NATIONAL HUMAN GENOME RESEARCH INSTITUTE Division of Intramural Research

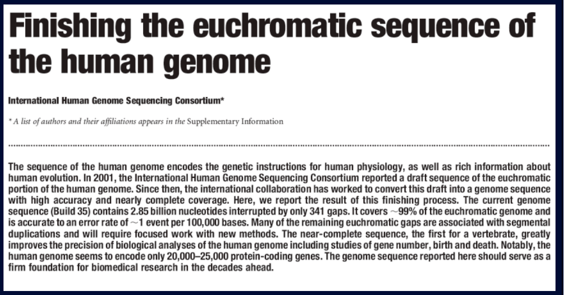
Current Topics in Genome Analysis Fall 2006

Week 5 (Part 2): Detection and Characterization of Non-Coding Functional Elements

Elliott H. Margulies, Ph.D.

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES | NATIONAL INSTITUTES OF HEALTH | genome.gov/DIR

Sequencing Complete



Finishing the euchromatic sequence of the human genome

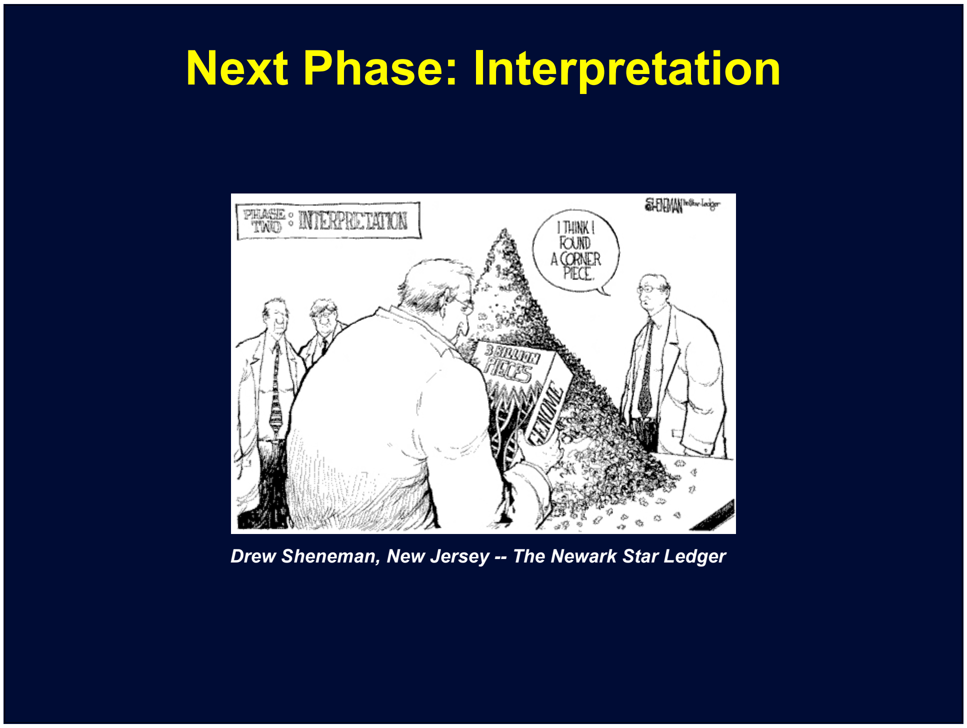
International Human Genome Sequencing Consortium*

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

The sequence of the human genome encodes the genetic instructions for human physiology, as well as rich information about human evolution. In 2001, the International Human Genome Sequencing Consortium reported a draft sequence of the euchromatic portion of the human genome. Since then, the international collaboration has worked to convert this draft into a genome sequence with high accuracy and nearly complete coverage. Here, we report the result of this finishing process. The current genome sequence (Build 35) contains 2.85 billion nucleotides interrupted by only 341 gaps. It covers ~99% of the euchromatic genome and is accurate to an error rate of ~1 event per 100,000 bases. Many of the remaining euchromatic gaps are associated with segmental duplications and will require focused work with new methods. The near-complete sequence, the first for a vertebrate, greatly improves the precision of biological analyses of the human genome including studies of gene number, birth and death. Notably, the human genome seems to encode only 20,000–25,000 protein-coding genes. The genome sequence reported here should serve as a firm foundation for biomedical research in the decades ahead.

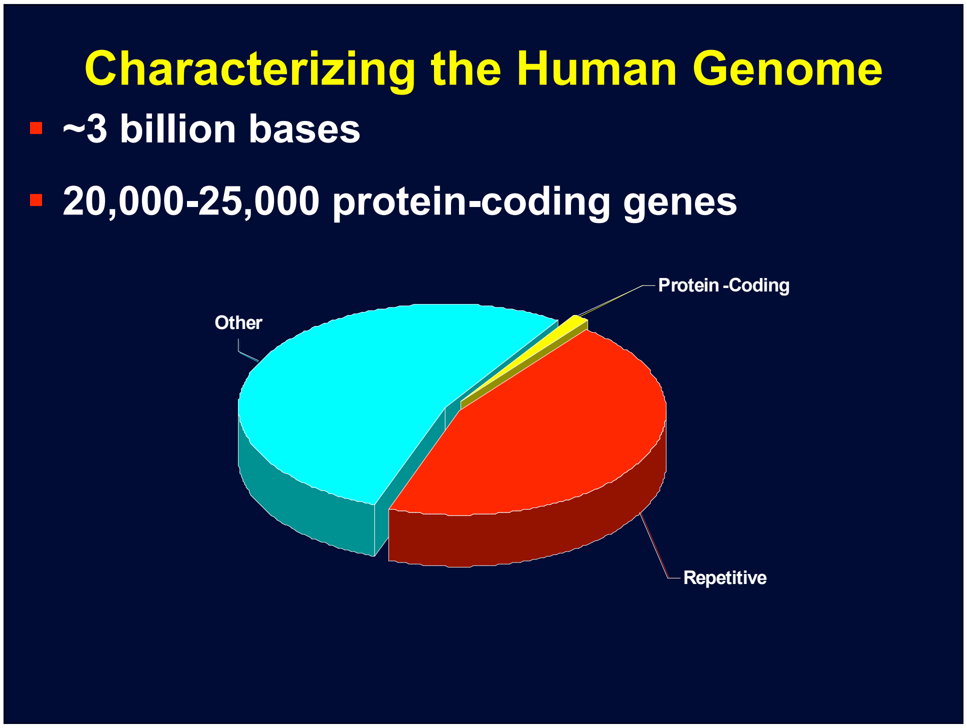
International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431: 931-945.



Next Phase: Interpretation

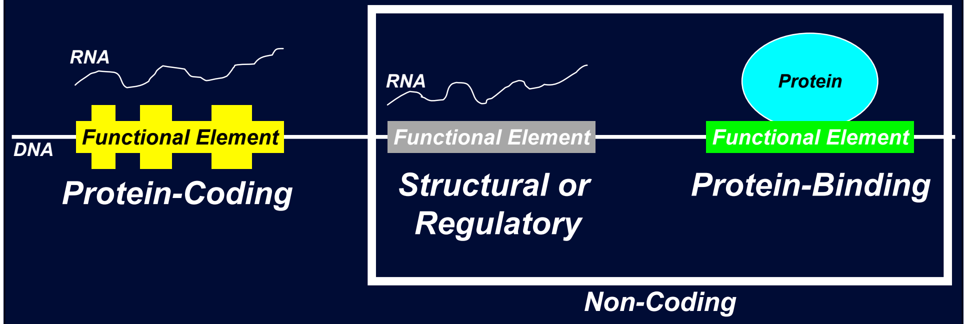
Drew Sheneman, New Jersey -- The Newark Star Ledger



Characterizing the Human Genome

- ~3 billion bases
- 20,000-25,000 protein-coding genes

What are Genomic Functional Elements?



- DNA sequences that either encode for some functioning unit (i.e. RNA) or that bind to proteins that perform some function

Non-coding Functional Elements

- Critical for gene regulation
- Maintain/Modify chromatin structure
- Candidate regions for human disease mutations
- Better understanding of human biology
- Changes in gene regulation rather than gene structure might be more influential in evolution (King & Wilson, 1975)

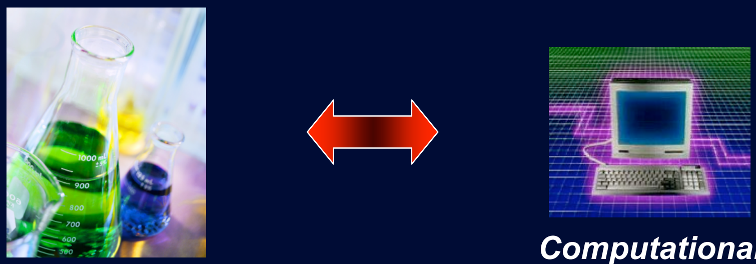
King MC & Wilson AC (1975) Evolution at two levels in humans and chimpanzees. *Science* 188: 107-116

Identifying Functional Elements

- We understand the “language” of coding sequences (i.e., protein-coding genes)
  - Exons and introns
  - Triplet code
  - Complementary datasets (i.e., ESTs, cDNAs)
- The language of non-coding functional elements is poorly understood
  - We don’t know what to look for
  - Signal:Noise problem with short degenerate motifs

Multi-Disciplinary Approaches are Needed

- Find sequences that are likely functional *without prior knowledge of the function*
- Then characterize functions



