

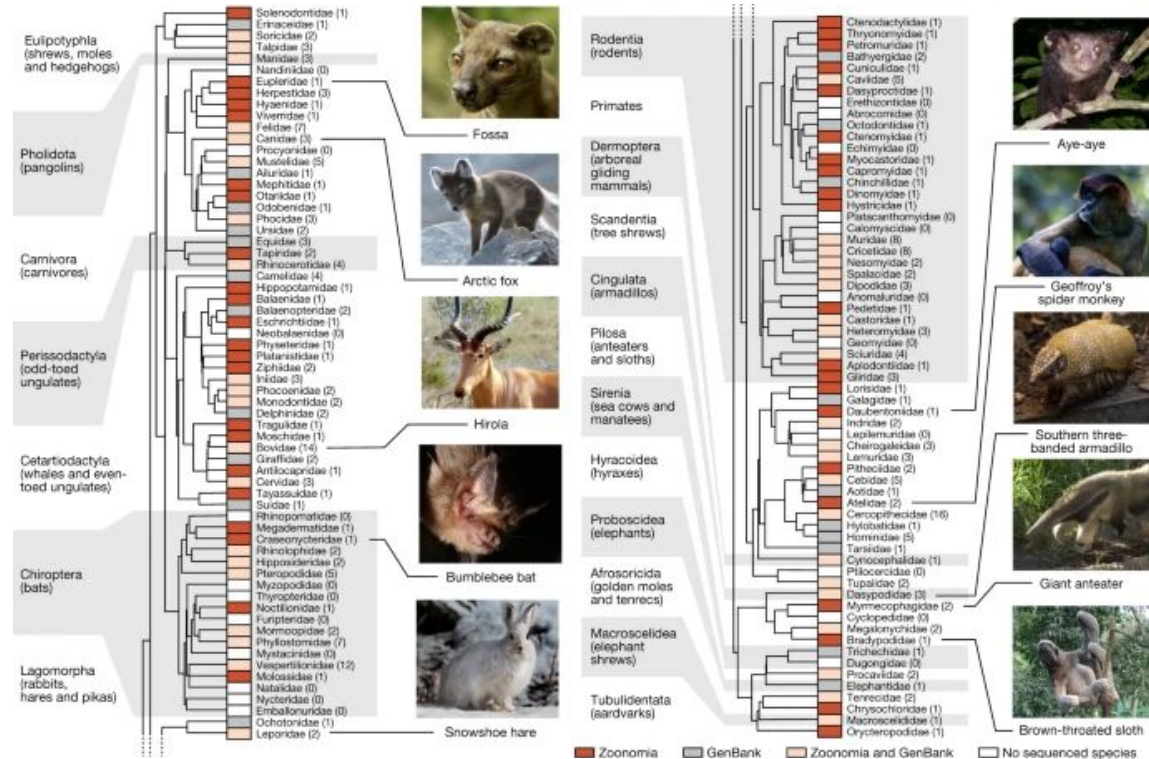
Fitting Evolutionary Matrices Using Modern Data: Final

Joshua Morin-Baxter, Talya Koschitzky, Robert Becker

Superspeed Review: Zoonomia

whole-genome alignment
from 240 species with
representatives from over
80% of mammalian
families, including 120
previously uncharacterized
species

spanning about 110 million
years of mammalian
evolution



Phylogenetic tree of the mammalian families in the Zoonomia Project alignment
<https://www.nature.com/articles/s41586-020-2876-6>

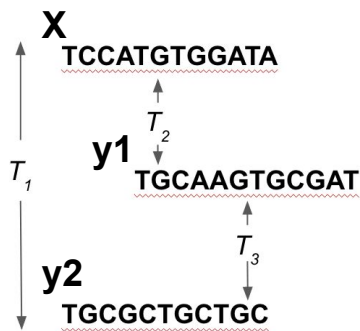
Superspeed Review: Our Project Outline

We plan to use the Zoonomia data to better fit evolutionary matrices

1. Get sequence alignments from Zoonomia data, research value of T for different pairs of species

2. Construct a likelihood function, with an initial value for A

3. Use optimization tools to find the value of A that optimizes the likelihood function



Likelihood for one x, y, T triplet:

$$f(x, y, \mathbf{A}, T) = \prod_i \mathbf{A}^T[x_i][y_i]$$

Multiply likelihoods from each x, y, t triplet to get overall likelihood function for optimization?

