



# What Made Us Human?

## A Case Study in Comparative Genomics

**Katherine S. Pollard**

UC Davis Genome Center & Dept of Statistics

<http://docpollard.com>

UCSF: November 30, 2006

© Copyright 2006, all rights reserved

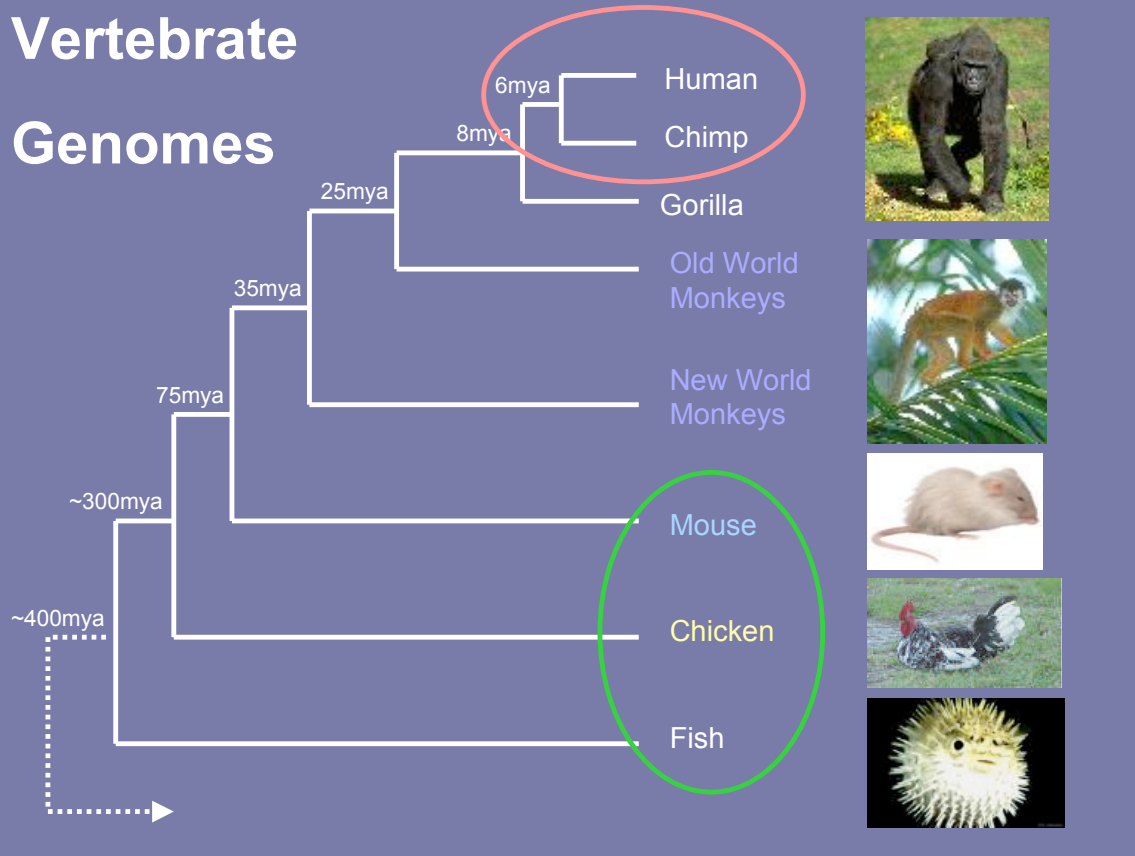
## Chimpanzee: *Pan troglodytes*

Our closest relative on the tree of life  
(MRCA 5-6 million years ago).

	Same	Different
<i>Disease</i>	HIV-1/SIV	AIDS, Alzheimer's
<i>Anatomy</i>	Opposable thumb	Hair, brain size
<i>Behavior</i>	Culture, tools	Agriculture
<i>Language</i>	Sign language	Spoken language

What genetic differences underlie the speciation and subsequent lineage-specific evolution?

# Vertebrate Genomes



## Comparative Genomics

- **Functional** regions are conserved
  - Genes: protein coding, RNA
  - Regulatory sequences: binding sites
  - Physical properties of the DNA: chromatin
- Differences record **evolutionary history**
  - Mutations → Substitutions
  - Loss and gain of DNA (indels)
  - Rearrangements, inversions, duplications

# Comparison to Human Genome

	Chimp	Mouse
<i>Genome size</i>	98.9%	89.7%
<i>Syntenic</i>	98.0%	89.4%
<i>Overall nucleotide identity</i>	95.2%	27.6%
<i>Aligned bases</i>	96.6%	40.0%
<i>Identity at aligned bases</i>	98.8%	69.0%
<i>Identity in genes</i>	99.3%	85.1%

