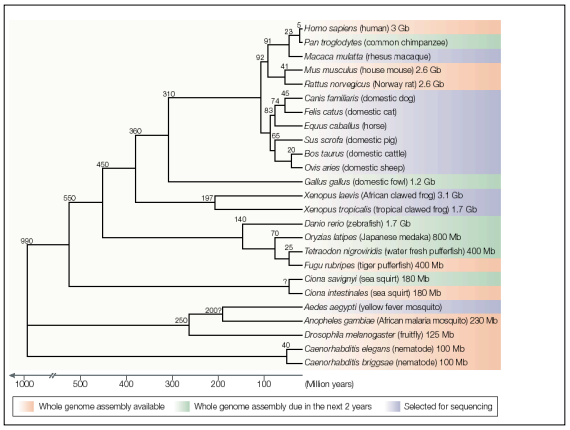
Project presentations

- All written projects due on Thursday March 10th.
- Student presentation will be on Thursday March 10th. All in J-280.
  - During lecture time Tuesday section will present
  - During lab hours Thursday section will present
    - E-mail me your power point presentation by Thursday March 10th at 11am.

Detecting conserved regions

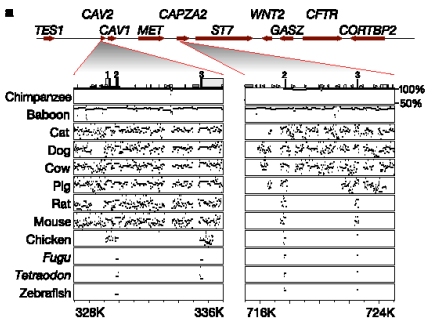
- Closely related species
- Divergent groups.



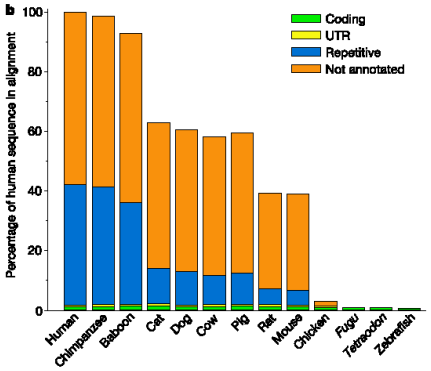
Phylogenetic footprinting

- Alignment of several divergent taxonomic groups.
- Generate percent identity plots (pips)
- Look for conserved regions across several levels.

Percent identity plots

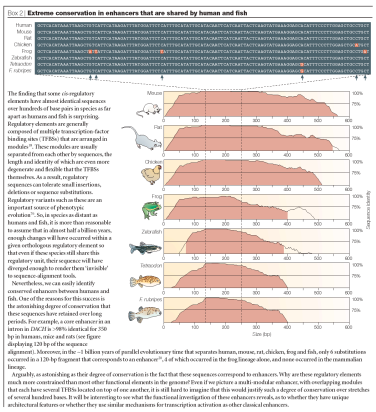


Different scales resolve



Multiple-species conserved sequences (MCSs)

- Calibrated so that 5% of sequence fell into MCS
- Found 1,194 elements conserved
- Average 58 base pairs
- 32% cover exons
  - Identifies 90.4% coding and 27% UTR
- 68% outside known exons
  - Are these elements functional?



Box 2 | **Extreme conservation in enhancers that are shared by human and fish**

The finding that some *cis*-regulatory elements have almost identical sequences over hundreds of base pairs in species as far apart as humans and fish is surprising. Regulatory elements are generally composed of multiple transcription-factor binding sites (TFBSs) that are arranged in modules[19]. These modules are usually separated from each other by sequences, the length and identity of which are even more degenerate and flexible that the TFBSs themselves. As a result, regulatory sequences can tolerate small insertions, deletions or sequence substitutions. Regulatory variants such as these are an important source of phenotypic evolution[19]. So, in species as distant as humans and fish, it is more than reasonable to assume that in almost half a billion years, enough changes will have occurred within a given orthologous regulatory element so that even if these species still share this regulatory unit, their sequence will have diverged enough to render them 'invisible' to sequence-alignment tools.