[Submitted on 22 Dec 2017 (v1), last revised 30 Jan 2017 (this version)]

Adam: A Method for Stochastic Optimization

Diederik P. Kingma, Jimmy Ba



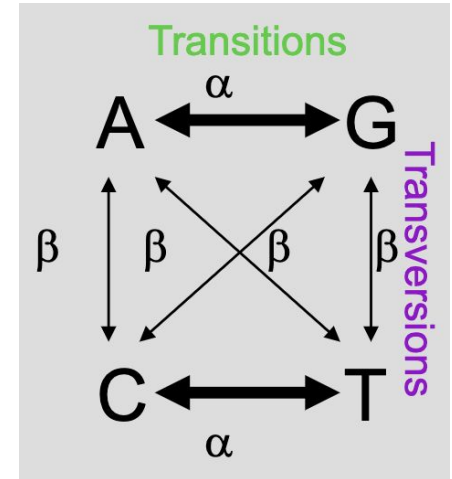
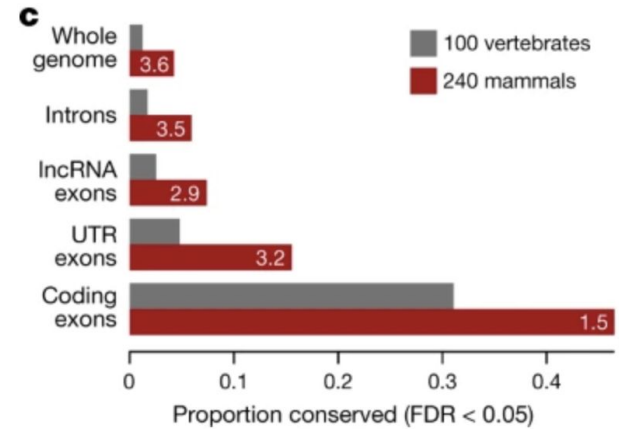
Our Hypothesis

Based on findings by Zoonomia project:

- **Expectation 1:** More mutations in lncRNA than UTR exons

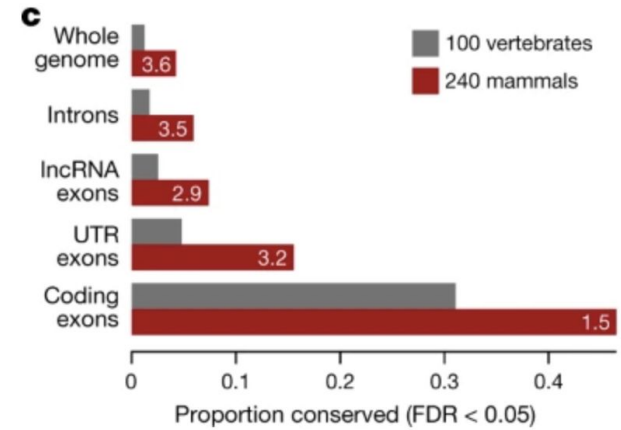
We'll expand this data by finding the conservation rates by base pair, rather than just general conservation.

- **Expectation 2:** Clear difference in the mutation rates between transitions and transversions, with transitions being more likely since these only require small changes to the ring structure.

