

# Highly Conserved Non-Coding Sequences in Vertebrates

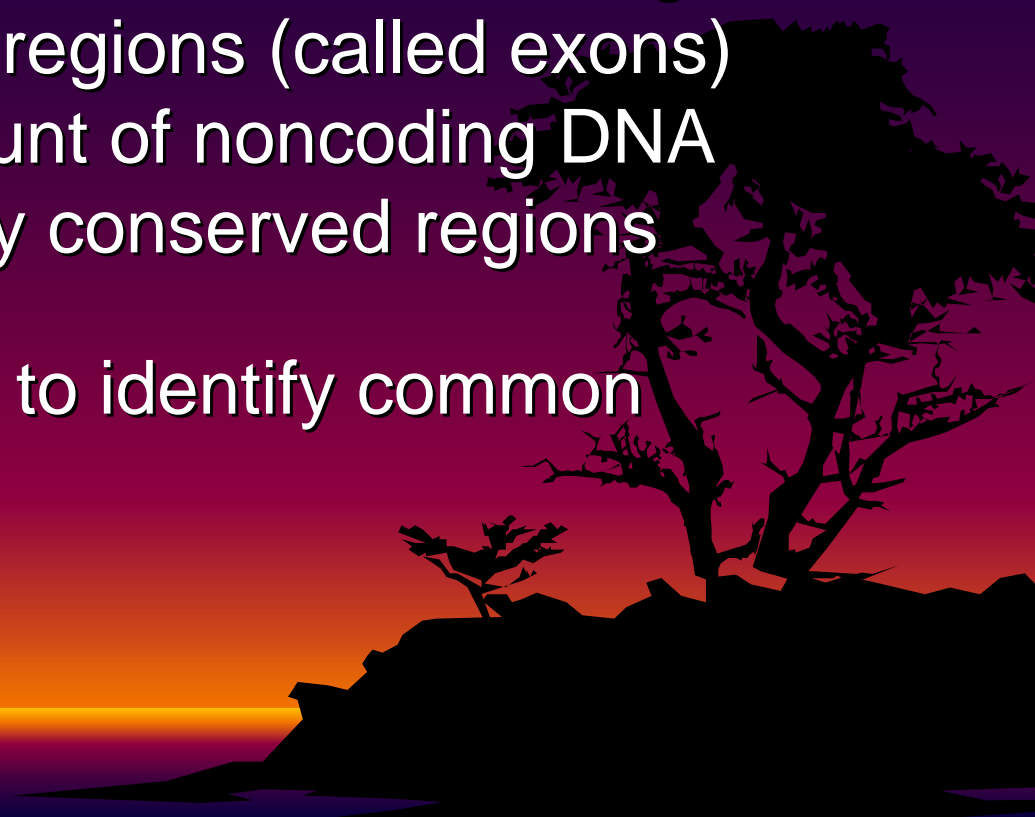
The background of the slide features a gradient from dark purple at the top to bright orange at the bottom, representing a sunset. Silhouettes of several trees are visible against this background, with the largest tree on the right side.

22.05.2006

*“... to turn our focus to those low hanging fruits that nature has already shown us to be important by keeping them untouched in our genome.”*

# Comparative genomics

- Comparative genomics is the analysis and comparison of genomes from different species
- sequence similarity, gene location, the length and number of coding regions (called exons) within genes, the amount of noncoding DNA in each genome, highly conserved regions
- Different organisms -> to identify common signatures



# Selecting species

- Closely related species: obscured functional elements
  - ✓ Species-specific functional elements
- Distantly related species: diverged functional elements
  - ✓ Only the most constrained functional elements can be identified
- “Steering a middle course” (species with moderate distance)
  - ✓ Non-uniform rate of evolution across genomic sequence