Nevertheless, we can easily identify conserved enhancers between humans and fish. One of the reasons for this success is the astonishing degree of conservation that these sequences have retained over long periods. For example, a core enhancer in an intron in *DACH* is >98% identical for 350 bp in humans, mice and rats (see figure displaying 320 bp of the sequence alignment). Moreover, in the ~1 billion years of parallel evolutionary time that separates human, mouse, rat, chicken, frog and fish, only 6 substitutions occurred in a 120-bp fragment that corresponds to an enhancer[20]. 4 of which occurred in the frog lineage alone, and none occurred in the mammalian lineage.

Arguably, as astonishing as their degree of conservation is the fact that these sequences correspond to enhancers. Why are these regulatory elements much more constrained than most other functional elements in the genome? Even if we picture a multi-modular enhancer, with overlapping modules that each have several TFBSs located on top of one another, it is still hard to imagine that this would justify such a degree of conservation over stretches of several hundred bases. It will be interesting to see what the functional investigation of these enhancers reveals, as to whether they have unique architectural features or whether they use similar mechanisms for transcription activation as other classical enhancers.

Only 6 substitutions in 120bp enhancer Over ~1 billion years

Are these MCSs functional?

- Fuse element to β-galactosidase gene.
  - Check expression in embryonic development.
- Mutagenesis of elements
  - Compare mouse and human

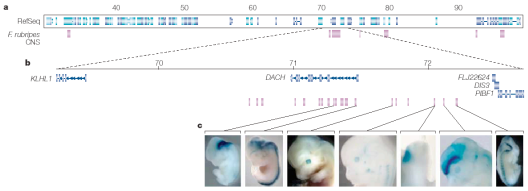


Figure 1 | **Architecture of human–*Fugu rubripes* conserved non-coding sequences in the human genome. a** | A 65-Mb segment of human chromosome 13 is shown that contains 145 well-characterized RefSeq genes (exons in blue). There are 51 human–*F. rubripes* conserved non-coding sequences (CNSs) in this region, which are distributed non-uniformly in clusters that contain 1–32 CNSs each (in purple). **b** | One cluster of human–*F. rubripes* CNS is illustrated in more detail. *DACH* — the only human gene in this region — is involved in key aspects of embryonic development. **c** | Testing some of the non-coding sequences that are conserved in humans and *F. rubripes* revealed that several of these elements correspond to enhancers in mouse embryos. In this assay, the sequence being tested is cloned upstream of a β-galactosidase reporter gene. If the cloned sequence is an enhancer, it will activate the reporter gene, which can be detected in an assay that stains the tissues that express β-galactosidase (in blue). Adapted with permission from REF. 20. © (2003) M. A. Nobrega and E. M. Rubin.

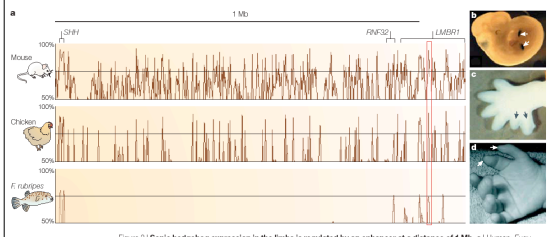


Figure 2 | **Sonic hedgehog expression in the limbs is regulated by an enhancer at a distance of 1 Mb. a** | Human–*Fugu rubripes* sequence comparisons, generated by VISTA, identify a conserved non-coding sequence in intron 5 of *LMBR1* (red box), which drives the expression of a reporter gene in a pattern that resembles the expression of sonic hedgehog (*SHH*) (arrows in **b**). Insertional mutagenesis in this region in mice results in preaxial polydactyly (arrows in **c**). In humans, mutations in this enhancer are also associated with preaxial polydactyly (arrows in **d**). Adapted with permission from REF. 21. © (2003) Oxford University Press and REF. 32 (1999) Elsevier Science Ltd.

