A Comparative Genomics Approach to Identifying Candidate Enhancers Associated with Mammalian Phenotypes

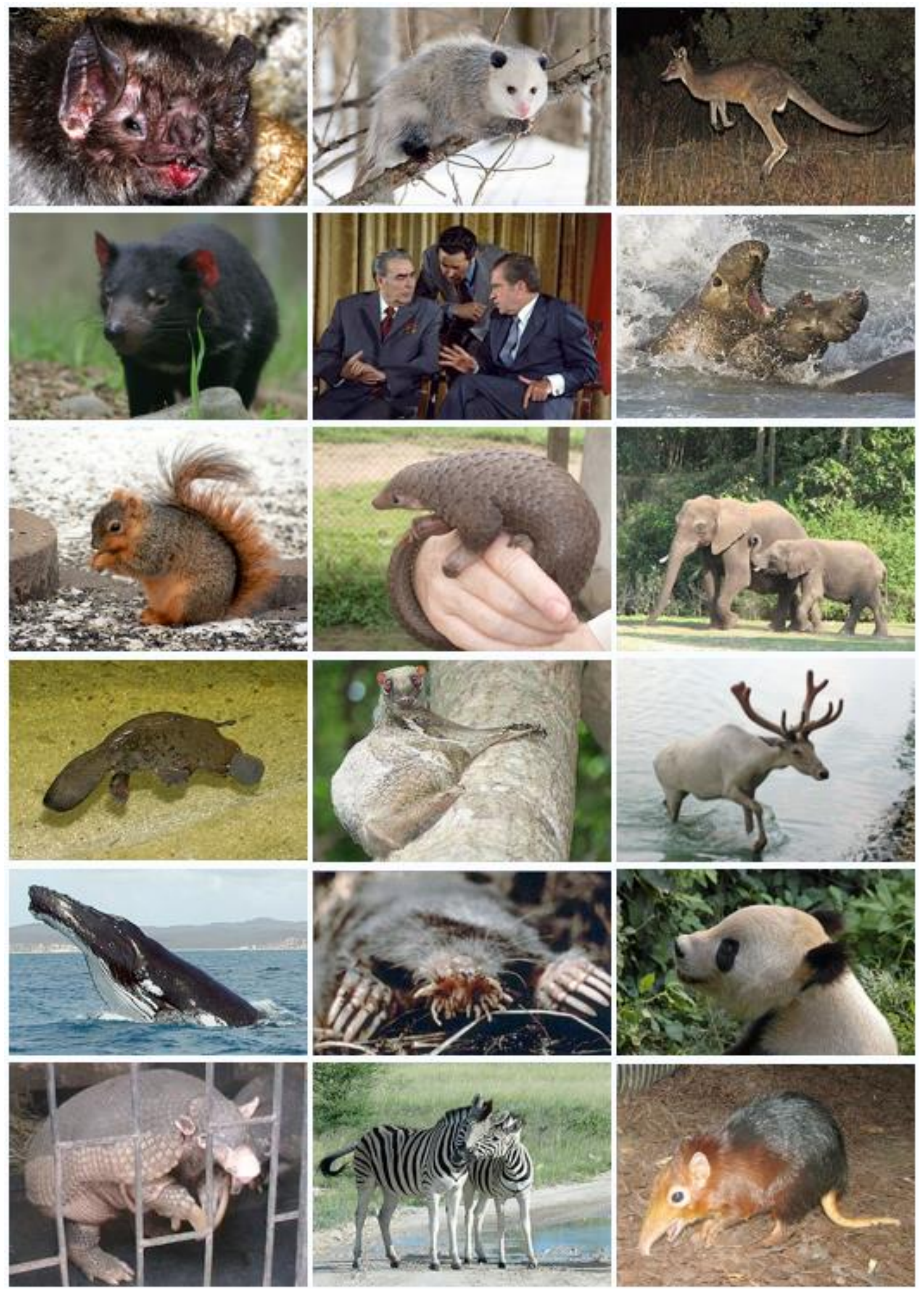
Computational Biology Department

Daniel E. Schaffer

Advisors: Dr. Irene Kaplow & Prof. Andreas Pfenning

Carnegie Mellon University School of Computer Science

Background



Source: commons.wikimedia.org/wiki/File:Mammal_Diversity_2011.png

Figure 1. Assorted mammals.

