RESEARCH ARTICLE SUMMARY

ZOONOMIA

Relating enhancer genetic variation across mammals to complex phenotypes using machine learning

Irene M. Kaplow*†, Alyssa J. Lawler†, Daniel E. Schäffer†, Chaitanya Srinivasan, Heather H. Sestili, Morgan E. Wirthlin, BaDoi N. Phan, Kavya Prasad, Ashley R. Brown, Xiaomeng Zhang, Kathleen Foley, Diane P. Genereux, Zoonomia Consortium, Elinor K. Karlsson, Kerstin Lindblad-Toh, Wynn K. Meyer, Andreas R. Pfenning*

The corresponding author

Carnegie Mellon University

Neuroscience Institute

About Us | People | Research | Academics

Neuroscience Institute › People › Faculty › Andreas Pfenning

Andreas Pfenning

Assistant Professor, Computational Biology and Neuroscience Institute

› Contact

Research Areas

Computational Neuroscience, Computational, Mathematical & Statistical Methods, Diseases & Disorders, Language & Reading, Motor Control, Systems Neuroscience

My motivation is to understand the principles that govern complex vertebrate behaviors and neurological disorders from a genetic and evolutionary perspective. I have a broad base of knowledge in computational biology, computer science, statistics, neurobiology, general biology, avian biology, genomics, genetics, personal genomics, and epigenetics, with a focus on analysis of stimulus-activated gene regulation in the brain.

I was born in Pittsburgh and grew up there in the Squirrel Hill neighborhood. My father's from Germany and I picked up the language by spending summers there. These days, I enjoy running, swimming, cross country skiing, and cooking with my wife, Mary. When we can, we spend time at York Beach in Maine or in the White Mountains of New Hampshire.

Background

Kaplow *et al. BMC Genomics* (2022) 23:291
https://doi.org/10.1186/s12864-022-08450-7

BMC Genomics

RESEARCH Open Access

Inferring mammalian tissue-specific regulatory conservation by predicting tissue-specific differences in open chromatin

Irene M. Kaplow[1,2*], Daniel E. Schäffer[1], Morgan E. Wirthlin[1,2], Alyssa J. Lawler[2,3], Ashley R. Brown[1,2], Michael Kleyman[1,2] and Andreas R. Pfenning[1,2,3*]

Abstract

Background: Evolutionary conservation is an invaluable tool for inferring functional significance in the genome, including regions that are crucial across many species and those that have undergone convergent evolution. Computational methods to test for sequence conservation are dominated by algorithms that examine the ability of one or more nucleotides to align across large evolutionary distances. While these nucleotide alignment-based approaches have proven powerful for protein-coding genes and some non-coding elements, they fail to capture conservation of many enhancers, distal regulatory elements that control spatial and temporal patterns of gene expression. The function of enhancers is governed by a complex, often tissue- and cell type-specific code that links combinations of transcription factor binding sites and other regulation-related sequence patterns to regulatory activity. Thus, function of orthologous enhancer regions can be conserved across large evolutionary distances, even when nucleotide turnover is high.

Results: We present a new machine learning-based approach for evaluating enhancer conservation that leverages the combinatorial sequence code of enhancer activity rather than relying on the alignment of individual nucleotides. We first train a convolutional neural network model that can predict tissue-specific open chromatin, a proxy for enhancer activity, across mammals. Next, we apply that model to distinguish instances where the genome sequence would predict conserved function versus a loss of regulatory activity in that tissue. We present criteria for systematically evaluating model performance for this task and use them to demonstrate that our models accurately predict tissue-specific conservation and divergence in open chromatin between primate and rodent species, vastly outperforming leading nucleotide alignment-based approaches. We then apply our models to predict open chromatin at orthologs of brain and liver open chromatin regions across hundreds of mammals and find that brain enhancers associated with neuron activity have a stronger tendency than the general population to have predicted lineage-specific open chromatin.

Conclusion: The framework presented here provides a mechanism to annotate tissue-specific regulatory function across hundreds of genomes and to study enhancer evolution using predicted regulatory differences rather than nucleotide-level conservation measurements.

Keywords: Gene expression evolution, Open chromatin prediction, Machine learning, Enhancers

*Correspondence: ikaplow@cs.cmu.edu; apfenning@cmu.edu
[1] Department of Computational Biology, Carnegie Mellon University, Pittsburgh, PA, USA
Full list of author information is available at the end of the article

2022

RESEARCH ARTICLE SUMMARY

ZOONOMIA

Relating enhancer genetic variation across mammals to complex phenotypes using machine learning

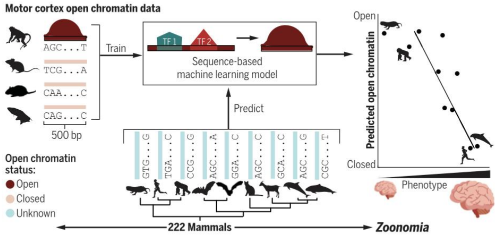
Irene M. Kaplow*†, Alyssa J. Lawler†, Daniel E. Schäffer†, Chaitanya Srinivasan, Heather H. Sestili, Morgan E. Wirthlin, BaDoi N. Phan, Kavya Prasad, Ashley R. Brown, Xiaomeng Zhang, Kathleen Foley, Diane P. Genereux, Zoonomia Consortium, Elinor K. Karlsson, Kerstin Lindblad-Toh, Wynn K. Meyer, Andreas R. Pfenning*

INTRODUCTION: Diverse phenotypes, including large brains relative to body size, group living, and vocal learning ability, have evolved multiple times throughout mammalian history. These shared phenotypes may have arisen repeatedly by means of common mechanisms discernible through genome comparisons.

RATIONALE: Protein-coding sequence differences have failed to fully explain the evolution of multiple mammalian phenotypes. This suggests that these phenotypes have evolved at least in part through changes in gene expression, meaning that their differences across species may be caused by differences in genome sequence at enhancer regions that control gene expression in specific tissues and cell types. Yet the enhancers involved in phenotype evolution are largely unknown. Sequence conservation–based approaches for identifying such enhancers are limited because enhancer activity can be conserved even when the individual nucleotides within the sequence are poorly conserved. This is due to an overwhelming number of cases where nucleotides turn over at a high rate, but a similar combination of transcription factor binding sites and other sequence features can be maintained across millions of years of evolution, allowing the function of the enhancer to be conserved in a particular cell type or tissue. Experimentally measuring the function of orthologous enhancers across dozens of species is currently infeasible, but new machine learning methods make it possible to make reliable sequence-based predictions of enhancer function across species in specific tissues and cell types.

RESULTS: To overcome the limits of studying individual nucleotides, we developed the Tissue-Aware Conservation Inference Toolkit (TACIT). Rather than measuring the extent to which individual nucleotides are conserved across a region, TACIT uses machine learning to test whether the function of a given part of the genome is likely to be conserved. More specifically, convolutional neural networks learn the tissue- or cell type–specific regulatory code connecting genome sequence to enhancer activity using candidate enhancers identified from only a few species. This approach allows us to accurately associate differences between species in tissue or cell type–specific enhancer activity with genome sequence differences at enhancer orthologs. We then connect these predictions of enhancer function to phenotypes across hundreds of mammals in a way that accounts for species' phylogenetic relatedness. We applied TACIT to identify candidate enhancers from motor cortex and parvalbumin neuron open chromatin data that are associated with brain size relative to body size, solitary living, and vocal learning across 222 mammals. Our results include the identification of multiple candidate enhancers associated with brain size relative to body size, several of which are located in linear or three-dimensional proximity to genes whose protein-coding mutations have been implicated in microcephaly or macrocephaly in humans. We also identified candidate enhancers associated with the evolution of solitary living near a gene implicated in separation anxiety and other enhancers associated with the evolution of vocal learning ability. We obtained distinct results for bulk motor cortex and parvalbumin neurons, demonstrating the value in applying TACIT to both bulk tissue and specific minority cell type populations. To facilitate future analyses of our results and applications of TACIT, we released predicted enhancer activity of >400,000 candidate enhancers in each of 222 mammals and their associations with the phenotypes we investigated.

CONCLUSION: TACIT leverages predicted enhancer activity conservation rather than nucleotide-level conservation to connect genetic sequence differences between species to phenotypes across large numbers of mammals. TACIT can be applied to any phenotype with enhancer activity data available from at least a few species in a relevant tissue or cell type and a whole-genome alignment available across dozens of species with substantial phenotypic variation. Although we developed TACIT for transcriptional enhancers, it could also be applied to genomic regions involved in other components of gene regulation, such as promoters and splicing enhancers and silencers. As the number of sequenced genomes grows, machine learning approaches such as TACIT have the potential to help make sense of how conservation of, or changes in, subtle genome patterns can help explain phenotype evolution. ■



Tissue-Aware Conservation Inference Toolkit (TACIT) associates genetic differences between species with phenotypes. TACIT works by generating open chromatin data from a few species in a tissue related to a phenotype, using the sequences underlying open and closed chromatin regions to train a machine learning model for predicting tissue-specific open chromatin and associating open chromatin predictions across dozens of mammals with the phenotype. [Species silhouettes are from PhyloPic]

The list of author affiliations is available in the full article online.
*Corresponding author. Email: ikaplow@cs.cmu.edu (I.M.K.); apfenning@cmu.edu (A.R.P.)
†These authors contributed equally to this work.