Experimental Wet-lab Research

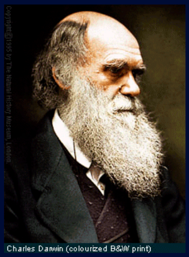
Computational Analyses

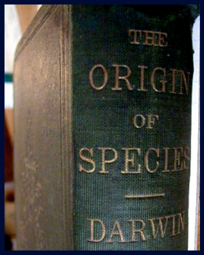
Comparative Genomics to Decode the Genome

```
TGCCGCGGAACTTTTCGGCTCTCTAAGGCTGTATTTTGATATACGAAAGGCACATTTTCCTTCCCTTTTCAAAATGCACCTTGCAAACGTAACAG
GAACCCGACTAGGCANAYOUGFINDAMEGGGAGGAGGAGGAAGGCAGGCTCCGGGGAAGCTGGTGGCAGCGGGTCCTGGGTCTGGCGGACCCTGA
CGCGAAGGAGGGTCTAGGAAGCTCTCCGGGGAGCCGGTTCTCCCGCCGGTGGCTTCTTCTGTCCTCCAGCGTTGCCAACTGGACCTAAAGAGAGG
CCGCGACTGTCGCCCACCTGCGGGATGGGCCTGGTGCTGGGCGGTAAGGACACGGACCTGGAAGGAGCGCGCGCGAGGGAGGGAGGCTGGGAGTC
AGAATCGGGAAAGGGAGGTGCGGGGCGGCGAGGGAGCGAAGGAGGAGAGGAGGAAGGAGCGGGAGGGGTGCTGGCGGGGGTGCGTAGTGGGTGGA
GAAAGCCGCTAGAGCAAATTTGGGGCCGGACCAGGCATHISAISAIMPORTANTGSTUFFGTGAAGGCGGGGGAAAGAGCAAAAGGAAGGGGTGG
TGTGCGGAGTAGGGGTGGGTGGGGGGAATTGGAAGCAAATGACATCACAGCAGGTCAGAGAAAAAGGGTTGAGCGGCAGGCACCCAGAGTAGTAG
GTCTTTGGCATTAGGAGCTTGAGCCCAGACGGCCCTAGCAGGGACCCCAGCGCCCGAGAGACCATGCAGAGGTCGCCTCTGGAAAAGGCCAGCGT
TGTCTCCAAACTTTTTTTCAGGTGAGAAGGTGGCCAACCGAGCTTCSUPERCALAFRAGALISTICEXPEALADOTIOUSAGTATGGGTTGGGTT
TGGGGTAAAGGAATAAGCAGTTTTTAAAAAGATGCGCTATCATTCATTGTTTTGAAAGAAAATGTGGGTATTGTAGAATAAAACAGAAAGCATTA
AGAAGAGATGGAAGAATGAACTGAAGCTGATTGAATAGAGAGCCACATCTACTTGCAACTGAAAAGTTAGAATCTCAAGACTCAAGTACGCTACT
ATGCACTTGTTTTATTTCATTTTTCTAAGAAACTAAAAATACTTGTTAATAAGTACCTAAGTATGGTTTATTGGTTTTCCCCCTTCATGCCTTGG
ACACTTGATTGTCTTCTTGGCACATACAGGTGCCATGCCTGCATATAGTAAGTGCTCAGAAAACATTTCTTGACTGAATTCAGCCAACAAAAATT
TTGGGGTAGGTAGAAAATATATGOTBLUEGTATTTATTGTTATGAGACTGGATATATCTAGTATTTGTCACAGGTAAATGATTCTTCAAAAATTG
AAAGCAAATTTGTTGAAATATTTATTTTGAAAAAAGTTACTTCACAAGCTATAAATTTTAAAAGCCATAGGAATAGATACCGAAGTTATATCCAA
CTGACATTTAATAAATTGTATTCATAGCCTAATGTGATGAGCCACAGAAGCTTGCAAACTTTAATGAGATTTTTTAAAATAGCATCTAAGTTCGG
AATCTTAGGCAAAGTGTTGTTAGATGTAGCACTTCATATTTGAAGTGTTCTTTGGATATTGCATCTACTTTGTTCCTGTTATTATACTGGTGTGA
ATGAATGAATAGGTACTGCTCTCTCTTGGGACATTACTTGACACATAATTACCCAATGAATAAGCATACTGAGGTATCAAAAAAGTCAAATATGT
TATAAATAGCTCATATITMADEGTHISTSLIDEAONGMYABIRTHDAYGSEPTEMBERGTWENTYEIGHTHAGCATGTGCAGTTAATCCTGGAAC
TCCGGTGCTAAGGAGAGACTGTTGGCCCTTGAAGGAGAGCTCCTCCCTGTGGATGAGAGAGAAGGACTTTACTCTTTGGAATTATCTTTTTGTGT
TGATGTTATCCACCTTTTGTTACTCCACCTATAAAATCGGCTTATCTATTGATCTGTTTTCCTAGTCCTTATAAAGTCAAAATGTTAATTGGCAT
AAATTATAGACTTTTTTTAGCAGAGAACTTTGAGGAACCTAAATGCCAACCAGTCTAAAAATGCAGTTTTCAGAAGAATGAATATTTCATGGATA
GTTCTAAATACTAATGAACTTTAAAATAGCTTACTATTGATCTGTCAAAGTGGGTTTTTATATAATTTTCTTTTTACAAATCACCTGACACATTT
AATATAGGTTAAAAAATGCTATCAGGCTGGTTTGCAAAGAAAATGTATTACAAAGGCTGCTAAGEEKSAMAKEAGOODCHUSBANDSTGTTCTCC
AAAATATTTCATAAGGTGCTTTAAGAATAGGTATGTTTTTAAAAGTTAAGTTCCTACTATTTATAGGAACTGACAATCACCTAAAATACCAATGA
TTACAAACTTCCTTCTGGCCTTCTGGACTGCAATTCTAAAAGTGTAAAAAACATATTTTCTGCATTAAGTTAGGCAGTATTGCTTAGTTTTCAAA
GTGGTAGGCTTTGGAGTCAGATTATTTTGATTCAGATCCTACATCTACTGTTTAGTAGCTCTGTTGCCTGAGGCAGGTCCCTTAACATCTCTGTG
TGTGACTTGACCTTTAAAAONETDAYALEFTIESTWILLARULEATHEGWORLDTATGAATGTGAAAAGTTAGCCTAATGTTAACTGCTATTATT
ATGGATTACCATATTTTCACATTCATCACAGTACATGCACCTTGTTAATATAAGATGCTCAATTCATCTTTGAGTATAATTTTGTGACTCTCAAT
CTGGATATGCAATGAGTGGGCCTGTATGAGAATTTAATTTATGAAAAATTGTGTTTCACATGGCCTTACCAGATATACAGGAAACACGTCACATG
TTTCTATTGTATGTTGTTAAATGCCTTAGAATTTAACTTTCTGAATAGGATCCCTTCAGTTTGAGAGTCATAAAAGAGTAAAATTATTATGGTAT
```

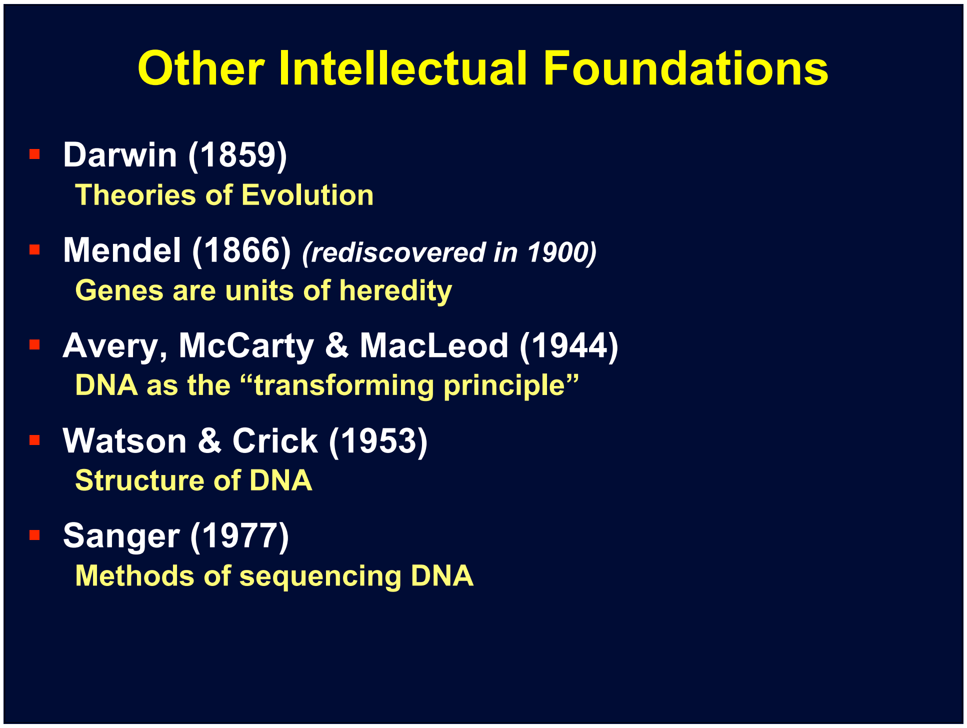
>target1:1308901-1311845

Charles Darwin



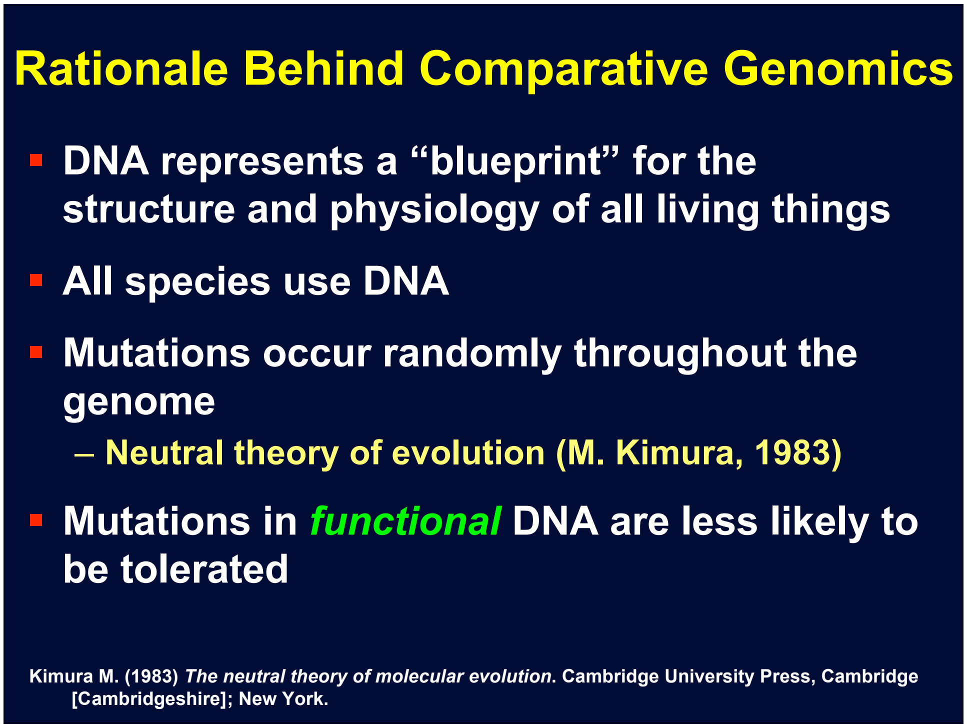


- Served as *naturalist* on a British science expedition around the world (1831 -- 1836)
- *The Origin of Species* (1859)
  - All species evolved from a single life form
  - “Variation” within a species occurs randomly
  - Natural selection
  - Evolutionary change is gradual