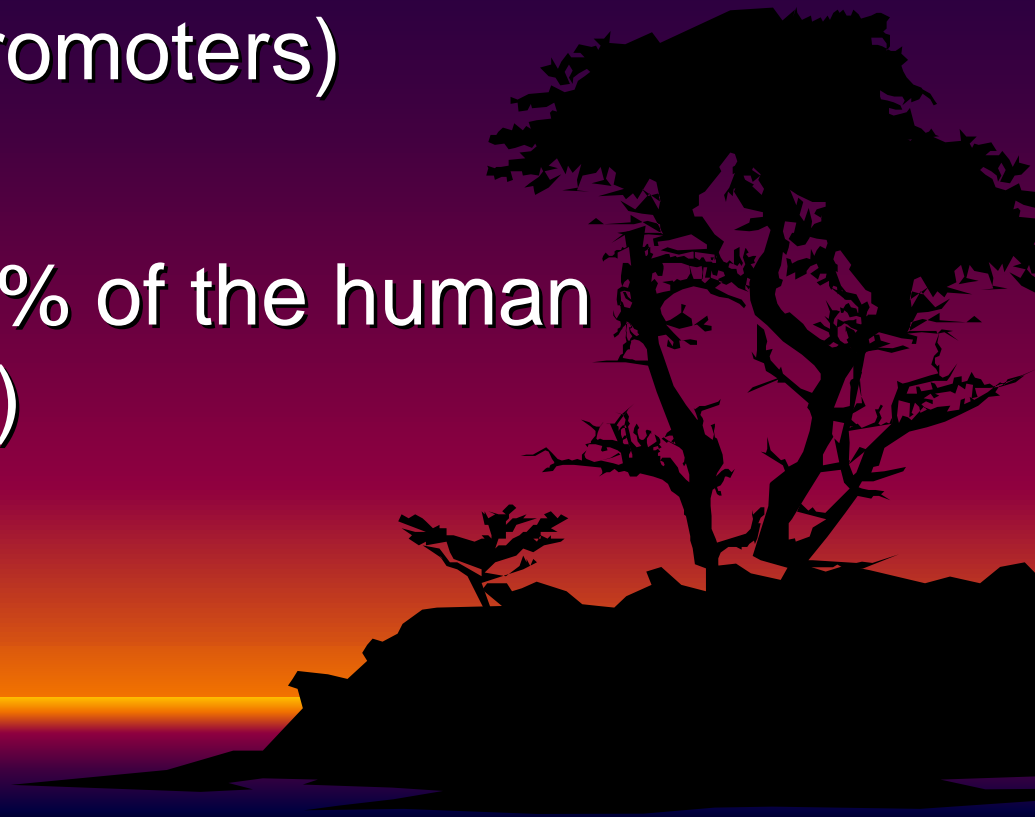
**Table 1.** URLs for Accessing Precomputed Whole-Genome Alignments and Their Analysis

Server or Browser	Genomes Covered	URL
EnteriX	enteric bacteria	<a href="http://bio.cse.psu.edu/">http://bio.cse.psu.edu/</a>
VISTA Genome Browser	human, mouse, rat	<a href="http://pipeline.lbl.gov/">http://pipeline.lbl.gov/</a>
UCSC Genome Browser	mammals, worms, zoo <sup>a</sup>	<a href="http://genome.ucsc.edu/">http://genome.ucsc.edu/</a>
Ensembl	mammals, fish, insects, worms	<a href="http://www.ensembl.org/">http://www.ensembl.org/</a>
GALA	human, mouse, rat	<a href="http://bio.cse.psu.edu/">http://bio.cse.psu.edu/</a>

<sup>a</sup>Data from multiple alignments of 13 vertebrate genome sequences homologous to the human *CFTR* region (Thomas et al. 2003) are included.

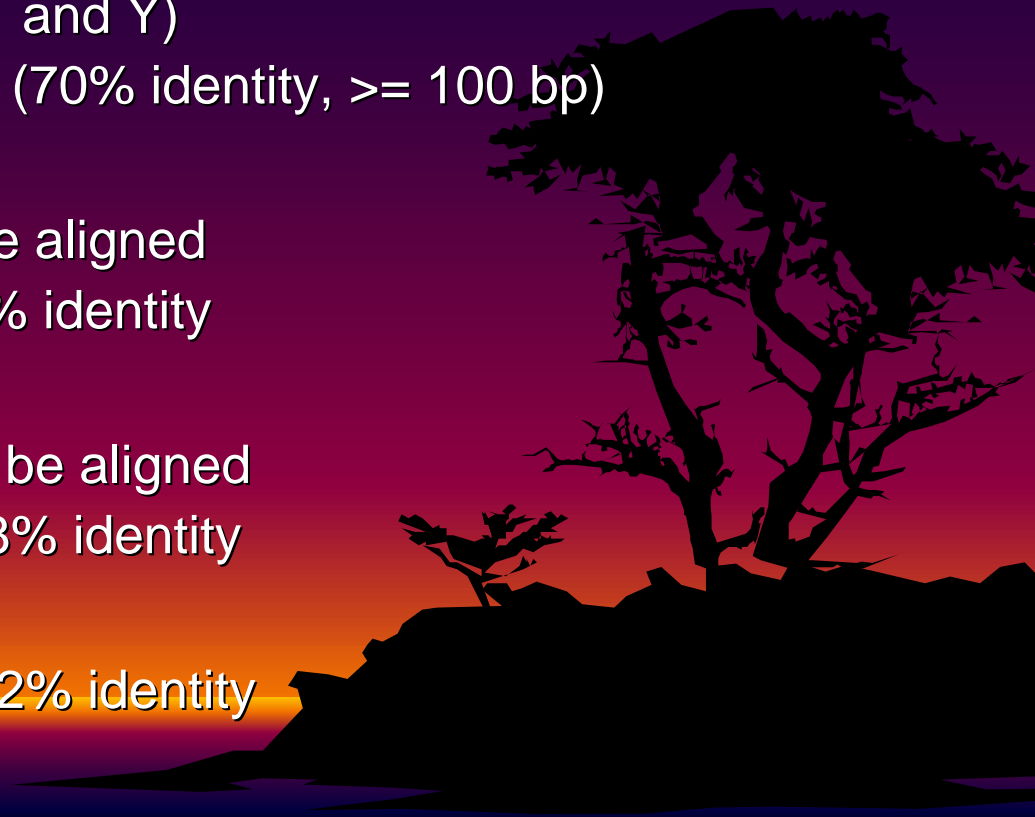
# Non-coding regions

- Intergenic and intragenic (introns)
- Regulation of transcription at the DNA level (enhancers, promoters)
- “gene deserts” – 25% of the human genome (> 500 kbp)



