Insights into mammalian evolution from large-scale comparative genomics

Matt Christmas
Institute of medical biochemistry and microbiology
Uppsala Universitet
matthew.christmas@imbim.uu.se
Undergrad

Prof. Andrew Lowe

Prof. Matthew Webster

Prof. Kerstin Lindblad-Toh

Post doc/researcher/ Trying to make a career out of it!
Zoonomia

Using the largest alignment of mammal genomes to date to explore mammalian evolution, disease, and conservation
Placental mammals – a diverse group

- 4 major subdivisions
- 19 extant orders
- >6,000 species
- ~100 million years of evolution
What can a comparative genomics approach tell us?

Large multi-species genome alignments can be used to reveal answers to these questions

The Zoonomia Project
The Zoonomia project

A comparative genomics multitool for scientific discovery and conservation

https://doi.org/10.1038/s41586-020-2876-6
Received: 17 April 2019
Accepted: 27 July 2020
Published online: 11 November 2020
Open access

The Zoonomia Project is investigating the genomics of shared and specialized traits in eutherian mammals. Here we provide genome assemblies for 131 species, of which all but 9 are previously uncharacterized, and describe a whole-genome alignment of 240 species of considerable phylogenetic diversity, comprising representatives from more than 80% of mammalian families. We find that regions of reduced genetic diversity are more abundant in species at a high risk of extinction, discern signals of evolutionary selection at high resolution and provide insights from individual reference genomes. By prioritizing phylogenetic diversity and making data available quickly and without restriction, the Zoonomia Project aims to support biological discovery, medical research and the conservation of biodiversity.
Genomes relative to year and phylogenetic relationships

Genomes for 675 mammal species relative to the Mammalia phylogenetic tree of 5911 living species shows the disproportionate representation of large-bodied and high-latitude species. Shown is the consensus timescaled phylogeny from Upham et al. (4) and genome data downloaded from NCBI on 9 February 2023.

Science, 380 (6643), .
DOI: 10.1126/science.add2209
Mammalian genome evolution

Ancestral mammal chromosomes
~180 Mya
19 autosomes + sex chromosomes

Ancestral therian (placentals + marsupials) chromosomes
17 autosomes + sex chromosomes
96 chromosomal rearrangements over 18 My

Other great apes have 23 autosomes.
Fusion of two ancestral chromosomes in human lineage

124 chromosomal rearrangements over 53 My
Genome size variation and TEs

Mean genome size = 2.67 Mb
Mean % TE = 46%

- All species
- Cetartiodactyla

- Aardvark
- Screaming Hairy Armadillo
- Brazilian guinea pig
- star-nosed mole

Genome size (Mb)

TE genome %
Genome size variation and TEs

Insights into mammalian TE diversity through the curation of 248 genome assemblies
Zoonomia genomes

- 241 placental mammal genomes (240 species), representing >80% of mammalian families
- At least one long-range assembly (contig N50 > 20 kb and scaffold N50 > 10 Mb) per order
- Reference-free Cactus alignment
- Single-base measures of evolutionary constraint (phyloP)
Avoiding reference bias

10bp deletion

Missed in the alignment

Reference genome
Reference-free alignment

- All against all
- Can identify lineage-specific variation
Using comparative genomics to study genome evolution

Accelerated evolution in certain lineages

Neutrally evolving non-functional regions

Evolutionary constraint

Consequences of this can be utilized to identify functional genomic regions via multi-species alignments
<table>
<thead>
<tr>
<th>Constraint</th>
<th>Neutral</th>
<th>Acceleration</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATCTAACAGATTAGA</td>
<td>ATCTAACAGATTAGA</td>
<td>ATCTAACAGATTAGA</td>
</tr>
<tr>
<td>ATCTAACAGATTAGA</td>
<td>ATCTAACAGATTAGA</td>
<td>ATCTAACAGATTAGA</td>
</tr>
<tr>
<td>ATCTAACAGATTAGA</td>
<td>ATCTAACAGATTAGA</td>
<td>ATCTAACAGATTAGA</td>
</tr>
<tr>
<td>ATCTAACAGATTAGA</td>
<td>ATCTAACAGATTAGA</td>
<td>ATCTAACAGATTAGA</td>
</tr>
<tr>
<td>ATCTAACAGATTAGA</td>
<td>ATCTAACAGATTAGA</td>
<td>ATCTAACAGATTAGA</td>
</tr>
<tr>
<td>ATCTAACAGATTAGA</td>
<td>ATCTAACAGATTAGA</td>
<td>ATCTAACAGATTAGA</td>
</tr>
</tbody>
</table>
More species, more power