Other Intellectual Foundations

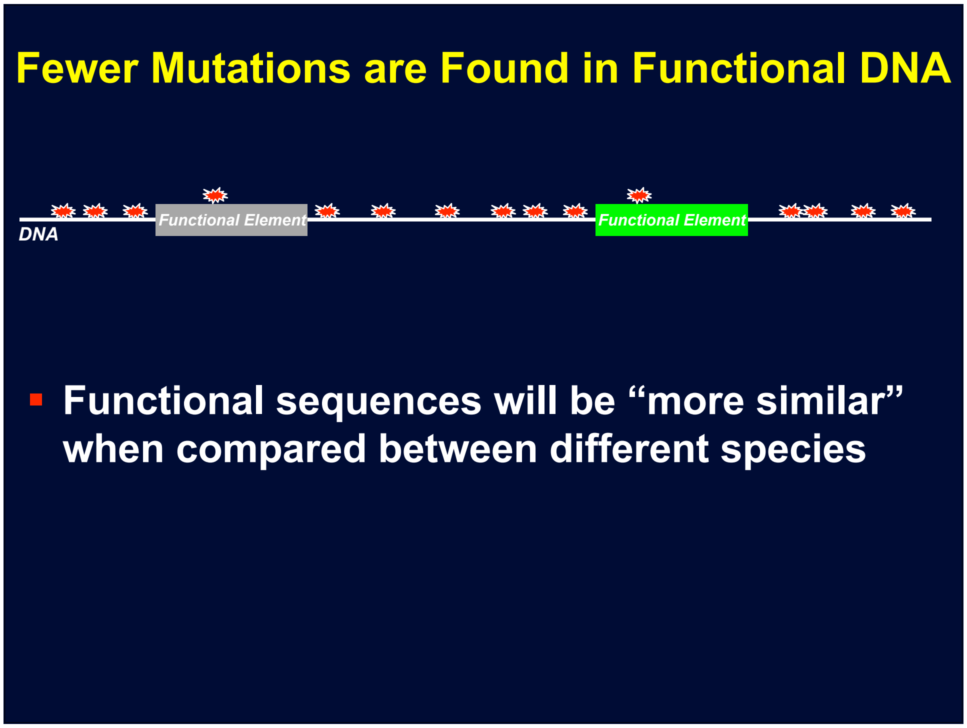
- Darwin (1859)
  Theories of Evolution
- Mendel (1866) *(rediscovered in 1900)*
  Genes are units of heredity
- Avery, McCarty & MacLeod (1944)
  DNA as the “transforming principle”
- Watson & Crick (1953)
  Structure of DNA
- Sanger (1977)
  Methods of sequencing DNA



Rationale Behind Comparative Genomics

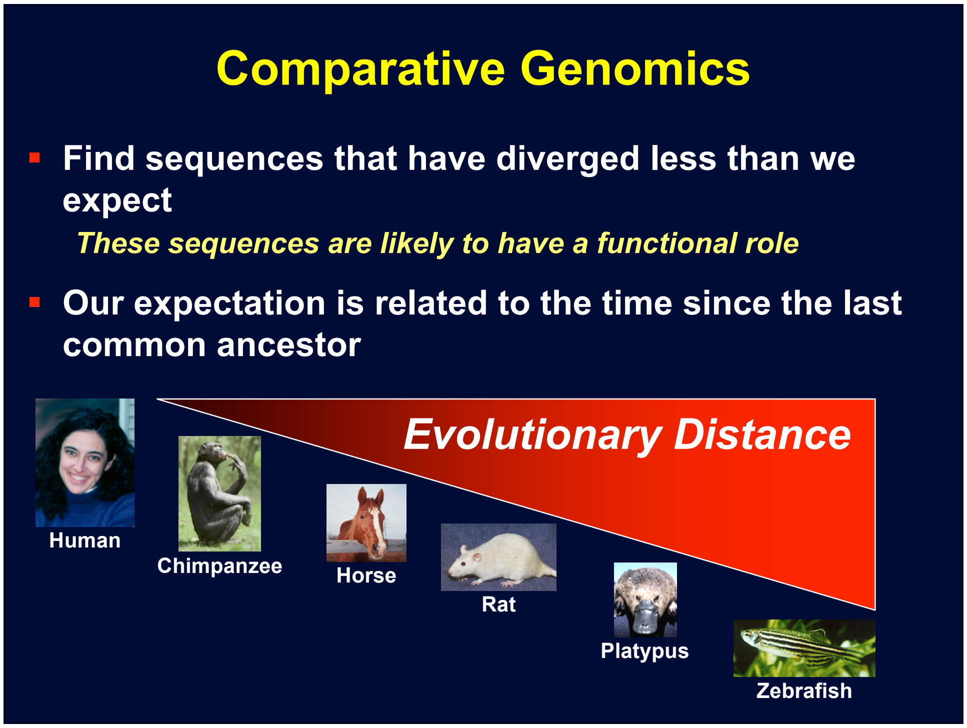
- DNA represents a “blueprint” for the structure and physiology of all living things
- All species use DNA
- Mutations occur randomly throughout the genome
  – Neutral theory of evolution (M. Kimura, 1983)
- Mutations in *functional* DNA are less likely to be tolerated

Kimura M. (1983) *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge [Cambridgeshire]; New York.



Fewer Mutations are Found in Functional DNA

- Functional sequences will be “more similar” when compared between different species

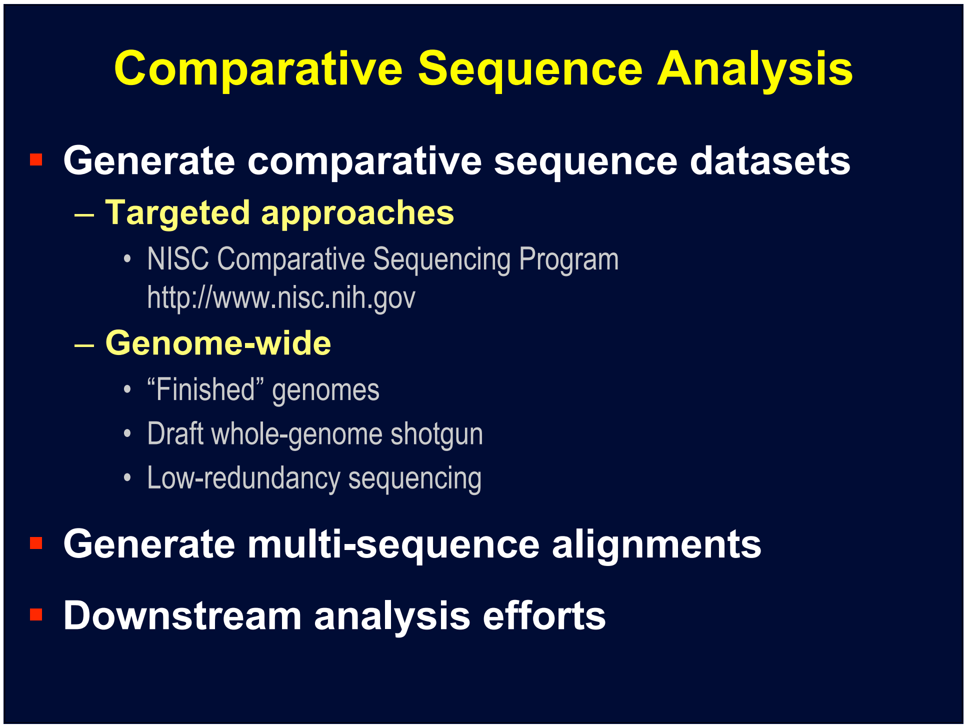


Comparative Genomics

- Find sequences that have diverged less than we expect

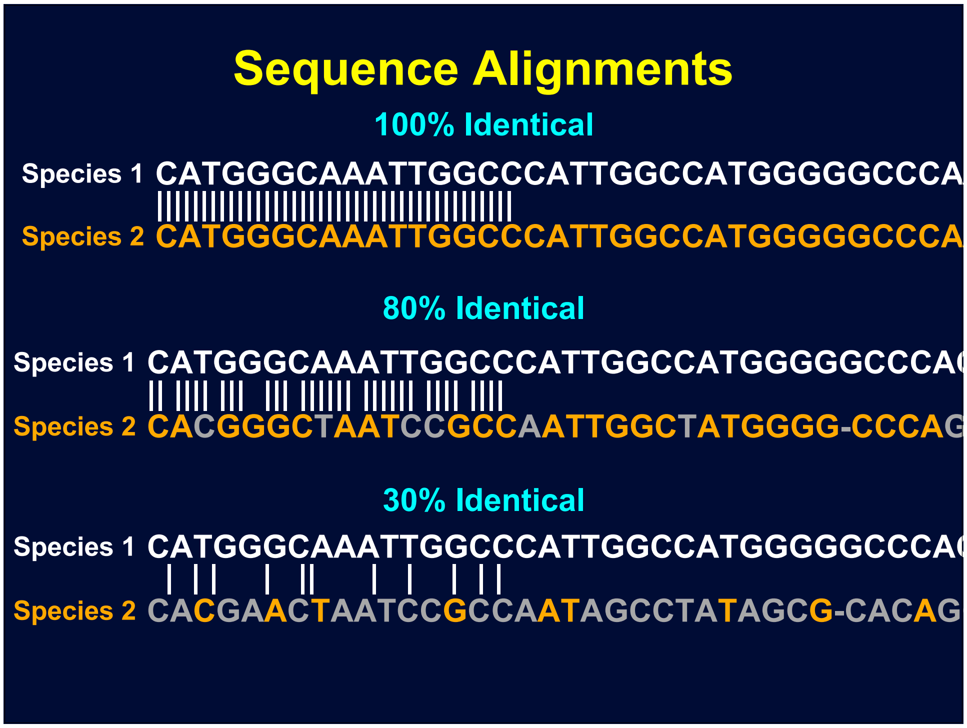
  These sequences are likely to have a functional role

- Our expectation is related to the time since the last common ancestor



Comparative Sequence Analysis

- Generate comparative sequence datasets
  - Targeted approaches
    - NISC Comparative Sequencing Program http://www.nisc.nih.gov
  - Genome-wide
    - "Finished" genomes
    - Draft whole-genome shotgun
    - Low-redundancy sequencing
- Generate multi-sequence alignments
- Downstream analysis efforts



Sequence Alignments

100% Identical

```
Species 1 CATGGGCAAATTGGCCCATTGGCCATGGGGGCCCA
          |||||||||||||||||||||||||||||||||||
Species 2 CATGGGCAAATTGGCCCATTGGCCATGGGGGCCCA
```

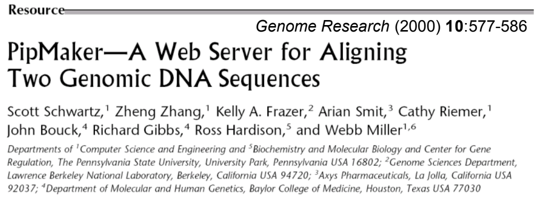
80% Identical

```
Species 1 CATGGGCAAATTGGCCCATTGGCCATGGGGGCCCA
          || |||| |||  ||| |||||| |||||| ||||
Species 2 CACGGGCTAATCCGCCAATTGGCTATGGGG-CCCAG
```

30% Identical

```
Species 1 CATGGGCAAATTGGCCCATTGGCCATGGGGGCCCA
           | ||   |  ||    |  |   | | |
Species 2 CACGAACTAATCCGCCAATAGCCTATAGCG-CACAG
```

Tools for Aligning Genomic Sequences (Targeted Regions)