# Species comparison

- Human vs. Mouse
  - ✓ 40% of the genomes can be aligned
  - ✓ ~5% of the human genome is evolved slower than the neutral rate
  - ✓ 481 ultraconserved elements (100% identity,  $\geq 200$  bp) in all chromosomes (except 21 and Y)
  - ✓  $>1000$  conserved elements (70% identity,  $\geq 100$  bp)
- Human vs. Chicken
  - ✓ ~4% of the genomes can be aligned
  - ✓ 97% of 481 align with 95.7% identity
- Human vs. Fugu
  - ✓ ~1.8% of the genomes can be aligned
  - ✓ 67.3% of 481 align with 76.8% identity
- Human vs. Dog
  - ✓ 99.2% of 481 align with 99.2% identity



# Human and Chimp Ancestors Might Have Interbred?

- The earliest known ancestors of modern humans might have reproduced with early chimpanzees to create a hybrid species, a new genetic analysis suggests



# Regulation of genes

- Flanking genes: involved in early developmental tasks (*DACH*, *PAX6*, etc.)

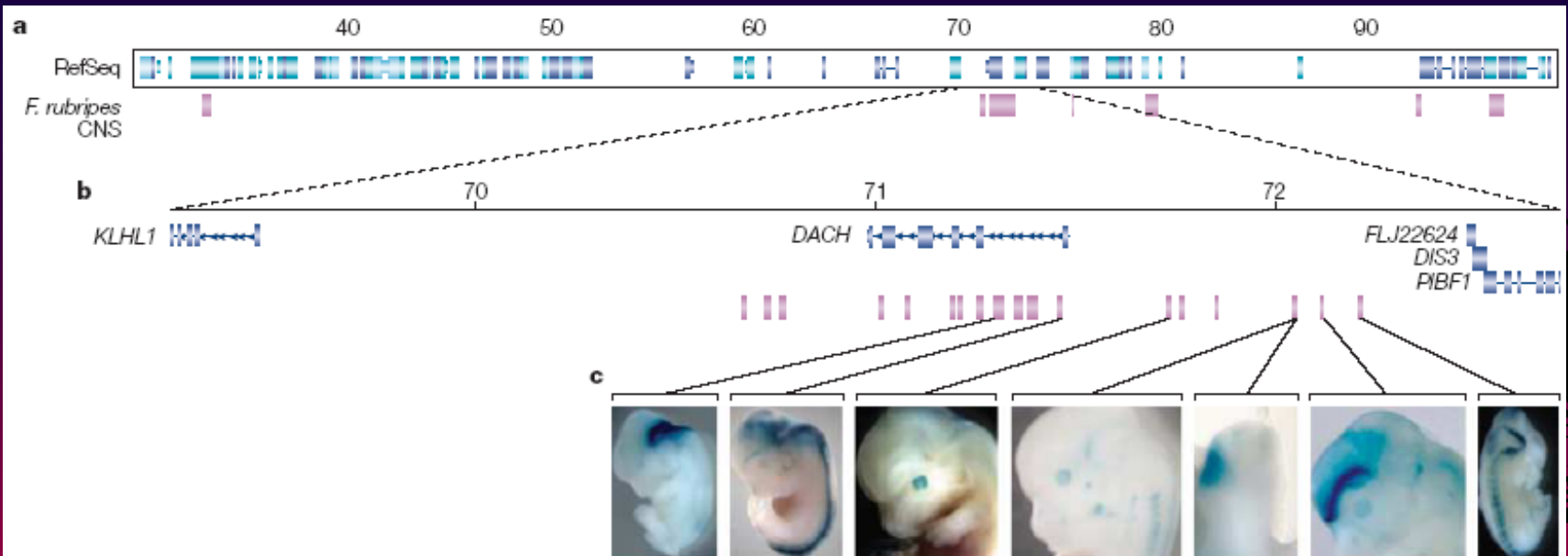


Figure 1 | **Architecture of human-*Fugu rubripes* conserved non-coding sequences in the human genome.** **a** | A 65-Mb segment of human chromosome 13 is shown that contains 145 well-characterized RefSeq genes (exons in blue). There are 51 human-*F. rubripes* conserved non-coding sequences (CNSs) in 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  $\beta$ -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  $\beta$ -galactosidase (in blue).