- Mammals are very diverse (**Fig. 1**)
- Many differences are likely due to changes in gene regulation between species[1,2]
- Enhancers are small DNA sequences that regulate gene activity in specific tissues[3]
  - Bound by transcription factors
- Goal: correlate their activity with phenotypes
  - Sequence similarity is insufficient
  - Instead, use models to predict activity Open chromatin regions (OCRs) are a proxy for enhancers in a species and tissue

References

[1] M. King, A. Wilson. *Science*. **188**, 107–16 (1975).

[2] C. Y. McLean, P. L. Reno, A. A. Pollen, A. I. Bassan, *et al*. *Nature*. **471**, 216–19 (2011).

[3] D. Villar, P. Flicek, D. T. Odom. *Nat. Rev. Genet.* **15**, 221–33 (2014).

[4] I. M. Kaplow, D. E. Schäffer, M. E. Wirthlin, A. J. Lawler, *et al*. *BMC Genom.* **23**, 291 (2022).

[5] I. M. Kaplow, A. J. Lawler, D. E. Schäffer, C. Srinivasen, *et al*. In preparation.

[6] C. Srinivasan, B. N. Phan, A. J. Lawler, E. Ramamurthy, *et al*. *J. Neurosci*. **41**, 9008–30 (2021).

[7] M. Wirthlin, I. M. Kaplow, A. J. Lawler, J. He, *et al*. *bioRxiv*. 356733 (2020).

[8] M. E. Wirthlin, Z. Zhang, I. M. Kaplow, D. E. Schäffer, *et al*. In preparation.

[9] Z. Yao, H. Liu, F. Xie, S. Fischer, *et al*. *Nature*. **598**, 103–10 (2021).

[10] T. E. Bakken, N. L. Jorstad, Q. Hu, B. B. Lake, *et al*. *Nature*. **598**, 111–19 (2021).

[11] Zoonomia Consortium. *Nature*. **587**, 240–45 (2020).

[12] J. Armstrong, G. Hickey, M. Diekhans, I. T. Fiddes, *et al*. *Nature*. **587**, 246–51 (2020).

[13] L. S. T. Ho, C. Ané. *Syst. Biol.* **63**, 397–408 (2014).

[14] E. Saputra, A. Kowalczyk, L. Cusick, N. Clark, M. Chikina. *Mol. Biol. Evol.* **38**, 3004–21 (2021).

[15] J. R. Burger, M. A. George, C. Leadbetter, F. Shaikh. *J. Mammal.* **100**, 276–83 (2019).

[16] S. Herculano-Houzel. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 10661–68 (2012).

[17] C. Y. McLean, R. L. Reno, A. A. Pollen, A. I. Bassan, *et al*. *Nature*. **471**, 216–19 (2011).

[18] M. Parrish, T. Ott, C. Lance-Jones, G. Schuetz, *et al*. *Mol. Cell. Biol.* **24**, 7102–12 (2004).

[19] P. Giusti-Rodríguez, L. Lu, Y. Yang, C. A. Crowley, *et al*. *bioRxiv*. 406330 (2019).

[20] D. Jeong, D. Lozano Casasbuenas, A. Gengatharan, K. Edwards, *et al*. *Cell. Rep*. **33**, 108257 (2021)

[21] L. Tan, W. Ma, H. Wu, Y. Zheng, *et al*. *Cell*. **184**, 741–58 (2021).

[22] J. den Hoed, E. de Boer, N. Voisin, A. J. M. Dingemans, *et al*. *Am. J. Hum. Genet.* **108**, 346–56 (2021).

