Human Genetic Uniqueness

Comp. Bio. C293: Lunch Seminar
Wednesday, 03 October 2018

Nima Hejazi &
Hector Roux de Bézieux

What makes human unique?

Aristote thought it’s the hand and opposable thumbs that made human unique

Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution

2014

Melissa J Hubisz and Katherine S Pollard

Outline

1. Motivation and reminders

2. Definition of Human accelerated regions(HAR)

3. Timescale of HAR

4. Characteristics of HAR

5. Limitations and criticisms

6. Future direction

Reminder on phylogenetic trees

Assumptions:

- Constant effective population size
- Neutral mutations and therefore constant mutation rate

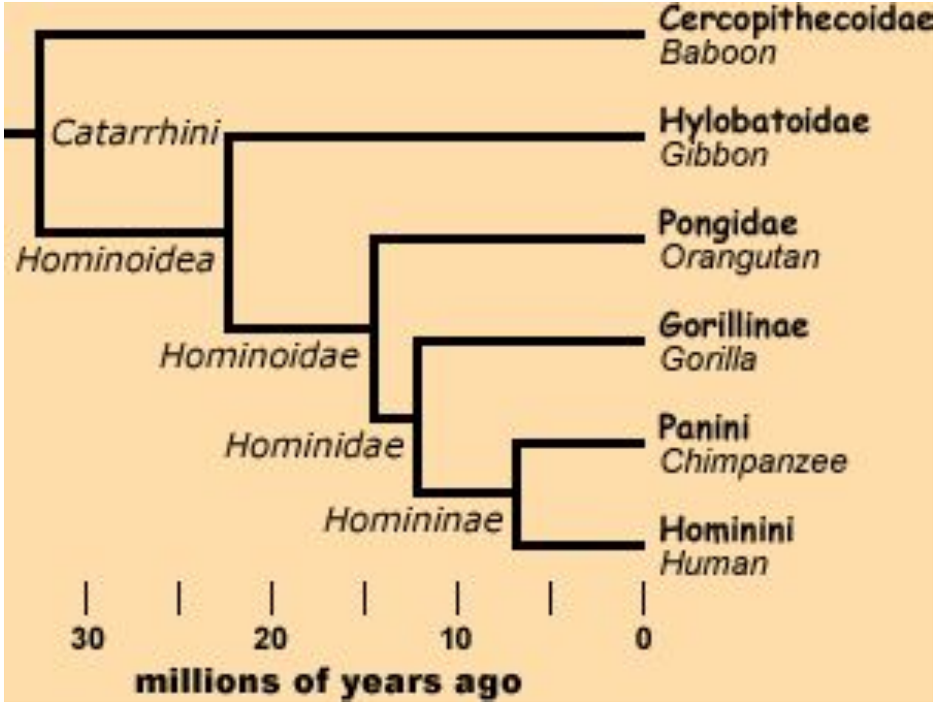
You can build trees for individuals and for species, and you can infer the rate μ

Reminder on phylogenetic trees

Assumptions:

- Constant effective population size
- Neutral mutations and therefore constant mutation rate

You can build trees for individuals and for species, and you can infer the rate μ



Motivation

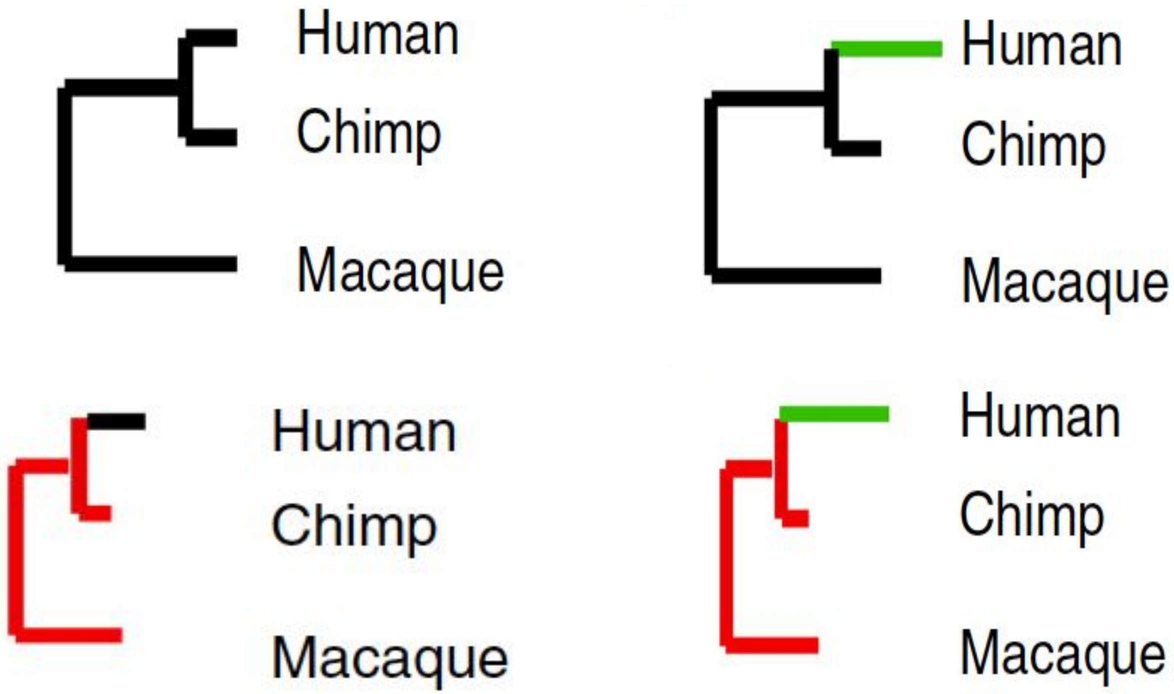
Assumptions:

- Constant effective population size
- Neutral mutations and therefore constant mutation rate

You can build trees for individuals and for species, and you can infer the rate μ

Is μ constant across the genome and across times? Of course not: selective pressure

Impact of different μ on the tree shape



Human Accelerated Regions (HAR)

- Defined broadly as “regions with drastically increased substitutions rates

- HAR regions are associated with a given time:
  - Specific to a lineage: human species
  - Specific to a comparison: human versus close apes

- Regions with newly high positive selection, or less negative selection.

How to identify HAR’s

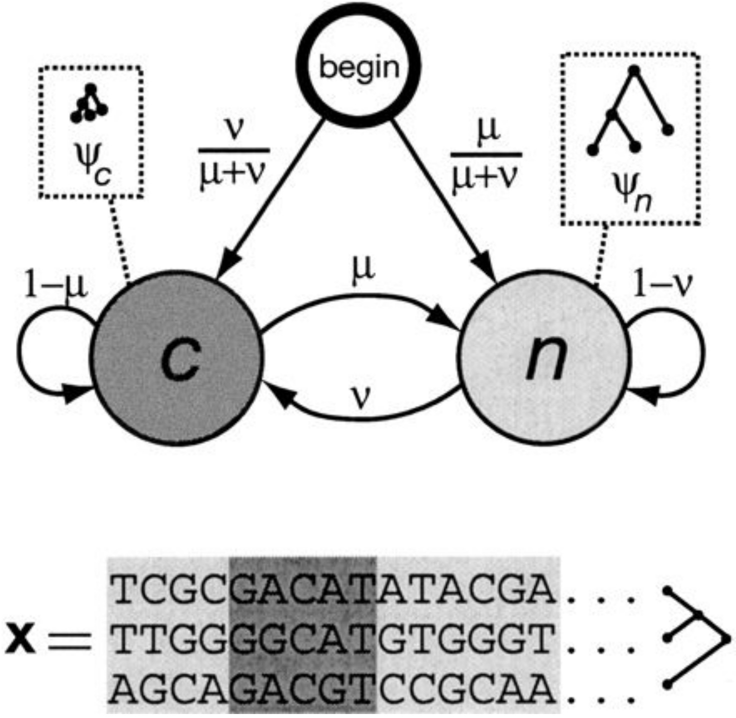
- Use a phylogenetic tree (either known or computed)

- Compare sequences and look at increased divergence
  - Can be proteins, rRNA, … with more or less complex distance functions
  - Here, DNA sequences (only tool for genome wide analysis)

- 100 bp long sequences

Use phastCons to obtain conserved regions

- Use the tree and the sequence alignment to compute conserved regions (phastCons)



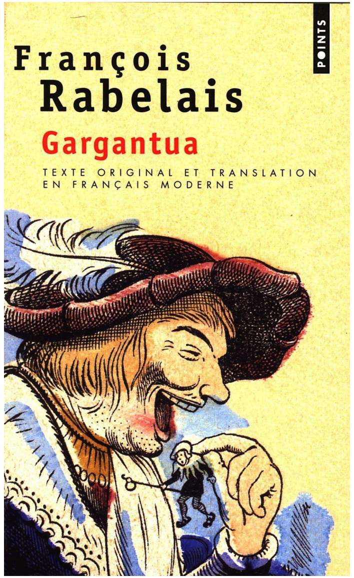
E. H. Margulies, M. Blanchette, NISC Comparative Sequencing Program, D. Haussler, and E. D. Green, 2003

Use phyloP to obtain HAR’s

- Use the tree and the sequence alignment to compute conserved regions (phastCons)
- On those conserved regions, estimate a null distribution for substitutions (for each region, using all sequences).
- Compute the actual number of mutation in every regions
- Get a p-value.

Siepel, Pollard, and Haussler (2006)

What makes human unique?



“Le rire est le propre de l’homme”
(laughter is mankind’s province)

Gargantua, Rabelais (1534)

Non-coding HAR

- 2701 ncHAR representing 96.6% of all HAR found in the genome
- Mean substitution in human of 1.7 per 100 bp, compared to 0.2 in other species for those regions
- Higher than other conserved regions
- Higher than flanking regions(which tend to be conserved as well)
- 3 times the neutral rate

Timescale of HAR: when did mutation occur?

