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

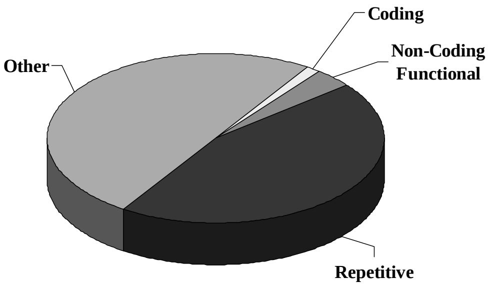
International Human Genome Sequencing Consortium*

* *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 … **nods. 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**

[redacted]

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

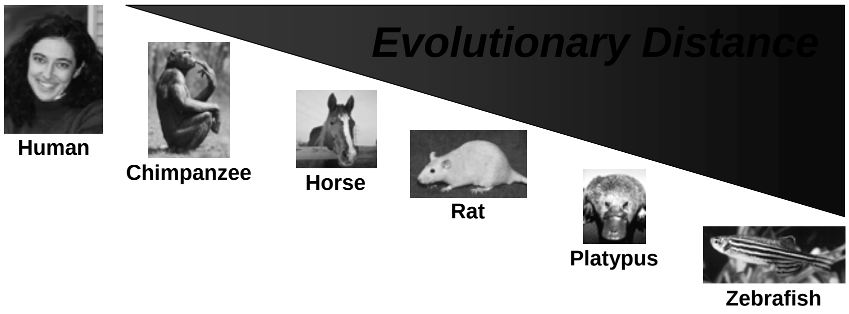
- DNA represents a “blueprint” for structure and physiology of all living things
- All species use DNA
- Mutations in *functional* DNA are less likely to be tolerated

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
```

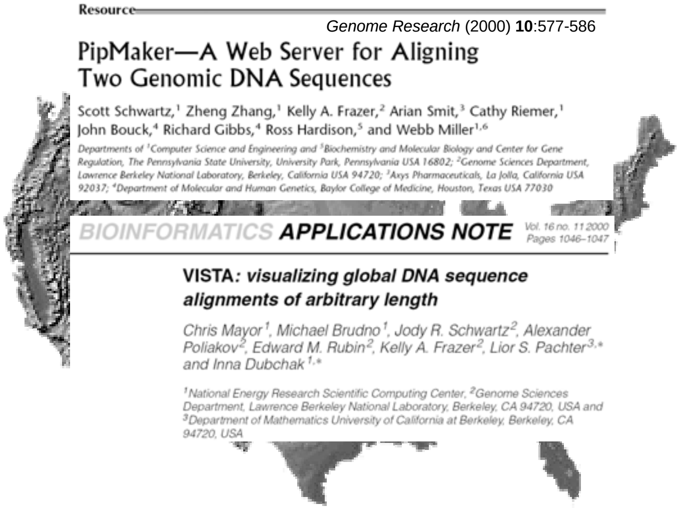
80% Identical

```
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 [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

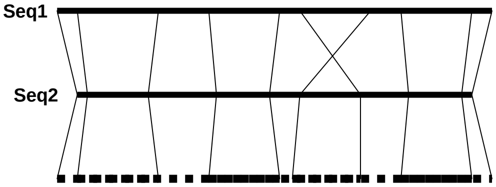
VISTA*: visualizing global DNA sequence alignments of arbitrary length*

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,*]

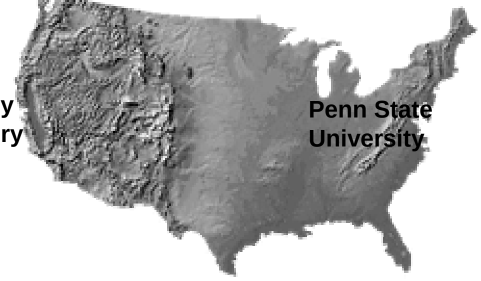
[1]National Energy Research Scientific Computing Center, [2]Genome Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA and [3]Department of Mathematics University of California at Berkeley, Berkeley, CA 94720, USA

PipMaker *vs.* VISTA

- **Visualization**
- **Alignment Strategy**
  - – **VISTA: `avid`**
  - – **PipMaker: `blastz`**



- **East Coast – West Coast**

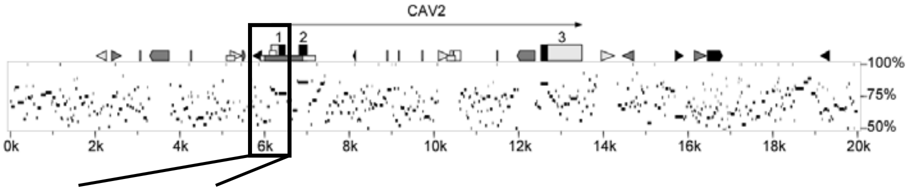


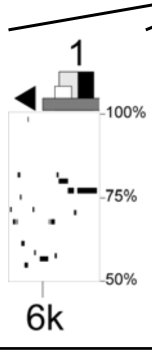
Lawrence Berkeley National Laboratory

Penn State University

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

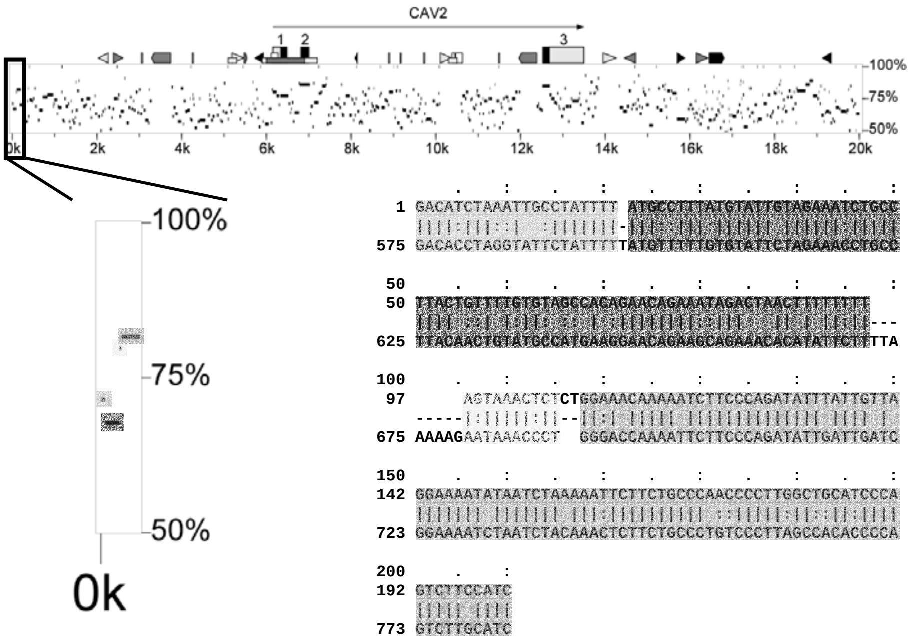
- **P**ercent **I**dentity **P**lot





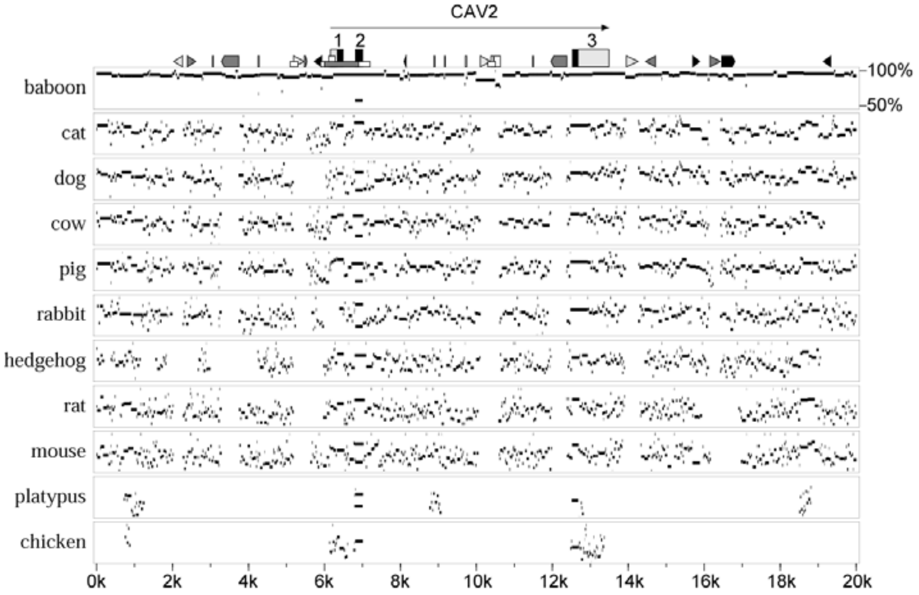
- **X-axis is the reference sequence**
- **Horizontal lines represent gap-free alignments**

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



```
          .    :    .    :    .    :    .    :    .    :
  1 GACATCTAAATTGCCTATTTT ATGCCTTTATGTATTGTAGAAATCTGCC
    ||||:|||::|   :|||||||-|||::|||:|||||| |||||||:|||||
575 GACACCTAGGTATTCTATTTTTATGTTTTTGTGTATTCTAGAAACCTGCC

 50       .    :    .    :    .    :    .    :    .    :
 50 TTACTGTTTTGTGTAGCCACAGAACAGAAATAGACTAACTTTTTTTT
    ||||  ::| |:||: :: | :|||||||||::|||   || | ||:||---
625 TTACAACTGTATGCCATGAAGGAACAGAAGCAGAAACACATATTCTTTTA

100       .    :    .    :    .    :    .    :    .    :
 97      AGTAAACTCTCTGGAAACAAAAATCTTCCCAGATATTTATTGTTA
    -----|:|||||:||--||:| ||||| ||||||||||||||| |||| |
675 AAAAGAATAAACCCT  GGGACCAAAATTCTTCCCAGATATTGATTGATC

150       .    :    .    :    .    :    .    :    .    :
142 GGAAAATATAATCTAAAAATTCTTCTGCCCAACCCCTTGGCTGCATCCCA
    ||||||| ||||||| |||:|||||||||| ::|||||:||::||:||||
723 GGAAAATCTAATCTACAAACTCTTCTGCCCTGTCCCTTAGCCACACCCCA

200       .    :
192 GTCTTCCATC
    ||||| ||||
773 GTCTTGCATC
```

MultiPipMaker



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

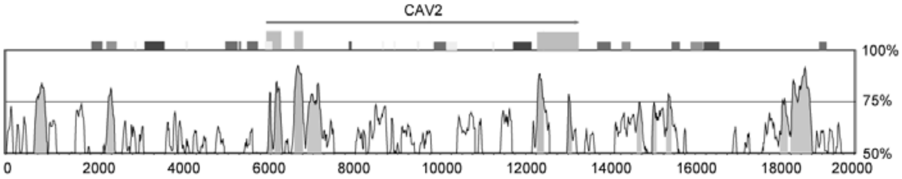
VISUALIZATION TOOLS FOR ALIGNMENTS

- **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**

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

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

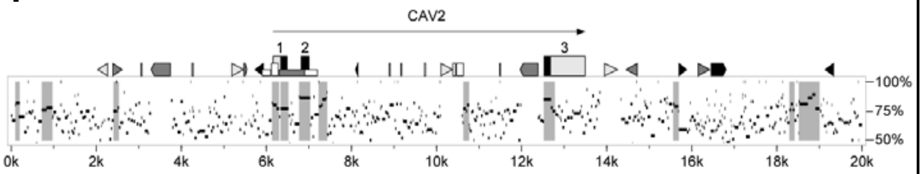
VISTA



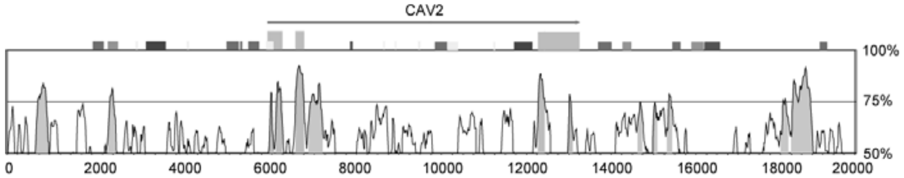
- **Percent Identity is plotted from:**
  - **100 base windows**
  - **Moved every 15 bases**
- **Colored regions meet certain alignment criteria**
  - **>100 bp >75% Identity**

What’s Your Preference?