Cite this article as I. M. Kaplow *et al.*, *Science* **380**, eabm7993 (2023). DOI: 10.1126/science.abm7993

READ THE FULL ARTICLE AT https://doi.org/10.1126/science.abm7993

2023

RESEARCH ARTICLE SUMMARY

ZOONOMIA

Vocal learning–associated convergent evolution in mammalian proteins and regulatory elements

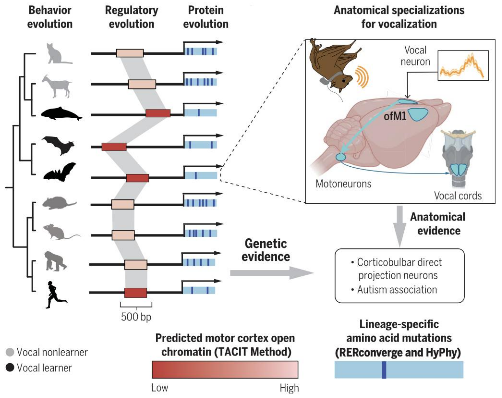
Morgan E. Wirthlin†, Tobias A. Schmid†, Julie E. Elie†, Xiaomeng Zhang, Amanda Kowalczyk, Ruby Redlich, Varvara A. Shvareva, Ashley Rakuljic, Maria B. Ji, Ninad S. Bhat, Irene M. Kaplow, Daniel E. Schäffer, Alyssa J. Lawler, Andrew Z. Wang, BaDoi N. Phan, Siddharth Annaldasula, Ashley R. Brown, Tianyu Lu, Byung Kook Lim, Eiman Azim, Zoonomia Consortium, Nathan L. Clark, Wynn K. Meyer, Sergei L Kosakovsky Pond, Maria Chikina, Michael M. Yartsev*‡, Andreas R. Pfenning*‡

INTRODUCTION: Vocal production learning ("vocal learning"), or the ability to modify vocalizations according to the social environment, forms the basis of human speech production. Among the Boreoeutherian mammals, this trait has evolved independently in four different lineages: humans, bats, cetaceans, and pinnipeds. In vertebrates, the evolution of vocal learning behavior has been associated with the evolution of brain anatomical features, including cortical long-range projection neurons (e.g., songbirds and humans). Moreover, neural circuits for the production of learned vocalization display convergent evolution in patterns of gene expression.

RATIONALE: Despite evidence for the convergent evolution of vocal learning at the behavioral, anatomical, and gene expression levels in vertebrates, the genetic underpinnings of vocal learning and human speech in mammals are poorly understood. New machine learning approaches and the newly sequenced mammalian genomes of the Zoonomia Consortium provide the foundation to rigorously study this question. The repeated evolution of vocal learning across mammals allows us to determine which parts of the genome are significantly associated with the behavior.

RESULTS: First, we studied convergent evolution in protein-coding regions using the RERconverge and HyPhy methods to find 200 significantly associated genes. The genes that tend to be under higher constraint in vocal learning mammals are enriched for genes involved in human autism. However, the vast majority of genes are driven by signals from only one or two clades of vocal learning mammals, suggesting that a large component of the genetic basis for the trait may lie instead in the convergent evolution of regulatory elements. To explore that hypothesis, we performed an anatomical and functional characterization of the Egyptian fruit bat motor cortex. We identified a subregion of the motor cortex that is implicated in vocal production and directly projects to the motoneurons controlling the bat's larynx. This allowed us to profile candidate regulatory elements active in this vocalization-associated subregion of the motor cortex by measuring open chromatin. These open chromatin regions and 222 mammalian genomes of the Zoonomia Consortium served as input to the Tissue-Aware Conservation Inference Toolkit (TACIT) machine learning approach, which was applied to find 50 candidate regulatory elements whose predicted motor cortex open chromatin measurements across mammals are highly correlated with the presence of vocal learning behavior. Many of these open chromatin regions were near genes associated with autism, and they tended to overlap with open chromatin specific to the long-range projection neurons that have been implicated in the evolution of vocal learning.

CONCLUSION: Although it is impossible to know which parts of the genome evolved for human speech production, we are able to use the repeated evolution of a component of that behavior, vocal learning, to find significantly associated genes and noncoding regions. Our results demonstrate that the presence of vocal learning behavior in a given clade leads to weak selective pressure across a broad range of genes and stronger selective pressure across a smaller number of motor cortex noncoding regions. These genes and noncoding regions show an association with autism, which suggests that there are shared regulatory networks for vocal and social behavior that tend to adapt in similar ways when a lineage evolves vocal learning behavior. More broadly, our results suggest that the evolutionary history of selective pressures across a location in the genome can provide insight into what that region might influence human behavior. ■



Finding vocal learning–associated regions of the mammalian genome. We compared the evolution of vocal learning behavior to the evolution of coding and noncoding elements of the genome, leveraging anatomical, electrophysiological, and epigenomic experiments in the Egyptian fruit bat orofacial motor cortex (ofM1). We show convergent evidence of the importance of long-range projection neurons and autism-associated gene networks.

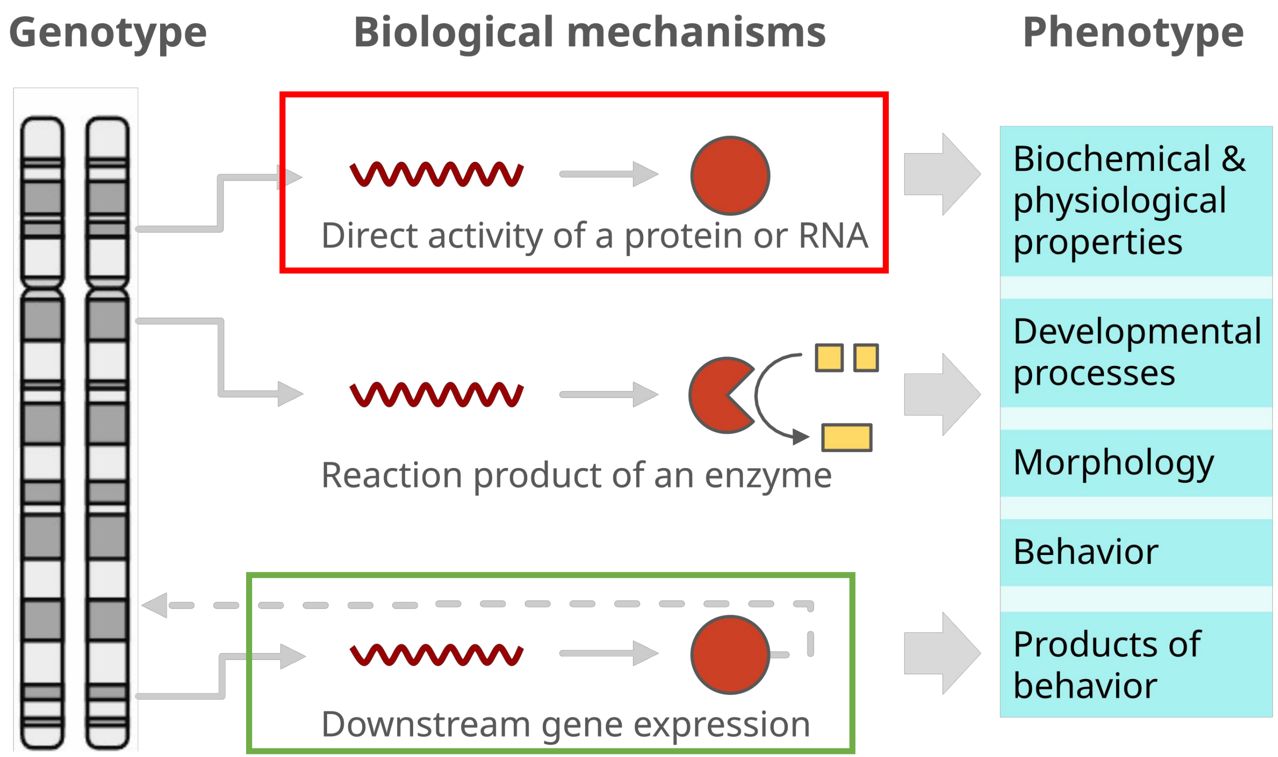
*Corresponding author. Email: myartsev@berkeley.edu (M.M.Y.); apfenning@cmu.edu (A.R.P.)
†These authors contributed equally to this work.
‡These authors contributed equally to this work.