Comparison with Neanderthals, Denisovans, and apes (for ncHAR):

- 7.1% of mutations are human specific

- 2.7% are shared

- More likely to be from before human-divergence than the rest of the genome (conserved and flanking regions)

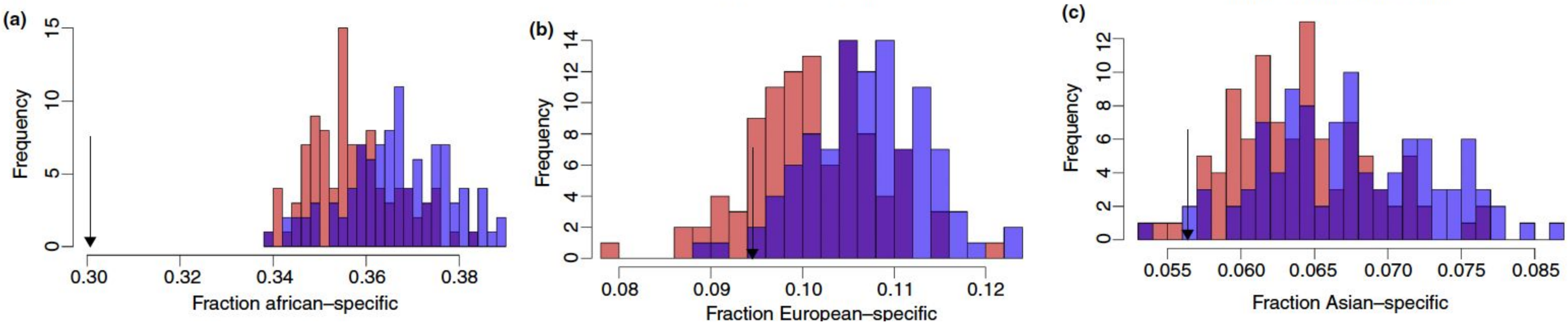
“depletion of accelerated evolution in the past 1 million years of human evolution compared to earlier in our lineage.”

Emergence of polymorphisms across species

Polymorphic rates in autosomal ncHAR from 54 modern human

- Most ncHAR mutations are fixed.
- Same as flanking regions, higher than conserved regions
- Much more archaic than other polymorphic sites

Emergence of polymorphisms in populations



🟥 flanking regions

🟦 conserved regions

→ ncHAR regions

ncHAR polymorphism are less population specific than others

They appeared before divergence

Why is there more polymorphism in HAR?

- Does the test identify polymorphisms as HAR? It does not seems so
- Past positive selection that increased some frequencies
- Relaxation of constraints in the past

➔ Future work is needed

What makes human unique?

“The human animal differs from the lesser primates in his passion for lists”

H Allen Smith

Characterization of HAR

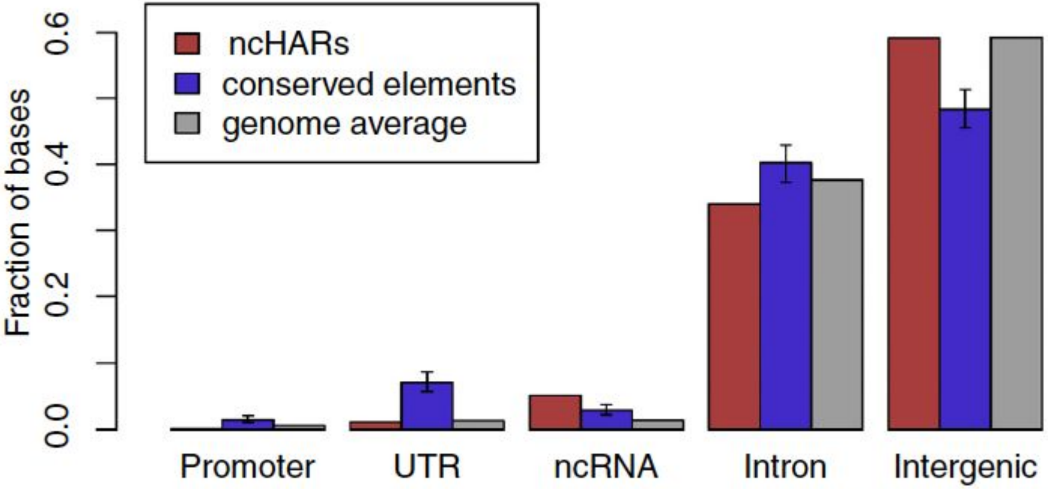
- Rate is 3 times higher than in neutral selection model: evidence for positive selection

- Drive the difference between chimp and humans

Location in the genome

- More likely near developmental genes, transcription factors and genes expressed in the central nervous system