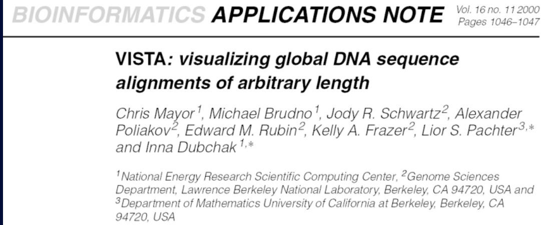
Resource

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



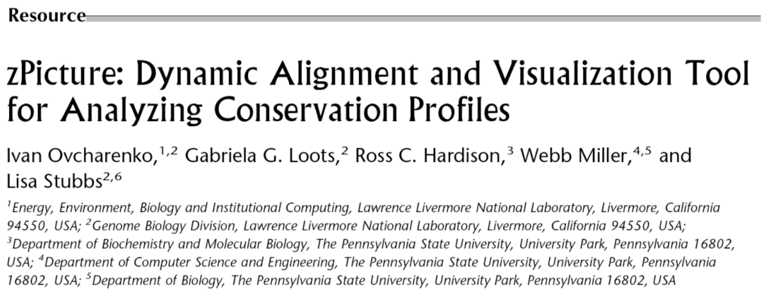
BIOINFORMATICS ***APPLICATIONS NOTE*** Vol. 16 no. 11 2000 Pages 1046–1047

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

Resources for Targeted Sequence Analysis



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

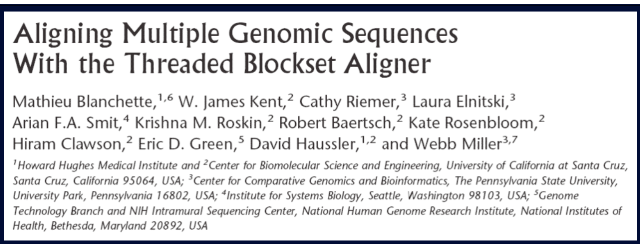
***Genome Research*, 2004, 14(3):472-7**

DCODE.org Comparative Genomics Center
comparing genomes to decipher the code of gene regulation

`http://www.dcode.org/`

Genome-wide Multi-sequence Alignments

- This is not a “solved problem”
- Significant challenges:
  - Finding the correct sequences to align
  - Not all sequences should align
  - Dealing with insertions/deletions
  - Handling duplications and rearrangements
  - Missing data challenges (i.e., sequencing gaps)

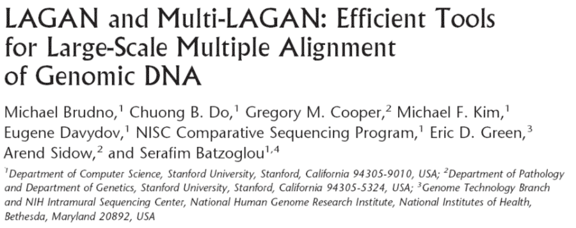


Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner

Mathieu Blanchette,[1,6] W. James Kent,[2] Cathy Riemer,[3] Laura Elnitski,[3] Arian F.A. Smit,[4] Krishna M. Roskin,[2] Robert Baertsch,[2] Kate Rosenbloom,[2] Hiram Clawson,[2] Eric D. Green,[5] David Haussler,[1,2] and Webb Miller[3,7]

[1]Howard Hughes Medical Institute and [2]Center for Biomolecular Science and Engineering, University of California at Santa Cruz, Santa Cruz, California 95064, USA; [3]Center for Comparative Genomics and Bioinformatics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; [4]Institute for Systems Biology, Seattle, Washington 98103, USA; [5]Genome Technology Branch and NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

Genome Research **(2004) 14:708-715**

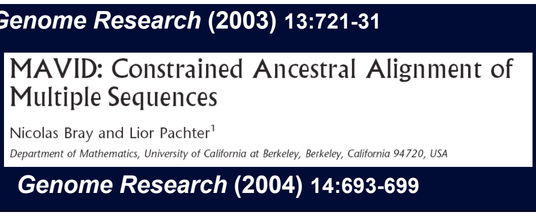


LAGAN and Multi-LAGAN: Efficient Tools for Large-Scale Multiple Alignment of Genomic DNA

Michael Brudno,[1] Chuong B. Do,[1] Gregory M. Cooper,[2] Michael F. Kim,[1] Eugene Davydov,[1] NISC Comparative Sequencing Program,[1] Eric D. Green,[3] Arend Sidow,[2] and Serafim Batzoglou[1,4]

[1]Department of Computer Science, Stanford University, Stanford, California 94305-9010, USA; [2]Department of Pathology and Department of Genetics, Stanford University, Stanford, California 94305-5324, USA; [3]Genome Technology Branch and NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

Genome Research **(2003) 13:721-31**



MAVID: Constrained Ancestral Alignment of Multiple Sequences

Nicolas Bray and Lior Pachter[1]

Department of Mathematics, University of California at Berkeley, Berkeley, California 94720, USA

Genome Research **(2004) 14:693-699**

Genome Browsers

UCSC Genome Bioinformatics

http://genome.ucsc.edu

e! project Ensembl

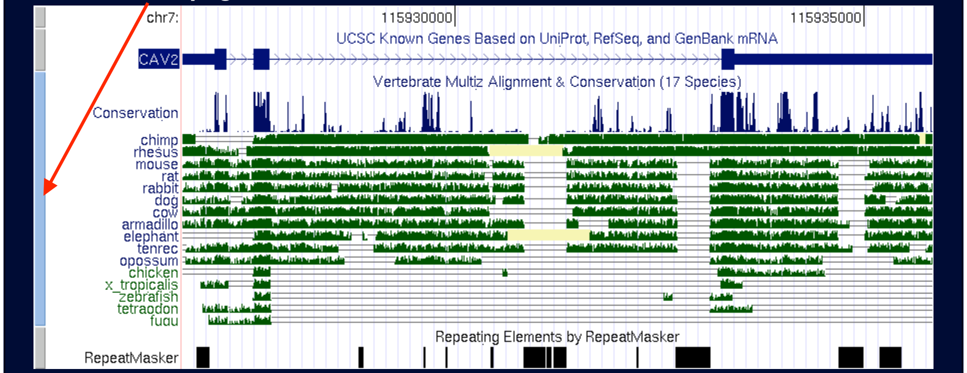
http://www.ensembl.org

NCBI Map Viewer

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

Multi-sequence Alignments at UCSC

Click here for track details page



Chaining Alignments

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

The Challenge:

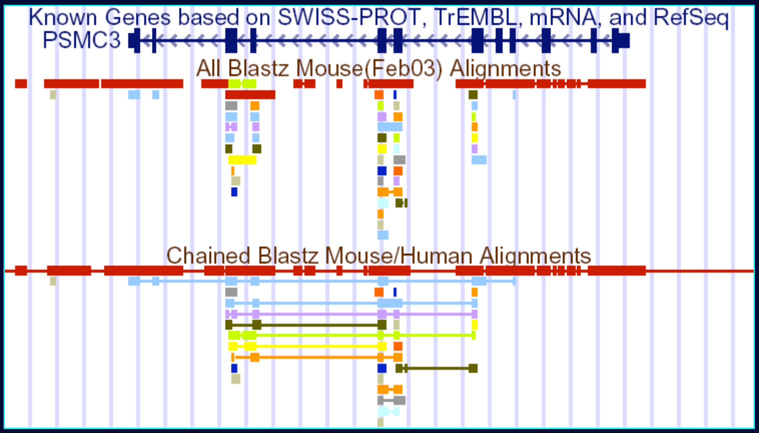
- Local alignments tend to break at transposon insertions, inversions, duplications, etc.
- Global alignments tend to force non-homologous bases to align.

The Solution:

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

Slide (though modified) Courtesy of Jim Kent

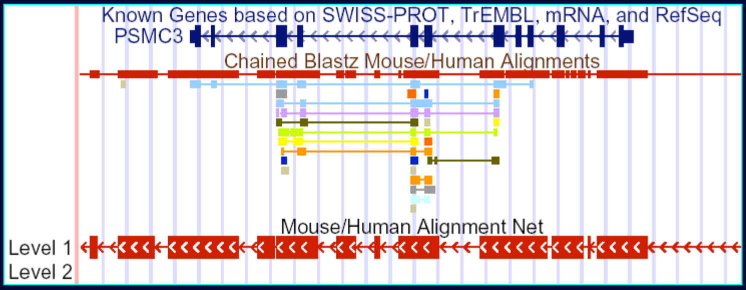
Chains join together related local alignments