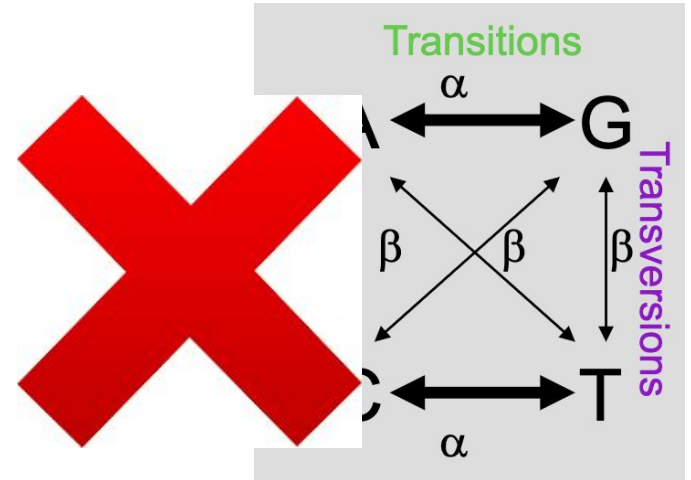


Our Hypothesis Revisited

More mutations in lncRNA than 5'-UTR exons



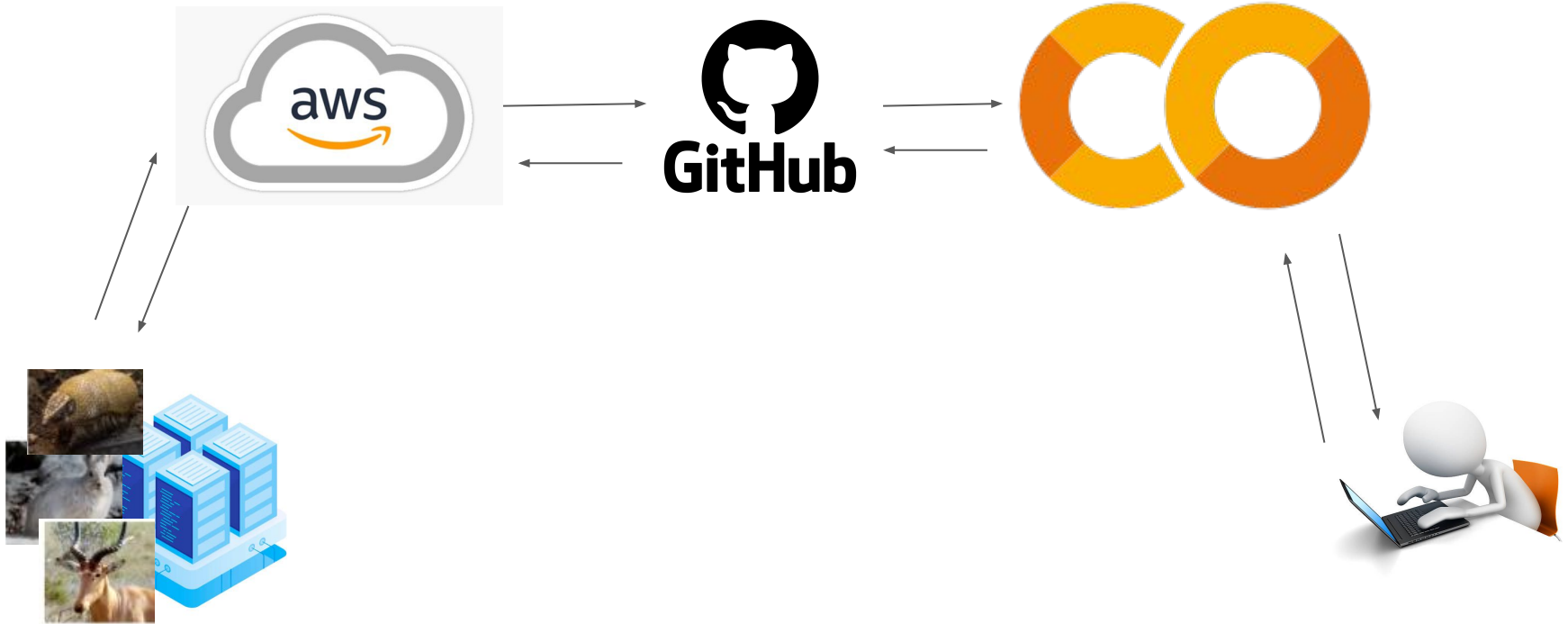
Clear difference in transversions and transitions



Sequences of interest: The Data



Sequences of interest: The Data



Sequences of Interest: Genomic Regions

- **5'-UTR**: NCBI human genome track export 5'-UTR exons (GRCh38)
- **lncRNA**: LNCipedia (compendium of human noncoding RNA) high-confidence set