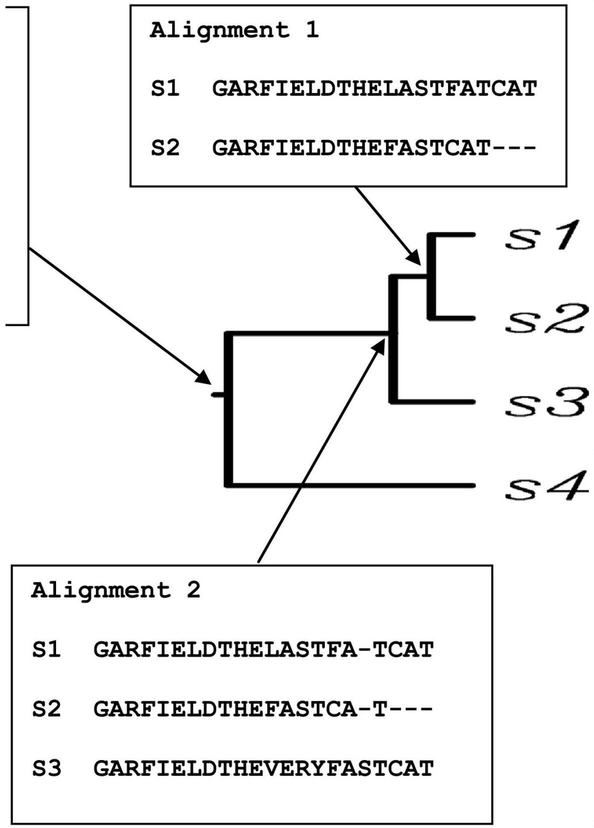
Cite this article as M. Wirthlin *et al.*, *Science* **383**, eabn3263 (2024). DOI: 10.1126/science.abn3263

READ THE FULL ARTICLE AT https://doi.org/10.1126/science.abn3263

2024

Rethinking the genetic basis of traits

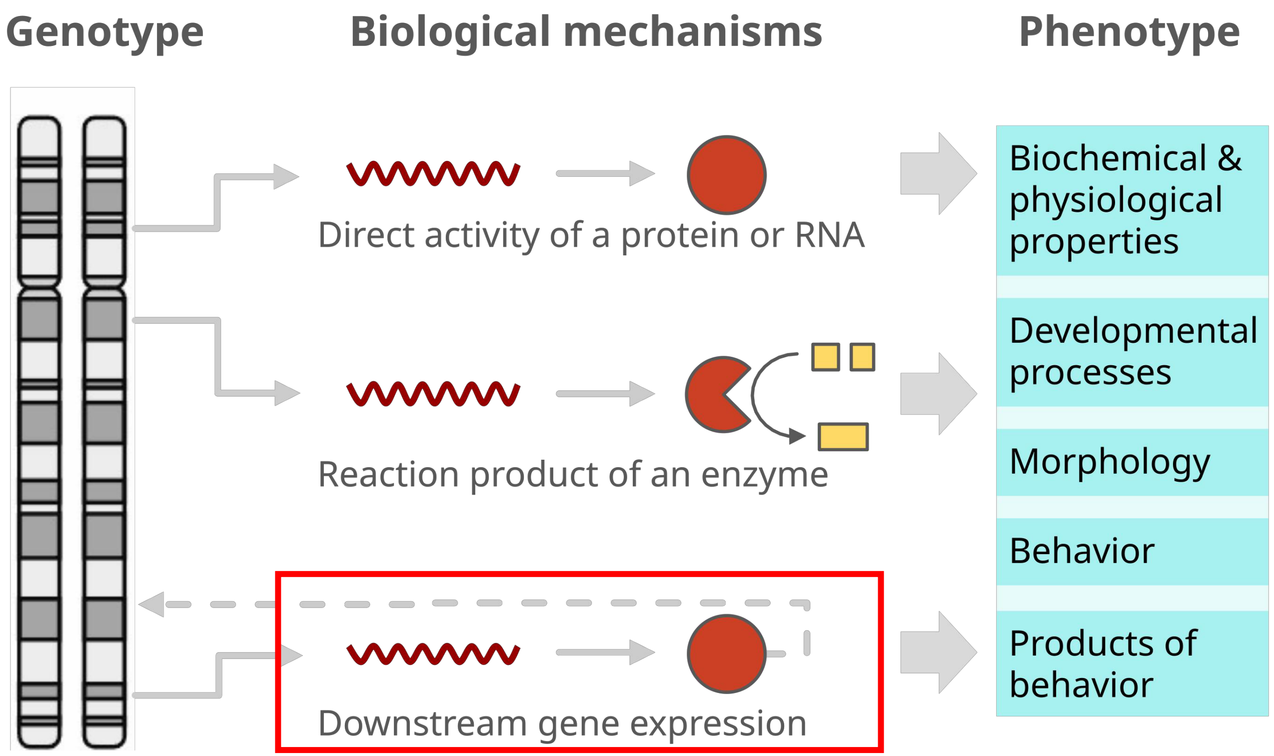


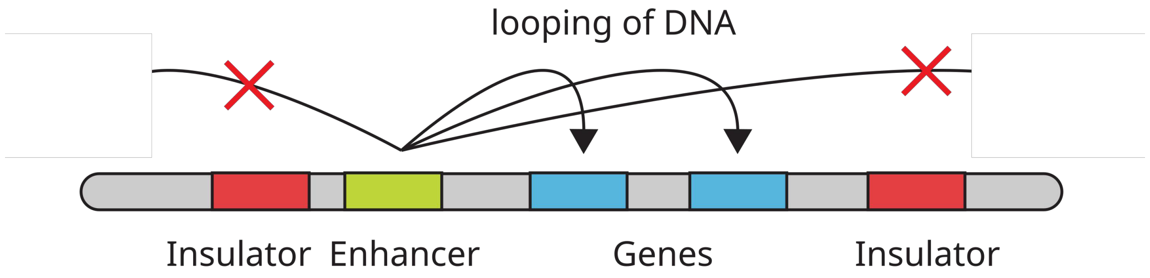


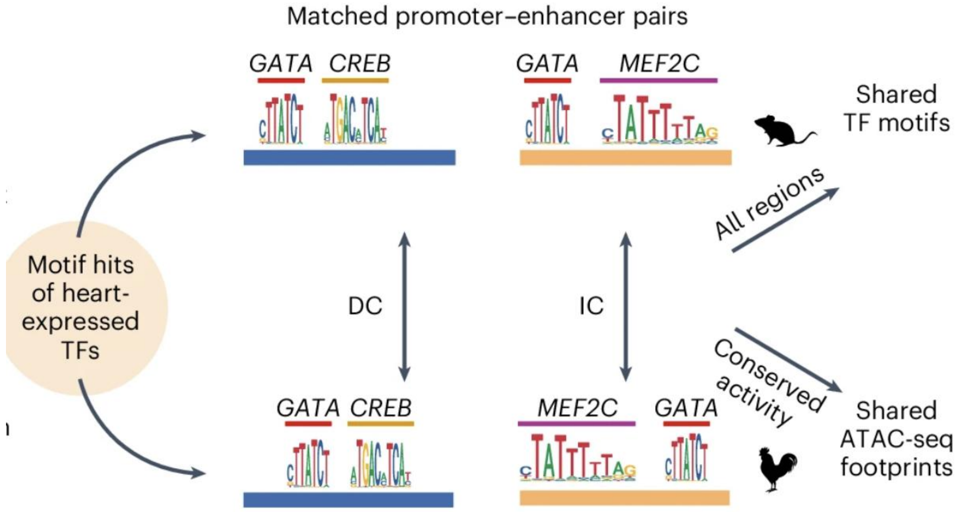
Biston betularia morpha *typica*, the standard light-colored peppered moth

B.betularia morpha *carbonaria*, the melanic form, illustrating discontinuous variation