[23] E. Bayram, Y. Topcu, P. Karakaya, U. Yis, *et al. Eur. J. Paediatr. Neurol.*. **17**, 1–6 (2013).

The Tissue-Aware Conservation Inference Toolkit (TACIT)

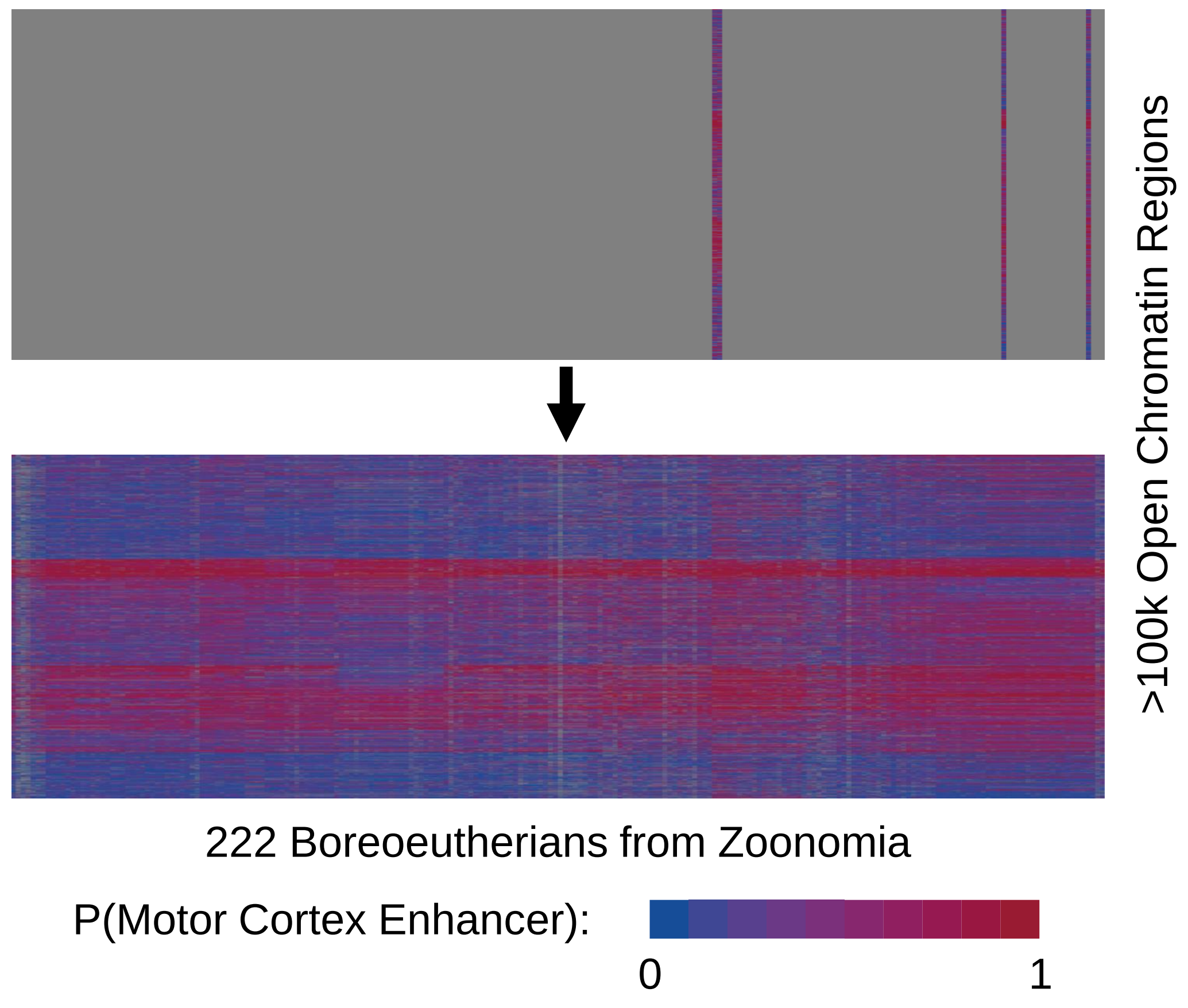
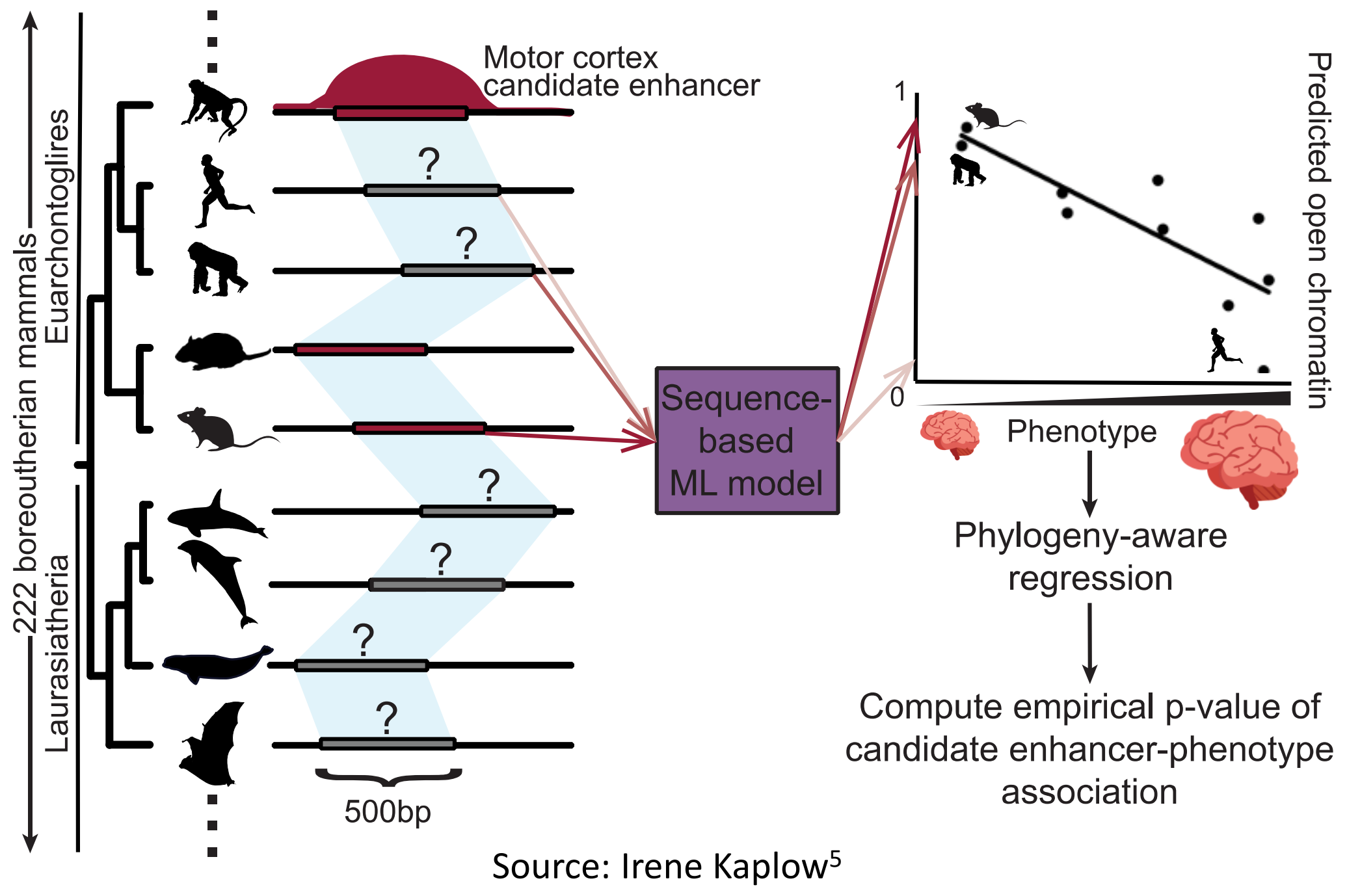


Figure 2. Predicting OCR Activity.