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

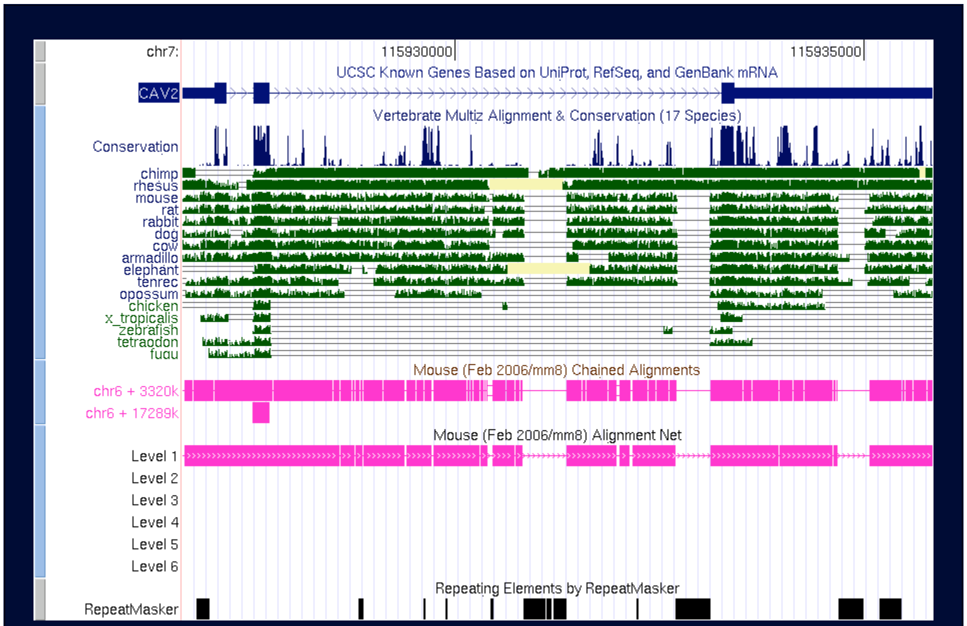
Slide Courtesy of Jim Kent

Net Alignments: Focus on Orthology

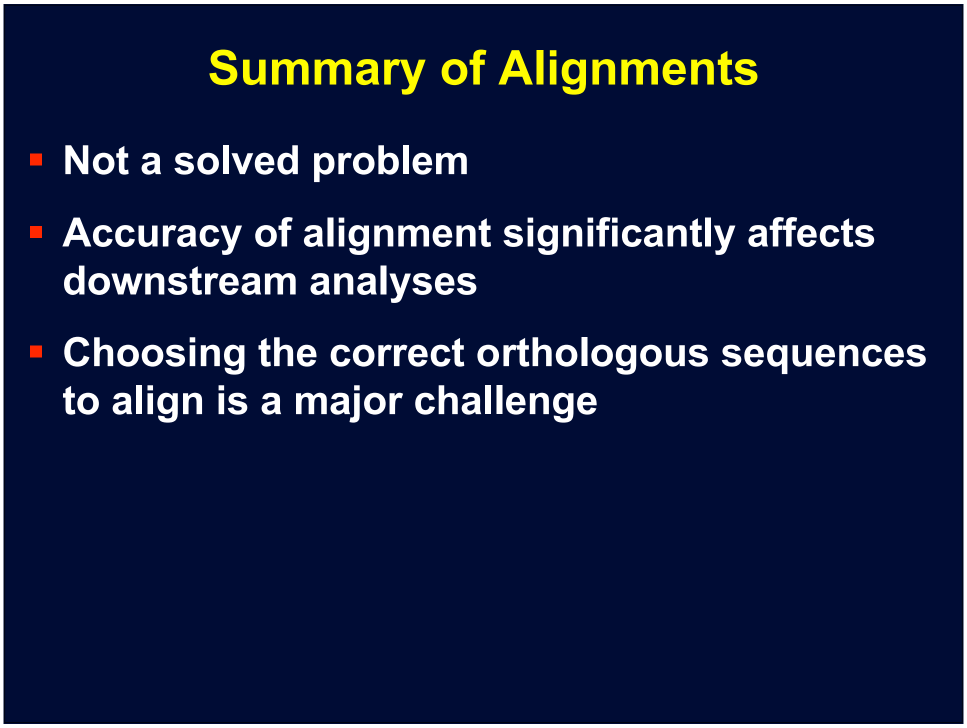

Known Genes based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
PSMC3
Chained Blastz Mouse/Human Alignments
Mouse/Human Alignment Net
Level 1
Level 2

- 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

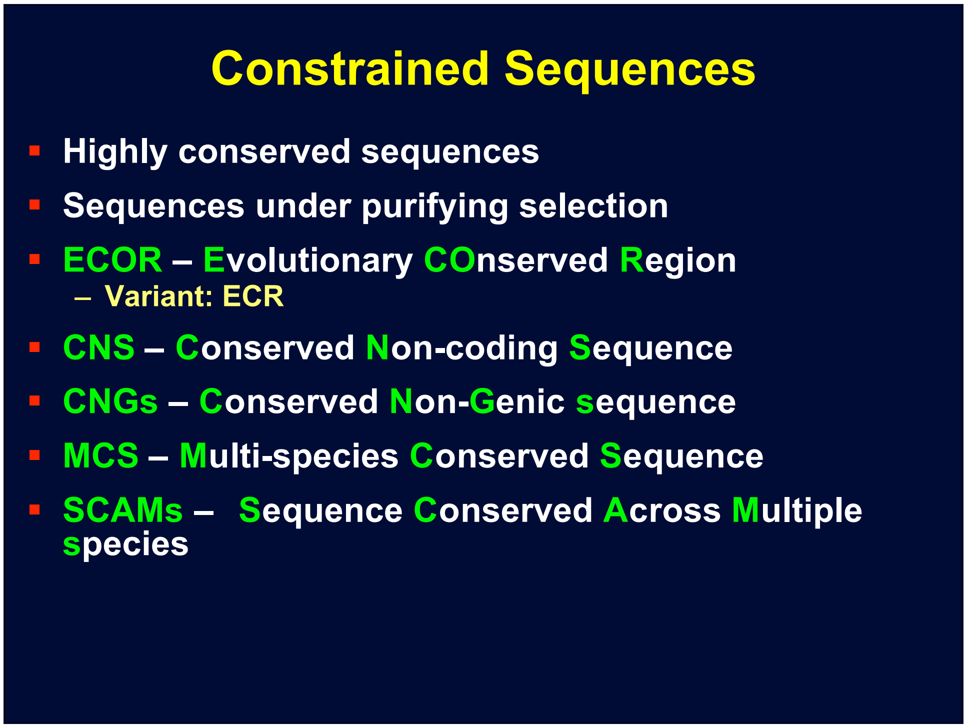

chr7:
115930000
115935000
UCSC Known Genes Based on UniProt, RefSeq, and GenBank mRNA
CAV2
Vertebrate Multiz Alignment & Conservation (17 Species)
Conservation
chimp
rhesus
mouse
rat
rabbit
dog
cow
armadillo
elephant
tenrec
opossum
chicken
x_tropicalis
zebrafish
tetraodon
fugu
Mouse (Feb 2006/mm8) Chained Alignments
chr6 + 3320k
chr6 + 17289k
Mouse (Feb 2006/mm8) Alignment Net
Level 1
Level 2
Level 3
Level 4
Level 5
Level 6
Repeating Elements by RepeatMasker
RepeatMasker

Click here for a more complicated example



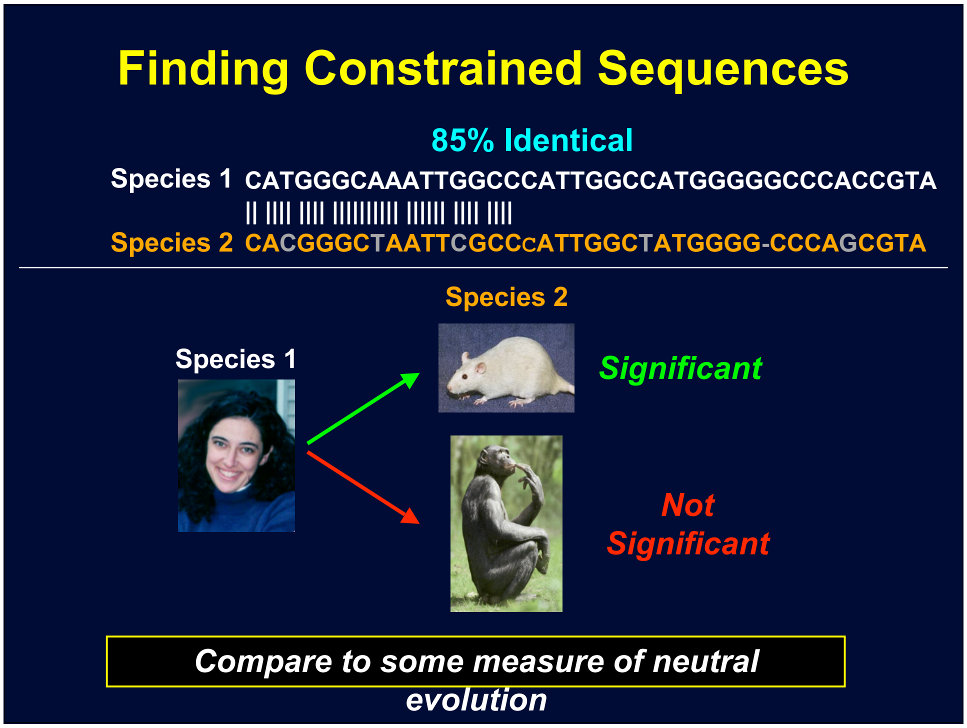
Summary of Alignments

- Not a solved problem
- Accuracy of alignment significantly affects downstream analyses
- Choosing the correct orthologous sequences to align is a major challenge



Constrained Sequences

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



Finding Constrained Sequences

85% Identical

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

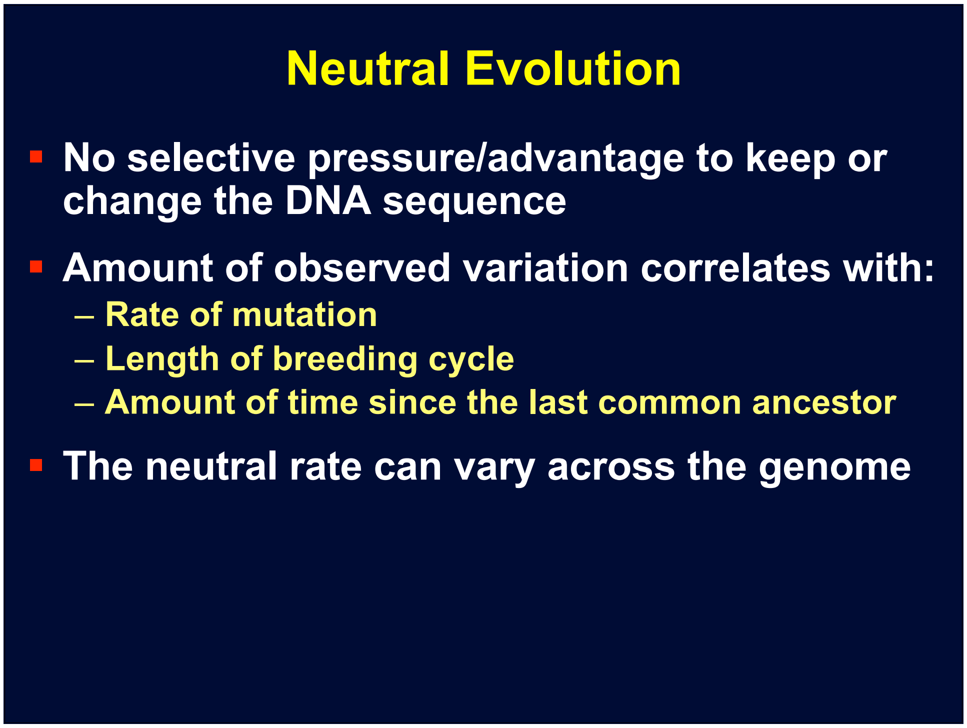
Species 2

Species 1

Significant

Not Significant

Compare to some measure of neutral evolution



Neutral Evolution

- No selective pressure/advantage to keep or change the DNA sequence
- Amount of observed variation correlates with:
  - Rate of mutation
  - Length of breeding cycle
  - 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

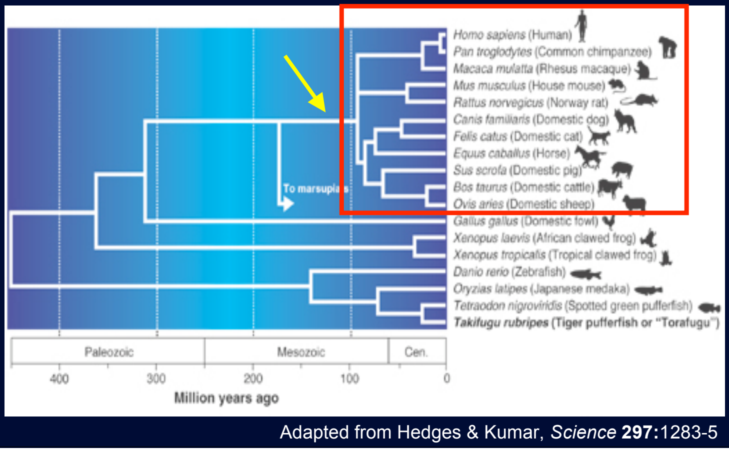
  Third position of codons which can be any base and code for the same amino acid