Many of conserved elements involved in development

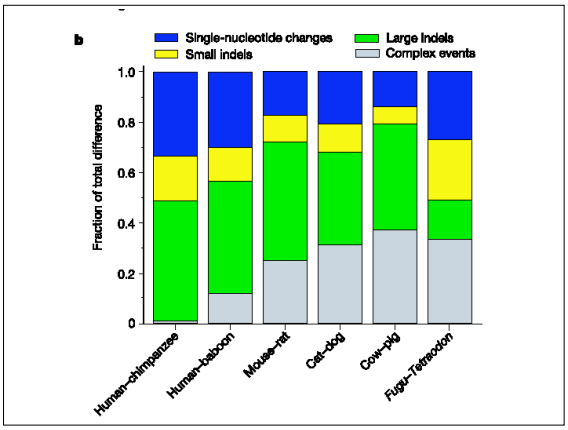
Table 1 | **Genes in proximity to highly conserved non-coding sequences***

Gene	Molecular function	Biological process	Reference
HOXB4	DNA-binding	Embryonic development	19
WNT1	Signal transducer	Embryonic development	79
SHH	Hydrolase and peptidase	Embryonic development	80,21
SCL (TAL1)	DNA-binding	Cell differentiation	29
SOX9	DNA-binding	Cell differentiation	81
DLL1	Protein-binding	Embryonic development	82
DLX1, -2, -5 and *-6*	DNA-binding	Embryonic development	24,31
HOXA1–13	DNA-binding	Embryonic development	83
HOXD cluster	DNA-binding	Embryonic development	84
DACH	Transcription factor	Embryonic development	20
NEUROG1	DNA-binding	Embryonic development	25
HOXC8	DNA-binding	Embryonic development	38
OTX2	DNA-binding	Embryonic development	28
CTGF	Growth-factor signalling	Cell growth/proliferation	26
PAX6	DNA-binding	Embryonic development	85
RUNX2	DNA-binding	Skeletal development	86

*Between mammals and fish. The molecular function and biological process of each gene were obtained from the Gene Ontology Consortium database (see online links box).

What type of changes are observed across these groups

- Nucleotide differences
- Insertions/deletions
- Complex events

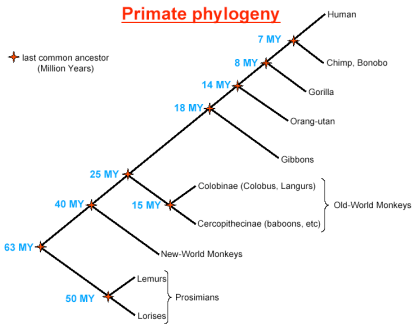


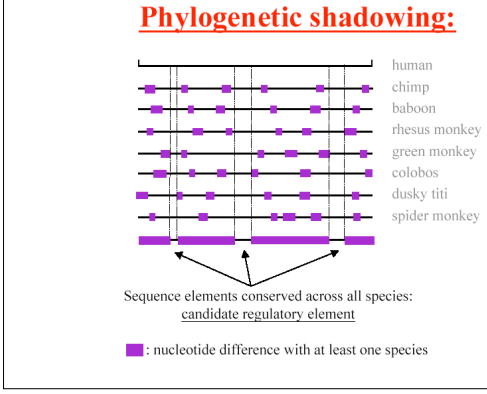
Problems with phylogenetic footprinting

- Alignment difficult between divergent species.
- Does not take into account phylogenetic tree information.
- Identification fairly arbitrary.
- Conservation must be across all lineages.
- Will miss lineage specific elements.

Phylogenetic shadowing

- Generate data from primates
  - Should be easier to align than divergent taxa
- Use phylogenetic information to look at rates of evolution across all lineages
  - Results in similar amount of divergence compared to human - mouse
- Classify sites as either rapidly evolving or slowly evolving
  - Statistical estimation versus arbitrary percent identity.





Amplification of target sequence from genomic DNA of ~20 primates (new- and old- world monkeys, hominoids)