Randomly sampled from these regions to reduce computation

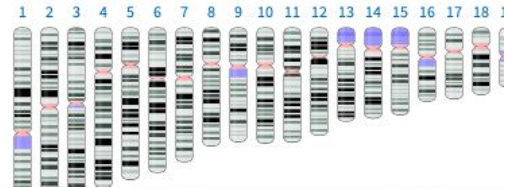
```
UTR    : 194431 regions of avg. len 202.579
lncRNA: 107039 regions of avg. len 23498.3
```

Assembly details

Name	GRCh38.p14
RefSeq accession	GCF_000001405.40
GenBank accession	GCA_000001405.29
Submitter	Genome Reference Consortium
Level	Chromosome
Category	Reference genome
Replaced by	GCF_000001405.25

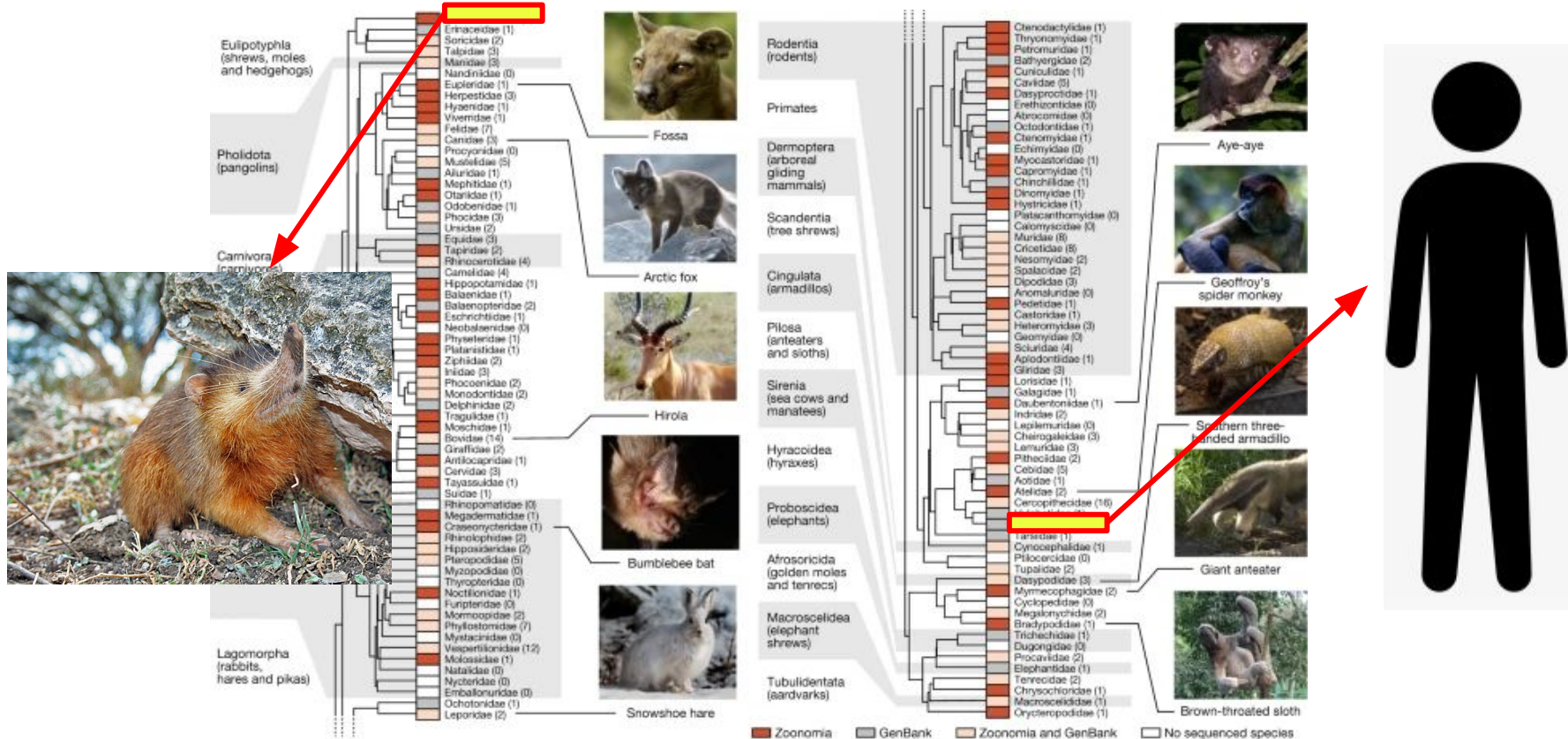
Annotation details

Annotation Release	110 
Release date	Apr 5, 2022



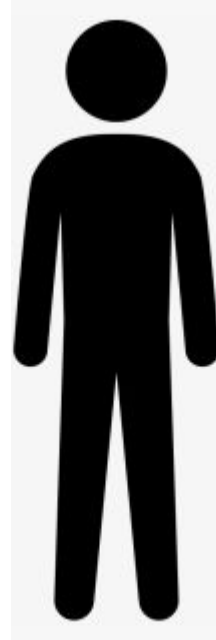
LNCipedia
version 5.2

Sequences of Interest: The Genomes



Sequences of Interest: The Genomes

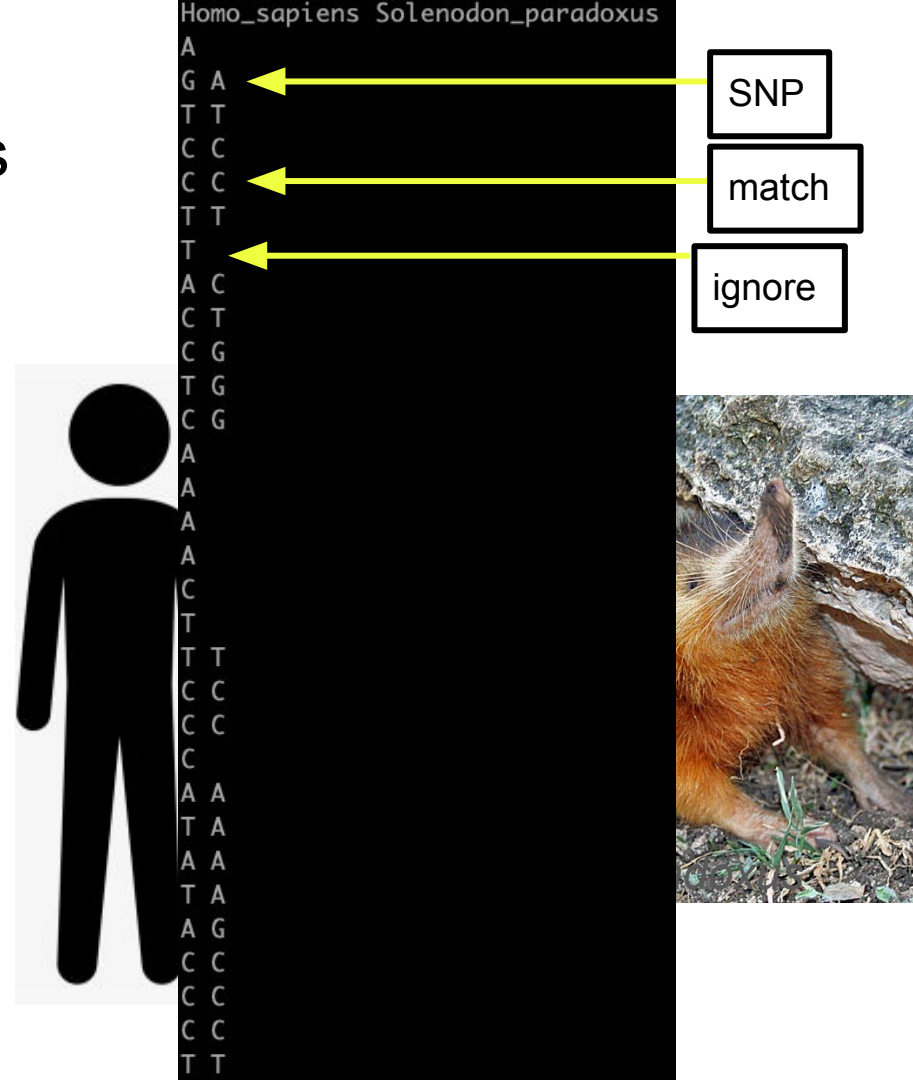
- 5'-UTR and lncRNA well-curated data widely available for humans
- Not available for *Solenodon paradoxus* (or most other animals)