# Substitution Rates

Much of the DNA in eukaryotic genomes is evolving at a **background** (neutral) rate:

human	T	A	A	A	T	G	C	A	C	T	A	T	G	A	A	A	A	A	T	A	A	A	C	A	A	G	C	A	C	A	A	A	A	A	C	A	C	
chimp	T	A	A	A	T	G	C	A	C	T	A	T	G	A	A	A	A	A	A	T	A	A	A	C	A	A	G	C	A	C	A	A	A	A	A	C	A	C
mouse	T	-	-	-	-	-	G	T	T	T	G	A	G	T	A	A	A	A	G	C	A	A	T	C	T	G	A	C	A	A	A	A	A	A	G	C	T	C
rat	T	-	-	-	-	-	G	T	T	T	G	A	G	T	A	A	A	A	G	C	A	A	T	C	A	A	A	T	A	A	A	A	A	A	G	C	T	C

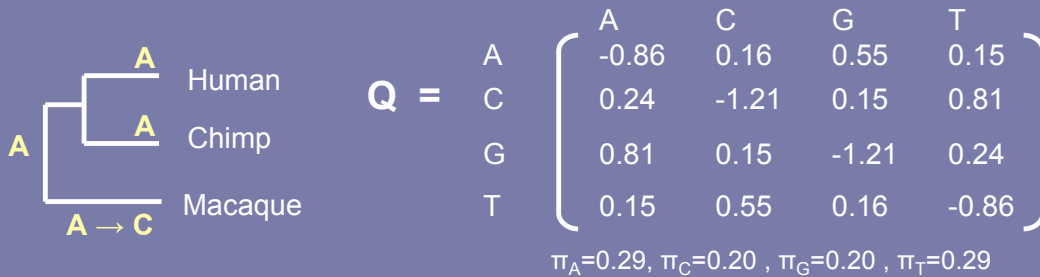
**Negative selection** on functional elements decreases the number of substitutions:

human	A	C	C	A	A	T	C	T	A	A	G	G	T	A	A	T	T	C	A	G	T	T	C	A	T	C	A	C	A	A	A	A	A	A	A	A
chimp	A	C	C	A	A	T	C	T	A	A	G	G	T	A	A	T	T	C	A	G	T	T	C	A	T	C	A	C	A	A	A	A	A	A	A	A
mouse	A	C	C	A	A	T	C	T	A	A	G	G	T	A	A	T	T	C	A	G	T	T	C	A	T	C	A	C	A	A	A	A	A	A	A	A
rat	A	C	C	A	A	T	C	T	A	A	G	G	T	A	A	T	T	C	A	G	T	T	C	A	T	C	A	C	A	A	A	A	A	A	A	A

Other forces increase substitutions...

- Positive selection
- Mutation rate increase

# Models for Molecular Evolution



$$P(t) = \exp(Qt) = \sum_i (Qt)^i / i!$$

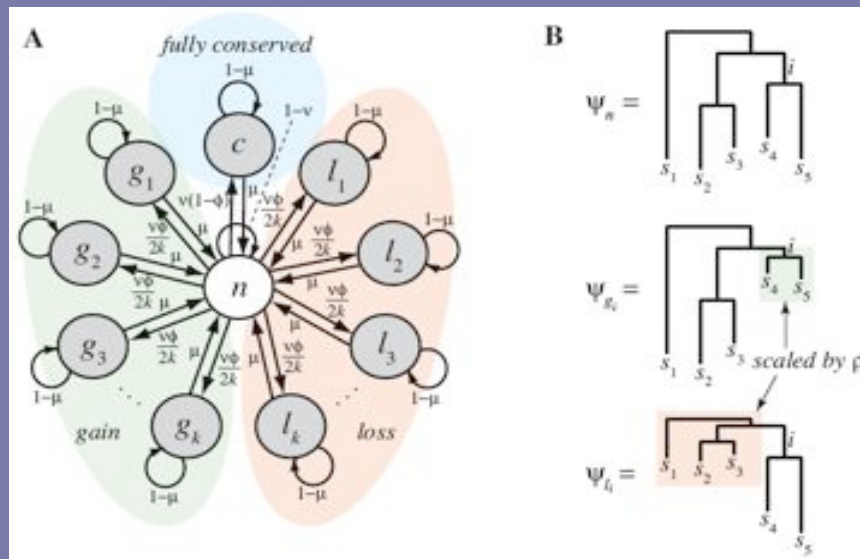
- Q is the instantaneous rate, and P(t) is the probability of a change in time t.
- Assumes substitutions follow a continuous-time **Markov model** (eg: REV, HKY85).
- Q typically scaled so that t = subs/site.