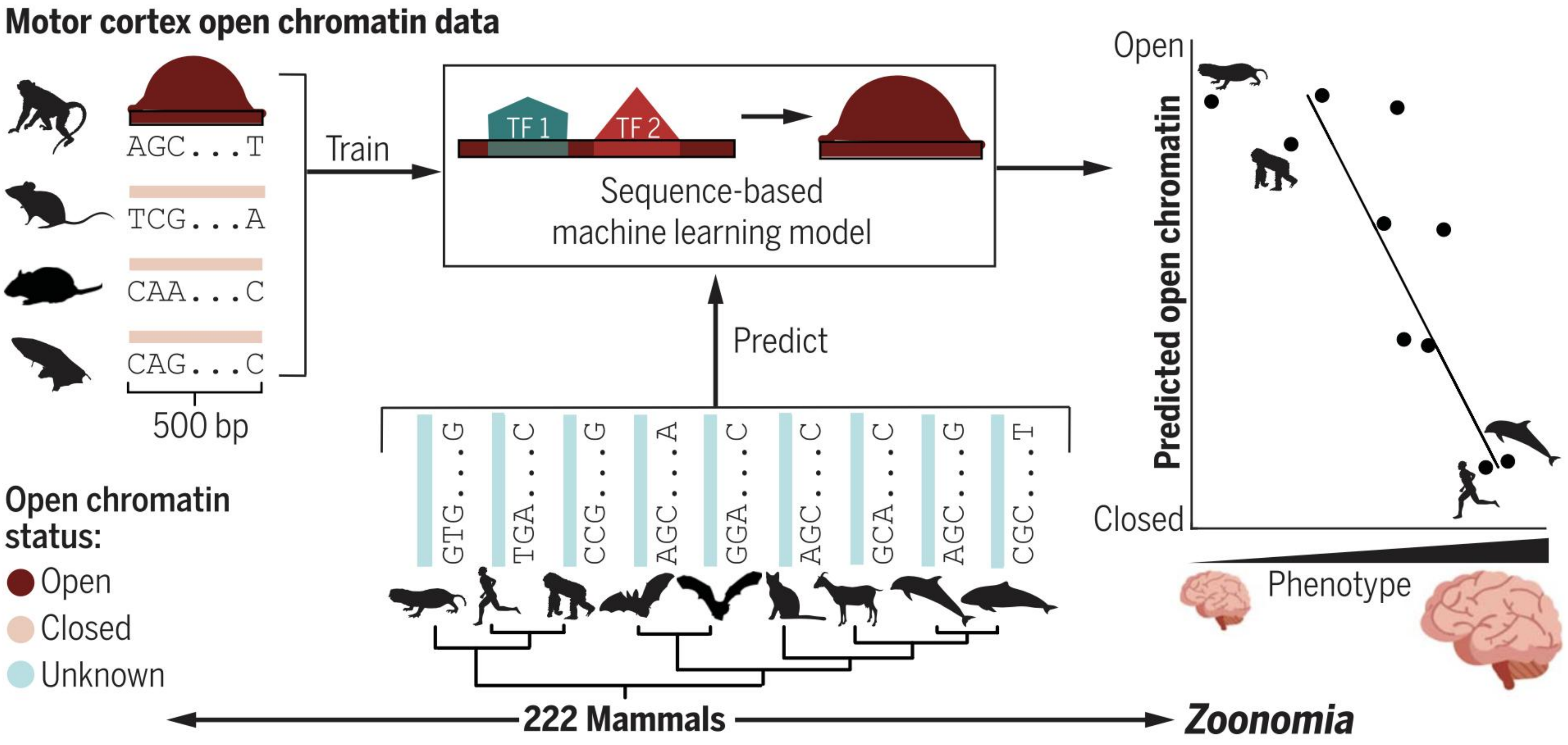
Rethinking the genetic basis of traits





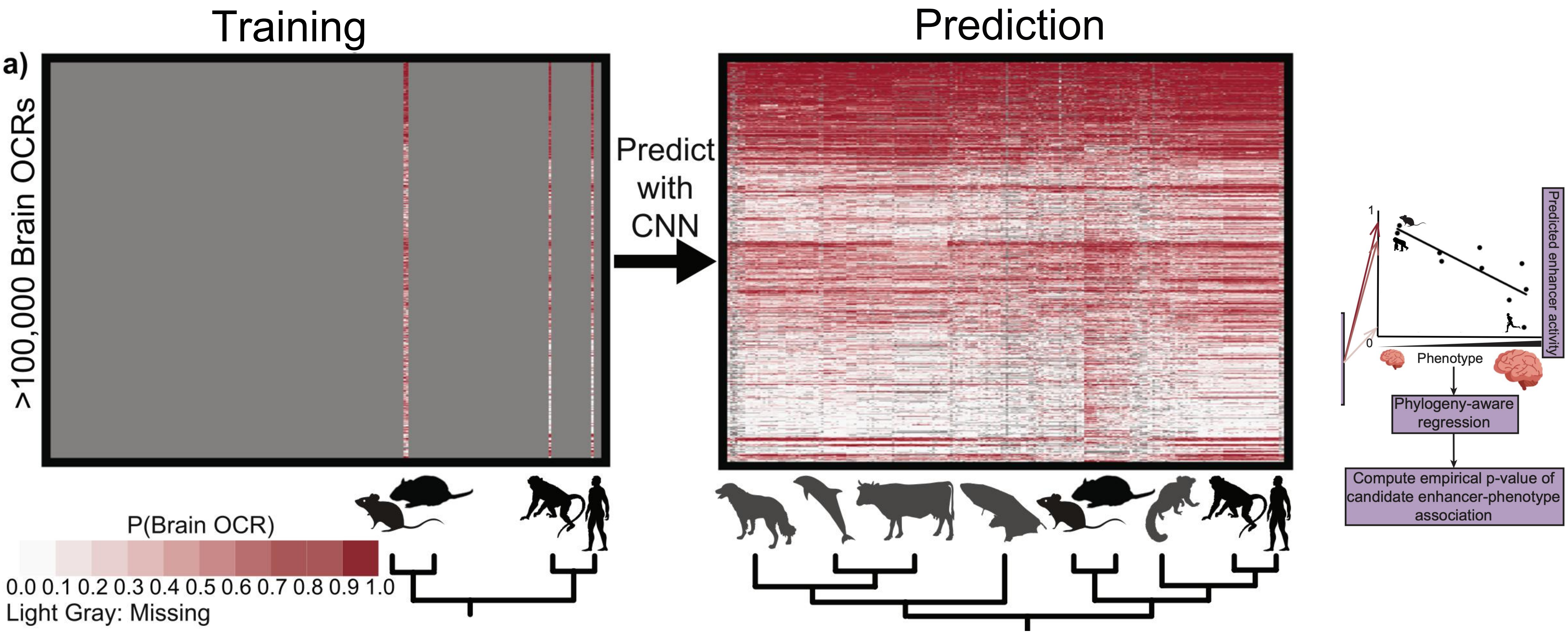


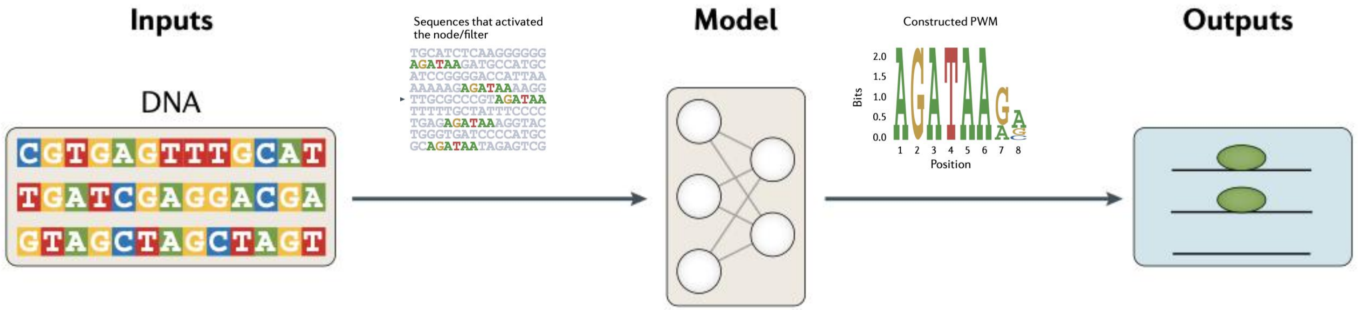
TACIT - Tissue-Aware Conservation Inference Toolkit

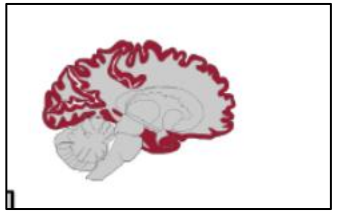


TACIT works by generating open chromatin data from a few species in a tissue related to a phenotype, using the sequences underlying open and closed chromatin regions to train a machine learning model for predicting tissue-specific open chromatin and associating open chromatin predictions across dozens of mammals with the phenotype.