Solenodon paradoxus

Sequences of Interest: halSnps

- Aligned curated 5'-UTR and lncRNA regions to *Solenodon paradoxus* via Zoonomia project
- Used SNP extraction tool to find SNPs (and not-SNPs) between genomes



Final A^T Matrices*

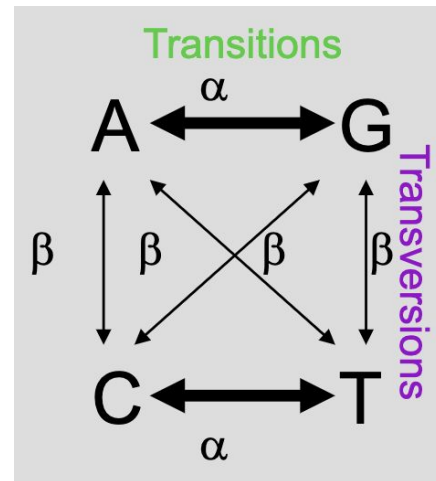


UTR (n=101)

	A	G	C	T
A	0.76	0.06	0.12	0.06
G	0.10	0.63	0.06	0.22
C	0.21	0.06	0.63	0.09
T	0.07	0.13	0.05	0.75

lncRNA (n=101 bp)

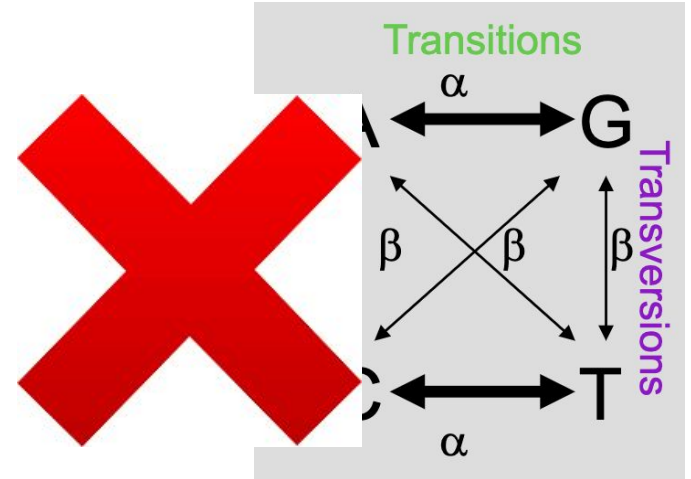
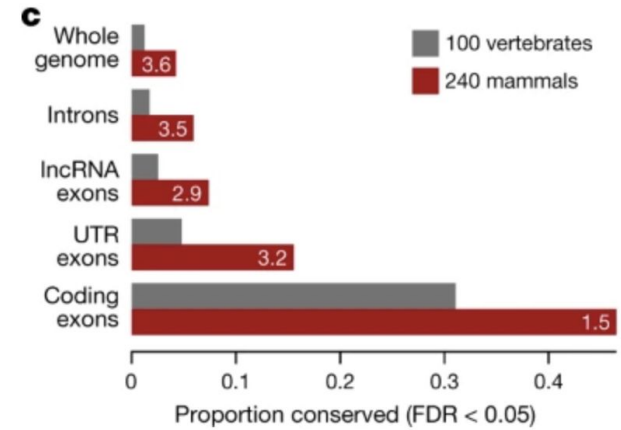
	A	G	C	T
A	0.76	0.05	0.12	0.07
G	0.10	0.61	0.05	0.24
C	0.23	0.05	0.62	0.10
T	0.07	0.12	0.05	0.77



*optimized for A^T instead of A because finding A proved computationally prohibitive and A^T still demonstrates our negative result

Conclusions

- 5'-UTR is more conserved than lncRNA, genome-wide
 - There is reason to believe this can be extended to most mammals
- However, there is little difference in the types of substitutions that are seen in these regions