# Lineage-Specific Evolution

human	T	G	T	C	A	G	C	T	G	A	A	A	T	G	A	T	G	G	G	C	G	T	A	G	A	C	G	C	A	C
chimp	T	A	T	C	A	A	C	T	G	A	A	A	T	T	A	T	A	G	G	T	G	T	A	G	A	C	A	C	A	T
mouse	T	A	T	C	A	G	C	T	G	A	A	A	T	T	A	T	A	G	G	T	G	T	A	G	A	C	A	C	A	T
rat	T	A	T	C	A	G	C	T	G	A	A	A	T	T	A	T	A	G	G	T	G	T	A	G	A	C	A	C	A	T

- Changes may occur in only part of the tree.
- Most methods for identifying **substitution rate variation** between genomic regions have little power to detect change in a subtree.
- We have developed two new approaches to this problem, each motivated by a different application...

# 1. DLESS: a phylogenetic HMM



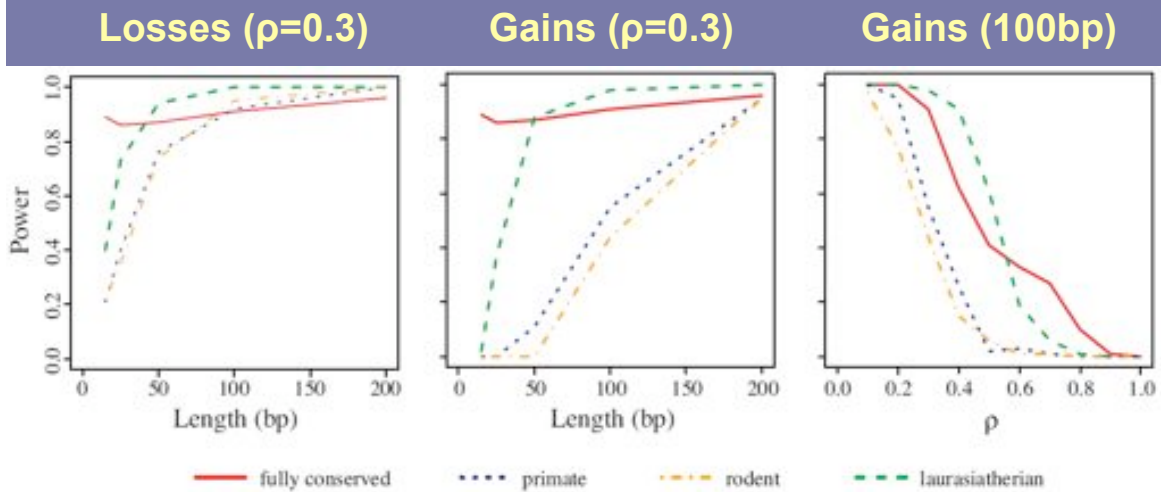
Parameters can be used for tuning or estimated, eg: MLEs.

- $\omega=1/\mu$  is expected length of conserved elements.
- $\gamma=v/(v+\mu)$  is expected fraction of bases in conserved elements.
- $\Phi$  is probability element is lineage-specific given it is conserved.

## Indel Model

- Ignoring alignment **gaps** omits some of the best evidence of lineage-specific effects.
- Instead...
  - Construct **indel history** (inferAncestors program)
  - Compute emission probabilities conditional on this history.
- phylo-HMM that **jointly** emits alignment columns and indel-history columns.
- Equivalent to multiplying usual emission probabilities by an indel history term.

# Simulations: Power of DLESS



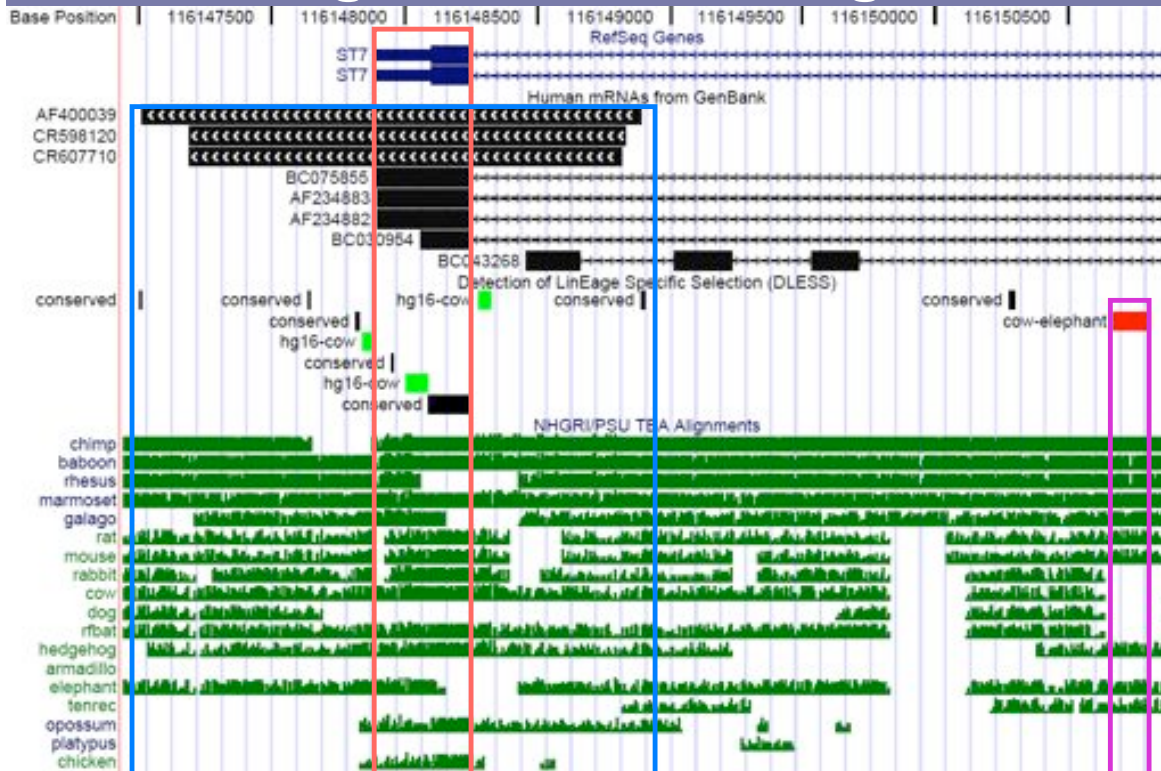
- 18 species phylogeny (ENCODE 4D sites);
- conserved elements (scale model by  $\rho$ ) embedded in 300bp of neutral sequence;
- power=proportion of 100 simulations where conserved element overlaps prediction.