# Detecting Non-coding regions

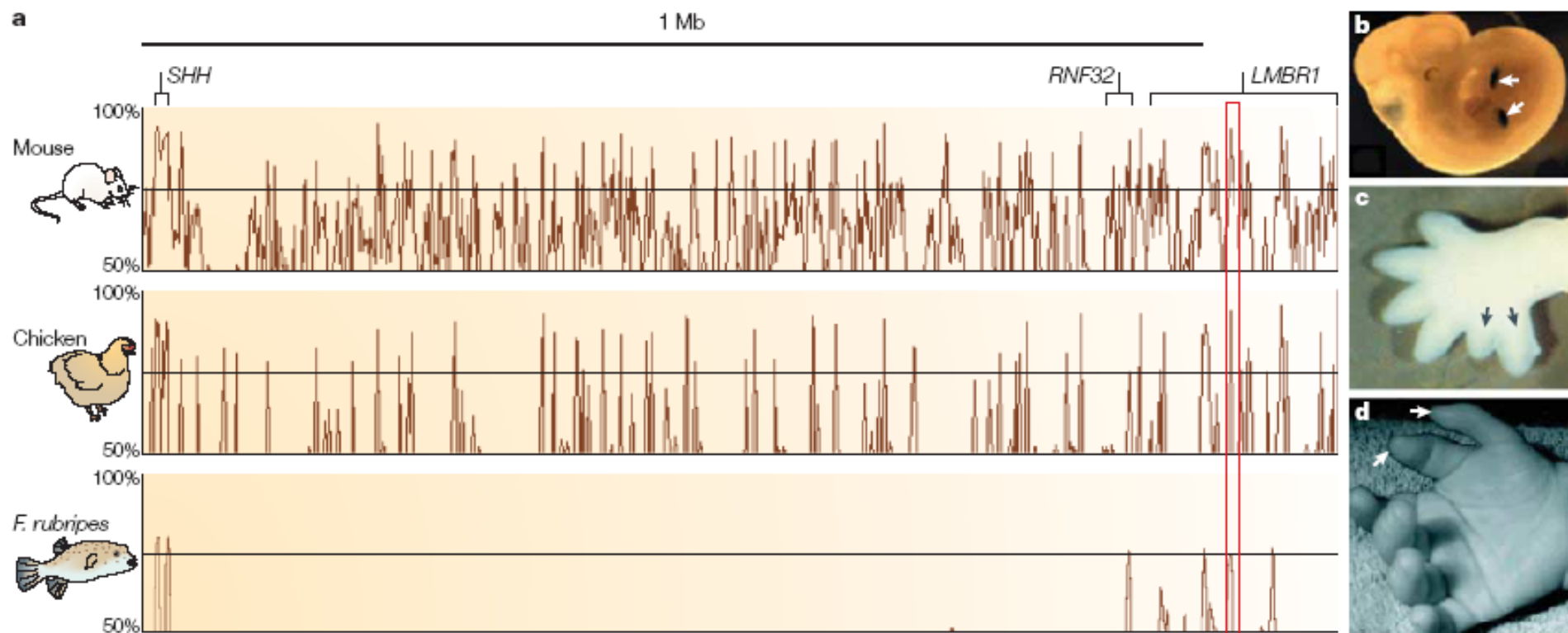


Figure 2 | **Sonic hedgehog expression in the limbs is regulated by an enhancer at a distance of 1 Mb.** a | Human-*Fugu rubripes* sequence comparisons, generated by *VISTA*, identify a conserved non-coding sequence in intron 5 of *LMBR1* (red box), which drives the expression of a reporter gene in a 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 REE 21. © (2003) Oxford University Press and REE 32 (1999) Elsevier Science Ltd.