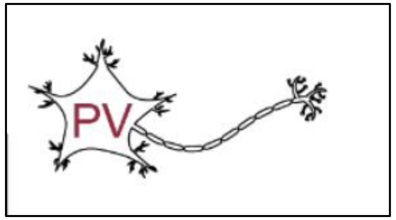
Model Training and Prediction





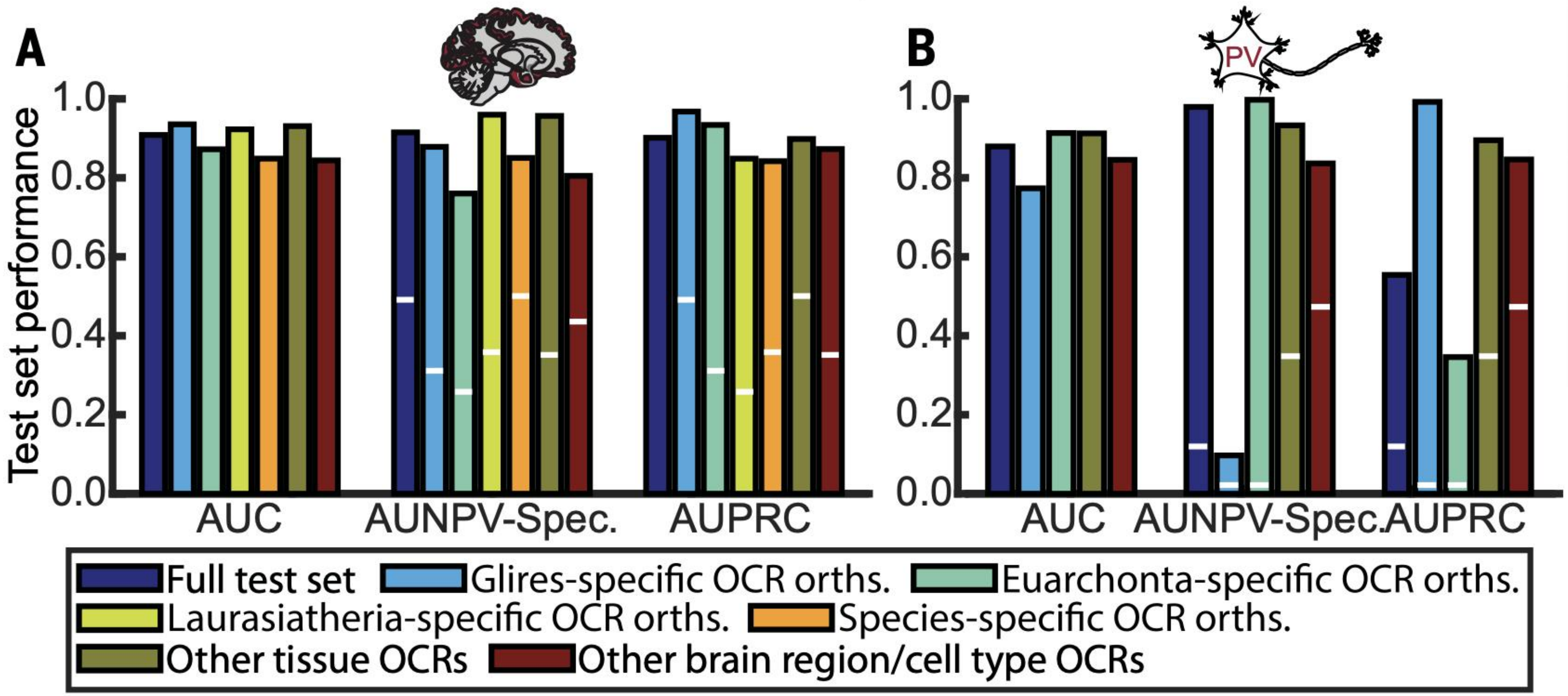


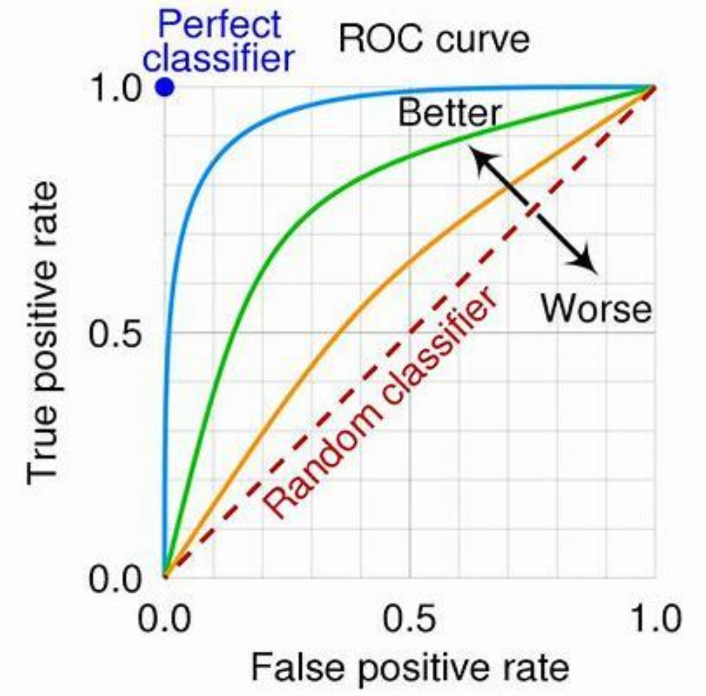
MultiSpeciesMotorC
ortexModel



MultiSpeciesPVM
odel

Performance evaluation - Does it Work?

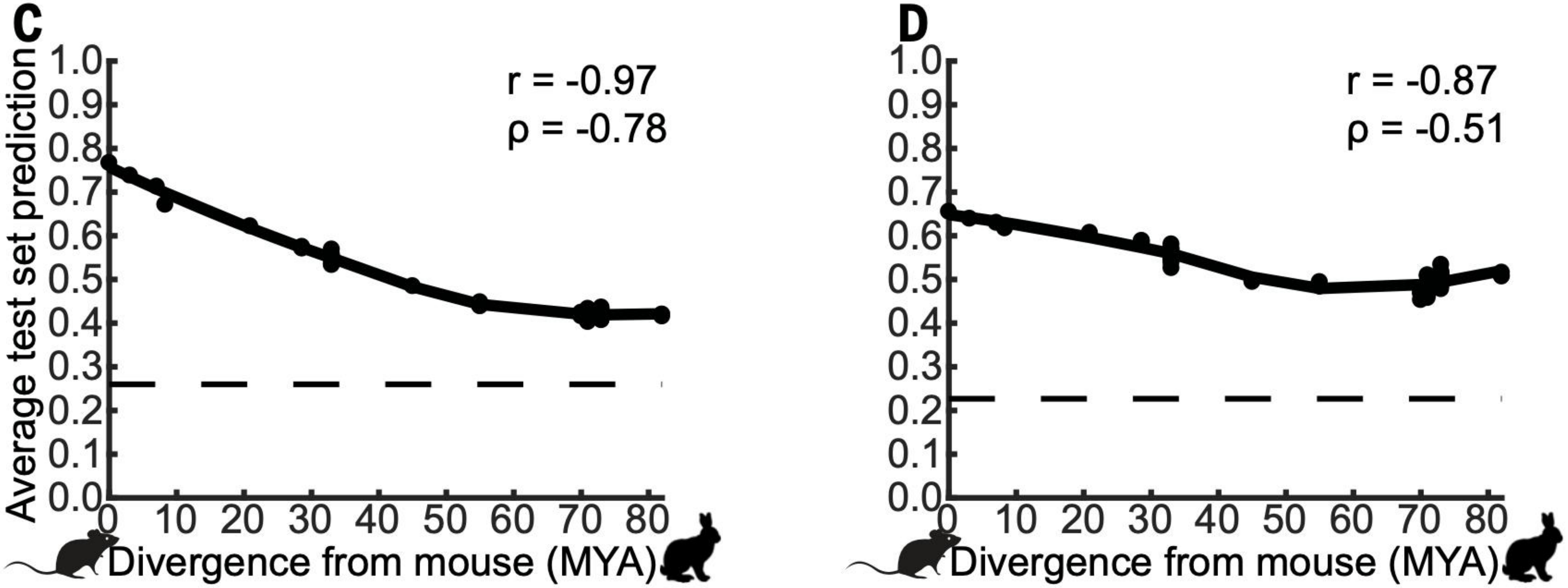




➢ **High Predictive Accuracy**

Both models demonstrate strong performance in predicting open chromatin status.