## Analysis of ENCODE Regions

- 19 species, 35Mb of aligned sequence.
- DLESS identifies 24,011 conserved elements.

	Coding sites	Non-coding sites	All sites
Fully conserved	53%	1.7%	3.4%
Lineage-specific losses	13%	0.9%	1.8%
Lineage-specific gains	4%	0.4%	0.5%
Total	70%	3.0%	5.7%

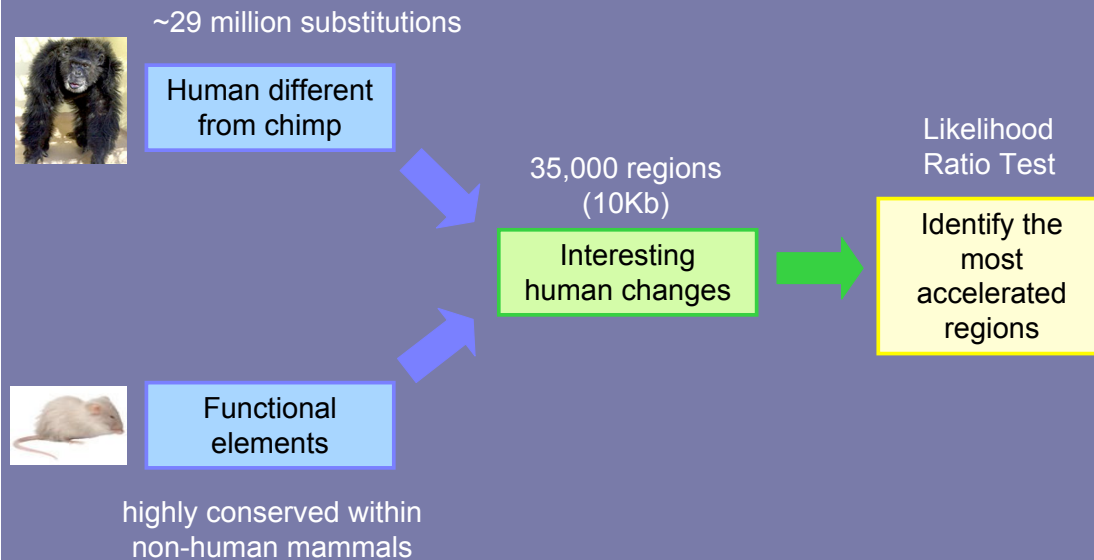
# ST7 gene in CFTR Region



## Summary: DLESS

- **Pros**
  - Element boundaries and lineages of interest need not be specified in advance.
  - Uses information in alignment gaps.
  - Good power on large lineages.
- **Cons**
  - Model is more sensitive to losses than gains, particularly short elements on small lineages.
  - Poor power to detect elements lost or gained on a small lineage (eg: human branch).