	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

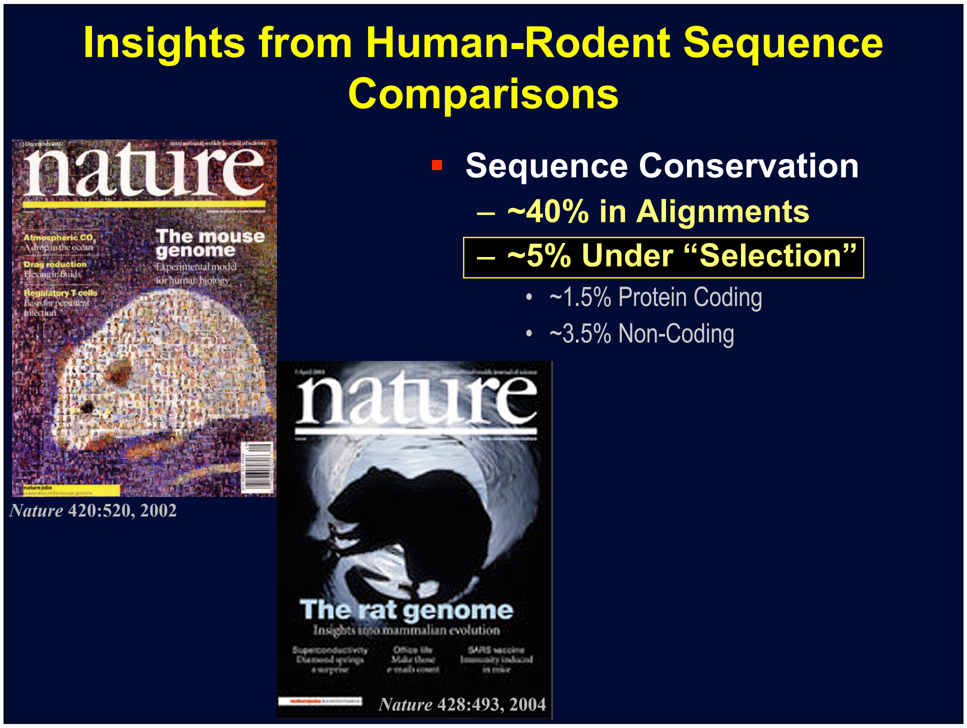
Types of Neutrally Evolving DNA

- 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

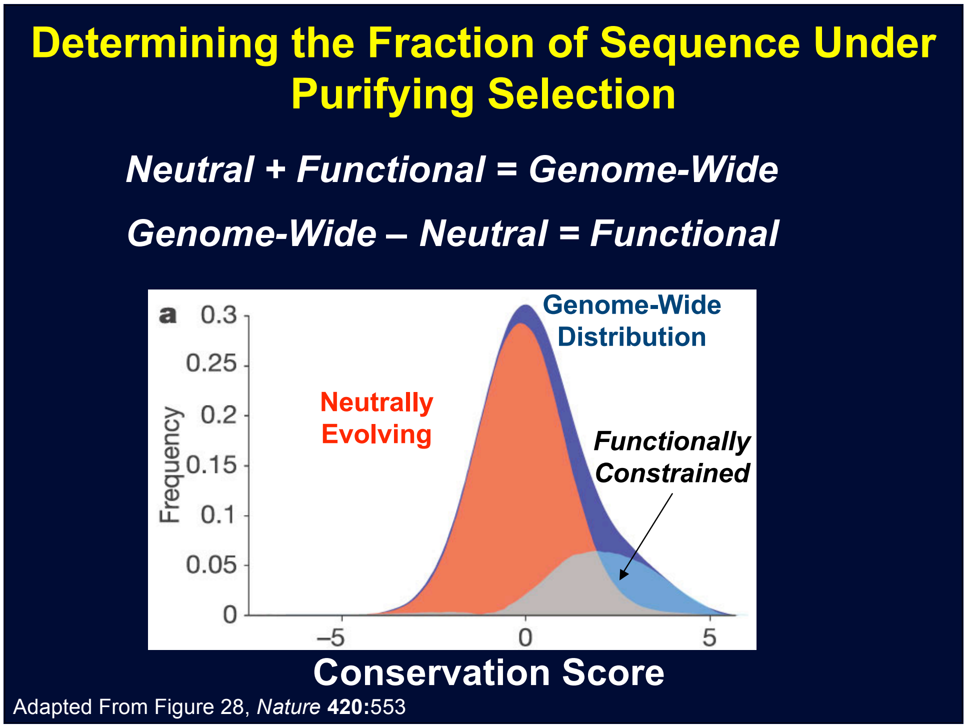


Insights from Human-Rodent Sequence Comparisons

Nature 420:520, 2002

Nature 428:493, 2004

- Sequence Conservation
  - ~40% in Alignments
  - ~5% Under "Selection"
    - ~1.5% Protein Coding
    - ~3.5% Non-Coding



Determining the Fraction of Sequence Under Purifying Selection

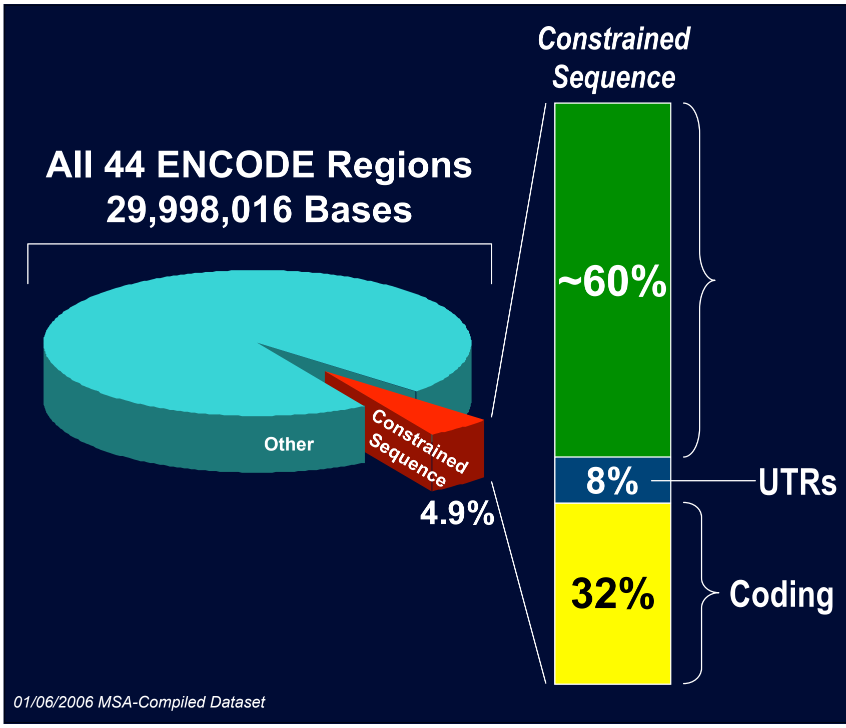
Neutral + Functional = Genome-Wide

Genome-Wide – Neutral = Functional

Conservation Score

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

All 44 ENCODE Regions
29,998,016 Bases

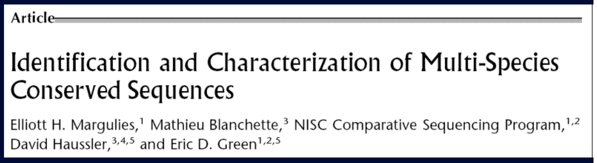


01/06/2006 MSA-Compiled Dataset

Measures of Sequence Conservation

Binomial-based Method

binCons



Article

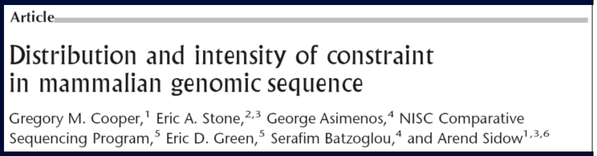
Identification and Characterization of Multi-Species Conserved Sequences

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**

Genomic Evolutionary Rate Profiling

GERP



Article

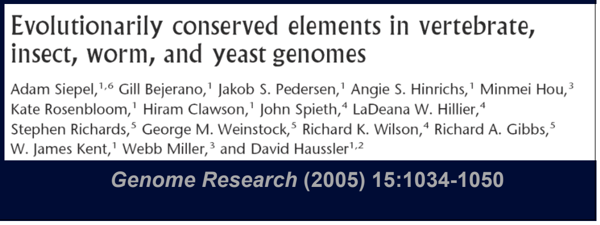
Distribution and intensity of constraint in mammalian genomic sequence

Gregory M. Cooper,[1] Eric A. Stone,[2,3] George Asimenos,[4] NISC Comparative Sequencing Program,[5] Eric D. Green,[5] Serafim Batzoglou,[4] and Arend Sidow[1,3,6]

Genome Research **(2005) 15:901-913**

PHylogenetic Analysis with Space/Time models

phastCons

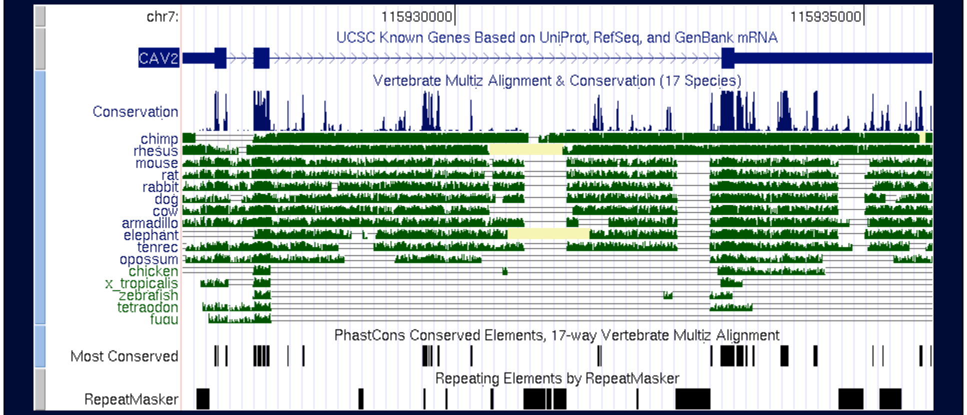


Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes

Adam Siepel,[1,6] Gill Bejerano,[1] Jakob S. Pedersen,[1] Angie S. Hinrichs,[1] Minmei Hou,[3] Kate Rosenbloom,[1] Hiram Clawson,[1] John Spieth,[4] LaDeana W. Hillier,[4] Stephen Richards,[5] George M. Weinstock,[5] Richard K. Wilson,[4] Richard A. Gibbs,[5] W. James Kent,[1] Webb Miller,[3] and David Haussler[1,2]