# Detecting Non-coding regions

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

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

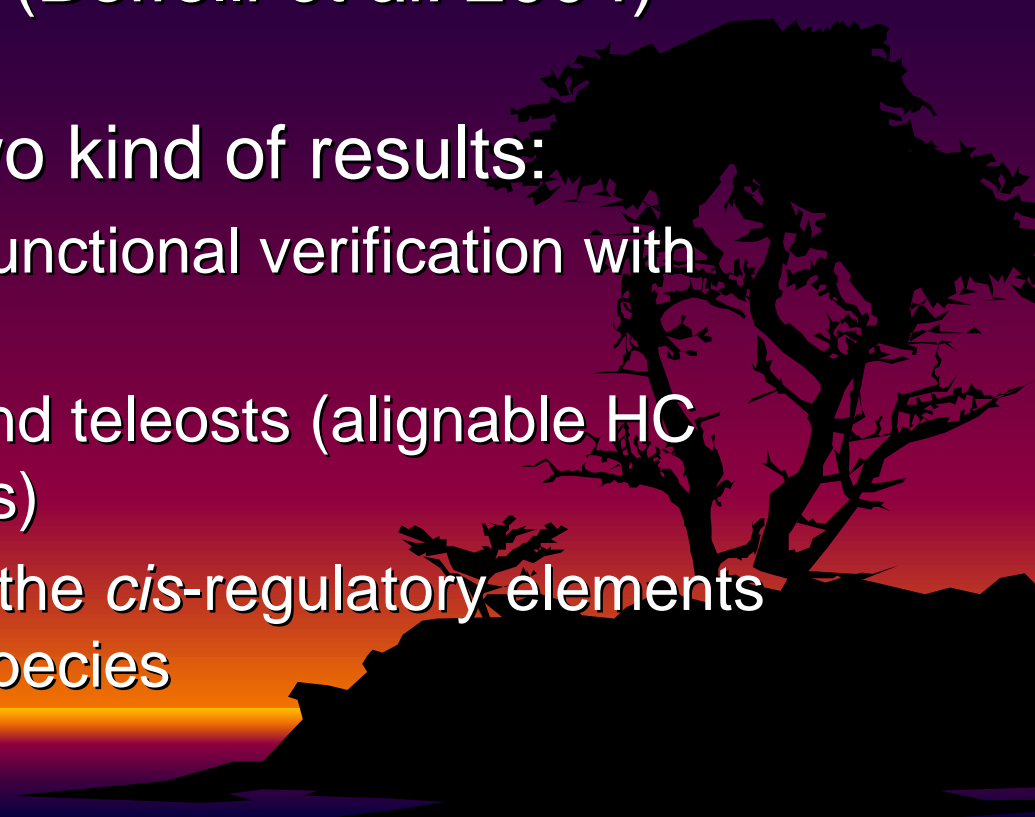
# Detecting Non-coding regions

- Why so slow evolution?
    - ✓ Maybe some structural mechanisms to prevent changes (in/del, subst, invers)
    - ✓ Frequency of SNPs is several-fold lower than in other genomic regions
  - Why only subset of the enhancers are detected in human-fish alignments?
    - ✓ Inadequate simple nucleotide-alignment tools
  - Why are the conserved non-coding regions clustered only around subset of genes among distant species?
    - ✓ crucial for basic vertebrate development
- 
- A silhouette of a tree is visible on the right side of the slide, set against a background of a sunset or sunrise with a gradient from orange to purple.

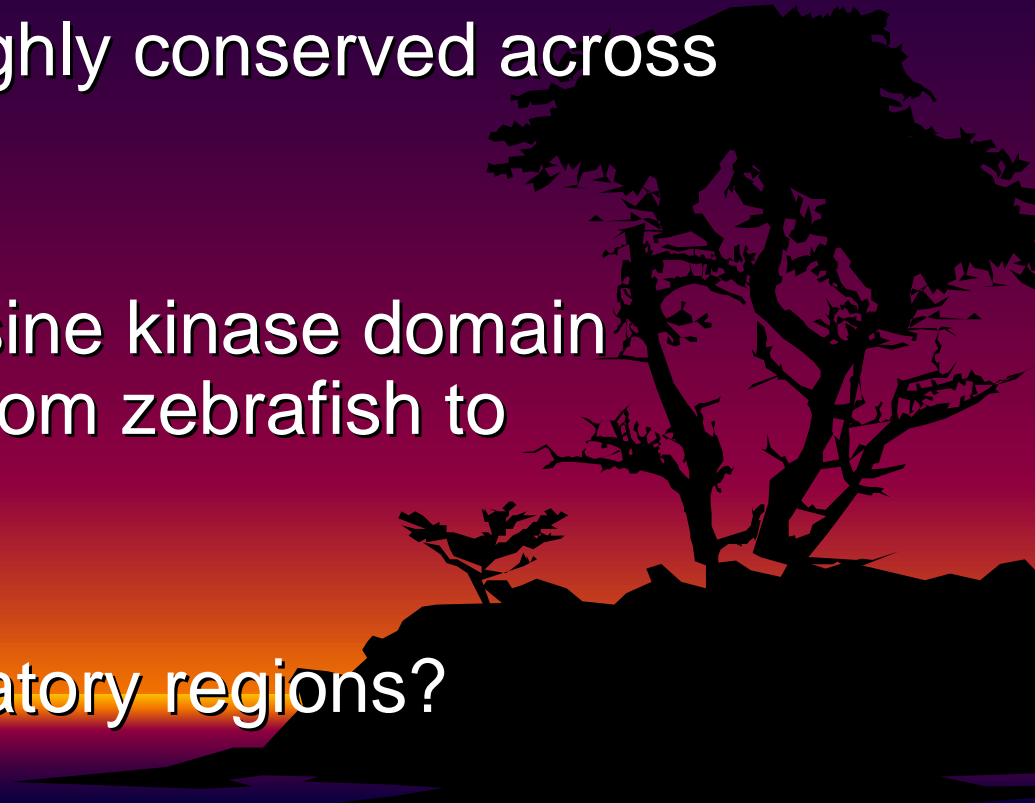
# Exploring the function

- Current hypothesis: sequences conserved over greater evolutionary distances are more likely to be functional than those conserved over lesser distances (Boffelli *et al.* 2004)
- Currently available two kind of results:
  - ✓ between mammals (functional verification with mice)
  - ✓ between mammals and teleosts (alignable HC non-coding sequences)

Those do not show that the *cis*-regulatory elements will function on both species