- More likely to be a coding region than average in the genome but less than other conserved regions



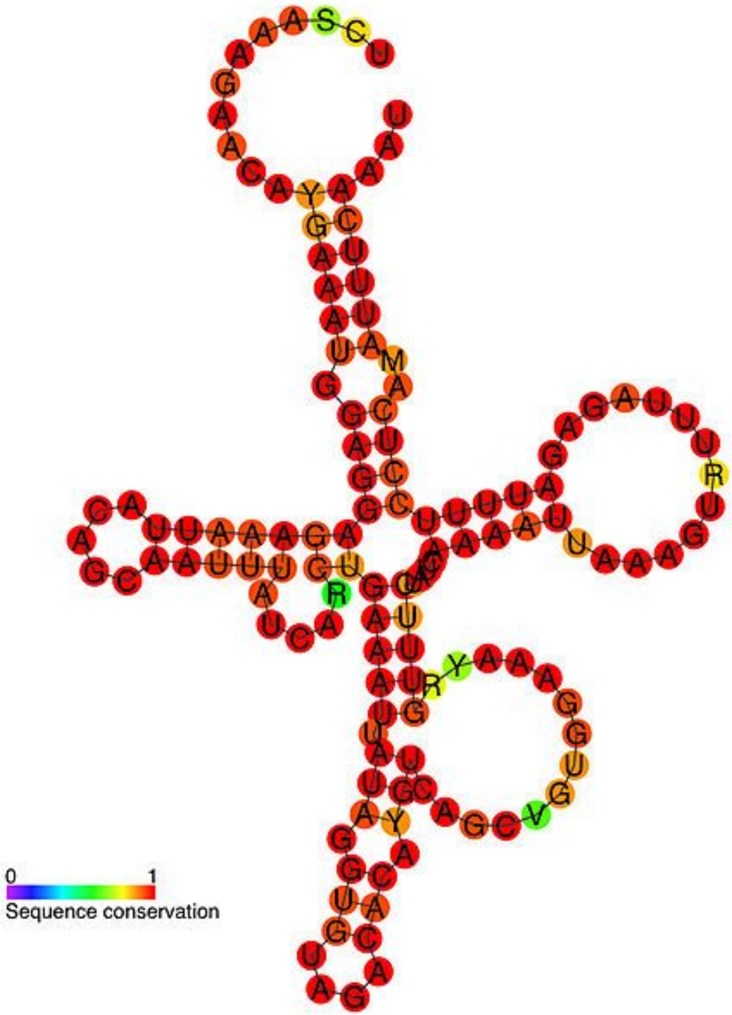
Current Opinion in Genetics & Development

HAR functions

- Non-coding RNAs including HAR1. HAR1A plays a role in development during 8th and 16th week, HAR1B is expressed in the brain.

- In general, gene expression enhancer in embryogenesis

- Some drive human-specific embryogenesis

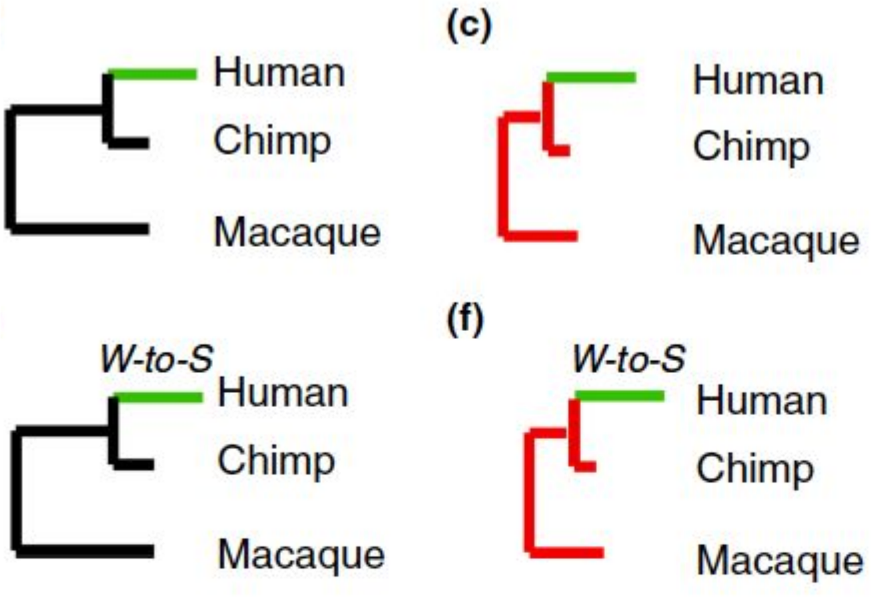


GC bias and HAR

HAR have more A/T to G/C substitutions than usual.