Genome Research **(2005) 15:1034-1050**

Constrained Sequences Available from UCSC


chr7:
115930000
115935000
UCSC Known Genes Based on UniProt, RefSeq, and GenBank mRNA
CAV2
Vertebrate Multiz Alignment & Conservation (17 Species)
Conservation
chimp
rhesus
mouse
rat
rabbit
dog
cow
armadillo
elephant
tenrec
opossum
chicken
x_tropicalis
zebrafish
tetraodon
fugu
PhastCons Conserved Elements, 17-way Vertebrate Multiz Alignment
Most Conserved
Repeating Elements by RepeatMasker
RepeatMasker

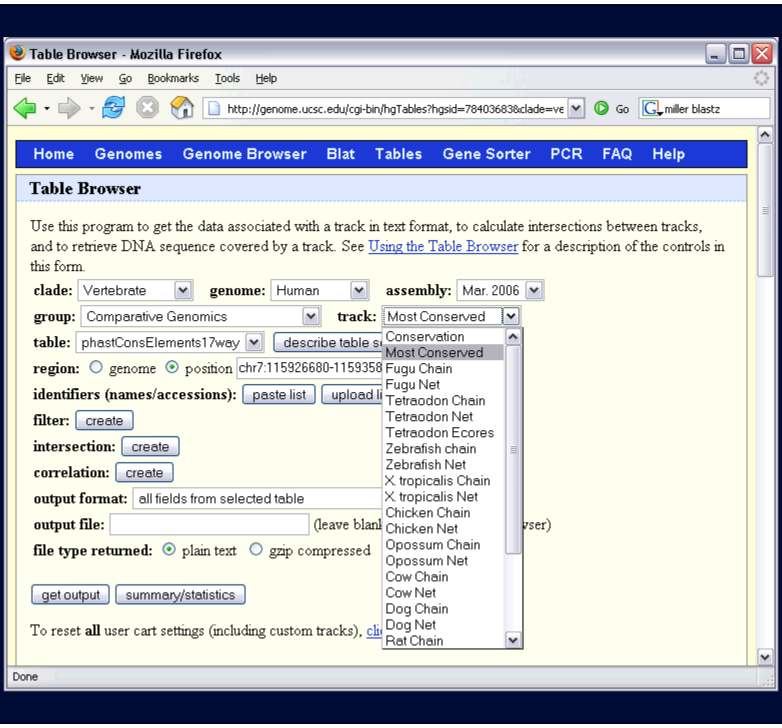


Table Browser - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://genome.ucsc.edu/cgi-bin/hgTables?hgsid=784036838clade=ve
Go
miller blastz
Home Genomes Genome Browser Blat Tables Gene Sorter PCR FAQ Help

Table Browser

Use this program to get the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. See Using the Table Browser for a description of the controls in this form.

clade: Vertebrate **genome:** Human **assembly:** Mar. 2006

group: Comparative Genomics **track:** Most Conserved

table: phastConsElements17way describe table sc

region: ○ genome ● position chr7:115926680-1159358

identifiers (names/accessions): paste list upload li

filter: create

intersection: create

correlation: create

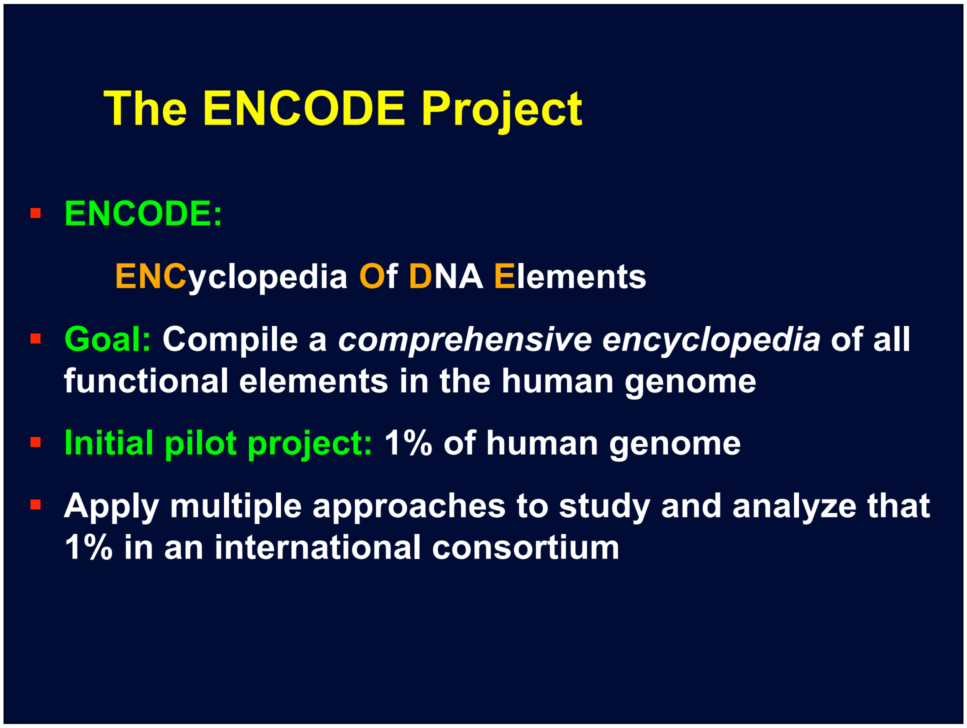
output format: all fields from selected table

output file: (leave blank ... wser)

file type returned: ● plain text ○ gzip compressed

get output summary/statistics

To reset **all** user cart settings (including custom tracks), cli

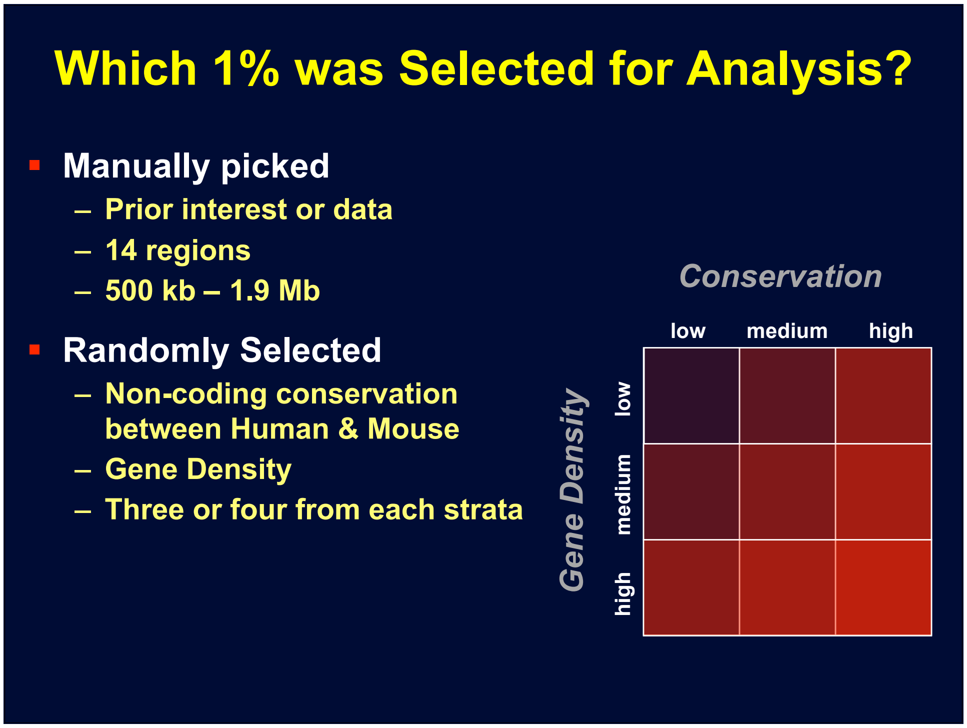


The ENCODE Project

- **ENCODE:**

  ENCyclopedia Of DNA Elements

- **Goal: Compile a *comprehensive encyclopedia* of all functional elements in the human genome**
- **Initial pilot project: 1% of human genome**
- **Apply multiple approaches to study and analyze that 1% in an international consortium**



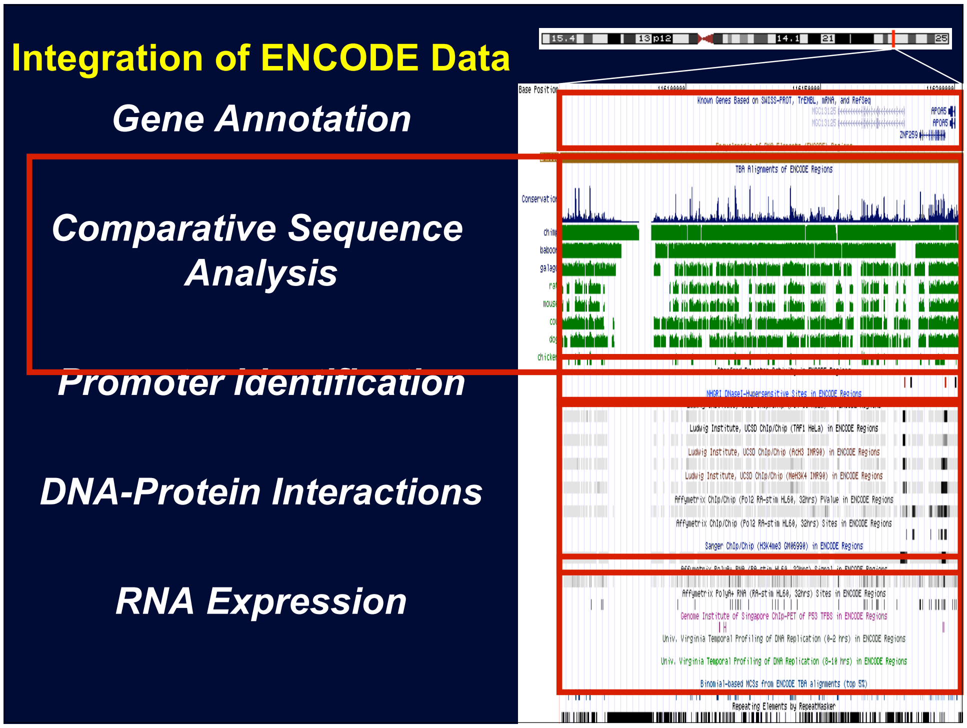
Which 1% was Selected for Analysis?