# Human vs Zebrafish

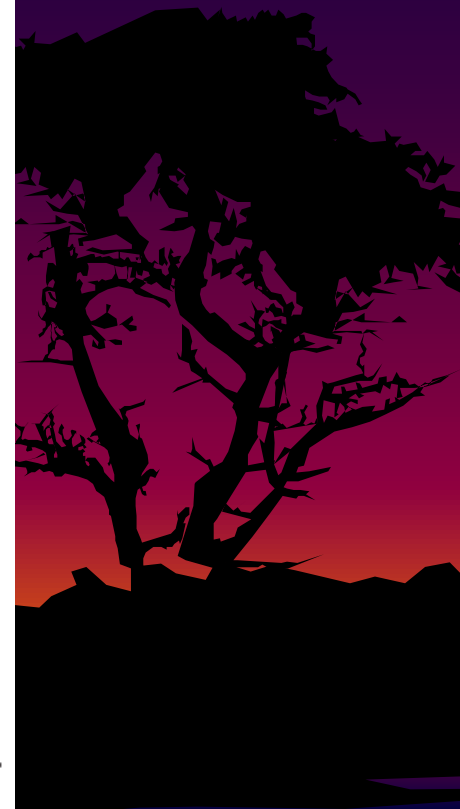
- Fisher *et al.* 2006 focused on the regulatory control of the RET receptor tyrosine kinase
  - RET expression is highly conserved across evolution
  - Exons encoding tyrosine kinase domain  $\geq 70\%$ ,  $\geq 100$  bp (from zebrafish to humans)
  - What about *cis*-regulatory regions?
- 
- A silhouette of a tree is visible on the right side of the slide, set against a background of a sunset or sunrise with a gradient from orange to purple.

# Expression of conserved sequence amplicons

**Table 1.** Noncoding sequences from zebrafish *ret* or human *RET* direct expression consistent with endogenous *ret*. The elements are described by their species of origin and distance in kilobases from the translation start site, and (i.e., ZCS-50, HCS+16). Abbreviations: CG, cranial ganglia; SC, spinal cord; PND, pronephric duct; IM, intermediate mesoderm; NTC, notochord; OLF, olfactory pit/placode; +, present.

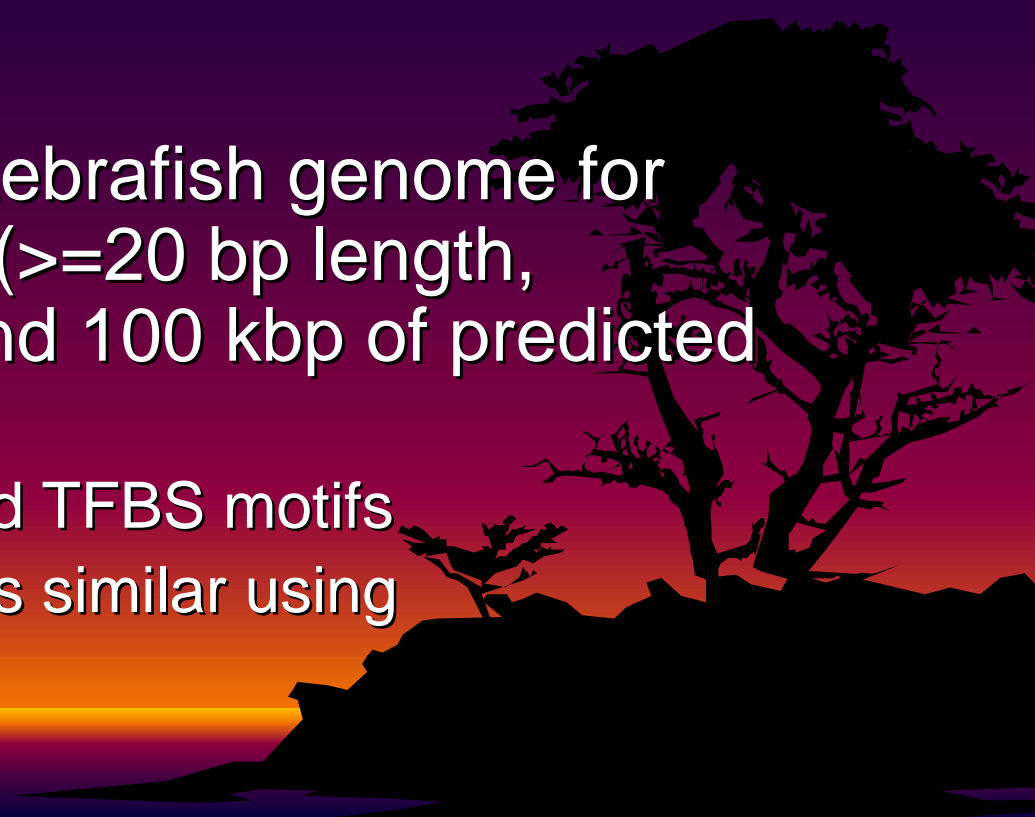
Constructs	Brain	SC	CG	ENS	NTC	OLF	Retina	Heart	IM/PND	Fin bud
ZCS-83	+	+	+			+				
ZCS-50	+	+	+		+			+		+
ZCS-36	+								+	
ZCS-34	+								+	
ZCS-31	+								+	
ZCS-19.7	+	+	+	+			+		+	
ZCS-14.7	+	+								
ZCS-9.5	+	+	+							
ZCS+7.6									+	
ZCS+35.5		+	+	+				+		
HCS-32	+	+	+					+		+
HCS-30									+	
HCS-23					+					
HCS-12		+							+	
HCS-8.7									+	
HCS-7.4									+	
HCS-5.2	+					+			+	
HCS+9.7				+					+	
HCS+16	+									
HCS+19	+	+								

\*Expression before 24 hours.