New tools to distinguish between the three models:

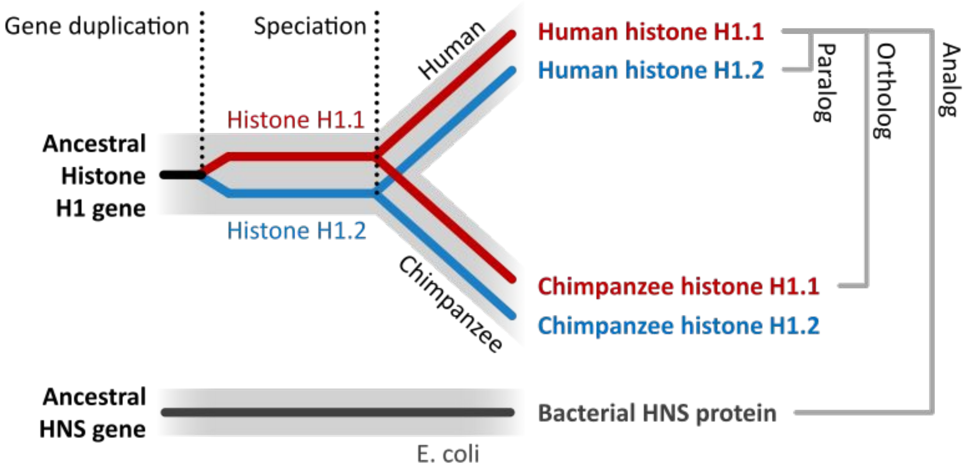
- 20% GC bias
- 20% relaxation of negative pressure
- 60% positive pressure



Limitations of HAR

- Most (90%) of differences between are structural variations

- Paralogous regions pose a problem in alignment and assembly: For now they have to be discarded



Key takeaways

- HAR are regions with lots of substitutions compared to other conserved regions
- In human versus apes, they tend to be quite old, driving the difference between hominids and apes
- Usually caused by positive pressure
- Usually implicated in development
- Only represent part of the differences between humans and apes

Thank you for listening

I am fond of pigs. Dogs look up to us. Cats look down on us. Pigs treat us as equals.

Winston Churchill

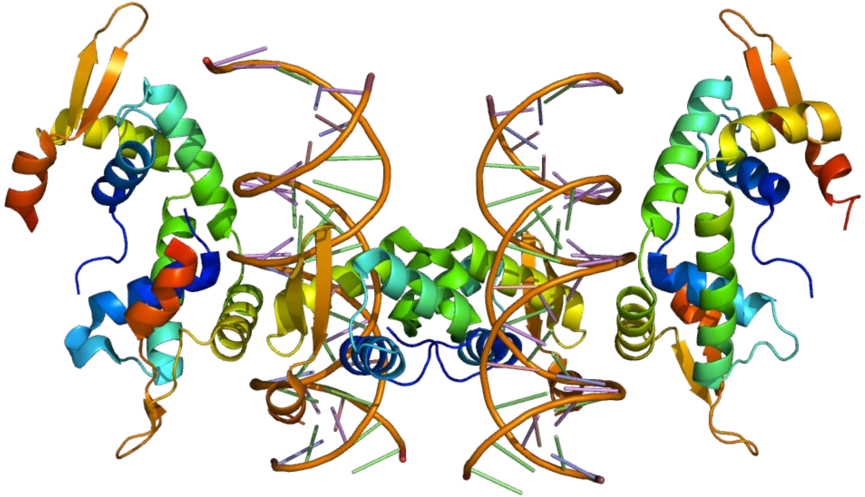
Molecular evolution of *FOXP2*, a gene involved in speech and language (Enard et al., *Nature*, 2002)

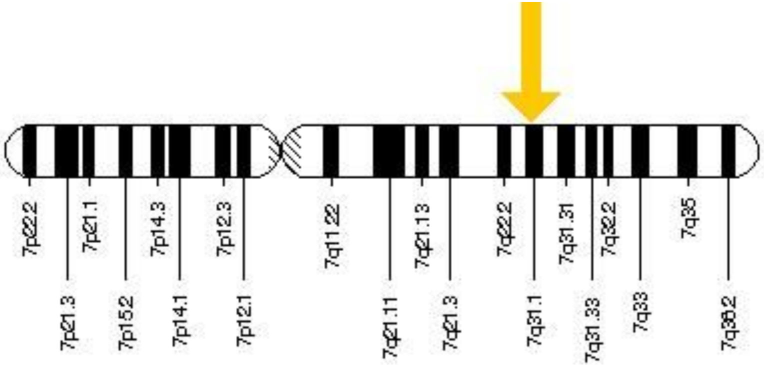
Outline

1. Molecular Biology of FOXP2
2. Comparative Genetics of FOXP2
3. Tracing Genetic History of FOXP2
4. Detection of a Selective Sweep
5. Disease Phenotypes and Evolution
6. Discussion / Conclusion

A Bit of Molecular Biology

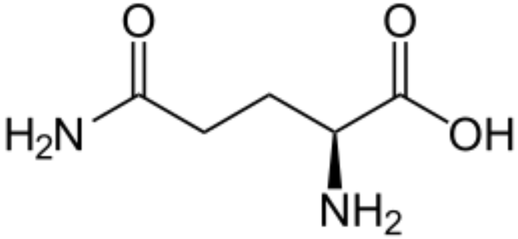
- FOXP2 - “forkhead box P2,” located on human chromosome 7 (7q31).
- Major splice form encodes a protein of 715 amino acids.
- Belongs to the forkhead class of transcription factors.