<table>
<thead>
<tr>
<th></th>
<th>29 mammals (2011)</th>
<th>240 mammals (2023)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total branch length</td>
<td>4.5</td>
<td>16.6</td>
</tr>
<tr>
<td>(substitutions/site)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probability of identical site</td>
<td>0.02</td>
<td>6x10^{-18}</td>
</tr>
<tr>
<td>(p)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expected identical sites</td>
<td>62 Mbp</td>
<td>191 bp</td>
</tr>
<tr>
<td>Constraint resolution</td>
<td>12 bp (p = 1 x 10^{-25})</td>
<td>Single base pair</td>
</tr>
<tr>
<td>Human genome constraint</td>
<td>&gt;5.5%</td>
<td>&gt;10.7%</td>
</tr>
</tbody>
</table>
Estimating evolutionary constraint and acceleration - phyloP

- **PhyloP** calculates constraint and acceleration p-values based on a multi-species alignment and a **model of neutral evolution** at individual nucleotides.

- Each base position in the alignment has a phyloP score: 0 = neutral, >0 = constraint, <0 = acceleration.

Running phyloP

• Generate neutral model file using phyloFit
  • Inputs: alignment file for neutral sites (.maf format from Cactus), tree topology

• Run phyloP
  • Inputs: alignment file and phylogenetic model from above
  • --method LRT – likelihood ratio test. How likely is the observed divergence given neutral model?
  • --mode CONACC gives positive p values/scores to indicate constraint and negative scores to indicate acceleration
Models of neutral evolution

• To understand sequence evolution over time we need **accurate models of neutral evolution**, i.e. sequence changes over time in the absence of selection

• Important to have accurate models of neutral evolution for
  • Inferring phylogeny (model assumption)
  • Divergence time estimates
  • Detecting constraint and acceleration

• Models of neutral evolution:
  • Four-fold degenerate sites
  • **Ancestral repeats**
Constraint scores reflect degeneracy in codons
Four-fold degenerate sites generally evolve neutrally, but it’s not all neutral...

18.6% of 4D sites are under constraint (FDR=0.05)

Constraint at 4D sites relates to gene regulation

Constraint in mammals

- Max score: 8.903 – all species aligned and identical, **highly constrained**
- **3.6 million perfectly conserved positions**
Distribution of constraint

- 3.26% of bases in the human genome identified as under significant constraint at FDR < 0.05 (phyloP>2.270)
- Form clusters – most (80%) are within 5bp of another constrained base
- Positions under constraint are less variable in humans (TOPMed SNPs)
- Constraint strongly enriched in coding sequence, moderately enriched in regulatory elements
- Single-base resolution of constraint scores evidenced by constraint in codons
Constraint in unannotated regions

- Nearly half of all constrained bases sit outside of annotations
- Identified >400,000 UNannotated Intergenic COnstrained RegioNs (UNICORNs)
zooUCEs

- Original set (2004): 481 segments > 200bp 100% conserved between human, mouse, and rat
- zooUCEs: ≥98% of species (235) are *aligned and identical*
- 4,552 zooUCEs ≥ 20bp. 753 overlap 318 original UCEs. 27 ≥ 100bp
- 69% are outside of protein coding exons
Genomic ‘hotspots’ of constraint

- Constraint measured in 100kb windows across human genome
- 53 bins with significant constraint (q<0.05)
- Includes all HOX clusters
- Large gene deserts with high constraint
So what’s all this constrained sequence doing if not coding for proteins?