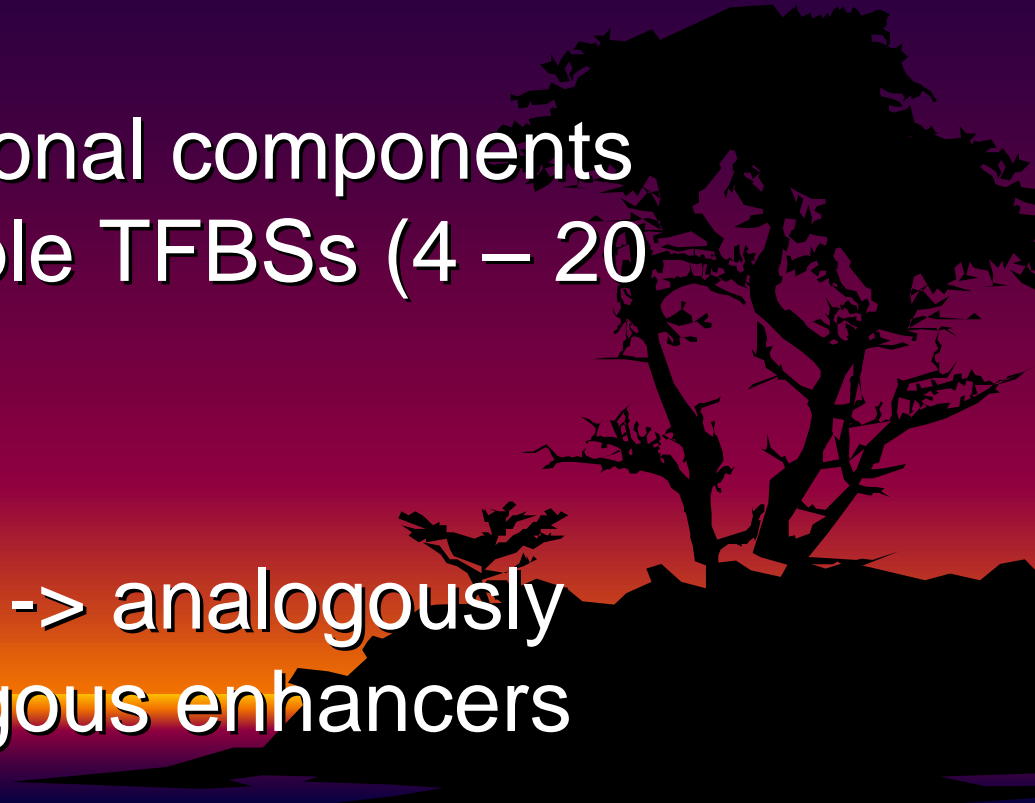
# Smaller elements

- Reduced alignment window size to 30 bp (with Suffle-LAGAN – designed to detect alignments in the presence of inversions and rearrangements)
  - ✓ No results
- Searched the entire zebrafish genome for homologies to HCSs ( $\geq 20$  bp length,  $\geq 70\%$  identity, around 100 kbp of predicted genes)
  - ✓ 41 out of 65 contained TFBS motifs
  - ✓ But the frequency was similar using nonconserved HCS

