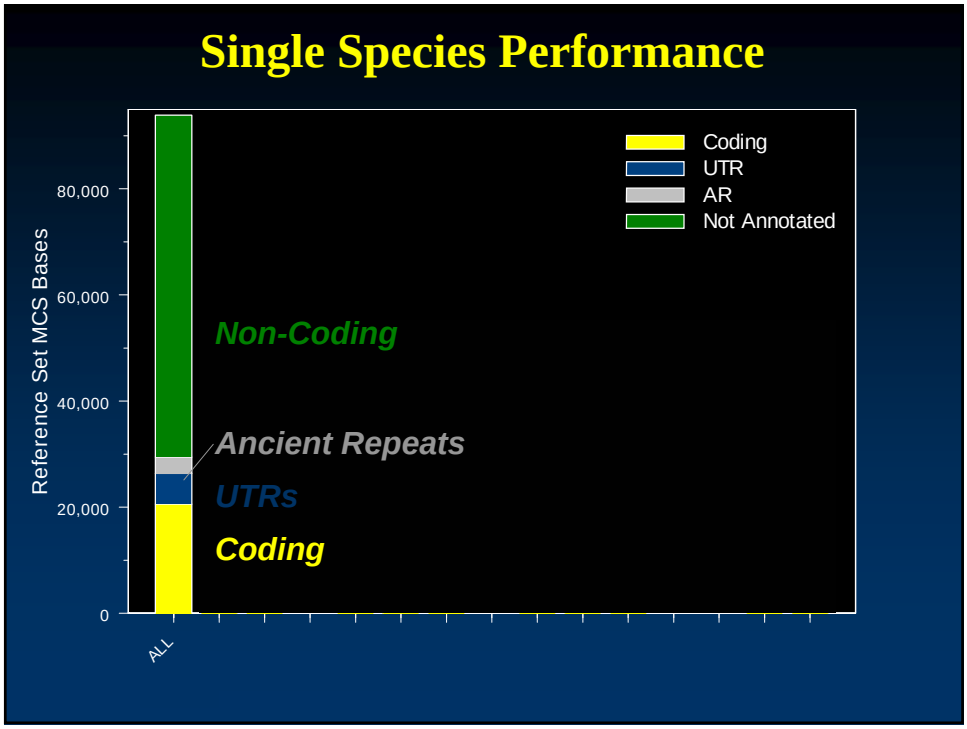
Marsupials

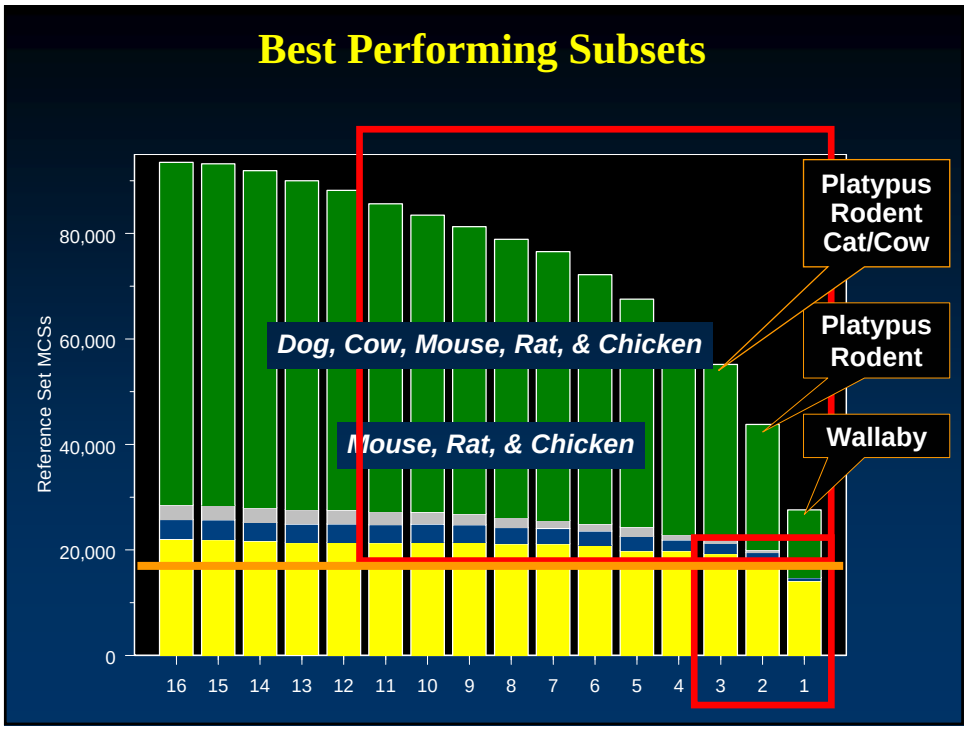
# MCS Overlap with Mouse Alignments



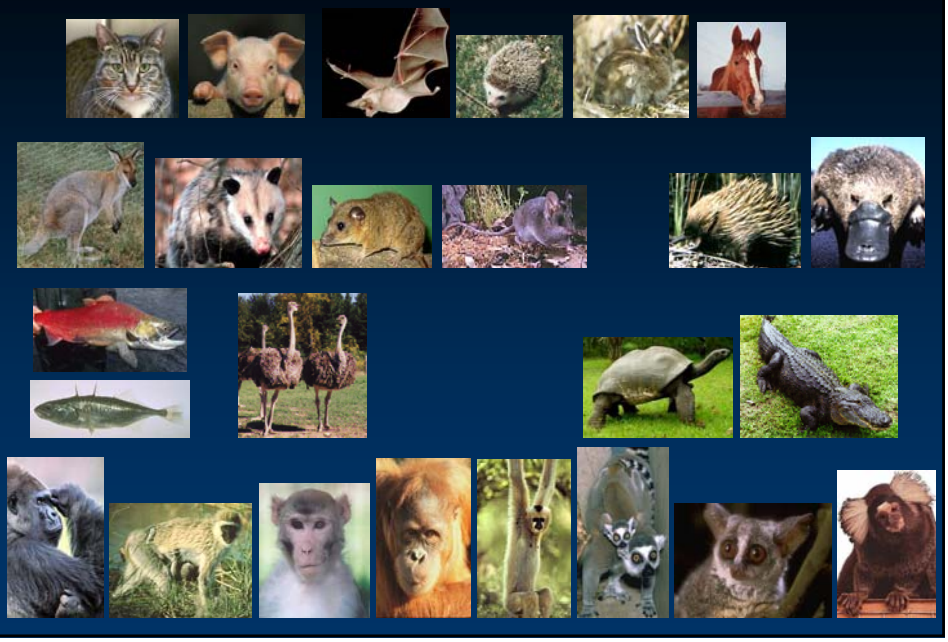

# Detection of MCSs with Different Species

- Investigating the Relative Contribution of Different Species' Sequences to MCS Detection using More Quantitative Approaches
- Re-Compute Conservation Score for All* Possible Subsets of Species
- Compare to a 'Reference Set' of MCSs
  - Generated with *All* Species
  - Surrogates for Conserved *Functional* Elements


Single Species Performance
Reference Set MCS Bases
80,000
60,000
40,000
20,000
0
ALL
Coding
UTR
AR
Not Annotated
Non-Coding
Ancient Repeats
UTRs
Coding


Best Performing Subsets
Reference Set MCSs
80,000
60,000
40,000
20,000
0
16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
Dog, Cow, Mouse, Rat, & Chicken
Mouse, Rat, & Chicken
Platypus Rodent Cat/Cow
Platypus Rodent
Wallaby

## More Species is Better



## MCS Detection and Sequence Quality

- To date, MCS detection has been with reasonably high-quality sequence
- What quality of sequence is desired for MCS detection — especially provided a set of high-quality reference sequences?