- **Manually picked**
  - **Prior interest or data**
  - **14 regions**
  - **500 kb – 1.9 Mb**
- **Randomly Selected**
  - **Non-coding conservation between Human & Mouse**
  - **Gene Density**
  - **Three or four from each strata**

Conservation

Gene Density

	low	medium	high
low			
medium			
high			



Integration of ENCODE Data

Gene Annotation

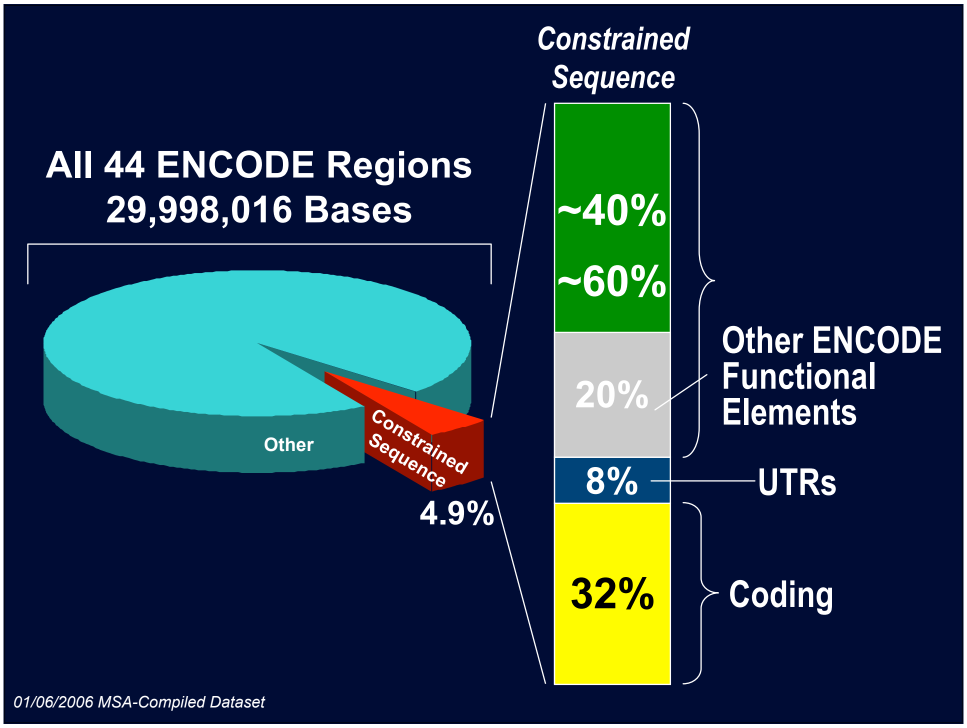
Comparative Sequence Analysis

Promoter Identification

DNA-Protein Interactions

RNA Expression

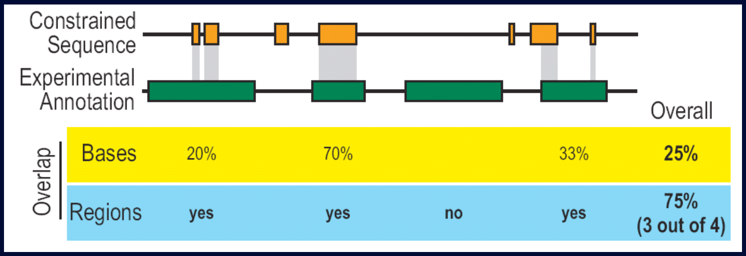
Base Position
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
APOA5
ZNF259
TBA Alignments of ENCODE Regions
Conservation
chimp
baboon
galago
rat
mouse
cow
dog
chicken
NHGRI DNaseI-Hypersensitive Sites in ENCODE Regions
Ludwig Institute, UCSD ChIp/Chip (TAF1 HeLa) in ENCODE Regions
Ludwig Institute, UCSD ChIp/Chip (AcH3 IMR90) in ENCODE Regions
Ludwig Institute, UCSD ChIp/Chip (MeH3K4 IMR90) in ENCODE Regions
Affymetrix ChIp/Chip (Pol2 RA-stim HL60, 32hrs) PValue in ENCODE Regions
Affymetrix ChIp/Chip (Pol2 RA-stim HL60, 32hrs) Sites in ENCODE Regions
Sanger ChIp/Chip (H3K4me3 GM06990) in ENCODE Regions
Affymetrix PolyA+ RNA (RA-stim HL60, 32hrs) Sites in ENCODE Regions
Genome Institute of Singapore ChIp-PET of P53 TFBS in ENCODE Regions
Univ. Virginia Temporal Profiling of DNA Replication (0-2 hrs) in ENCODE Regions
Univ. Virginia Temporal Profiling of DNA Replication (8-10 hrs) in ENCODE Regions
Binomial-based MCSs from ENCODE TBA alignments (top 5%)
Repeating Elements by RepeatMasker



All 44 ENCODE Regions
29,998,016 Bases

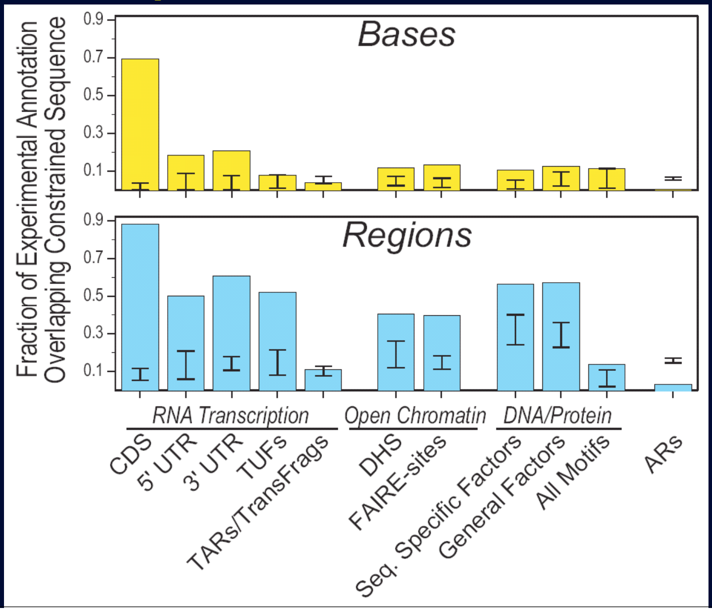
01/06/2006 MSA-Compiled Dataset

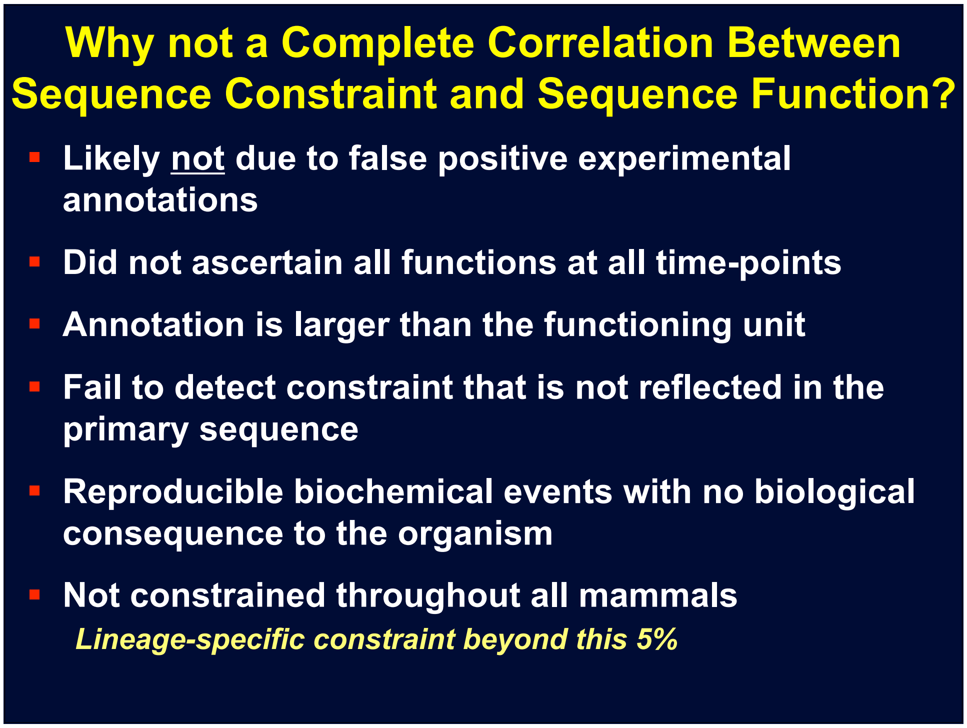
Assessing the Overlap between Constrained Sequences and Experimental Annotations



Overlap					Overall
Bases	20%	70%		33%	**25%**
Regions	**yes**	**yes**	**no**	**yes**	**75%** **(3 out of 4)**

Overlap between Constrained Sequences and Experimental Annotations

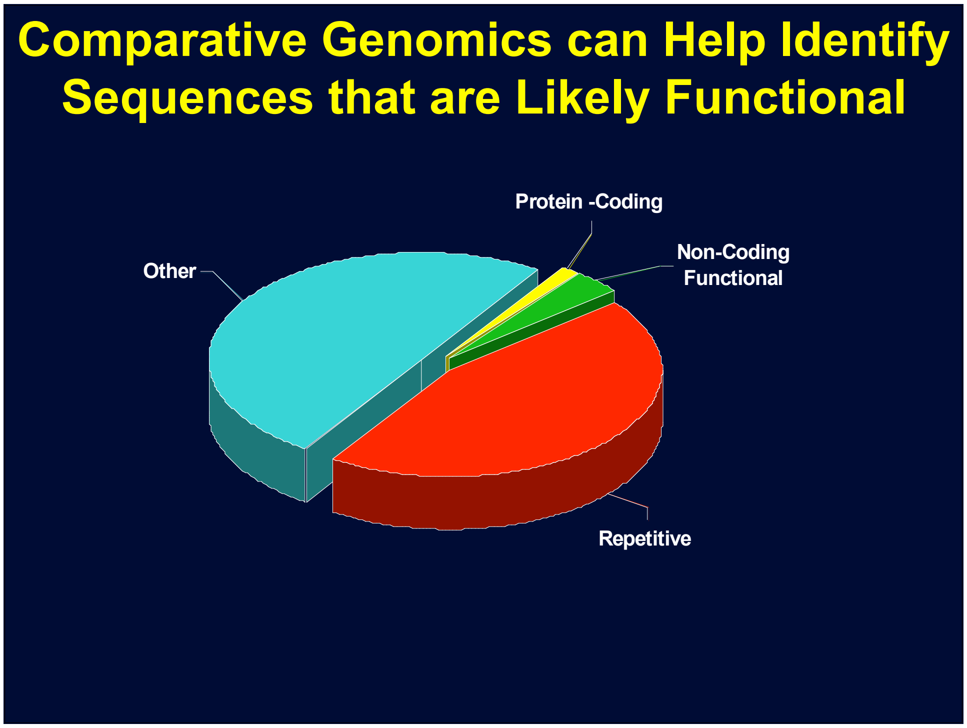




Why not a Complete Correlation Between Sequence Constraint and Sequence Function?

- Likely not due to false positive experimental annotations
- Did not ascertain all functions at all time-points
- Annotation is larger than the functioning unit
- Fail to detect constraint that is not reflected in the primary sequence
- Reproducible biochemical events with no biological consequence to the organism
- Not constrained throughout all mammals

  Lineage-specific constraint beyond this 5%



Comparative Genomics can Help Identify Sequences that are Likely Functional