PipMaker

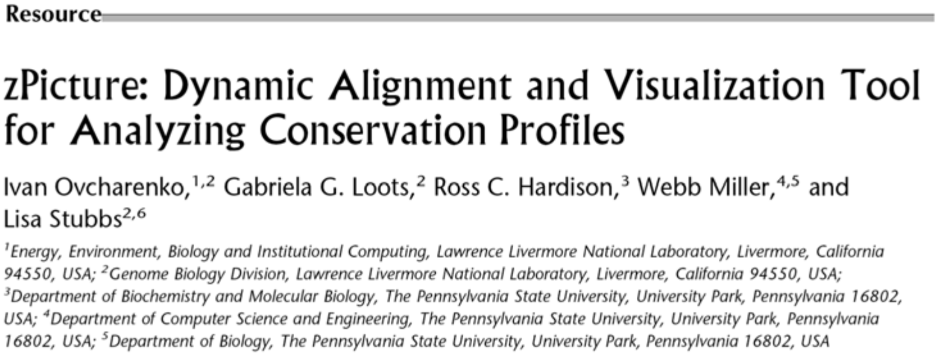


VISTA



East & West Coast Unite

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



Resource

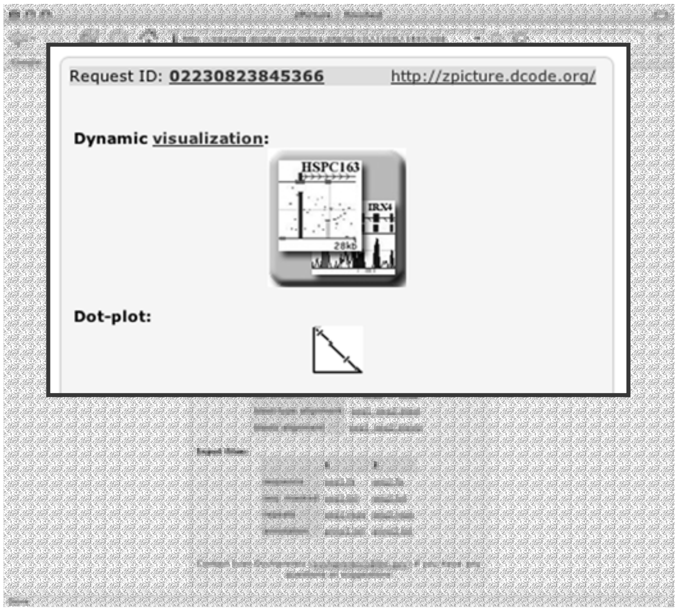
zPicture: Dynamic Alignment and Visualization Tool for Analyzing Conservation Profiles

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

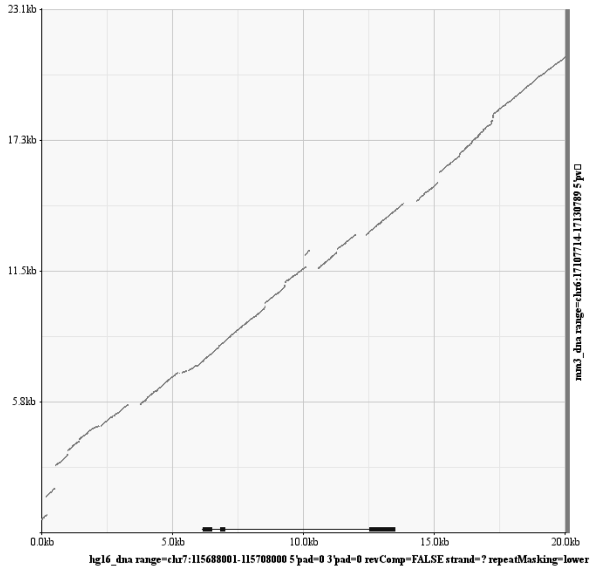


Request ID: 02230823845366 http://zpicture.dcode.org/

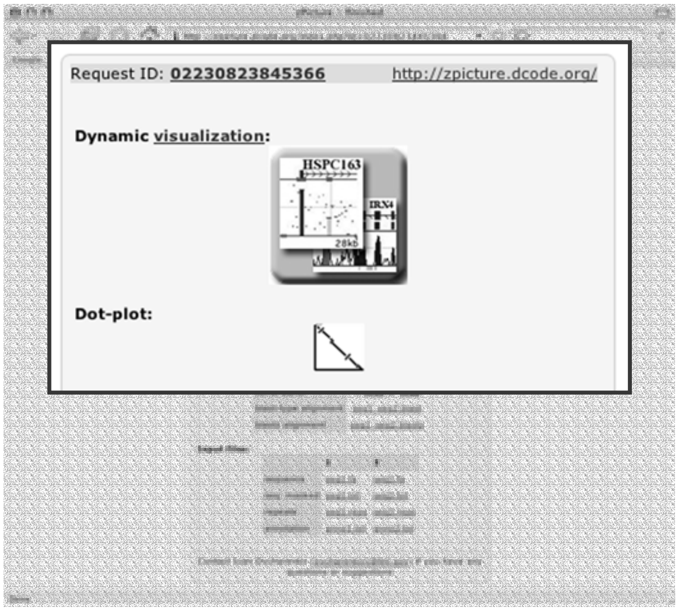
Dynamic visualization:

Dot-plot:

zPicture: Dot Plot View



zPicture Output Summary Page



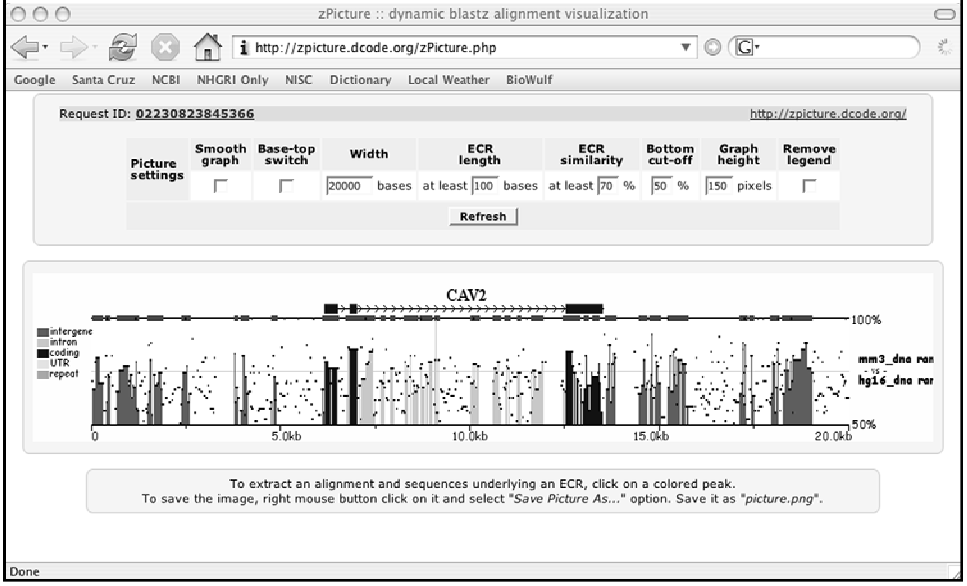
Request ID: 02230823845366 http://zpicture.dcode.org/

Dynamic visualization:

Dot-plot:

Dynamic Visualization Options

PipMaker-style



zPicture :: dynamic blastz alignment visualization

http://zpicture.dcode.org/zPicture.php

Request ID: 02230823845366 http://zpicture.dcode.org/

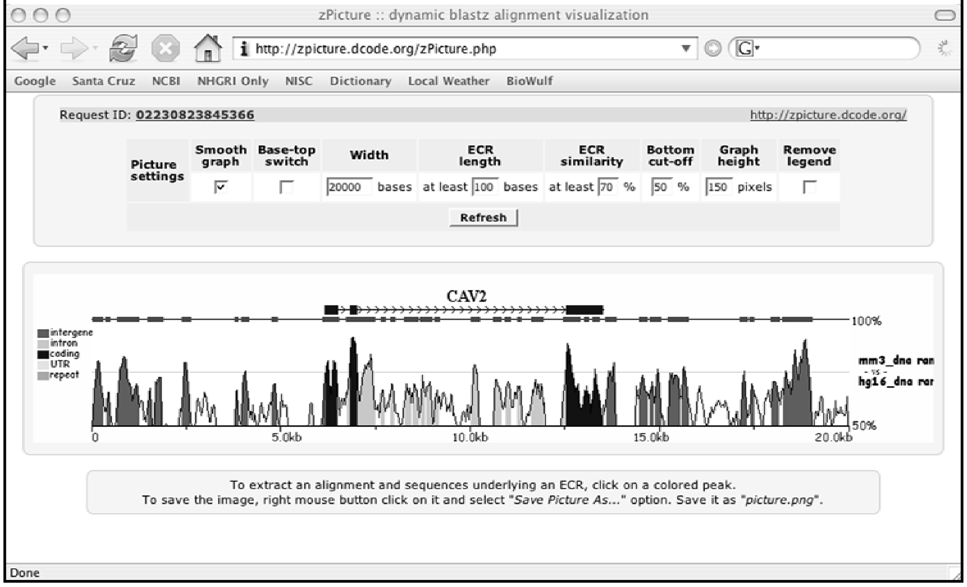
Picture settings	Smooth graph	Base-top switch	Width	ECR length	ECR similarity	Bottom cut-off	Graph height	Remove legend
	□	□	20000 bases	at least 100 bases	at least 70 %	50 %	150 pixels	□

To extract an alignment and sequences underlying an ECR, click on a colored peak.
To save the image, right mouse button click on it and select "*Save Picture As...*" option. Save it as "*picture.png*".

Done

Dynamic Visualization

VISTA-style

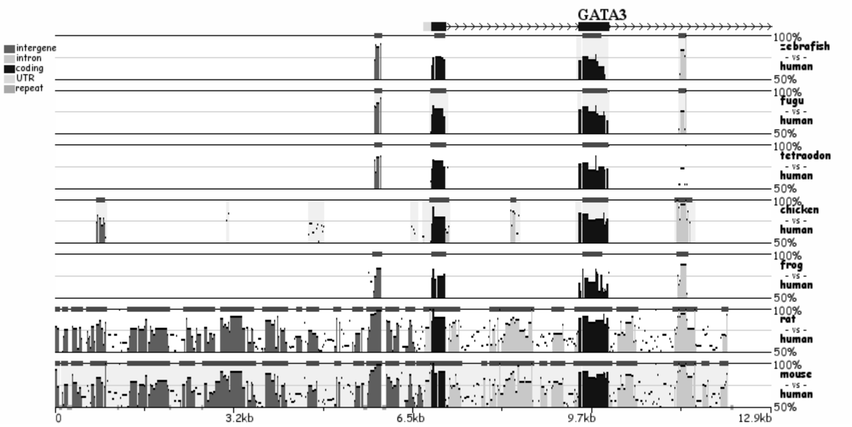


Request ID: 02230823845366 http://zpicture.dcode.org/

Picture settings	Smooth graph	Base-top switch	Width	ECR length	ECR similarity	Bottom cut-off	Graph height	Remove legend
	☑	□	20000 bases	at least 100 bases	at least 70 %	50 %	150 pixels	□

To extract an alignment and sequences underlying an ECR, click on a colored peak.
To save the image, right mouse button click on it and select "*Save Picture As...*" option. Save it as "*picture.png*".