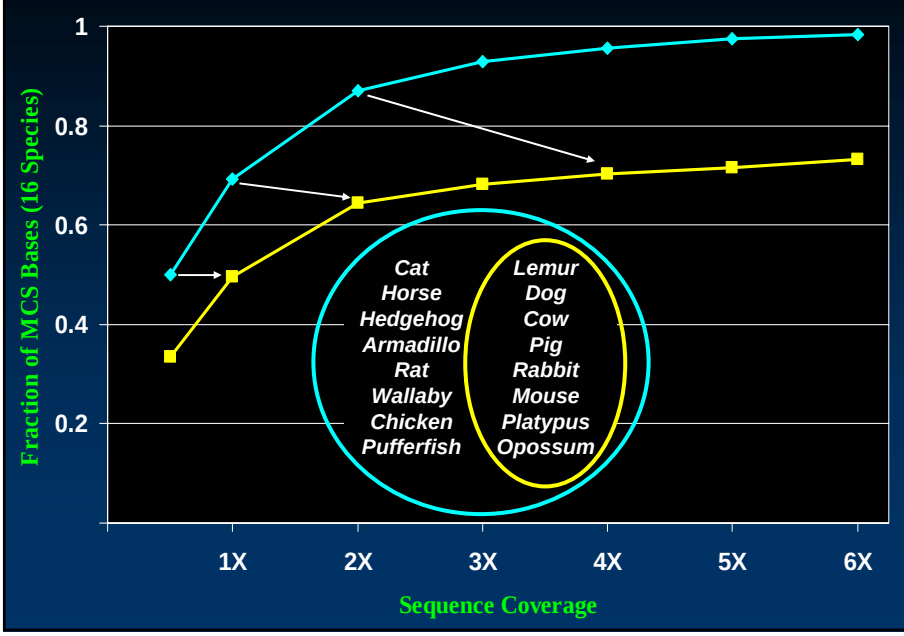
## MCS Detection and Sequence Quality

- What Tradeoffs are encountered between sequence coverage *vs.* number of species?

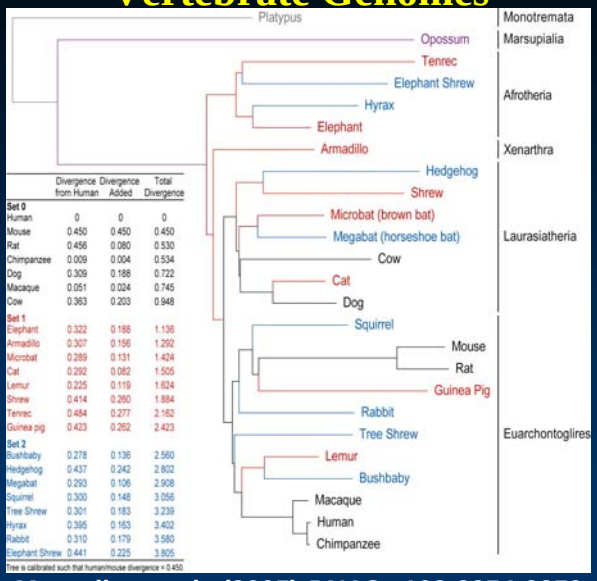
***Our Data Set Provides a Unique Opportunity to Explore these Questions***

1) Re-create 0.5X, 1X, 2X… Read-Coverage Datasets
2) Analyze for MCSs
3) Compare to "Finished" MCSs

## Sequence Coverage vs. MCS Detection



# Low-Redundancy Sequencing of Multiple Vertebrate Genomes




| | Divergence from Human | Divergence Added | Total Divergence |
|---|---|---|---|
| **Set 0** | | | |
| Human | 0 | 0 | 0 |
| Mouse | 0.450 | 0.450 | 0.450 |
| Rat | 0.456 | 0.080 | 0.530 |
| Chimpanzee | 0.009 | 0.004 | 0.534 |
| Dog | 0.309 | 0.188 | 0.722 |
| Macaque | 0.051 | 0.024 | 0.745 |
| Cow | 0.363 | 0.203 | 0.948 |
| **Set 1** | | | |
| Elephant | 0.322 | 0.188 | 1.136 |
| Armadillo | 0.307 | 0.156 | 1.292 |
| Microbat | 0.289 | 0.131 | 1.424 |
| Cat | 0.292 | 0.082 | 1.505 |
| Lemur | 0.225 | 0.119 | 1.624 |
| Shrew | 0.414 | 0.260 | 1.884 |
| Tenrec | 0.484 | 0.277 | 2.162 |
| Guinea pig | 0.423 | 0.262 | 2.423 |
| **Set 2** | | | |
| Bushbaby | 0.278 | 0.136 | 2.560 |
| Hedgehog | 0.437 | 0.242 | 2.802 |
| Megabat | 0.293 | 0.106 | 2.908 |
| Squirrel | 0.300 | 0.148 | 3.056 |
| Tree Shrew | 0.301 | 0.183 | 3.239 |
| Hyrax | 0.395 | 0.163 | 3.402 |
| Rabbit | 0.310 | 0.179 | 3.580 |
| Elephant Shrew | 0.441 | 0.225 | 3.805 |

Tree is calibrated such that human/mouse divergence = 0.450

Margulies et al., (2005) *PNAS*, 102:3354-3359