## 2. Likelihood Ratio Test



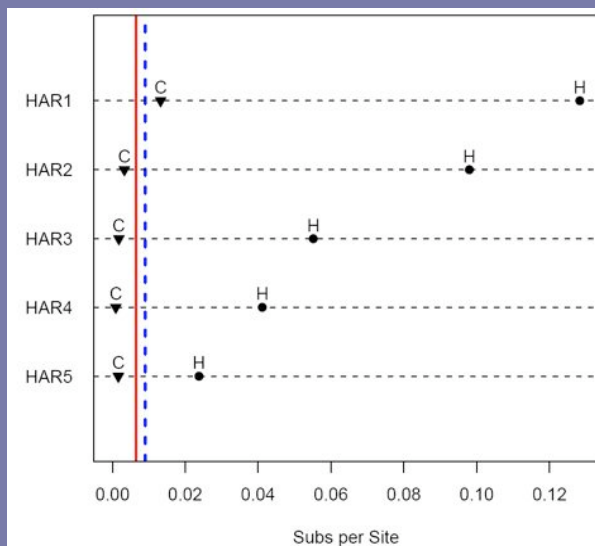
## Likelihood Ratio Test (LRT)

- Model with parameter for acceleration in human vs. model without this parameter
- Fitting issues: use genome-wide REV model and scale it for each region
- P-values by simulation from the no acceleration model (FDR adjusted)
- Omit branches used to define regions → 12 vertebrates plus chimp-human ancestor

## 202 Human Accelerated Regions

- LRT adjusted  $p < 0.1$  (49 with  $p < 0.05$ )
  - Human substitution rate  $>$  chimp rate
  - Also exceeds background (neutral) rate
- Highly **conserved** in vertebrates
  - 99% present in dog with 97% identity,
  - 35% present in fish with 73% identity.
- Mostly **non-coding**: 66% intergenic, 32% intronic, 1.5% coding.
- Nearby genes involved in **transcriptional regulation**.

## HAR1-5: Extreme Acceleration



Background rates from ENCODE 4d sites

red = genome-wide

blue = last chromosome band

- LRT FDR adjusted  $p < 0.0005$ .
- Human substitution rate is **7X** chimp.
- Human rate also exceeds neutral rates, even for subtelomeres.
- Index of dispersion = **9.23** ( $p = 0.018$ ).

# Resequencing HAR1-5

- Five primates (macaque, spider, orang, gorilla, chimp) **agree with chimp assembly**
- Human diversity panel (24 member)
  - HAR1-4: changes **fixed**
  - HAR5: 2 polymorphic changes, 7 fixed

## HAR1

```

assembly tGt caGct gaaat Gat GggCgt agacGcaCgt cagcGgCggaat Ggt t t ct at caaaat Gaa agt Gt t
human    tGt caGct gaaat Gat GggCgt agacGcaCgt cagcGgCggaat Ggt t t ct at caaaat Gaa agt Gt t
chimp    t At caAct gaaat Tat AggTgt agacAcaTgt cagcAgTggaat Agt t t ct at caaaat Taa agt At t
gorilla  t At caAct gaaat Tat AggTgt agacAcaTgt cagcAgTggaat Agt t t ct at caaaat Taa agt At t
orang    t At caAct gaaat Tat AggTgt agacAcaTgt cagcAgTggaat Agt t t ct at caaaat Taa agt At t
macaque  t At caGct gaaat Tat AggTgt agacAcaTgt cagcAgTggaat Agt t t ct at caaaat Taa agt At t

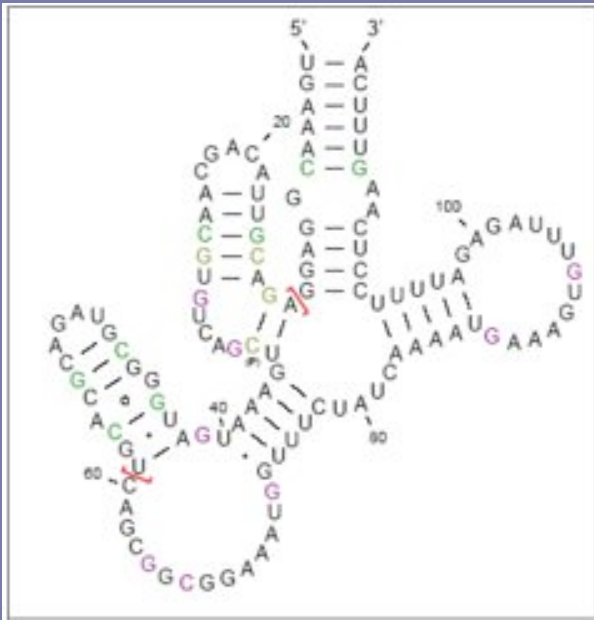
assembly agagatt t t cct caaGt t t caaat GaGgcgaat ccc
human    agagatt t t cct caaGt t t caaat GaGgcgaat ccc
chimp    agagatt t t cct caaAt t t caaat TaTgcgaat ccc
gorilla  agagatt t t cct caaAt t t caaat TaTgcgaat ccc
orang    agagatt t t cct caaAt t t caaat TaTgcgaat ccc
macaque  agagatt t t cct caaAt t t caaat TaTgcgaat ccc
    
```

# HAR1



- 2 mRNAs from human hippocampus
- Reverse strand alt spliced EST from testes
- 118bp with conserved secondary structure