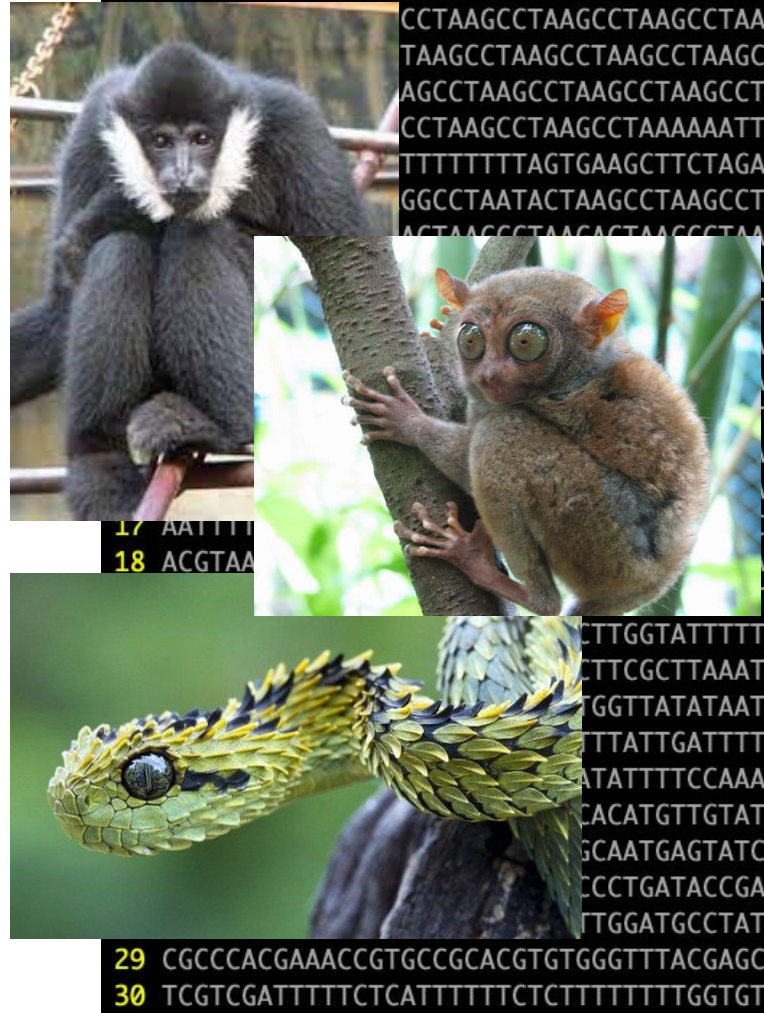
Future Direction

Major limitation: only optimized over small part of the genome and only between two species*

Optimize for A^T instead of A

Can be extended to compare other regions of the genome and many other species

*result was recapitulated in several other species and with many different random samples, data not shown



Questions

Timeline and Milestones

	Sun	Mon	Tue	Wed	Thurs	Fri	Sat
3/27-4/2	Research data availability						
4/3-4/9	Extract relevant initial data from the Zoonomia project						
4/10-4/16	Fit an evolutionary matrix using sampled Zoonomia data						
4/17-4/23	Find additional corresponding sequences and fit matrices on these as well						
4/24-4/30	Combine results to obtain a best fitting matrix, repeat for different regions of interest						
5/1-5/7	Analyze results, create graphics, and work on final report						
5/8-5/9	Finalize report and submit						

The Algorithm

- Wrote our log likelihood function that takes in a single pair of sequences x and y and represents the probability that x evolves to y in time T given evolutionary matrix A
- Chose library for optimization **scipy.optimize**
- Chose Trust-Region Constrained Algorithm (**method='trust-constr'**)*: Allows us to maximize our log-likelihood function subject to constraints
- Constraint 1: The rows of A must sum to 1
- Constraint 2: Each entry in A must be between 0 and 1 (a probability)