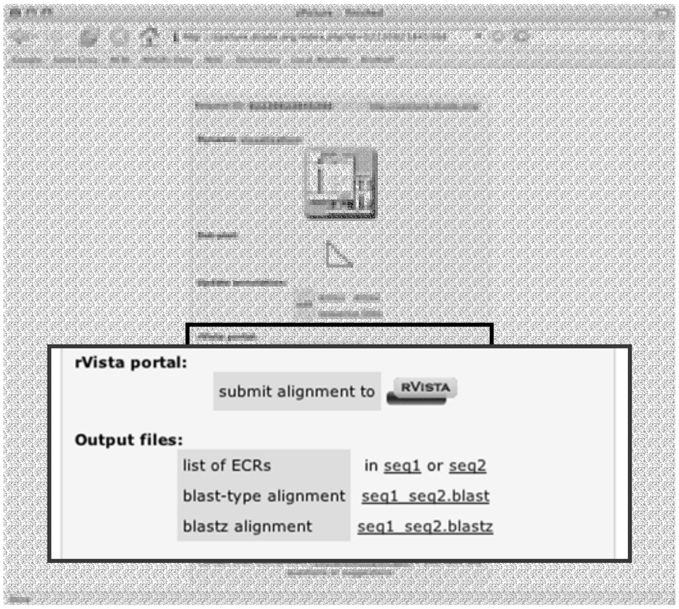
The Multi-Species version of zPicture



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



rVista portal:

submit alignment to

Output files:

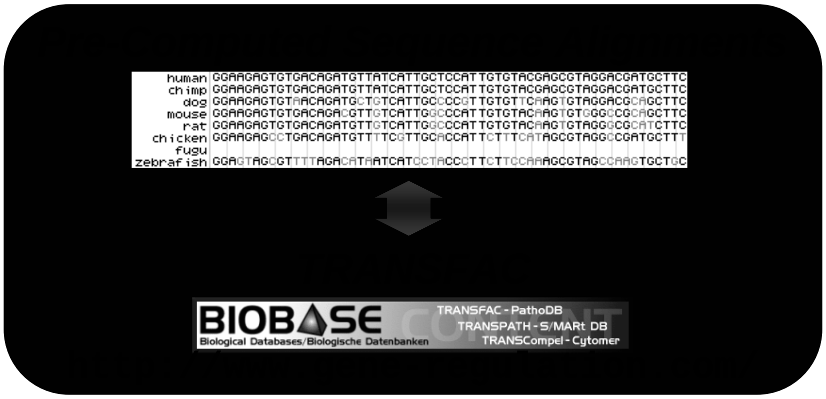
list of ECRs	in seq1 or seq2
blast-type alignment	seq1_seq2.blast
blastz alignment	seq1_seq2.blastz

Are there any transcription factor binding sites in my alignment?

http://www.dcode.org - Comparative Genomics Center at Lawrence Livermore National Laboratory

rVista 2.0

http:// rvista . dcode . org



```
    human GGAAGAGTGTGACAGATGTTATCATTGCTCCATTGTGTACGAGCGTAGGACGATGCTTC
    chimp GGAAGAGTGTGACAGATGTTATCATTGCTCCATTGTGTACGAGCGTAGGACGATGCTTC
      dog GGAAGAGTGTAACAGATGCTGTCATTGCCCCGTTGTGTTCAAGTGTAGGACGCAGCTTC
    mouse GGAAGAGTGTGACAGACGTTGTCATTGGCCCATTGTGTACAAGTGTGGGCCGCAGCTTC
      rat GGAAGAGTGTGACAGATGTTGTCATTGGCCCATTGTGTACAAGTGTAGGGCGCATCTTC
  chicken GGAAGAGCCTGACAGATGTTTTCGTTGCACCATTCTTTCATAGCGTAGGCCGATGCTTT
     fugu
zebrafish GGAGTAGCGTTTTAGACATAATCATCCTACCCTTCTTCCAAAGCGTAGCCAAGTGCTGC
```

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

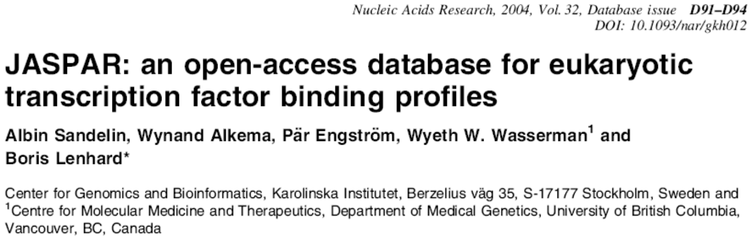
TRANSFAC

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**

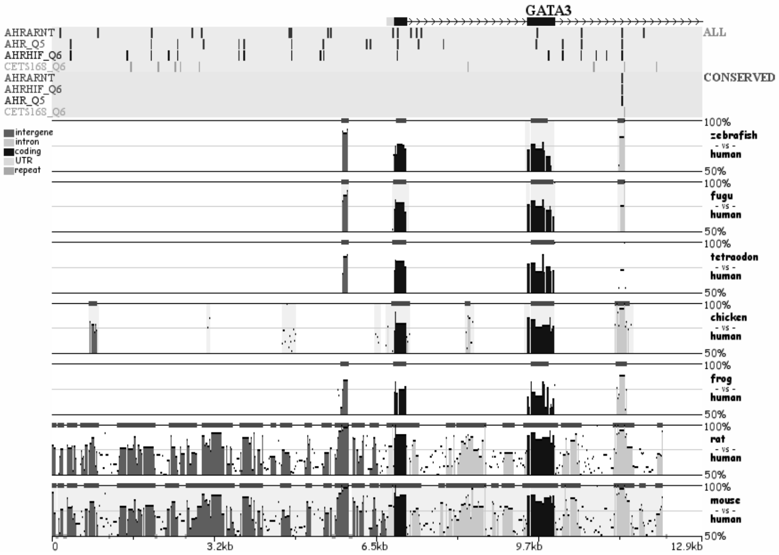
`http://jaspar.cgb.ki.se/`

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***

[redacted]

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

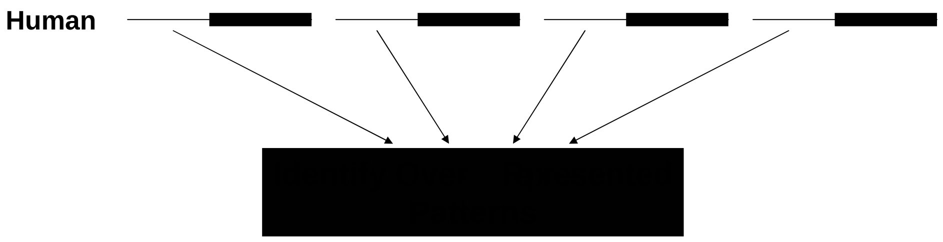
Outline

- 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?

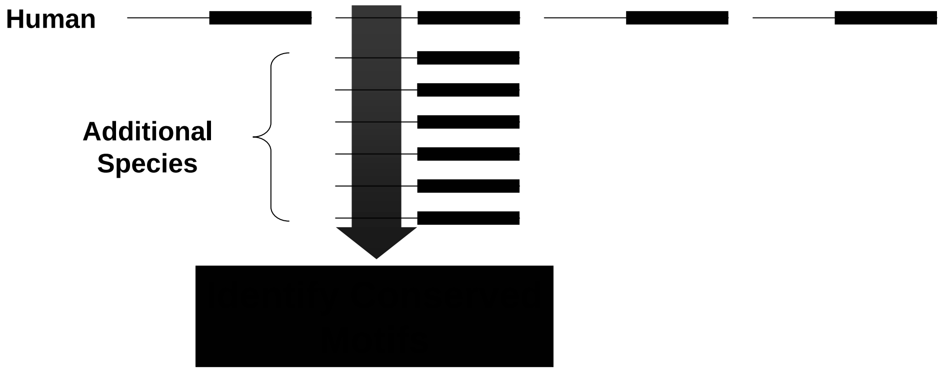
Coordinately Regulated Genes



Phylogenetic Footprinting

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

Orthologous Genes



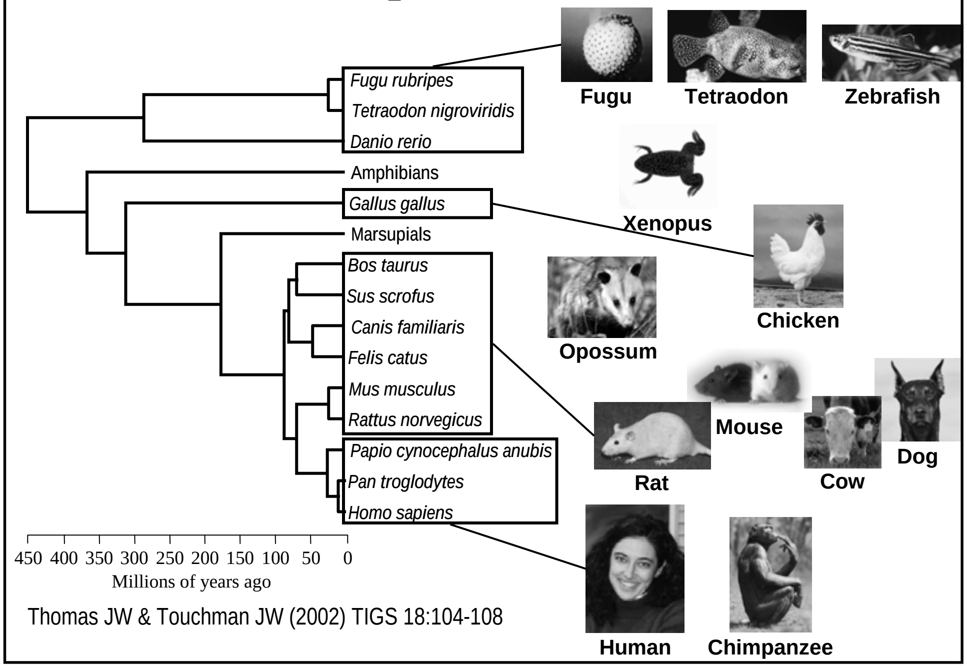
Summary of Phylogenetic Footprinting Tools

- **FootPrinter –** <u>http://bio.cs.washington.edu/software.html</u>
  - 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:**
  - <u>Box 1</u> of Ureta-Vidal et al. (2003) *Nature Reviews Genetics* **4:**251-262
- **Bayesian Approaches (and home of the Gibbs sampler)**
  - <u>http://www.wadsworth.org/resnres/bioinfo/</u>
- **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

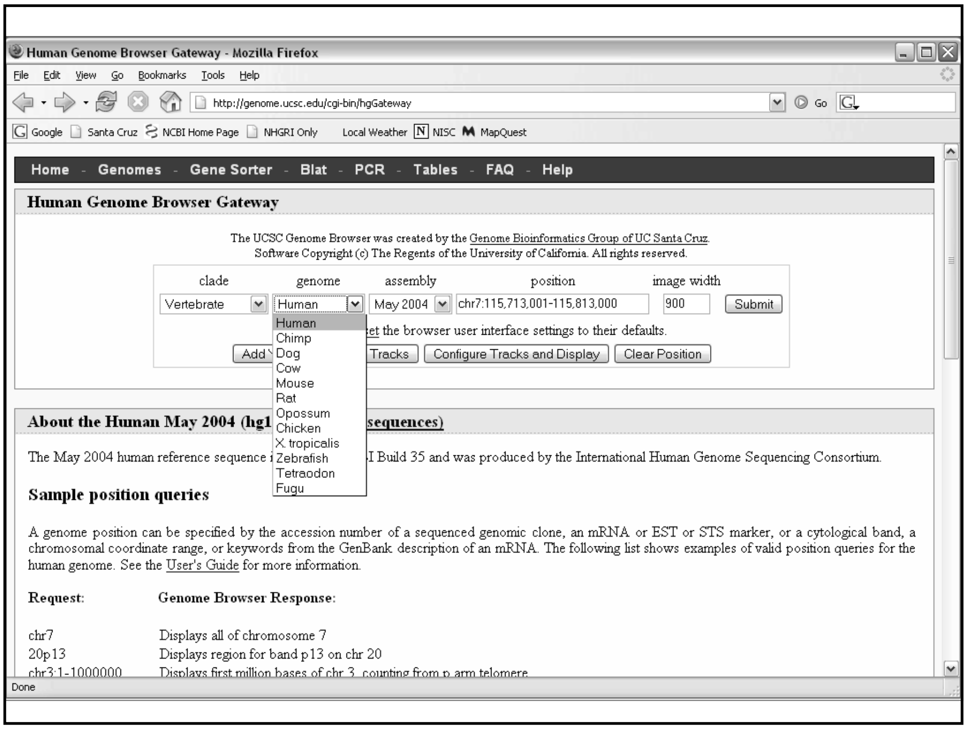
UCSC Genome Bioinformatics

e! project Ensembl

http://www.ensembl.org

NCBI Map Viewer

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


Human Genome Browser Gateway - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://genome.ucsc.edu/cgi-bin/hgGateway
Go
Google Santa Cruz NCBI Home Page NHGRI Only Local Weather NISC MapQuest
Home - Genomes - Gene Sorter - Blat - PCR - Tables - FAQ - Help
Human
Chimp
Dog
Cow
Mouse
Rat
Opossum
Chicken
X. tropicalis
Zebrafish
Tetraodon
Fugu
Submit
Add Y
Tracks
Configure Tracks and Display
Clear Position
Done

Human Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.

clade	genome	assembly	position	image width
Vertebrate	Human	May 2004	chr7:115,713,001-115,813,000	900

et the browser user interface settings to their defaults.