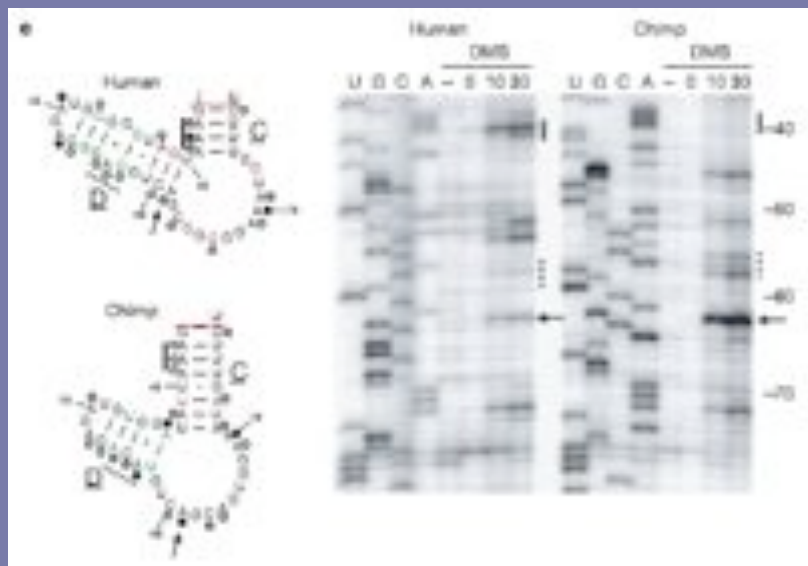
# HAR1: a novel RNA gene



green=compensatory transitions,  
yellow=compensatory transversions,  
purple=substitutions in unpaired regions.

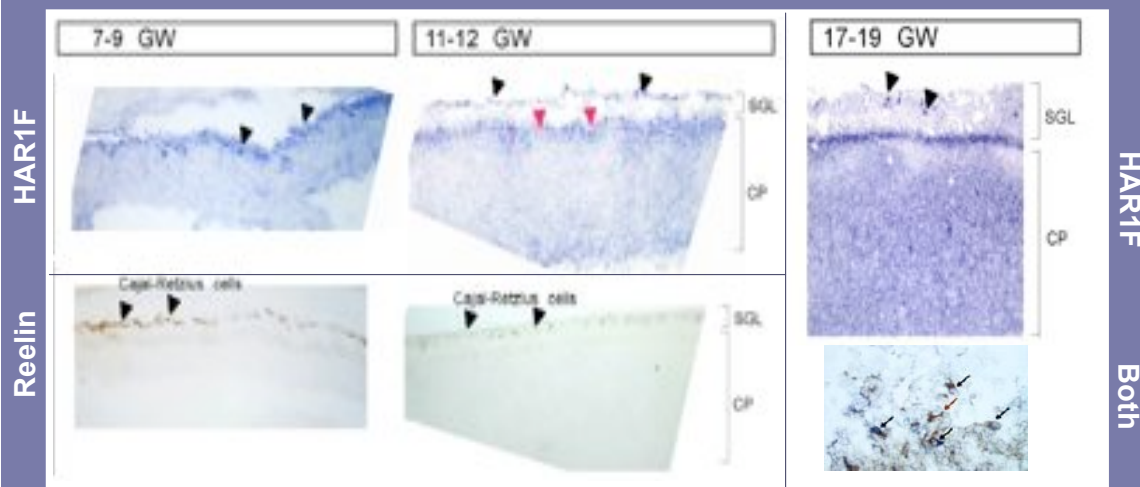
- 10 **compensatory** changes ( $p=4e-8$ ), plus 2 in platypus
- 8 changes in **loops**
- ORFs in transcripts
  - too fast evolving
  - novel proteins
- No convincing miRNA precursors

## DMS Probing of HAR1 Structure



Human changes altered/strengthened the secondary structure.

# HAR1F Embryonic Expression



- Expressed in neocortex from 7 to 19 GW
- Co-expressed with **Reelin** in Cajal-Retzius cells in subpial granular layer and cortical plate

## What Shaped the HARs?

- **Mutation Rate Change/Bias**
  - Hard to test, though estimates of  $\theta$  high in HAR1
- **Drift** (loss of a functional element)
  - Faster substitution rates than neutral (201/202 HARs exceed neutral rate vs. 33/202 in chimp)
- **Directional Selection**
  - scale is about right (~5Kb)
  - conflicting evidence in SNP data
- **Recombination**
  - error prone repair of DSBs
  - BGC to help drive the GC alleles to fixation

# Acknowledgements

**David Haussler** - University of California, Santa Cruz

## Wet Lab

Sofie Salama

Sol Katzman

Bryan King

Courtney Onodera

## Dry Lab

Jakob Pedersen

Adam Siepel (Cornell Univ.)