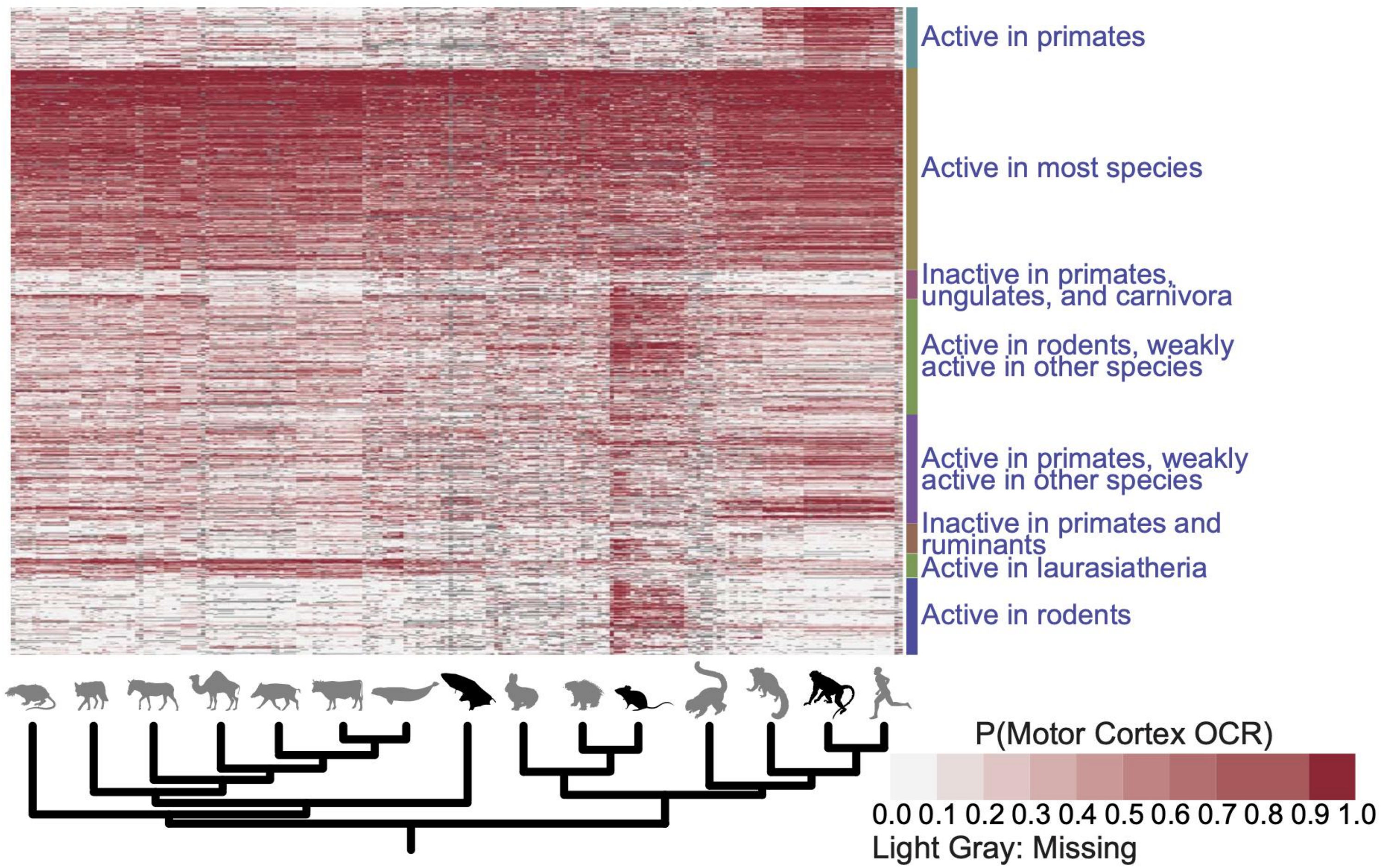
Performance evaluation



- **Phylogenetic Signal**

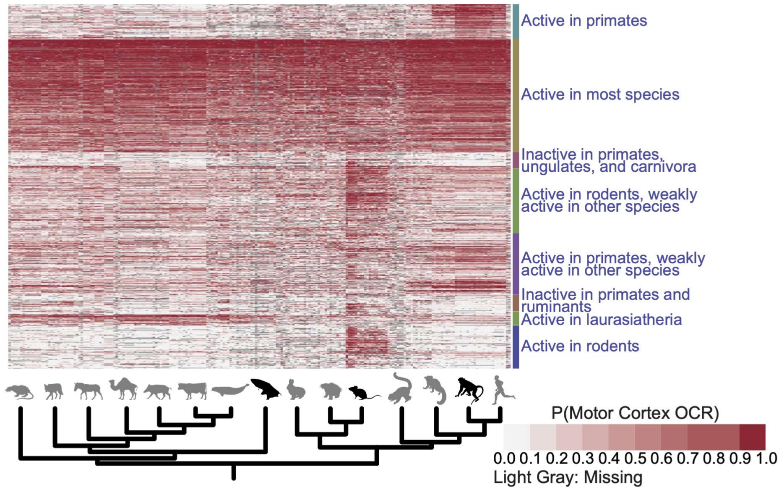
The models' predictions reflect evolutionary relationships.

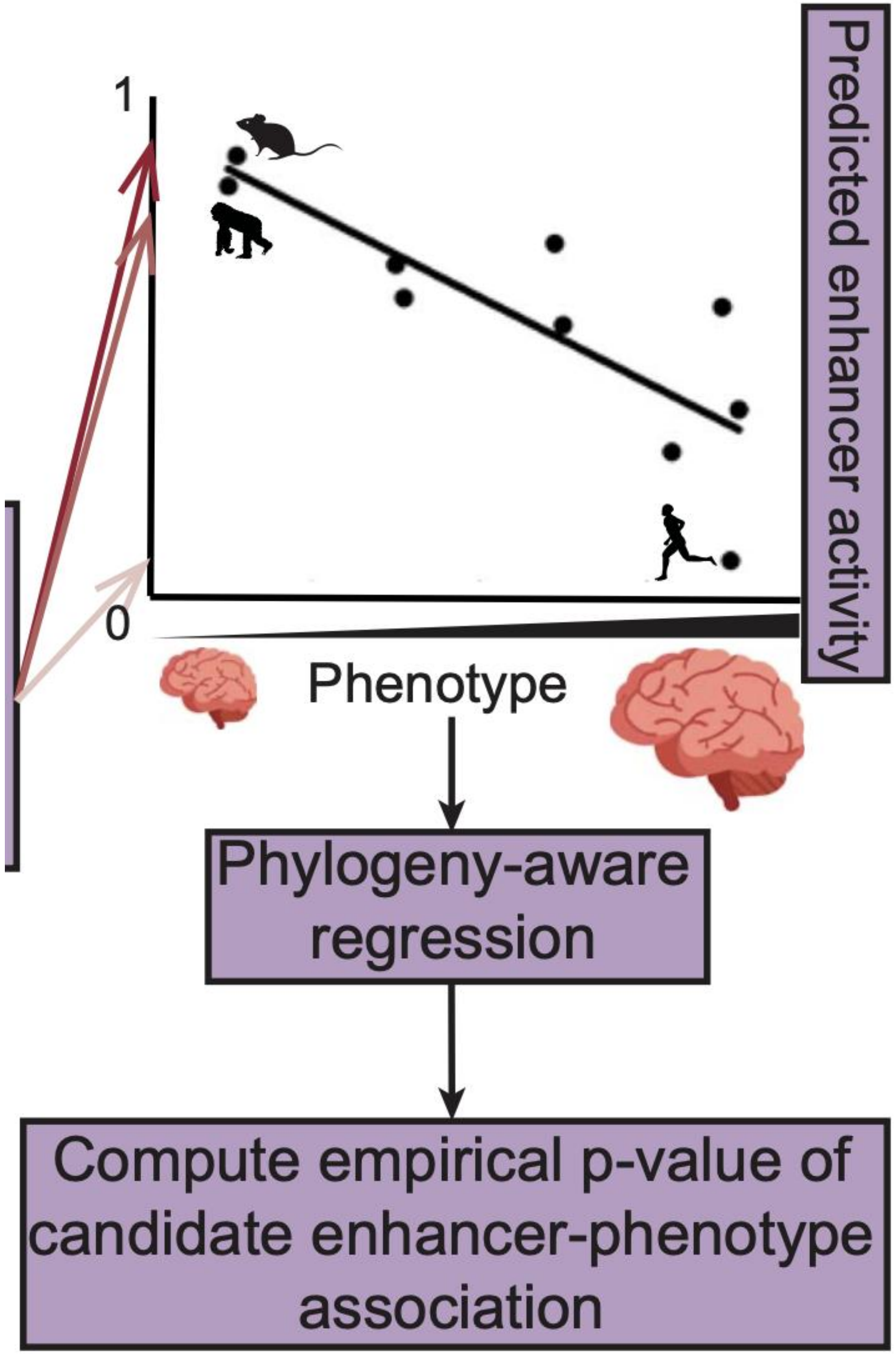
Prediction Across Species



- **Predicted Activity Landscape**
  1. Clustering by Activity Pattern
  2. Phylogenetic Ordering of Species