About the Human May 2004 (hg1 sequences)

The May 2004 human reference sequence i I Build 35 and was produced by the International Human Genome Sequencing Consortium.

Sample position queries

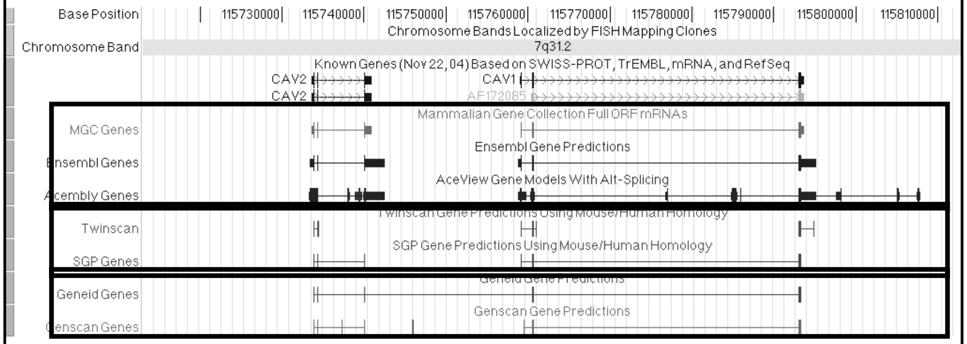
A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, or a cytological band, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the User's Guide for more information.

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
20p13	Displays region for band p13 on chr 20
chr3:1-1000000	Displays first million bases of chr 3, counting from p arm telomere

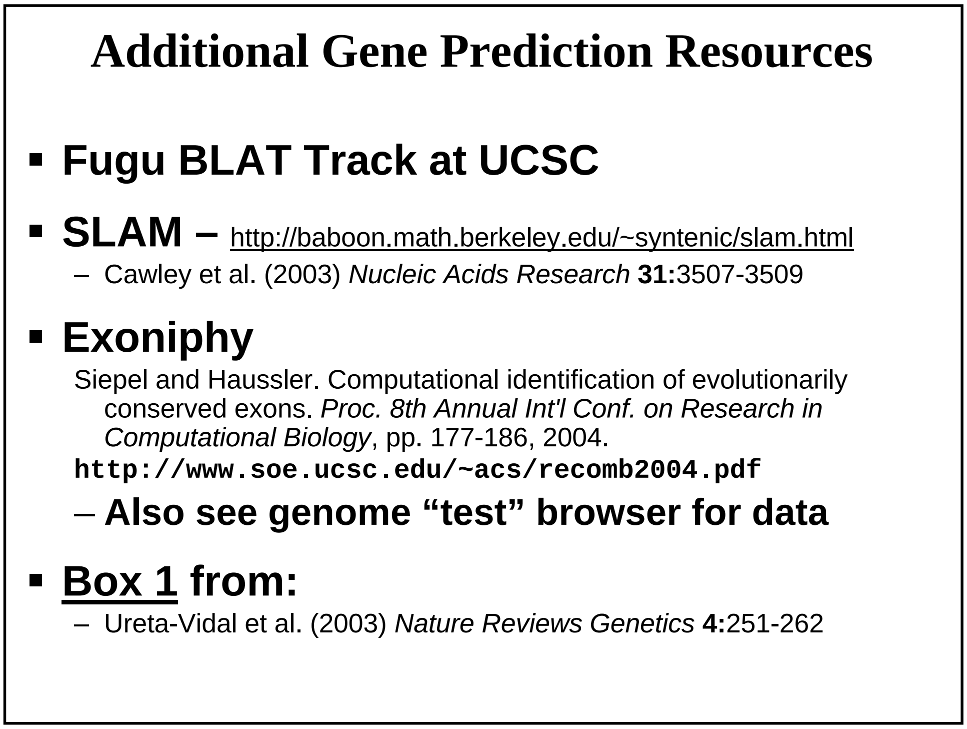
Outline

- 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


Base Position
115730000
115740000
115750000
115760000
115770000
115780000
115790000
115800000
115810000
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band
7q31.2
Known Genes (Nov 22, 04) Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
CAV2
CAV2
CAV1
AF172085
Mammalian Gene Collection Full ORF mRNAs
MGC Genes
Ensembl Gene Predictions
Ensembl Genes
AceView Gene Models With Alt-Splicing
Acembly Genes
Twinscan Gene Predictions Using Mouse/Human Homology
Twinscan
SGP Gene Predictions Using Mouse/Human Homology
SGP Genes
Geneid Gene Predictions
Geneid Genes
Genscan Gene Predictions
Genscan Genes

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



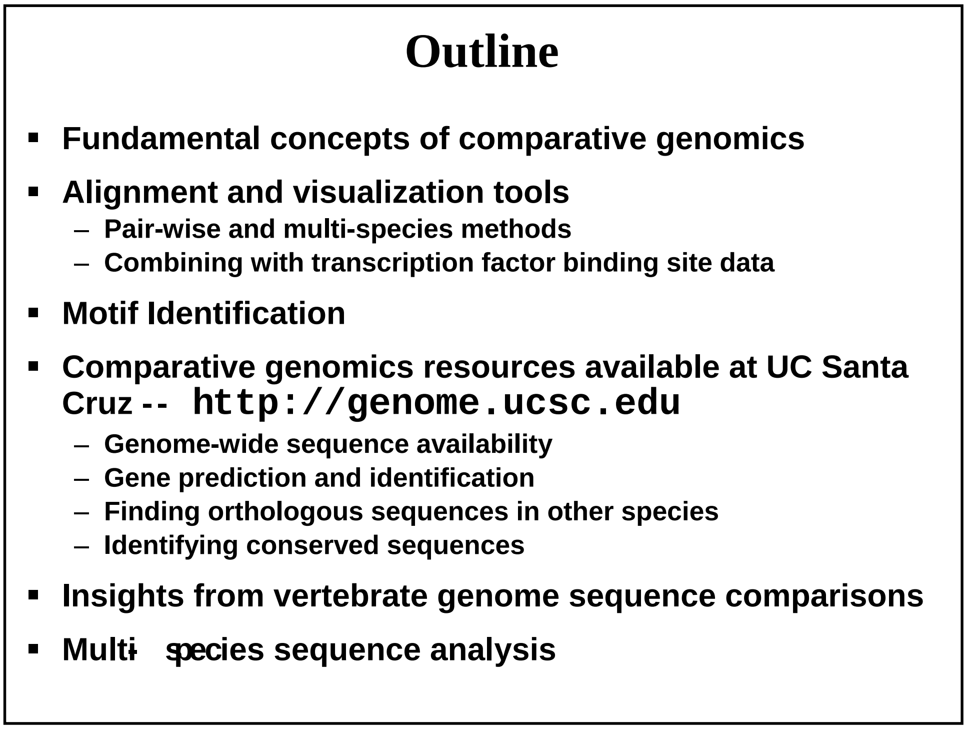
Additional Gene Prediction Resources

- **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:

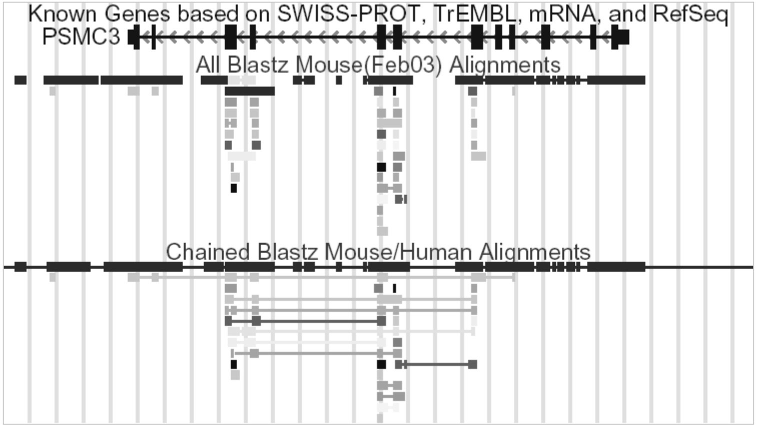
- **Local alignments tend to break at transposon insertions, inversions, duplications, etc.**
- **Global alignments tend to force non-homologous bases to align.**

The Solution:

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

Slide (though modified) Courtesy of Jim Kent

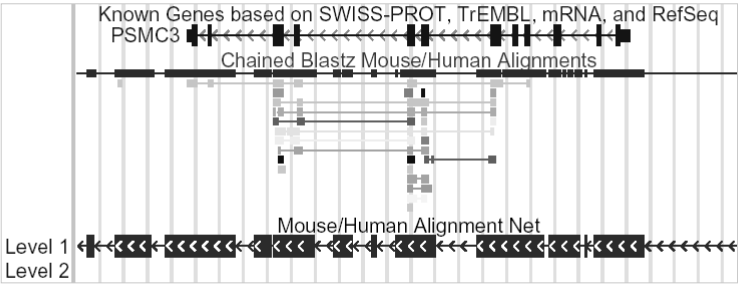
Chains join together related local alignments


Known Genes based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
PSMC3
All Blastz Mouse(Feb03) Alignments
Chained Blastz Mouse/Human Alignments

Protease Regulatory Subunit 3

Slide Courtesy of Jim Kent

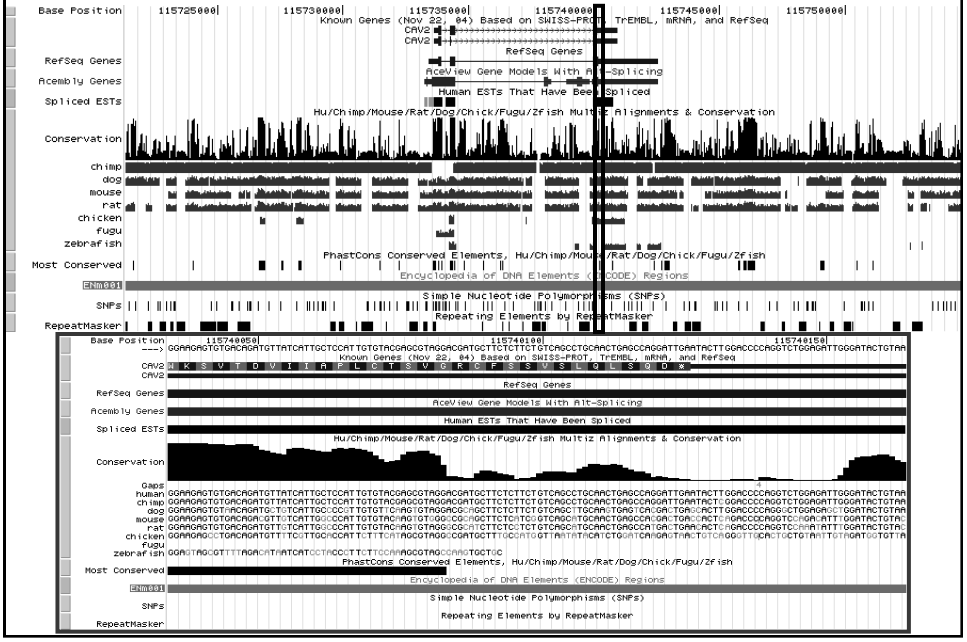
Net Alignments: Focus on Orthology


Known Genes based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
PSMC3
Chained Blastz Mouse/Human Alignments
Mouse/Human Alignment Net
Level 1
Level 2