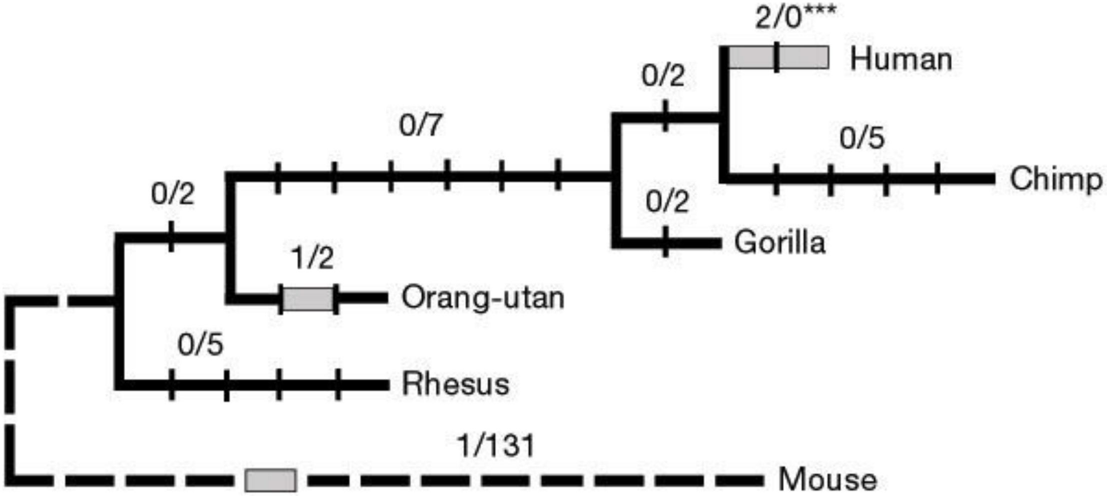




A Bit More Molecular Biology

- Contains glutamine-rich region consisting of two polyglutamine tracts.
- These regions have been shown to have elevated mutation rates.
- In FOXP2, lengths of these stretches differ for all studied taxa.





FOXP2 and language disorders

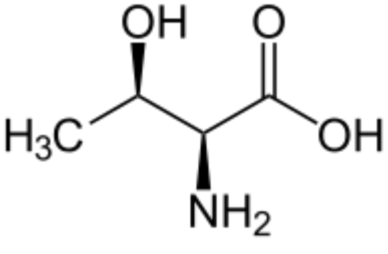
- Polyglutamine tract variation does not co-segregate with language disorder.
- The most common mutation in FOXP2 results in severe speech impairment known as developmental verbal dyspraxia.
- FOXP2 appears to be required for proper brain and lung development - in mice, knockout studies result in mice with impaired vocalizations.
- FOXP2 is highly expressed in areas of the brain known to be involved in language and speech development, including the basal ganglia and inferior frontal cortex.

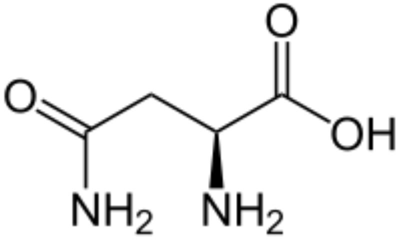
Comparative Genetics of FOXP2

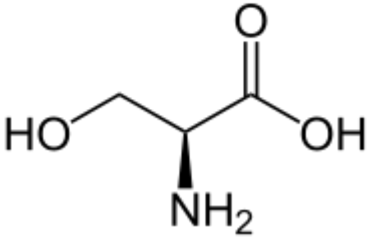
- Polyglutamine tract variation does not co-segregate with language disorder.
- Only 3 amino acid differences with FOXP2 protein orthologue in mouse.
- Among 5% most conserved proteins based on comparison with 1,880 human-rodent gene pairs.
- Chimpanzee, Rhesus macaque, and gorilla FOXP2 proteins are all exactly identical, with 1 difference from mouse and 2 from humans.
- Orang-utan FOXP2 shows 2 differences from mouse and 3 from humans.

Comparative Molecular Genetics of FOXP2

- Evidence shows that 2 of 3 amino acid differences between humans and mice occurred in the humans after separation from chimpanzee common ancestor.
- Both such differences occur in exon 7 of FOXP2 gene, the first being a Thr to Asn change (position 303) and the second a Asn to Ser switch (position 325).