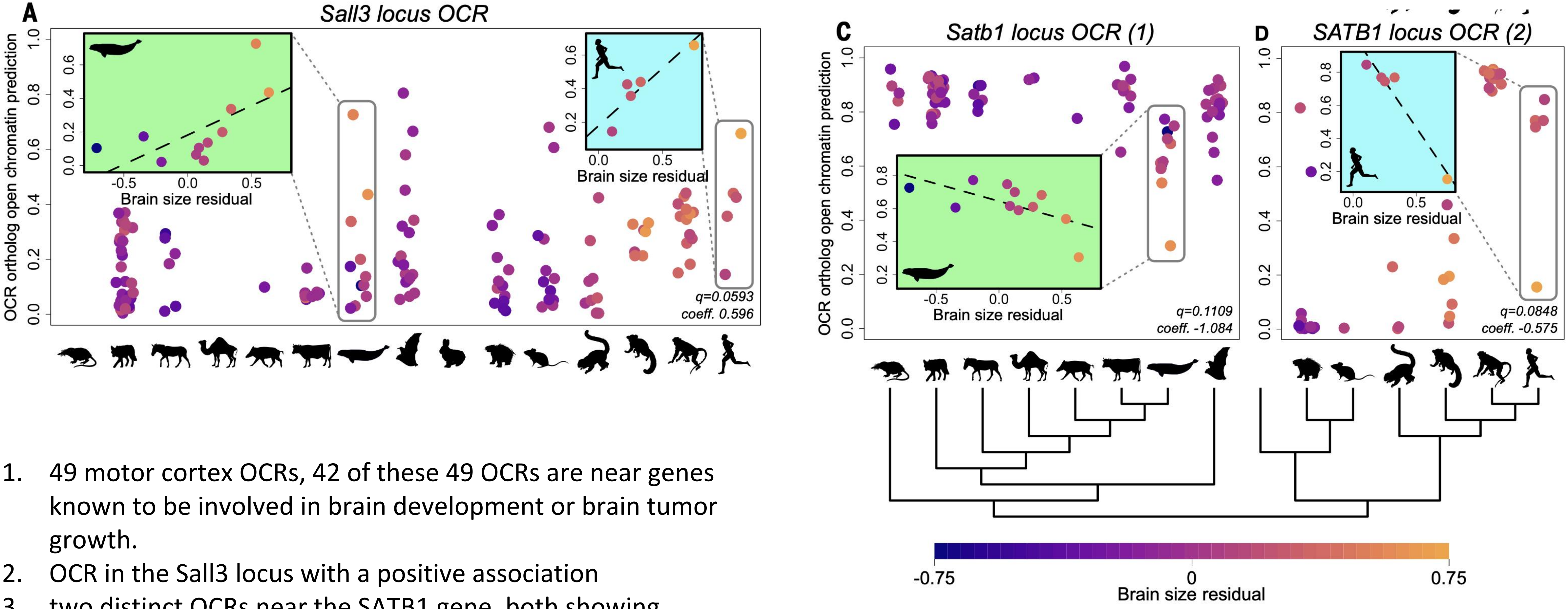
Linking to Phenotypes



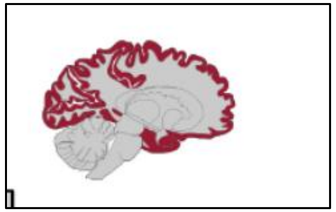


These cross-species predictions of enhancer activity are then statistically associated with specific phenotypes (like brain size or social behavior)

Key Finding - Brain Size-Associated Motor Cortex OCRs

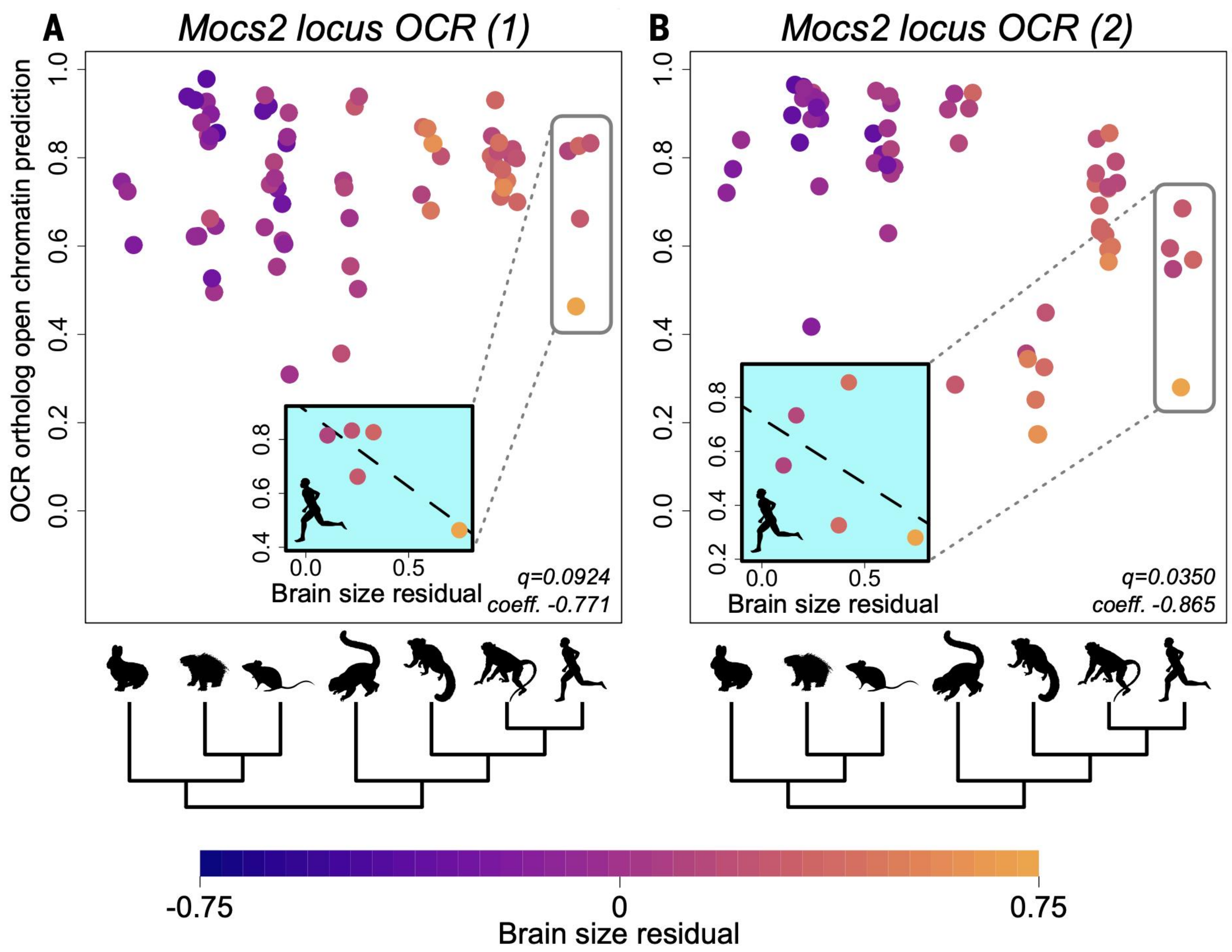


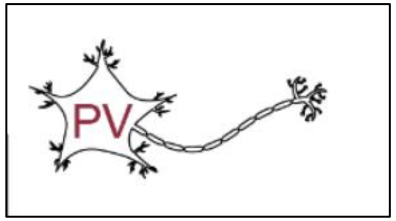
1. 49 motor cortex OCRs, 42 of these 49 OCRs are near genes known to be involved in brain development or brain tumor growth.
2. OCR in the Sall3 locus with a positive association
3. two distinct OCRs near the SATB1 gene, both showing negative associations with brain size residual



MultiSpeciesMotorCortexModel

Key Finding - Brain Size-Associated PV+ Interneurons OCRs





MultiSpeciesPVModel

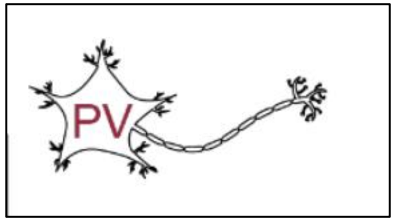
1. 15 OCRs in these neurons associated with brain size residual.
2. OCRs located near the Mocs2 gene exhibit a negative association between their predicted open chromatin and brain size residual

Genetic basis of social behavior

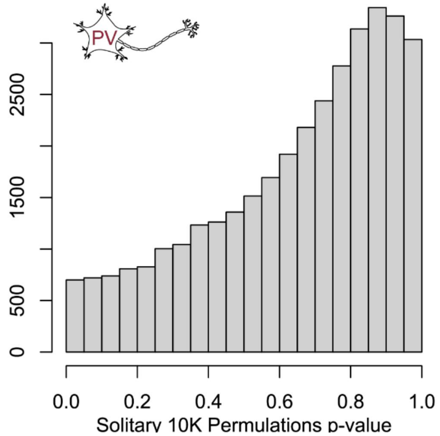
Animals that are solitary are those which have minimal interaction with other members of their species.