Andy Kern

## Genome Browser Team

**Pierre Vanderhaeghen** - University of Brussels, ULB

Nelle Lambert

Marie-Alexandra Lambot

Sandra Coppens



Clint

# References

1. KS Pollard, SR Salama, B King, et al. (2006). **Forces shaping the fastest evolving regions in the human genome.** *PLoS Genetics* 2: e168.
2. KS Pollard, SR Salama, N Lambert, et al. (2006). **An RNA gene expressed during cortical development evolved rapidly in humans.** *Nature* 443: 167-172.
3. A Siepel, KS Pollard, D Haussler (2006). **New methods for detecting lineage-specific selection.** *RECOMB'06*.
4. The Chimpanzee Sequencing and Analysis Consortium (2005). **Initial sequence of the chimpanzee genome and comparison with the human genome.** *Nature* 437: 69-87.

Software: **phast** library of C functions available by request from Adam Siepel [acs4@cornell.edu](mailto:acs4@cornell.edu). Coming soon in R.

## Future Work

- Functional characterization of HAR1F and HAR1R genes:
  - Interaction with Reelin or other proteins
  - Affect of human specific changes
  - Anti-sense regulation
- Investigation of other HARs
  - Nearly half are adjacent to a human disease gene, half adjacent to a transcription factor
  - Do human changes affect gene regulation?
- Accelerated evolution in other lineages

## phyloP: Assessing Significance

1. We derived a **general solution** for the null distribution of the number of substitutions at one site on a branch of length  $t$ .
2. From this we can compute the distribution of the number of substitutions in a phylogeny over an alignment of length  $L$ .
3. Computations use **convolutions** and **dynamic programming**.
4. For lineage-specific elements, this is a **bivariate** distribution (subtree, supertree).

# Alternative Methods

- **Relative rates test (RRT)**, chimp vs. human
  - use chimp-human ancestor to define regions
  - power issue
  - doesn't account for genomic context
- **K statistic**  $\epsilon (0,1)$  = likelihood over the RRT rejection region, supertree vs. human
  - uses a fitted REV model for conserved seqs
  - discrete (points just outside the rejection region get counted as 0)
- Top hits are insensitive to the method...

# HAR1 Orthologs

## HAR1 region

13 amniotes  
(1 copy each)

## HAR1F

primates  
dog  
cow

## HAR1R

chimp  
macaque



High divergence of primates from human HAR1F extends through first exon and splice site.

## HAR1 Adult Expression (qPCR)

Sample	HAR1F <sup>a</sup>	HAR1R <sup>a</sup>
cerebral cortex total RNA	1.0 <sup>b</sup> (0.77, 1.29)	0.095 (0.054, 0.17)
frontal lobe total RNA	1.2 (0.70, 2.0)	0.12 (0.067, 0.22)
temporal lobe total RNA	0.72 (0.55, 0.95)	0.049 (0.029, 0.083)
parietal lobe total RNA	0.77 (0.61, 0.98)	0.083 (0.052, 0.13)
occipital pole total RNA	1.10 (0.92, 1.31)	0.12 (0.065, 0.21)
insula total RNA	0.91 (0.62, 1.32)	0.078 (0.045, 0.13)
hippocampus total RNA	0.65 (0.44, 0.96)	0.051 (0.031, 0.087)
pons total RNA	0.51 (0.35, 0.76)	0.12 (0.094, 0.14)
medulla oblongata total RNA	0.39 (0.27, 0.56)	0.043 (0.025, 0.074)
fetal brain total RNA	0.14 (0.11, 0.18)	0.003 (0.002, 0.005)
brain total RNA	0.96 (0.71, 1.3)	0.024 (0.015, 0.039)
testes total RNA	0.12 (0.10, 0.15)	0.12 (0.047, 0.31)
thalamus polyA RNA <sup>c</sup>	4.9 (3.2, 7.5)	0.51 (0.27, 0.98)
hypothalamus polyA RNA <sup>c</sup>	4.8 (3.5, 6.6)	0.54 (0.30, 0.96)

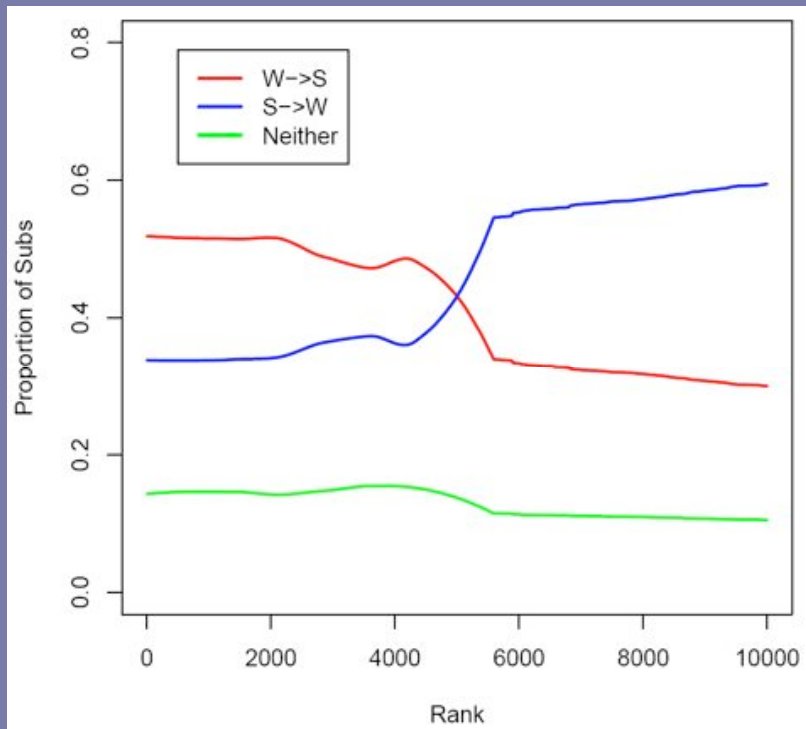
- HAR1F 10X HAR1R in adult brain, 50X in fetal brain.
- Equal levels in adult testes (HAR1Rb only in testes).
- Neither present in other tissues (adrenal, bladder, breast, colon, liver, pancreas, placenta, skeletal muscle, and thymus).