- 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


Base Position
115725000
115730000
115735000
115740000
115745000
115750000
Known Genes (Nov 22, 04) Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
CAV2
CAV2
RefSeq Genes
Acembly Genes
AceView Gene Models With Alt-Splicing
Spliced ESTs
Human ESTs That Have Been Spliced
Hu/Chimp/Mouse/Rat/Dog/Chick/Fugu/Zfish Multiz Alignments & Conservation
Conservation
chimp
dog
mouse
rat
chicken
fugu
zebrafish
PhastCons Conserved Elements, Hu/Chimp/Mouse/Rat/Dog/Chick/Fugu/Zfish
Most Conserved
Encyclopedia of DNA Elements (ENCODE) Regions
ENm001
Simple Nucleotide Polymorphisms (SNPs)
SNPs
Repeating Elements by RepeatMasker
RepeatMasker
Base Position
115740050
115740100
115740150
Known Genes (Nov 22, 04) Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
CAV2
CAV2
RefSeq Genes
AceView Gene Models With Alt-Splicing
Acembly Genes
Human ESTs That Have Been Spliced
Spliced ESTs
Hu/Chimp/Mouse/Rat/Dog/Chick/Fugu/Zfish Multiz Alignments & Conservation
Conservation
Gaps
human
chimp
dog
mouse
rat
chicken
fugu
zebrafish
PhastCons Conserved Elements, Hu/Chimp/Mouse/Rat/Dog/Chick/Fugu/Zfish
Most Conserved
Encyclopedia of DNA Elements (ENCODE) Regions
ENm001
Simple Nucleotide Polymorphisms (SNPs)
SNPs
Repeating Elements by RepeatMasker
RepeatMasker

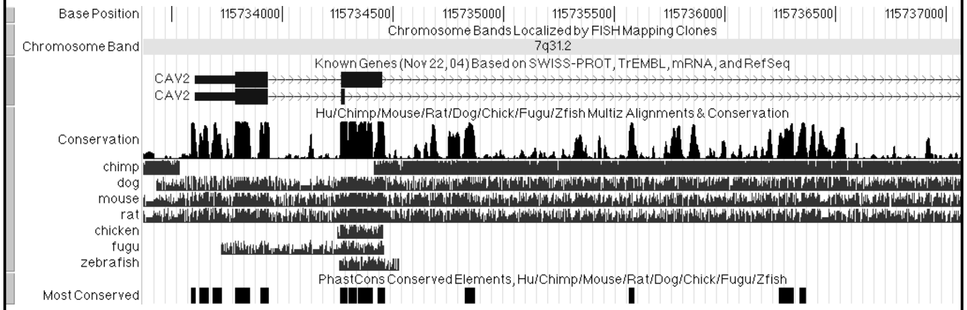
Conservation Score at UCSC

- **Displays evolutionary conservation based on a phylogenetic hidden Markov model**

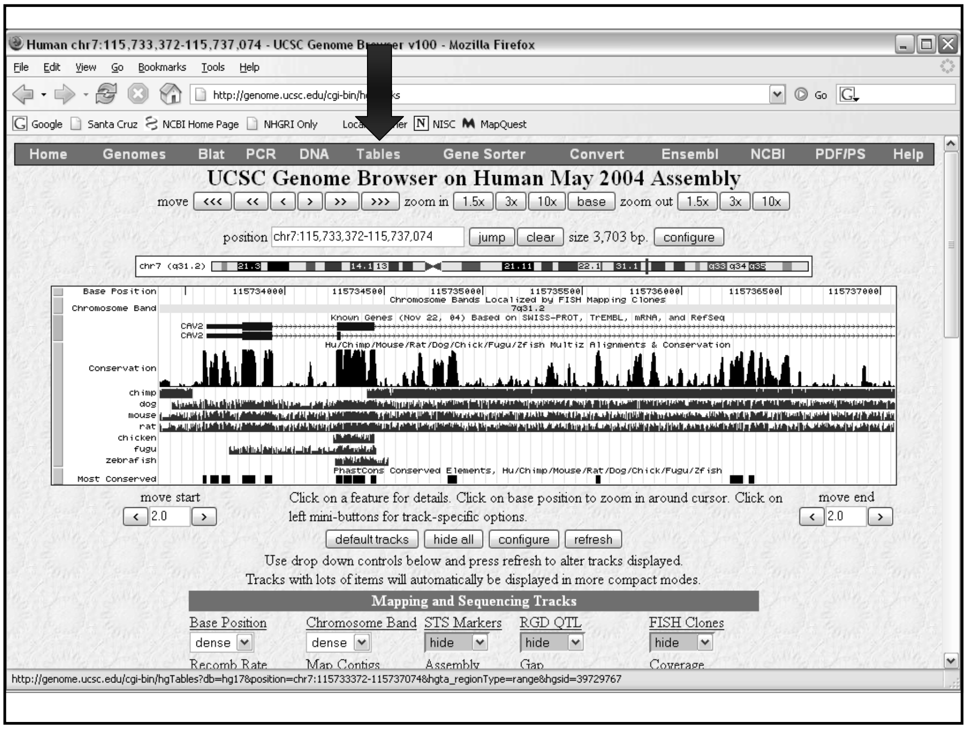
[redacted]

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

“Most Conserved” Track at UCSC


Base Position
115734000
115734500
115735000
115735500
115736000
115736500
115737000
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band
7q31.2
Known Genes (Nov 22, 04) Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
CAV2
CAV2
Hu/Chimp/Mouse/Rat/Dog/Chick/Fugu/Zfish Multiz Alignments & Conservation
Conservation
chimp
dog
mouse
rat
chicken
fugu
zebrafish
PhastCons Conserved Elements, Hu/Chimp/Mouse/Rat/Dog/Chick/Fugu/Zfish
Most Conserved

[redacted]



Human chr7:115,733,372-115,737,074 - UCSC Genome Browser v100 - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://genome.ucsc.edu/cgi-bin/hg...ks
Google Santa Cruz NCBI Home Page NHGRI Only Local Weather NISC MapQuest
Home Genomes Blat PCR DNA Tables Gene Sorter Convert Ensembl NCBI PDF/PS Help
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
jump clear configure
Base Position
Chromosome Band
Known Genes (Nov 22, 04) Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
CAV2
Hu/Chimp/Mouse/Rat/Dog/Chick/Fugu/Zfish Multiz Alignments & Conservation
Conservation
chimp
dog
mouse
rat
chicken
fugu
zebrafish
PhastCons Conserved Elements, Hu/Chimp/Mouse/Rat/Dog/Chick/Fugu/Zfish
Most Conserved
move start
move end
default tracks hide all configure refresh
Mapping and Sequencing Tracks
Base Position
Chromosome Band
STS Markers
RGD QTL
FISH Clones
Recomb Rate
Map Contigs
Assembly
Gap
Coverage

UCSC Genome Browser on Human May 2004 Assembly

position chr7:115,733,372-115,737,074 size 3,703 bp.

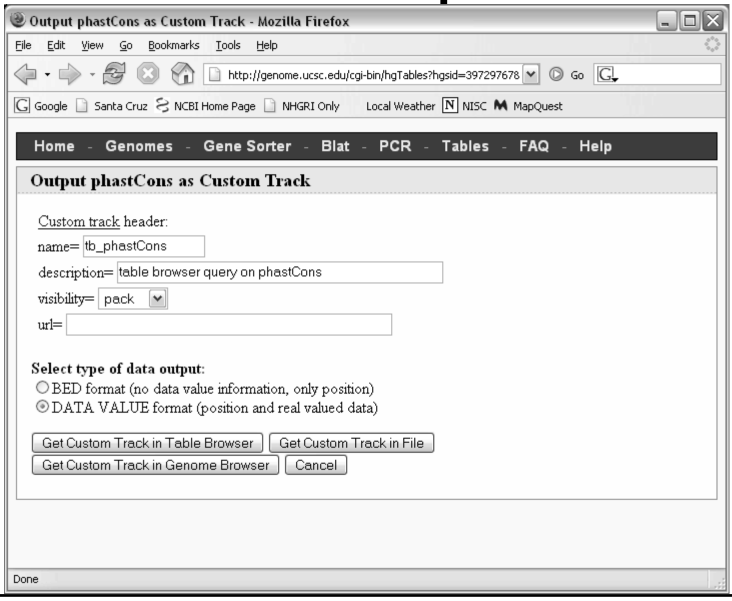
Click on a feature for details. Click on base position to zoom in around cursor. Click on left mini-buttons for track-specific options.

Use drop down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

Base Position: dense; Chromosome Band: dense; STS Markers: hide; RGD QTL: hide; FISH Clones: hide

http://genome.ucsc.edu/cgi-bin/hgTables?db=hg17&position=chr7:115733372-115737074&hgta_regionType=range&hgsid=39729767

Using the Table Browser to get Highly Conserved Sequences



Output phastCons as Custom Track - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://genome.ucsc.edu/cgi-bin/hgTables?hgsid=397297678
Google Santa Cruz NCBI Home Page NHGRI Only Local Weather NISC MapQuest
Home - Genomes - Gene Sorter - Blat - PCR - Tables - FAQ - Help
Get Custom Track in Table Browser
Get Custom Track in File
Get Custom Track in Genome Browser
Cancel
Done

Output phastCons as Custom Track

Custom track header:

name= tb_phastCons

description= table browser query on phastCons

visibility= pack

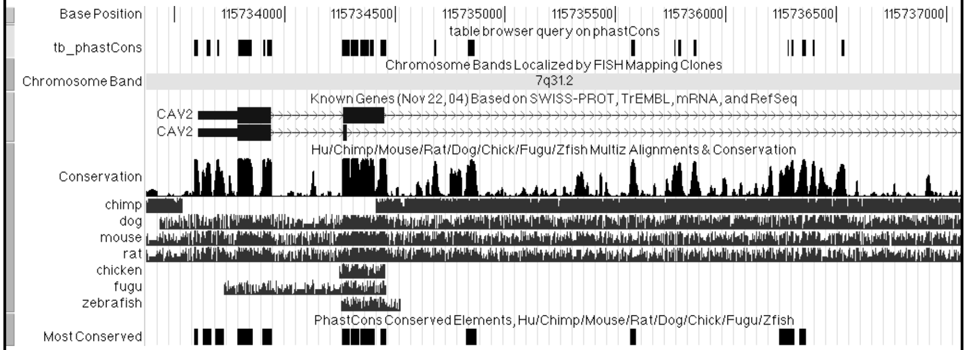
url=

Select type of data output:

○ BED format (no data value information, only position)

● DATA VALUE format (position and real valued data)

Using the Table Browser to get Highly Conserved Sequences


Base Position
115734000
115734500
115735000
115735500
115736000
115736500
115737000
table browser query on phastCons
tb_phastCons
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band
7q31.2
Known Genes (Nov 22, 04) Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
CAV2
CAV2
Hu/Chimp/Mouse/Rat/Dog/Chick/Fugu/Zfish Multiz Alignments & Conservation
Conservation
chimp
dog
mouse
rat
chicken
fugu
zebrafish
PhastCons Conserved Elements, Hu/Chimp/Mouse/Rat/Dog/Chick/Fugu/Zfish
Most Conserved