Investigating Protein Structure Variation

- Comparison of predicted protein structures for humans, chimpanzees, orang-utans, mice revealed human-specific change at position 325 creates potential target for phosphorylation by protein kinase C.
- Should be interpreted in light of prior work showing that phosphorylation of forkhead transcription factors may mediate transcriptional regulation.
- In particular, human-specific change in position 325 of FOXP2 may carry functional consequences relevant to speech and language development.

Tracing Genetic Changes in FOXP2

- Possible amino acid changes in FOXP2 are *fixed* among humans.
- 130 Myr of evolution: 1 AA change between mice and common ancestor of humans and chimpanzees.
- 4.6-6.2 Myr of evolution: 2 fixed AA changes in human lineage, compared with 0 in chimpanzees and other primate lineages (except 1 in orang-utan).
- Likelihood ratio test (for constancy of ratio of AA replacements): significant increase in human lineage (p-value < 0.001); no change in other lineages.
- Finding is consistent with positive selection on AA changes in humans but does not rule out human-specific relaxation of constraints on FOXP2.

Methodology: DNA Sequencing & Data Analysis

- Amplification by PCR, with sequencing of overlapping fragments of FOXP2 coding regions from first-strand cDNA, for all analysed species.
- Designed primers from human BAC sequence, with each nucleotide position read from both strands for reach individual.
- Sequence traces analyzed manually for polymorphic positions using the *DNAStar* package.
- Sequences aligned with *ClustalW*; statistics calculated with *DnaSP*.
- Coalescent simulations, based on a fixed number of segregating sites and no recombination, were used to obtain p-values for the D-statistic and H-statistic.

Detecting a Selective Sweep

- Selective sweep - “reduction of elimination of variation among nucleotides near a mutation in DNA,” due to fixation associated with positive selection.
- Sequenced segment of 14,063bp over introns 4, 5, 6 of the FOXP2 gene in 20 individuals from diverse populations.
- Sequenced same segment in chimpanzees (central and west Africa) and an orang-utan.
- Null hypothesis: no difference between number of low-frequency alleles in observed data versus prediction under neutral model of random-mating.
- Tajima’s D-statistic = -2.20, with p-value = 0.002, thus making occurrence under the neutral model implausible.

Methodology: Tajima’s D-Statistic

- Goal: distinguish between sequence of DNA evolving randomly and one evolving under a non-random process.
- Tajima's test identifies sequences that do not fit the neutral theory model at equilibrium between mutation and genetic drift.
- Tajima demonstrated by simulation that a Beta(0,1) distribution may be used to approximate the distribution of the test statistic D.
- Simulations generally used to compute a p-value associated with D.

$$E[\pi] = \theta = E\left[\frac{S}{\sum_{i=1}^{n-1}\frac{1}{i}}\right] = 4N\mu$$

$$D = \frac{d}{\sqrt{\hat{V}(d)}}$$

Methodology: Tajima’s D-Statistic

- D < 0: excess of low frequency polymorphisms relative to expectation under the neutral model (population bottleneck, selective sweep).
- D > 0: low levels of both low frequency and high frequency polymorphisms, suggesting decrease in population size and/or balancing selection.

$$E[\pi] = \theta = E\left[\frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}\right] = 4N\mu$$

$$D = \frac{d}{\sqrt{\hat{V}(d)}}$$

Detecting a Selective Sweep

- Tajima’s D-statistic = -2.20, with p-value = 0.002, thus making occurrence under the neutral model implausible.
- D-statistic is not robust to violations of assumption of no population growth (“random-mating population of constant size”); could lead to negative values of D throughout the genome
- Compared to sample of 313 genes from 164 chromosomes, FOXP2 has the second lowest value of the D-statistics (lowest D = -2.25).

Methodology: Modeling Selective Sweep

- Polymorphism data summarized by 2 parameters, with a summary likelihood approach used in the estimation of the fixation time T.
- Using coalescent simulations, the likelihood of T is estimated as proportion of of n simulated data sets, where difference between the observed data parameters and simulated parameters differ by a user-specified tolerance.
- Method requires 4 additional nuisance parameters, but these are treated as fixed since it’s computationally infeasible to estimate them (?)

Detecting a Selective Sweep

- Selective sweep could also be characterized by a heightened presence of derived alleles at high frequency than under standard neutral model.
- To compute an H-value, ancestral states of human variable positions were inferred from chimpanzee and orang-utan sequences.
- Resultant H-value = -12.24 (versus null value of zero), with p-value = 0.042; would be more extreme under model accommodating population growth.
- Evidence strongly supports the 2 human-specific AA substitutions in exon 7 of FOXP2 as candidate sites for being affected by selective sweep.

Methodology: Fay & Wu’s H-value

- Goal: distinguish between sequence of DNA evolving randomly and one evolving under positive selection (n.b., more specific than Tajima’s D).
- Frequently used in the identification of sequences that have experience selective sweeps.
- Computation of H uses both population polymorphism data and data from an outgroup species, allowing identification of ancestral state of allele prior to split in the relevant lineages.
- H < 0: suggests excess of high-frequency derived SNPs.
- H > 0: suggests deficit of moderate and high-frequency derived SNPs.