## Biased Gene Conversion (BGC)

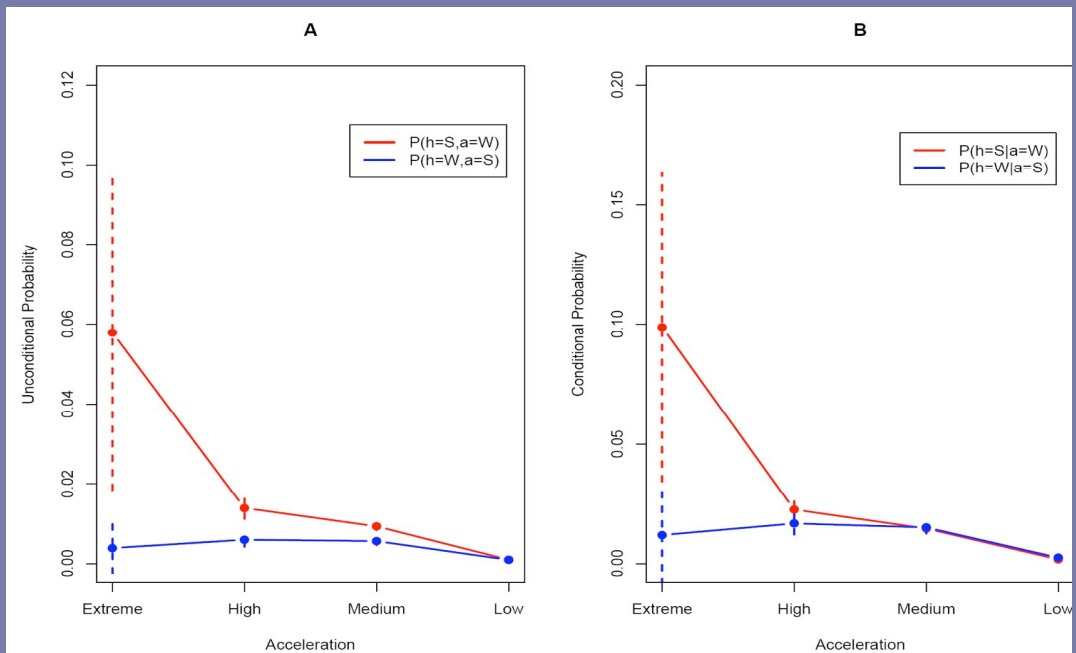
A recombination driven process that leads to increased fixation of GC alleles.

- HAR1-49 are 3x more likely to fall in **last band** of chromosomes vs. all conserved.
- 35/43 human subs in HAR1-5 are **W→S** (among most biased regions genome-wide).
- W→S biased windows ~1Kb.
- 57% W→S vs. 29% S→W in HAR1-202.
- Current **recombination rates** slightly high.

# W→S Bias



# W→S Bias



Bias is partially due to ancestral base composition.

## Directional Selection

- HAR1 & HAR2: Reduced polymorphism relative to divergence vs. surrounding 1Mb
    - HKA (Hey pvals) and a coalescent-based test
    - significantly accelerated blocks are ~5Kb.
    - ascertainment bias is an issue with dbSNP
  - Resequenced 6.5Kb around HAR1
    - no skew in the frequency spectrum
    - lacks a good (sequenced) control region
- Weak (small footprint) or >250,000 years ago

## What Forces Shaped the HARs?

- **Recombination**
  - error prone repair of DSBs
  - BGC to help drive the GC alleles to fixation
  - scale is > 200bp (mean track length in humans)
- **Selection for G+C content**
  - Bernardi isochore theory (scale is « 100Kb)
  - increased gene expression (Kudla et al.)
- **Selection for fitness-increasing changes**
  - scale is about right (~5Kb)
  - conflicting evidence in SNP data