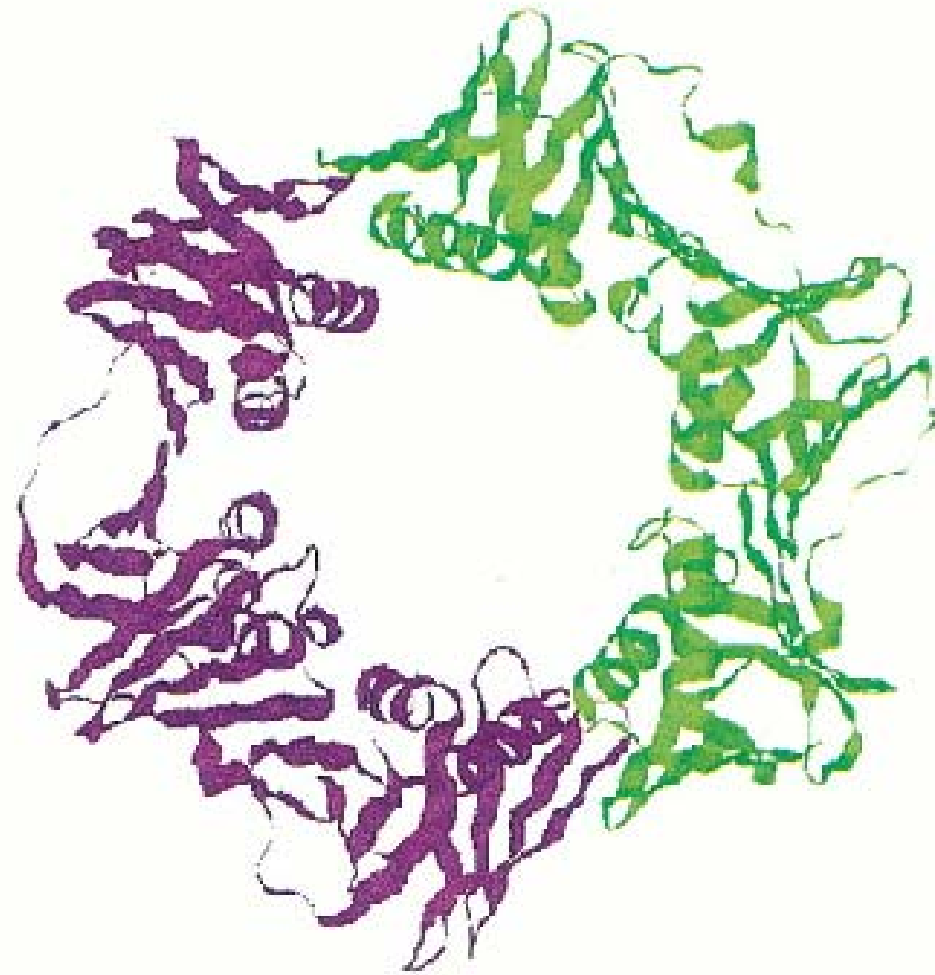
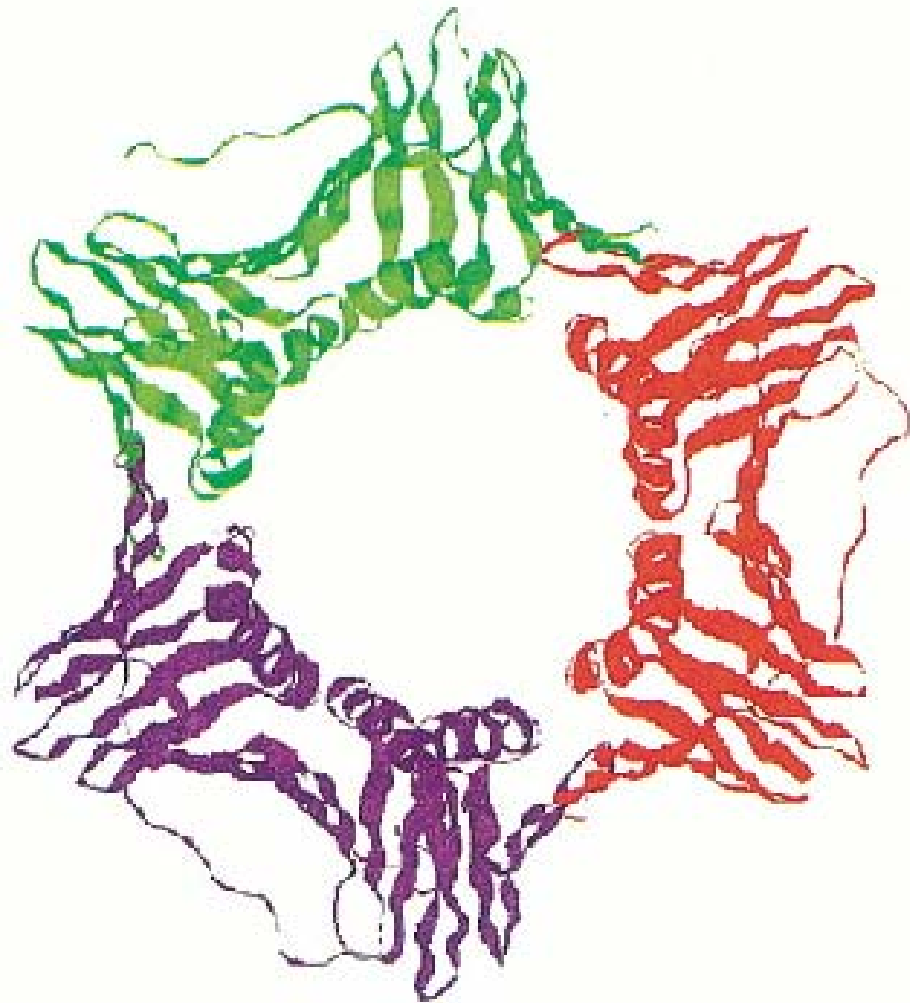




# Conclusions

- Developed efficient method to evaluate putative enhancer elements (in vivo) using Tol2 vectors
  - Responsible functional components are single or multiple TFBSs (4 – 20 bp)
  - Similar expression -> analogously acting, not orthologous enhancers
- 
- A silhouette of a tree is visible on the right side of the slide, set against a background of a sunset or sunrise with a gradient from purple to orange.

# <10% Sequence Identity



Yeast PCNA trimer  
(Krishna et al. 1994)

*E. coli*  $\beta$  subunit dimer  
(Kong et al. 1992)

Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS.  
Conservation of RET regulatory function from human to zebrafish  
without sequence similarity. *Science*. 2006 Apr 14;312(5771):276-9.