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

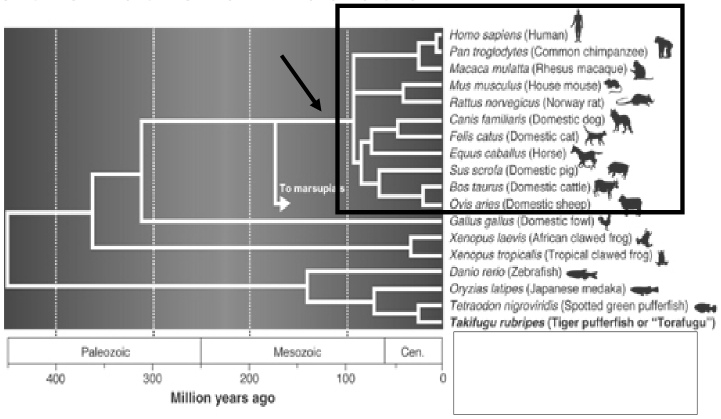
Types of Neutrally Evolving DNA

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

	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

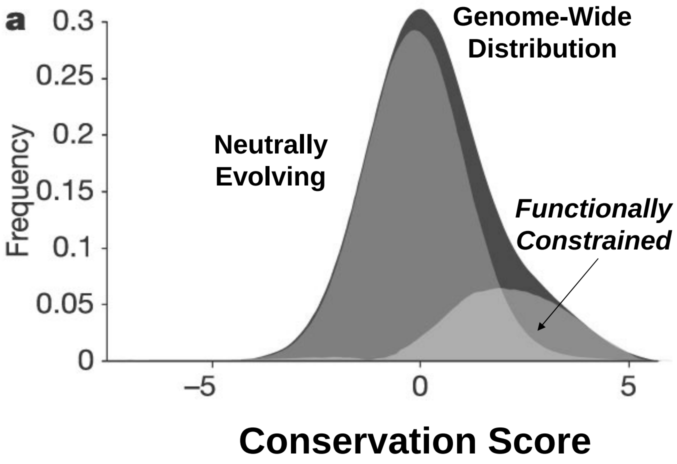
- **Ancestral Repeats**
  - **Ancient Relics of Transposons Inserted Prior to the Eutherian Radiation**


Homo sapiens (Human)
Pan troglodytes (Common chimpanzee)
Macaca mulatta (Rhesus macaque)
Mus musculus (House mouse)
Rattus norvegicus (Norway rat)
Canis familiaris (Domestic dog)
Felis catus (Domestic cat)
Equus caballus (Horse)
Sus scrofa (Domestic pig)
Bos taurus (Domestic cattle)
Ovis aries (Domestic sheep)
Gallus gallus (Domestic fowl)
Xenopus laevis (African clawed frog)
Xenopus tropicalis (Tropical clawed frog)
Danio rerio (Zebrafish)
Oryzias latipes (Japanese medaka)
Tetraodon nigroviridis (Spotted green pufferfish)
Takifugu rubripes (Tiger pufferfish or "Torafugu")
To marsupials
Paleozoic
Mesozoic
Cen.
400
300
200
100
0
Million years ago

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

Determining the Fraction of Sequence Under Purifying Selection

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



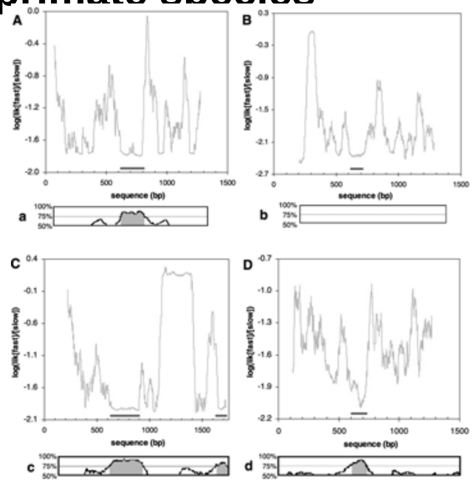
Outline

- 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

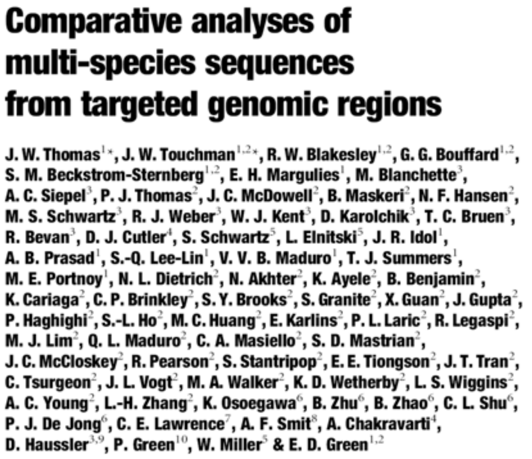
Phylogenetic Shadowing

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

- Identifying sequence *differences* between multiple primate species



Multi-Species Comparative Sequence Analysis

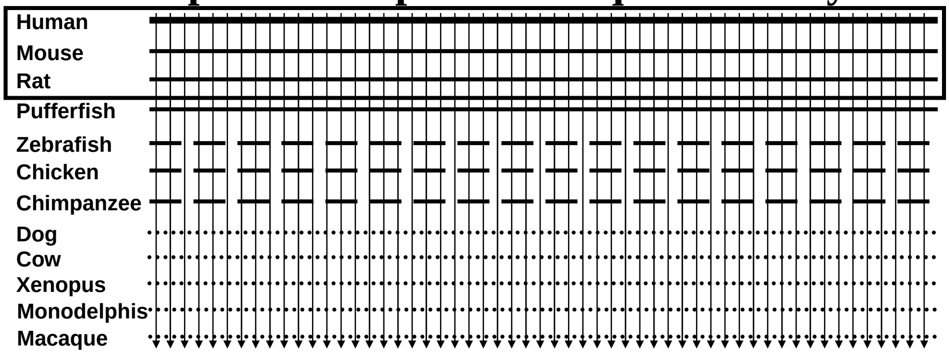


Comparative analyses of multi-species sequences from targeted genomic regions

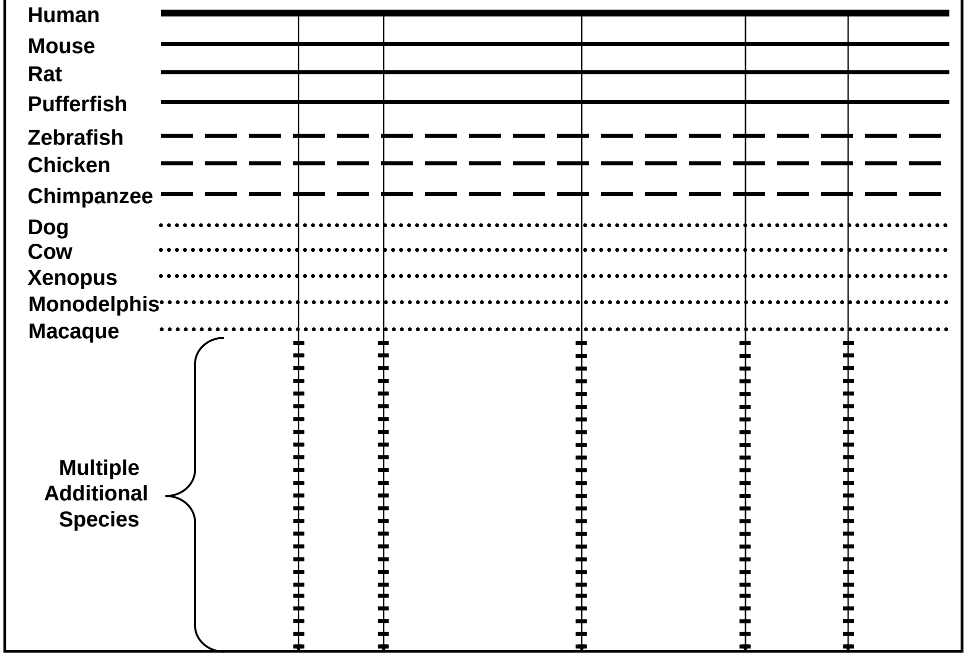
J. W. Thomas[1*], J. W. Touchman[1,2*], R. W. Blakesley[1,2], G. G. Bouffard[1,2], S. M. Beckstrom-Sternberg[1,2], E. H. Margulies[1], M. Blanchette[3], A. C. Siepel[3], P. J. Thomas[2], J. C. McDowell[2], B. Maskeri[2], N. F. Hansen[2], M. S. Schwartz[3], R. J. Weber[3], W. J. Kent[3], D. Karolchik[3],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

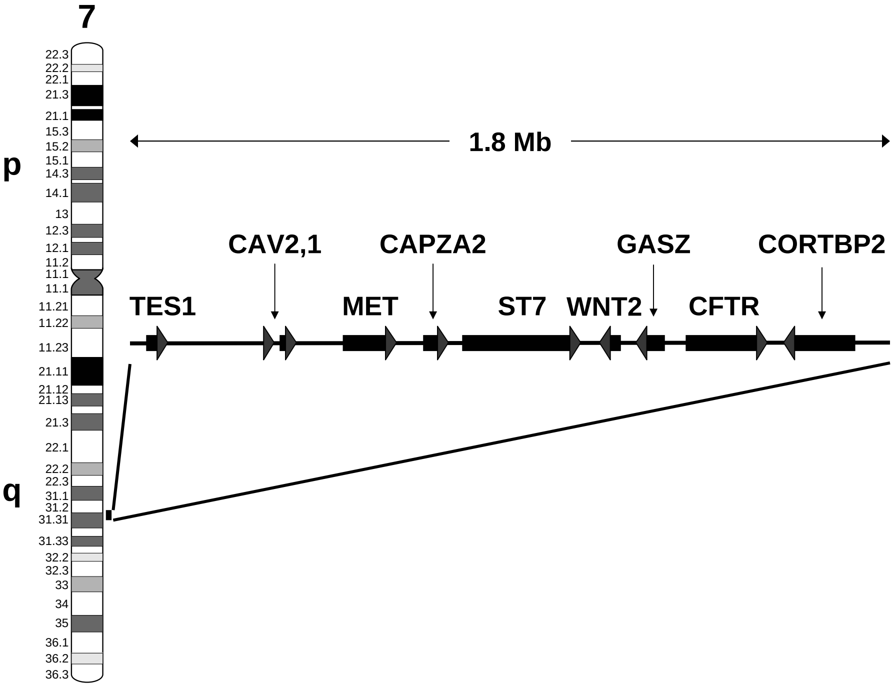
Nature 424:788, 2003

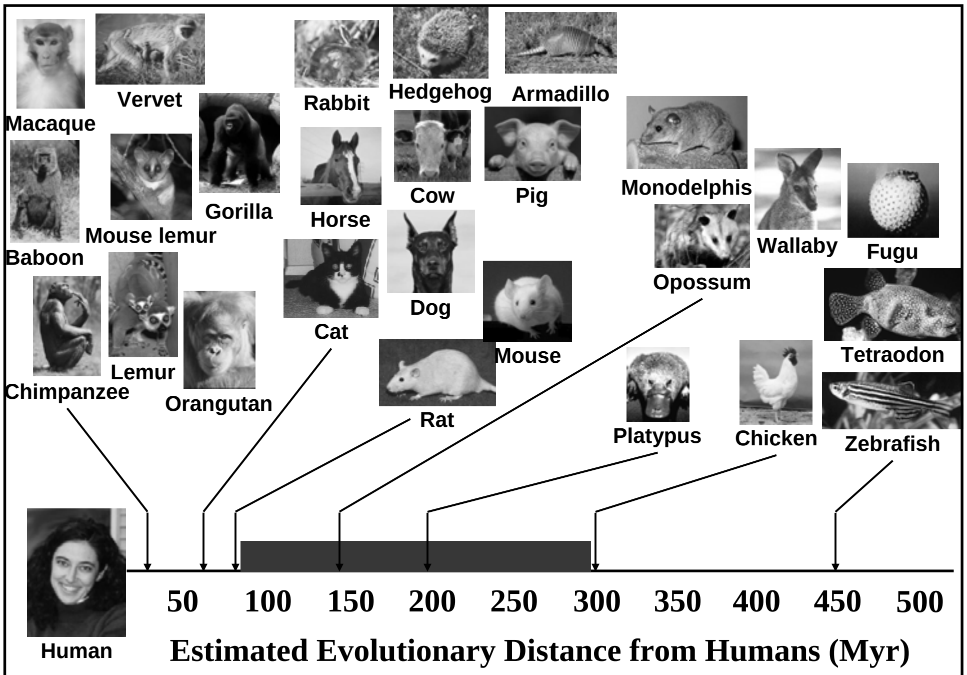
Multi-Species Comparative Sequence Analysis