↓

Direct sequence of the PCR product

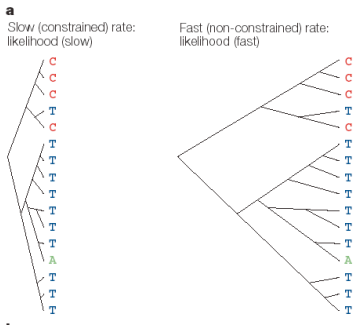
↓

Multiple sequence alignement with ClustalW:

```
Allen       CACCCAGNTAATTTTTGTATTTTTAGTAGAG---ACAGGGTTTCACTACGTTGGCCAGGC 264
Green       CACCCGGNTAATTTTTGTATTTTTAGTAGAG---ACAGGGTTTCACTACGTTGGCCAGGC 193
Human       CACCCGGCTAATTTTTGTATTTTTAGTAGAG---TCGGGGTTTCACTATGTTGGCCAGGC 398
Chimp       CACCCGGCTAATTTTTGTATTTTTAGTAGAG---TCGGGGTTTCACTATGTTGGCCAGGC 367
Orangutan   CGCCCGGCTAATTTTTGTATTTTTAGTAGAG---TCGGGATTTCACTATGTTGGCCAGGC 310
Colobus     CACCCGGATAATTTTTGTATTTTTAGTAGAG---ACGGGGTTTCACTACGTTGGCCAGGC 331
Douc        CACCCGGCTAATTTTTGTATTTTTAGTAGAG---ACGGGGTTTCACTACGTTGGCCAGGC 328
Francois    CACCCGGCTAATTTTTGTATTTTTAGTAGAG---ACGGGGTTTCACTACGTTGGCCAGGC 194
Drill       CACCCGGCTAATTTTTGTATTTTTAGTAGAG---ACAGGGTTTCACTACGTTGGCCAGGC 303
Mangabey    CACCCGGNTAATTTTTGTATTTTTAGTAGAG---ACAGGGTTTCACTACGTTGGCCAGGC 192
Owl         CATCCGGCTAATTTTTGTATTTTTAGTAAAGGCGACAGGATTTCACCATGTTGGCCAGGC 311
Squirrel    CACCCGGCTAAATTTTGTATTTTTAGTAGAGGCGACAGGATTTCACCATGTTGGCCAGGC 196
Tamarin     CACCTAAATAA-TTTTGTATTTTTAGAAAAGGCGACAGGATTTCACCACGTTGGCCAGGC 331
Titi        CACCTGGCTAATTTTTGTATTTTTAGTACAGGCAACAGGATTTCACCATGTTGGCCAGGC 322
Consensus   *  *    *** ************** * **     * ** ****** * ********** 
```

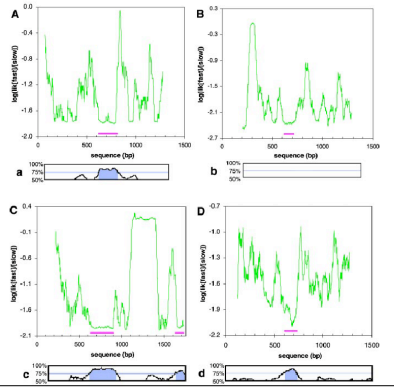
Statistics

- Classify sites based on likelihood of being fast of slow evolving
- Calculate log(like[fast]/[slow]) - look for more negative = conserved
  - log(0.5) = -0.3
  - log(2) = 0.3



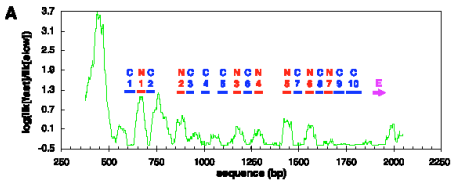
Exons

- Identify exons from regions with 4 known intron-exon boundries.



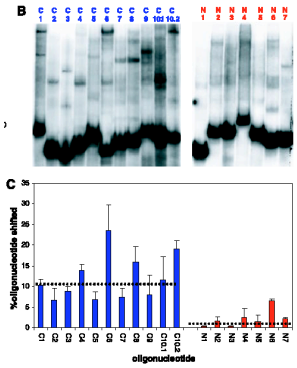
Prediction of regulatory regions in primate specific gene

- Discover conserved and nonconserved elements
- Test function of each by gel shift assay



Functional test

- Generate labeled primers of conserved and non-conserved regions
- Perform gel-shift assay to see if primers bind proteins.



Conclusions

- Closely related species can be useful in identification of conserved regions.
- Alignments are fairly robust
- Classifying sites using likelihood ration identifies exons and conserved regions effectively.

Which method should we use?

- Phylogenetic footprinting
  - Effective at identifying highly conserved features across divergent lineages
- Phylogenetic shadowing
  - Effective at getting taxon-specific elements.

Which species should we get genome sequences to increase power

- Marsupials?
- Reptiles?
- Fish?
- More mammals?
- More primates?