Disease phenotypes related to FOXP2

- Multiple difficulties with both expressive and receptive aspects of language and grammar; exact nature of deficit has not yet been exactly ascertained.
- Impairment of selection and sequencing of fine orofacial movements, an ability not present in great apes but typical of humans.

The evolution of human language

- **<u>Speculation</u>:** one or both human-specific AA substitutions in exon 7 of FOXP2 could affect control of orofacial movements and language proficiency.
- Under such a speculation, fixation of the human-specific variant of FOXP2 could be strongly involved in the evolution of human language.
- Estimates suggest that fixation of FOXP2 variant occurred in the last 200,000 years, concomitant with the emergence of anatomically modern humans.

Concluding Remarks

Future directions

- Most functional assays are done in transgenic mice or zebrafish models

- Need more high-throughput assays to test at scale

- Human uniqueness may come from mutations in the chimp lineage and not our.

Future directions

Human uniqueness may come from mutations in the chimp lineage and not our.

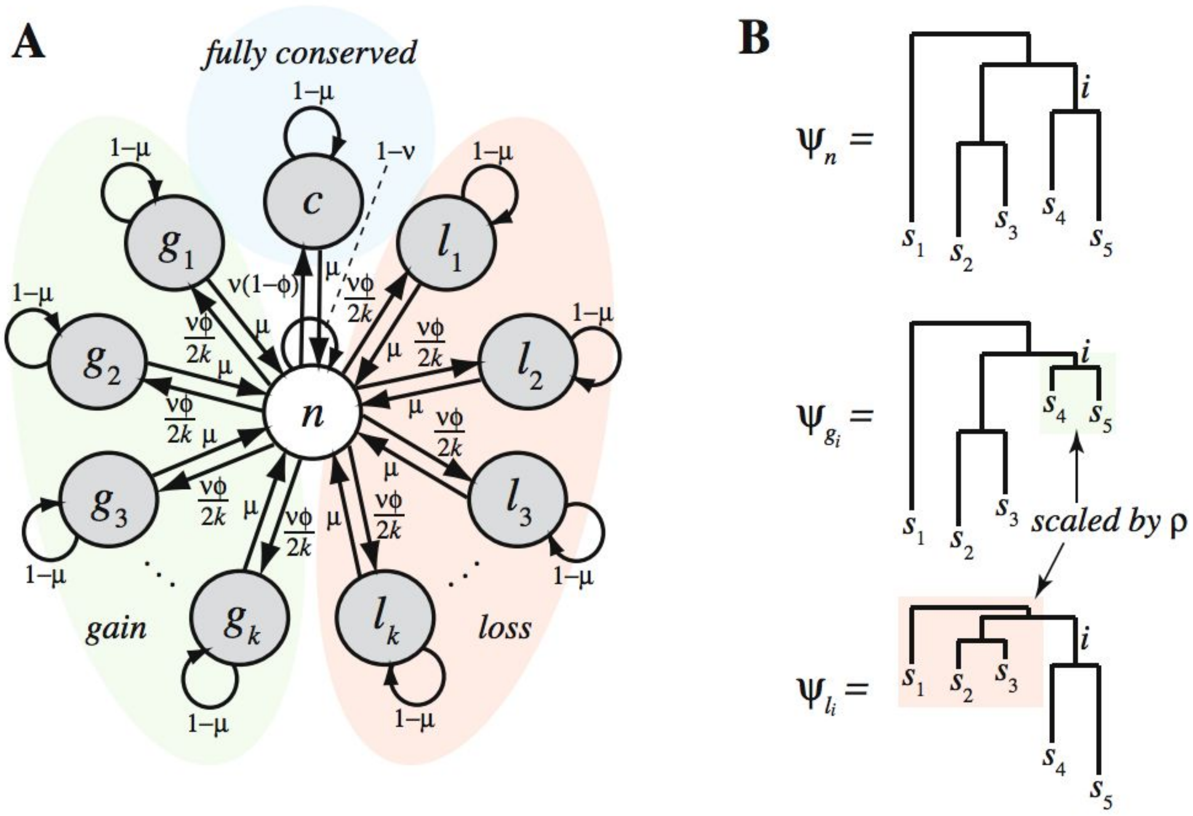
"Some people talk to animals. Not many listen though. That's the problem." – A.A. Milne

Thank you for listening

I am fond of pigs. Dogs look up to us. Cats look down on us. Pigs treat us as equals.

Winston Churchill

Supplementary of phastCons


A
fully conserved
gain
loss
c
n
g_1
g_2
g_3
g_k
l_1
l_2
l_3
l_k
1−μ
1−ν
μ
ν(1−ϕ)
$\frac{\nu\phi}{2k}$
B
$\Psi_n =$
$\Psi_{g_i} =$
$\Psi_{l_i} =$
i
s_1
s_2
s_3
s_4
s_5
scaled by ρ