1. DNA
2. Enhancer
3. Promoter
4. Gene
5. Transcription Activator Protein
6. Mediator Protein
7. RNA Polymerase

Mammalian evolution of human cis-regulatory elements and transcription factor binding sites

[Image of DNA and molecular structures]
Questions?

Break

30 mins
Using estimates of constraint to explore mammalian evolution
Exercise – how does the selective pressure on a sequence affect the inferred phylogeny?
Constrained sites

Order

- Afrosoricida
- Carnivora
- Cetartiodactyla
- Chiroptera
- Cingulata
- Dermoptera
- Eulipotyphla
- Hyracoidea
- Lagomorpha
- Macroscelidea
- Perissodactyla
- Pholidota
- Pilosa
- Primates
- Proboscidea
- Rodentia
- Scandentia
- Sirenia
- Tubulidentata
New phylogeny resolves some long-standing debates in mammalian evolution

- Based on ~500k near-neutrally evolving sites
- Supports the ‘long-fuse’ model of mammalian diversification

Methods:
SVD quartets – coalescent trees
IQ-TREE – maximum likelihood concatenated trees
Effects of selection when estimating divergence time

Take-home messages:

• Neutral is best – models assume this

• BUT with a massive amount of data (i.e. genome-wide), can still achieve accurate estimates

• Depends on the phylogenetic distance you are looking over
Phylogenomic discordance: gene tree ≠ species tree

Caused by:
• Incomplete lineage sorting
• Historical hybridization

Interordinal

Intraordinal (Simians)
Linking genotype to phenotype

Identifying the genomic changes that underlie specific traits
Lineage-specific acceleration

<table>
<thead>
<tr>
<th>Constraint</th>
<th>Neutral</th>
<th>Acceleration</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATCTAAGGACCCCGGAGTTAACCA</td>
<td>ATCTAAGGTCCGGGAAACCCCA</td>
<td>ATCTAAGGCACCAGGCGGCTGTTTT</td>
</tr>
<tr>
<td>ATCTAAGGACCCCGGAGTTAACCA</td>
<td>ATCTAAGGACCCAGGTAACCTTTTA</td>
<td>ATCTAAGGACCCAGGTAACCTTTTA</td>
</tr>
<tr>
<td>ATCTAAGGACCCCGGAGTTAACCA</td>
<td>ATCTAAGGACCCAGGTAACCTTTTA</td>
<td>ATCTAAGGACCCAGGTAACCTTTTA</td>
</tr>
<tr>
<td>ATCTAAGGACCCCGGAGTTAACCA</td>
<td>ATCTAAGGACCCAGGTAACCTTTTA</td>
<td>ATCTAAGGACCCAGGTAACCTTTTA</td>
</tr>
<tr>
<td>ATCTAAGGACCCCGGAGTTAACCA</td>
<td>ATCTAAGGACCCAGGTAACCTTTTA</td>
<td>ATCTAAGGACCCAGGTAACCTTTTA</td>
</tr>
</tbody>
</table>

Constraint:  
Neutral:  
Acceleration:
Identifying lineage-specific positive selection

Human accelerated regions

What/where might these be?

- 312 human accelerated regions (zooHARs) – many are neurodevelopmental enhancers with cell type-specific activity
- Overlap TADs containing human-specific structural variants
- Enrichment of hARs that affect 3D genome organisation
- ‘Rewiring’ of regulatory interactions between hARs and neurodevelopmental genes
Species-specific deletions

- Use the alignment to identify sites/sequences deleted in a single species
- Human-specific conserved deletions (hCONDELs)

PPP2CA
Detecting smells – a universal animal trait

Mmm, smells good!
Olfaction in mammals

- **Olfactory receptor (OR) genes**: G-protein-coupled receptors, contain several transmembrane α-helical domains. Detect odour molecules in the environment.