222 boreoutherian mammals
Euarchontoglires
Laurasiatheria
Motor cortex candidate enhancer
?
500bp
Sequence-based ML model
1
0
Predicted open chromatin
Phenotype
Phylogeny-aware regression
Compute empirical p-value of candidate enhancer-phenotype association

Source: Irene Kaplow[5]

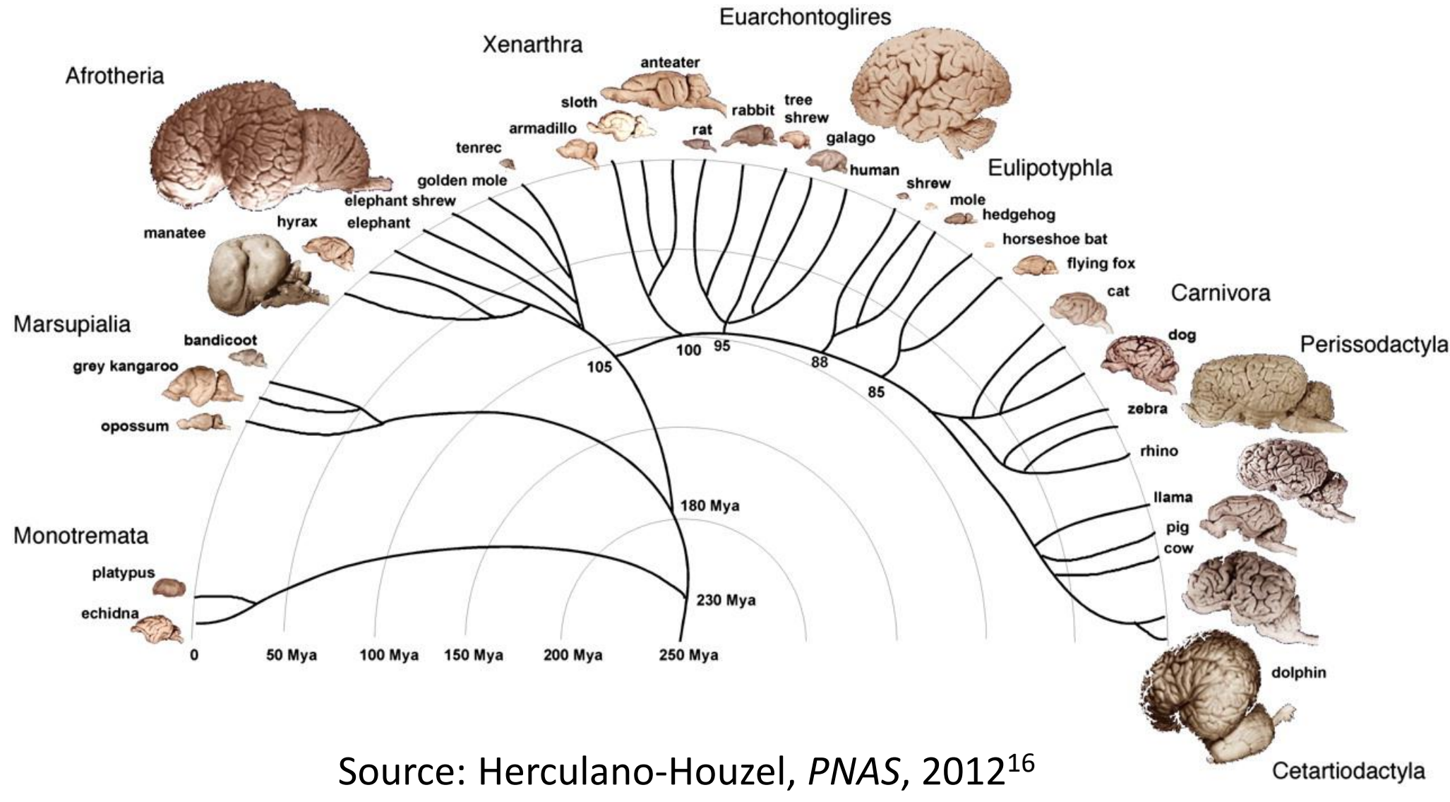
Figure 3. TACIT Overview

A model trained on OCRs from a few species was used to make predictions of OCR activity across orthologous sequences in 222 mammals. These predictions were then correlated with a phenotype (e.g. brain size), and significance was determined by comparison to simulated null phenotypes.

- Train machine learning models to predict enhancer activity in specific tissues
  - In this project: CNNs[4,5] on OCRs identified in brain regions[6,7,8] & cell types[9,10] of 2-5 species
- Predict enhancer activity across many species with aligned genomes (**Fig. 2-3**)
  - In this project: >200 mammals[11] in a Cactus alignment[12]
- Find correlations between predicted OCR activity and phenotype annotations (**Fig. 3**)
- Fit line (or logistic curve) accounting for phylogenetic relationships[13]
- Compute p-values by comparison with fit to null phenotype distribution
  - Phylogenetic permulations[14] preserve the tree topology of the phenotypes