Multi-Species Comparative Sequence Analysis

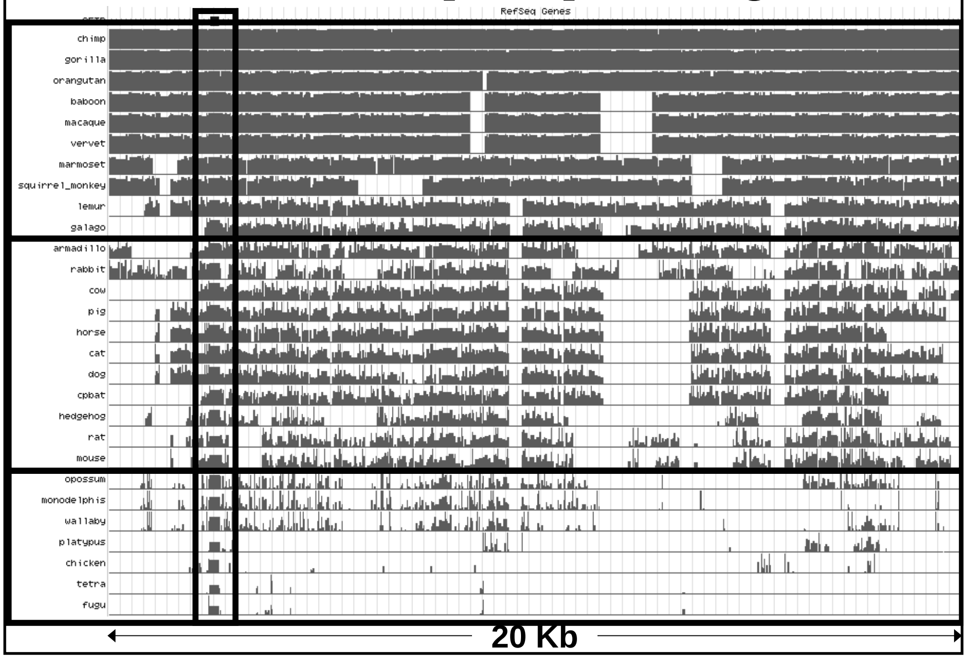






Adapted from Kumar & Hedges, Nature 1998

UCSC View of Multiple Sequence Alignments



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**

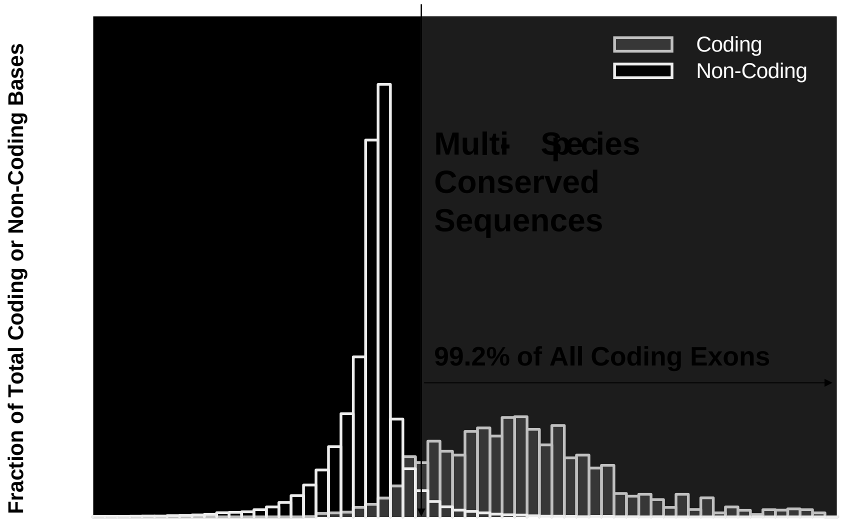
Human	`GCGGGGGCCTTCGGACCGCGCGGCG`
Cat	`iiiiiiiiiiimimiiimiiiimii`
Chicken	`m+miiiiiimimiiim++iiiiiim`
Chimpanzee	`iiiiiiiiiiiiiiiiiiiiiiiii`
Baboon	`iiiiiiiiiiimiiiiiiiiiiiii`
Dog	`iiiiiiiii+++++++++++++iii`
Cow	`immiiiimimmmiiiiiiiiiimii`
Pig	`iiimiiiiiimmimiiiiimiimii`
Rat	`im++++++++mimiiimmiiiimmm`
Mouse	`immiiimii+++miiimmiiiimmm`
Fugu	`-------------------------`
Tetraodon	`-------------------------`
Zebrafish	`-------------------------`

[redacted]

[redacted]

skip

Multi-Species Conservation Score Distribution



Multi-Species Conservation Score

Article

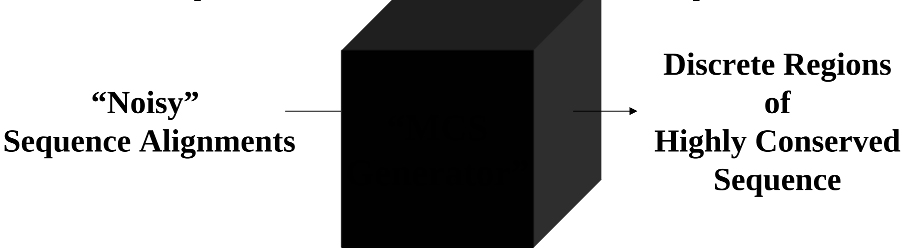
Identification and Characterization of Multi-Species Conserved Sequences

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

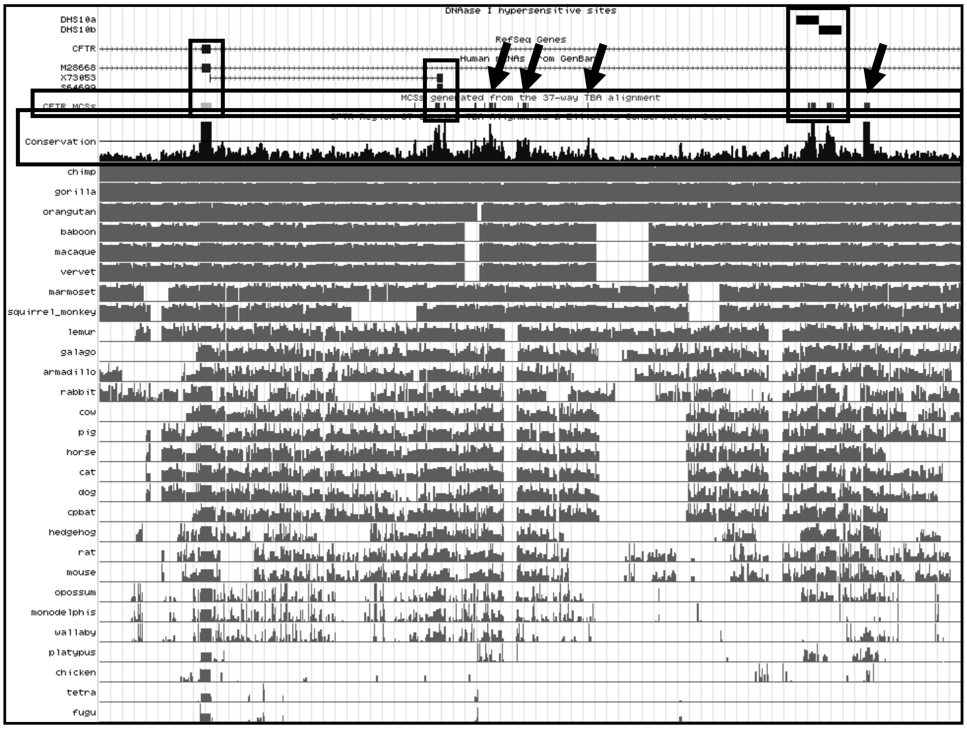
MCS

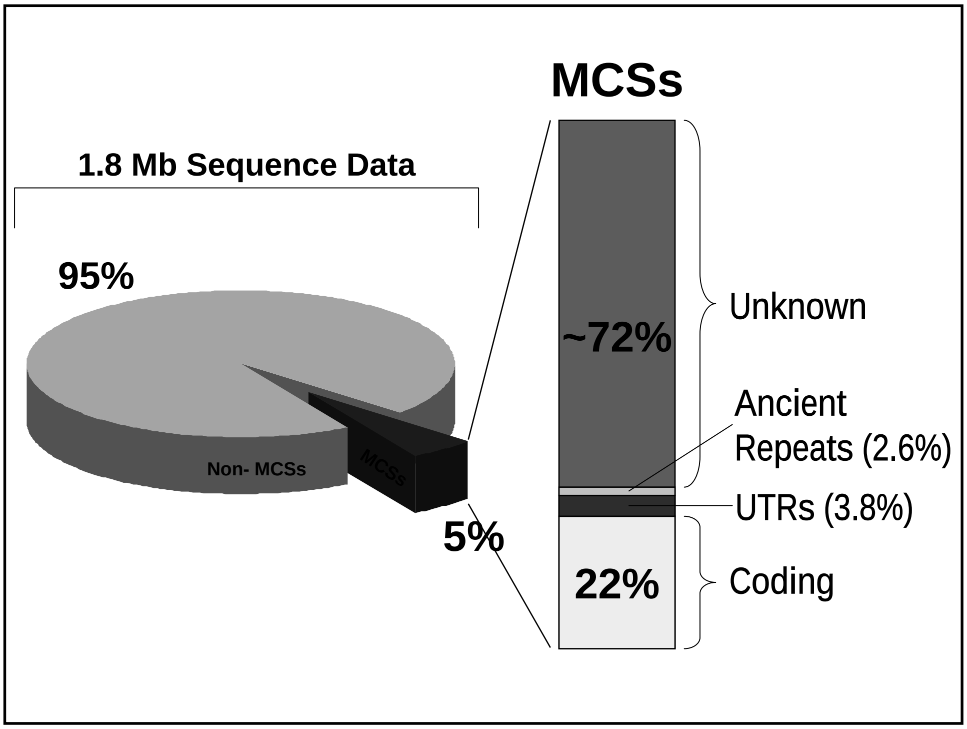
Multispecies Conserved Sequence

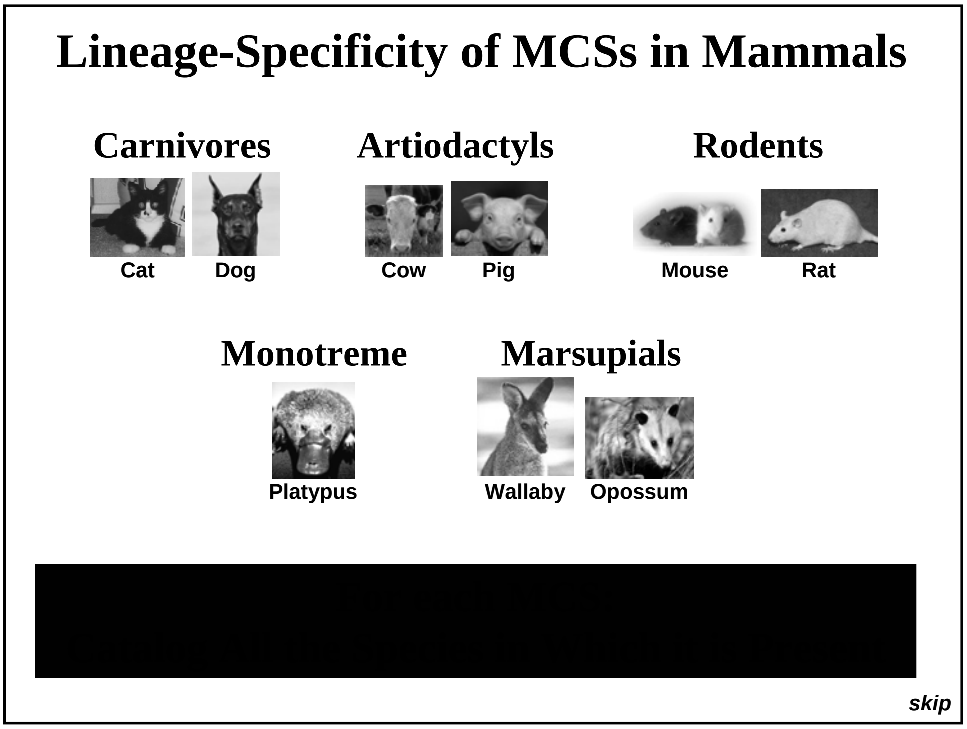


“Noisy”
Sequence Alignments

Discrete Regions
of
Highly Conserved
Sequence


DNase I hypersensitive sites
DHS10a
DHS10b
RefSeq Genes
CFTR
Human mRNAs from GenBank
M28668
X73053
S64699
MCSs generated from the 37-way TBA alignment
CFTR MCSs
CFTR Region 37 Species TBA Alignments & Elliott's Conservation Score
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