- Largest gene superfamily in vertebrate genomes

- Olfactory turbinals - extensive framework of delicate bones in nasal cavity that greatly enlarge the surface area available for conditioning inspired air, reducing water loss, and improving olfaction
Counting OR genes in mammals

- Olfactory receptor gene (OR) detection in 249 species
- African elephant has largest repertoire (1,765 functional genes)
- Killer whale smallest (24 functional genes)
- Bumblebee bat and North Pacific right whale – only 332 and 392 respectively
Correlating genotype and phenotype

• OR counts strongly correlated with olfactory turbinal counts

• Highlights potential for investigating multifactorial evolutionary responses to a selection pressure
Hibernation in mammals

- Physiological state of metabolic depression (torpor). Species capable of core temperature depression below 18°C for >24h
- Ancestral trait in mammals? Found in all deep lineages

<table>
<thead>
<tr>
<th>Minimum body temperature (°C)</th>
<th>Sleep</th>
<th>Daily heterothermy</th>
<th>Shallow torpor</th>
<th>Deep torpor</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.6-11.0</td>
<td>5-35°C</td>
<td>5-30°C</td>
<td>10-30°C</td>
<td>&lt;2°C</td>
</tr>
<tr>
<td>10.6-11.0</td>
<td>10-20%</td>
<td>10-30%</td>
<td>10-20%</td>
<td>&lt;2°C</td>
</tr>
<tr>
<td>Heart rate (% of active)</td>
<td>70-90%</td>
<td>10-30%</td>
<td>10-20%</td>
<td>&lt;2°C</td>
</tr>
<tr>
<td>Metabolic rate (% of BMR)</td>
<td>70-90%</td>
<td>3-50%</td>
<td>3-50%</td>
<td>1-4%</td>
</tr>
<tr>
<td>Blood pressure (% of active)</td>
<td>85%</td>
<td>50-70%</td>
<td>50-70%</td>
<td>1-10%</td>
</tr>
<tr>
<td>Daily activity level</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>None</td>
</tr>
<tr>
<td>Average energy savings</td>
<td>5-15%</td>
<td>60-70%</td>
<td>Low</td>
<td>None</td>
</tr>
<tr>
<td>Duration</td>
<td>&lt;24 h</td>
<td>&lt;24 h</td>
<td>&gt;24 h</td>
<td>&gt;24 h</td>
</tr>
<tr>
<td>Occurrence</td>
<td>Year-round</td>
<td>Year-round</td>
<td>Seasonal inducible</td>
<td>Seasonal</td>
</tr>
</tbody>
</table>

Shading refers to magnitude of parameter (i.e., more intense color represents higher body temperature or longer duration of time). Abbreviation: BMR, basal metabolic rate.
Hibernation in mammals

- Effects of the cold on the function of cellular and molecular processes requires specific adaptations to survive at the limitations of basic biology and physiology, e.g. transcription and translation can be halted.

- Arctic ground squirrel (*Spermophilus parryii*) can drop its core body temperature below 0°C – the lowest of any mammal.

- Protective mechanisms in the hibernating brain.

Testing for differential evolutionary rates associated with hibernation

- RERconverge identified 511 genes slower-evolving and 253 faster-evolving in hibernators
- Faster evolving genes enriched in GO pathways relating to synaptic transmission
- Slower evolving genes enriched in GO pathways relating to DNA repair – under higher constraint in hibernators
What about regulatory sequences?

• Tissue-Aware Conservation Inference Toolkit (TACIT) for associating \textit{cis}-regulatory elements with specific phenotypes

• Associates open chromatin predictions with phenotype annotations
Using genomes to inform conservation

RESEARCH ARTICLE SUMMARY

The contribution of historical processes to contemporary extinction risk in placental mammals

Conclusions

• Massive, multi-way whole genome alignments are now possible!
• Informative for identifying effects of selection on the genome
• Single-base measures of constraint and acceleration pinpoint functional sequences
• Identification of neutrally evolving sequences important for inferring accurate phylogenetic relationships and divergence times
• Comparative genomics can reveal the genetic underpinnings of traits
• Measures of genomic variation informative for extinction risk and conservation
Constraint in human TEs

- ~11% of constrained bases are within repeats in the human genome
- Enriched in DNA transposons and simple repeats
- Repeat classes depleted in constraint have been more recently active
- Most primate-specific TFBS overlap TEs

Greg Andrews, Zhiping Weng, UMass