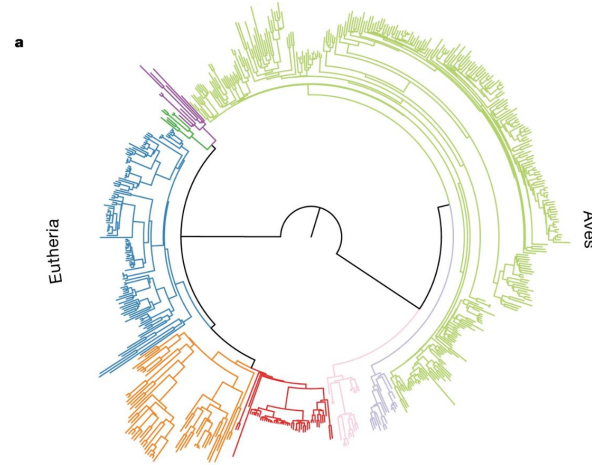
The Problem

- This optimization algorithm only works when A isn't raised to a high power (realized we weren't properly raising A to a power in initial testing).
- This added complexity makes it difficult for the algorithm to find an optimal solution in practice
- Tried to use other basic optimization methods (stochastic gradient descent, etc.) but don't work because constraints are important (otherwise all values of A will blow up)
- Tried solving for A_T and taking the T th root to get A , but often led to imaginary numbers in our matrix

*Lalee, Marucha, Jorge Nocedal, and Todd Plantega. 1998. On the implementation of an algorithm for large-scale equality constrained optimization. SIAM Journal on Optimization 8.3: 682-706.

Question slide: Alignment of genomes

“240 genomes aligned as part of a 600-way pan-amniote alignment using the Cactus alignment software: Rather than aligning to a single anchor genome, Cactus infers an ancestral genome for each pair of assemblies”



nature.com/articles/s41586-020-2871-y/figures/3

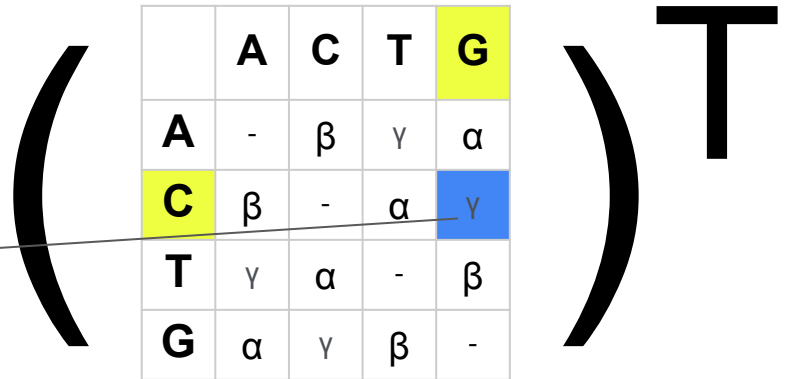
Question slide: Calculation of likelihood function

$$f(x, y, \mathbf{A}, T) = \prod_i \mathbf{A}^T[x_i][y_i]$$

x : TCC**A**TGTGGATA:
 y : TC**G**ATGTGGATA

$$f(x, y, A, T) = p_1 p_2 p_3 \cdots p_n$$

$$\mathbf{A}^T[C][G]$$



	A	C	T	G
A	-	β	γ	α
C	β	-	α	γ
T	γ	α	-	β
G	α	γ	β	-

Question slide: How we handle a gap

$$f(x, y, \mathbf{A}, T) = \prod_i \mathbf{A}^T[x_i][y_i]$$

TCCATGTG-GAT
| | | | |
TGCA-GTGCG-T

(

	A	C	T	G
A	-	β	γ	α
C	β	-	α	γ
T	γ	α	-	β
G	α	γ	β	-

)^T

Question slide: How to further optimize A?

Not sure yet but still working on it

$$f_1(x, y, \mathbf{A}, T) = \prod_i \mathbf{A}^T[x_i][y_i]$$

$$f_2(w, z, \mathbf{A}, T) = \prod_i \mathbf{A}^T[w_i][z_i]$$

$$f_k(a, b, \mathbf{A}, T) = \prod_i \mathbf{A}^T[a_i][b_i]$$

$$f = f_1 f_2 f_3 \cdots f_k$$

$\left(\begin{array}{c|c|c|c|c} & \mathbf{A} & \mathbf{C} & \mathbf{T} & \mathbf{G} \\ \hline \mathbf{A} & - & \beta & \gamma & \alpha \\ \mathbf{C} & \beta & - & \alpha & \gamma \\ \mathbf{T} & \gamma & \alpha & - & \beta \\ \mathbf{G} & \alpha & \gamma & \beta & - \end{array} \right)^T$

	A	C	T	G
A	-	β	γ	α
C	β	-	α	γ
T	γ	α	-	β
G	α	γ	β	-