- Study associated enhancers to provide insight into regulatory mechanisms governing phenotypes

Brain Size Results[5]

- Using brain size w/ body mass regressed out[15]
  - Large variation across mammals[16] (**Fig. 4**)
- Known to have evolved through regulatory sequence deletion in humans[17]
- 34 motor cortex and 13 parvalbumin-neuron OCRs with significant associations ($p_{FDR} < 0.05$)
- 41 of 47 near known neurodevelopmental genes



Source: Herculano-Houzel, *PNAS*, 2012[16]

Figure 3. Mammalian brain size.

Brain Size Highlights

- Positively-associated motor cortex OCR near *SALL3* (**Fig. 5A**)
  - *SALL3* regulates neuron maturation[18]
  - In human cortex, OCR is physically close to *SALL3* and not other genes[19]
- Negatively-associated motor cortex OCR near *LRIG1* (**Fig. 5B**)
  - *LRIG1* regulates neural precursor development[20]
  - OCR is physically close to *LRIG1* in both human and mouse cortices[19,21]
- Two negatively-associated motor cortex OCRs near the gene *SATB1* (**Fig. 5C-D**)
  - Mutations in *SATB1* cause abnormal brain size[22]
  - One physically close to *SATB1* in mouse cortex[21]
- Two negatively-associated parvalbumin-neuron OCRs near the gene *Mocs2*
  - Mutations in *Mocs2* also result in abnormally small brains[23]