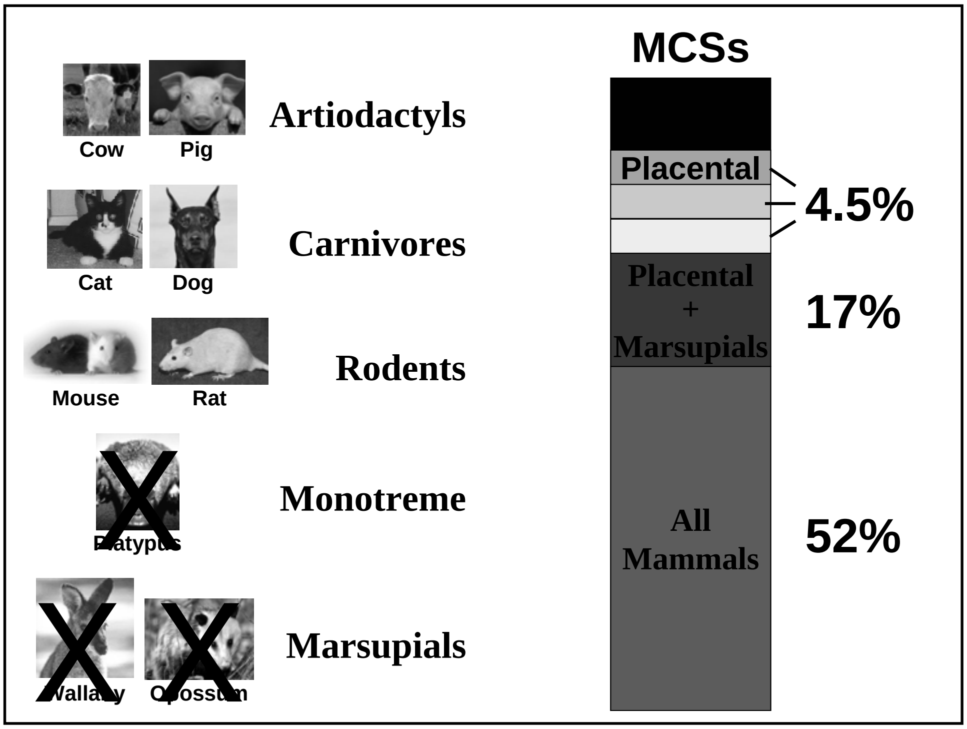

MCSs
1.8 Mb Sequence Data
95%
Non- MCSs
MCSs
5%
~72%
Unknown
Ancient Repeats (2.6%)
UTRs (3.8%)
22%
Coding



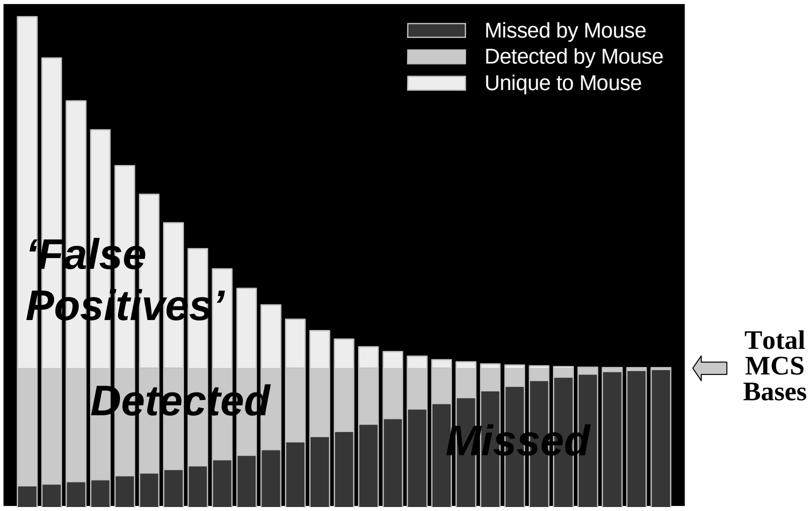
Lineage-Specificity of MCSs in Mammals

[redacted]

skip



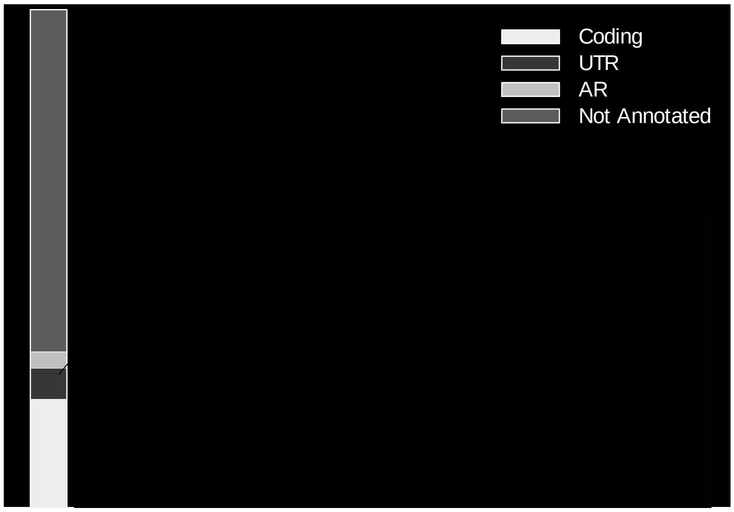
MCS Overlap with Mouse Alignments



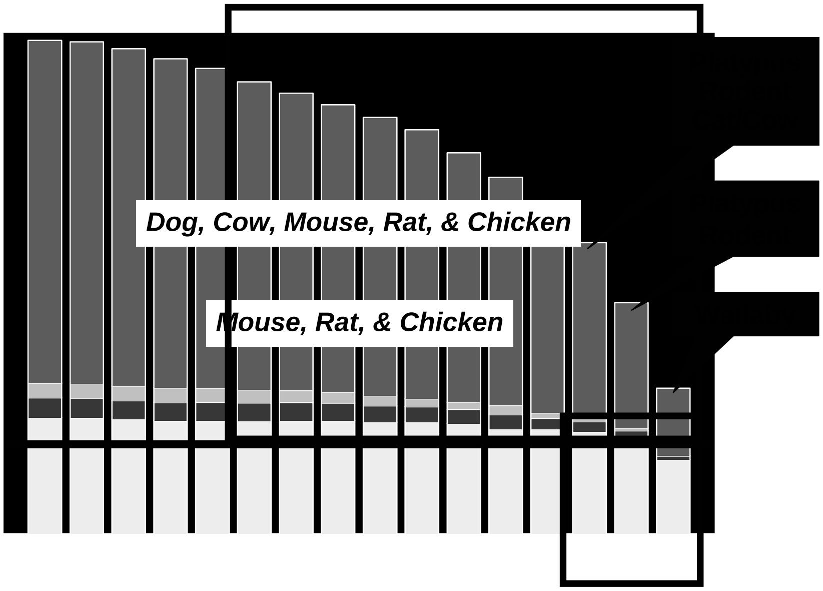
Detection of MCSs with Different Species

- 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

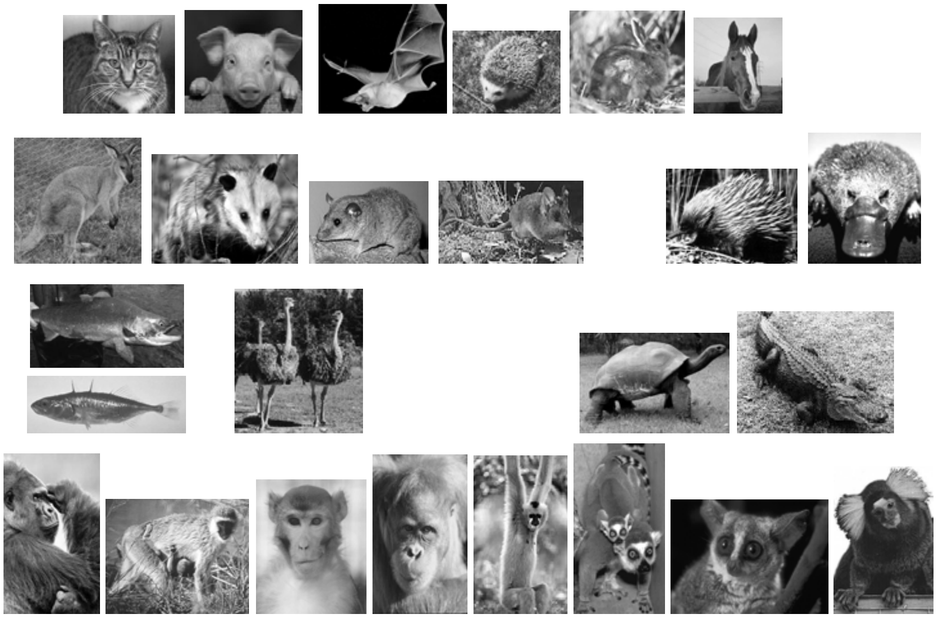
Single Species Performance



Best Performing Subsets



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?

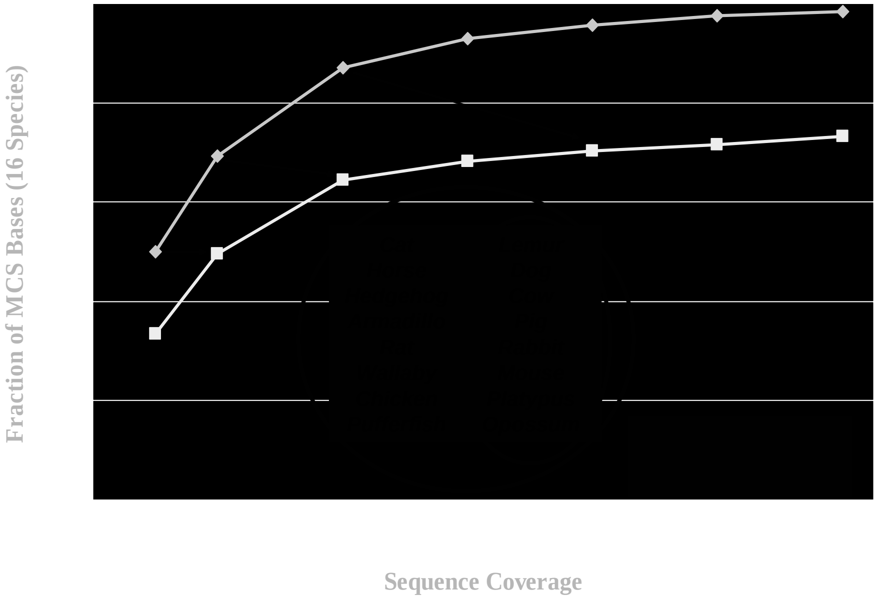
MCS Detection and Sequence Quality

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

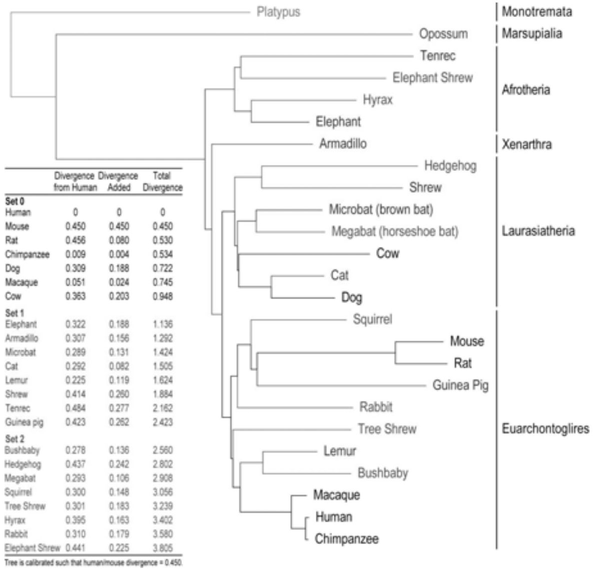
[redacted]

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