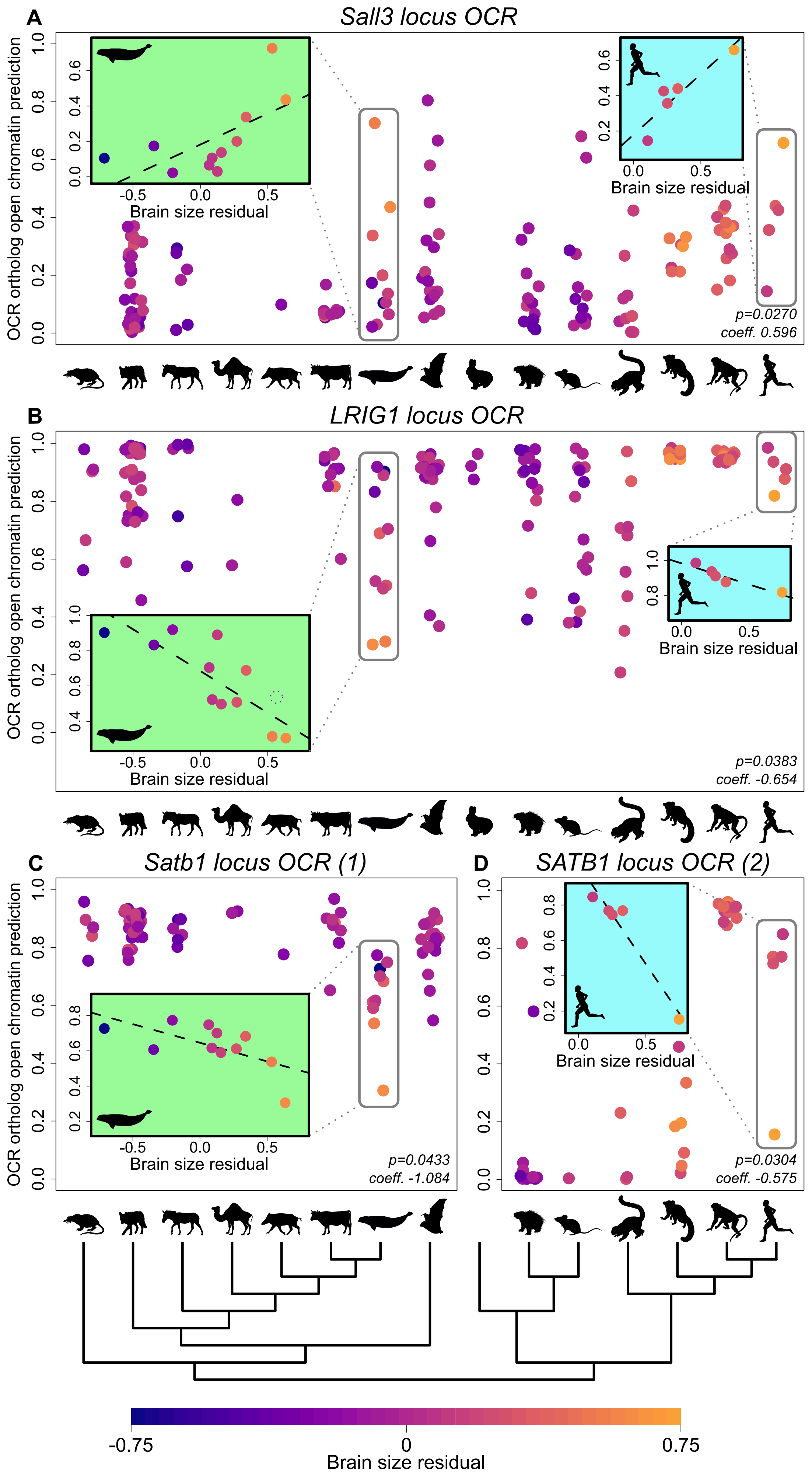


Figure 5. Selected motor cortex OCRs associated with brain size.

Each point represents one ortholog, grouped along the x-axis by clade as shown by the tree below. Associations in the Hominoid and Cetacean clades are shown in blue and green insets. Points are colored by brain size residual following the scale below.

Other Results

- Two OCRs in a key locus associated with social behavior in humans and mice[5]
- 53 OCRs associated with vocal learning[8]
- Extension: Train a CNN to predict whether OCRs are involved in response to neuron activation
  - Insufficient accuracy to use in associations

Acknowledgements

I would like to thank Irene Kaplow for her continued mentorship and several contributions, including training motor cortex models, Andreas Pfenning for his mentorship, Alyssa Lawler for her work on parvalbumin neurons, Morgan Wirthlin for her work on vocal learning, as well as the many other members of the Neurogenomics Lab and collaborators who worked on this project. I would also like to thank Elinor Karlsson, Kerstin Lindblad-Toh, and the other members of the Zoonomia Consortium. This work used the Extreme Science and Engineering Discovery Environment, through the Pittsburgh Supercomputing Center Bridges and Bridges-2 systems.