Social animals are those animals that interact highly with other animals, usually of their own species (conspecifics)

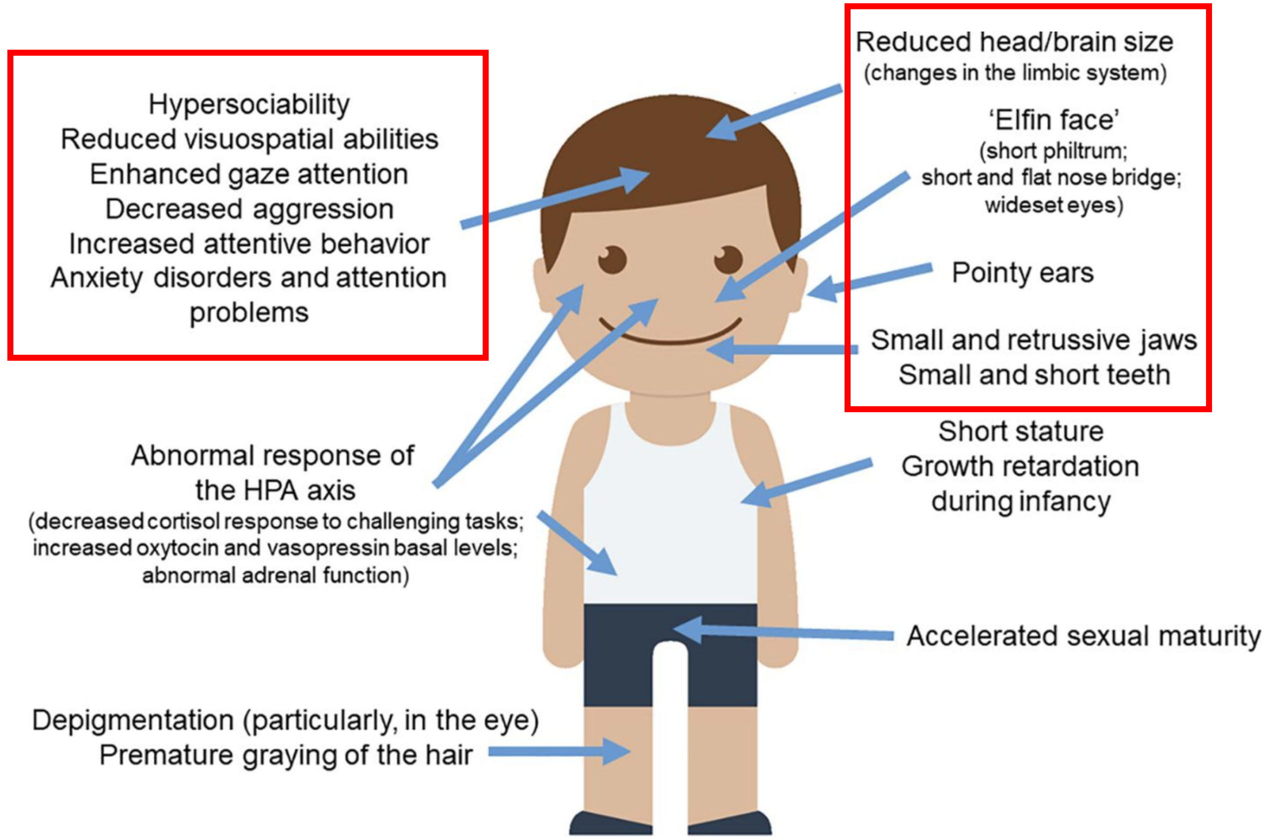
ASD - autism spectrum disorder - 自闭症
SCZ - schizophrenia - 精神分裂症



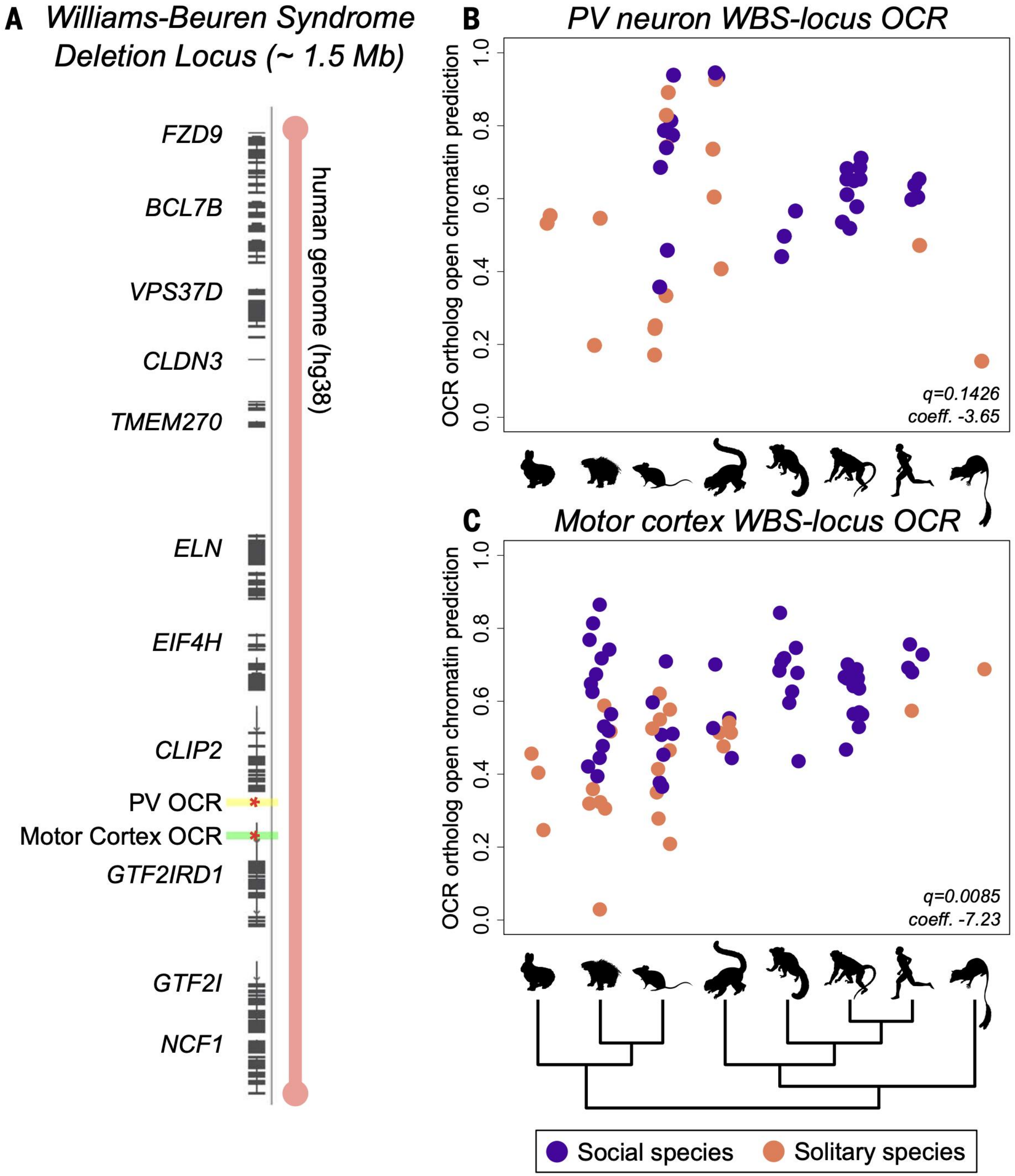
PV+ interneurons regulating social behaviors



Genetic basis of social behavior

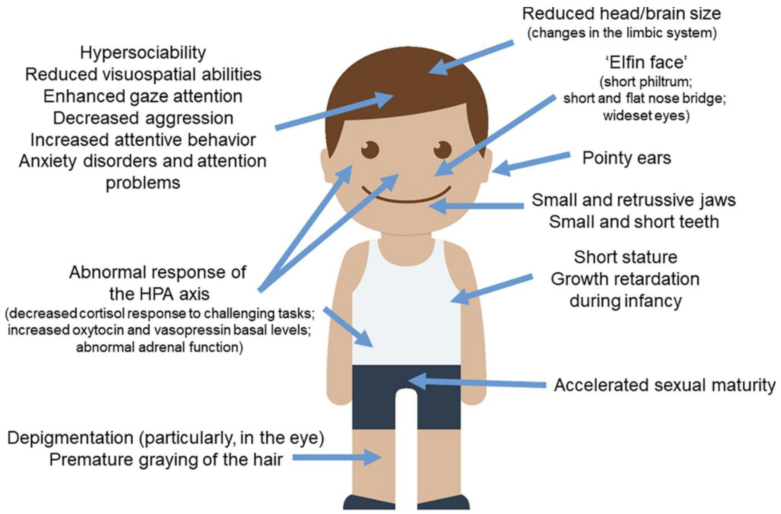


Williams-Beuren Syndrome (WBS)

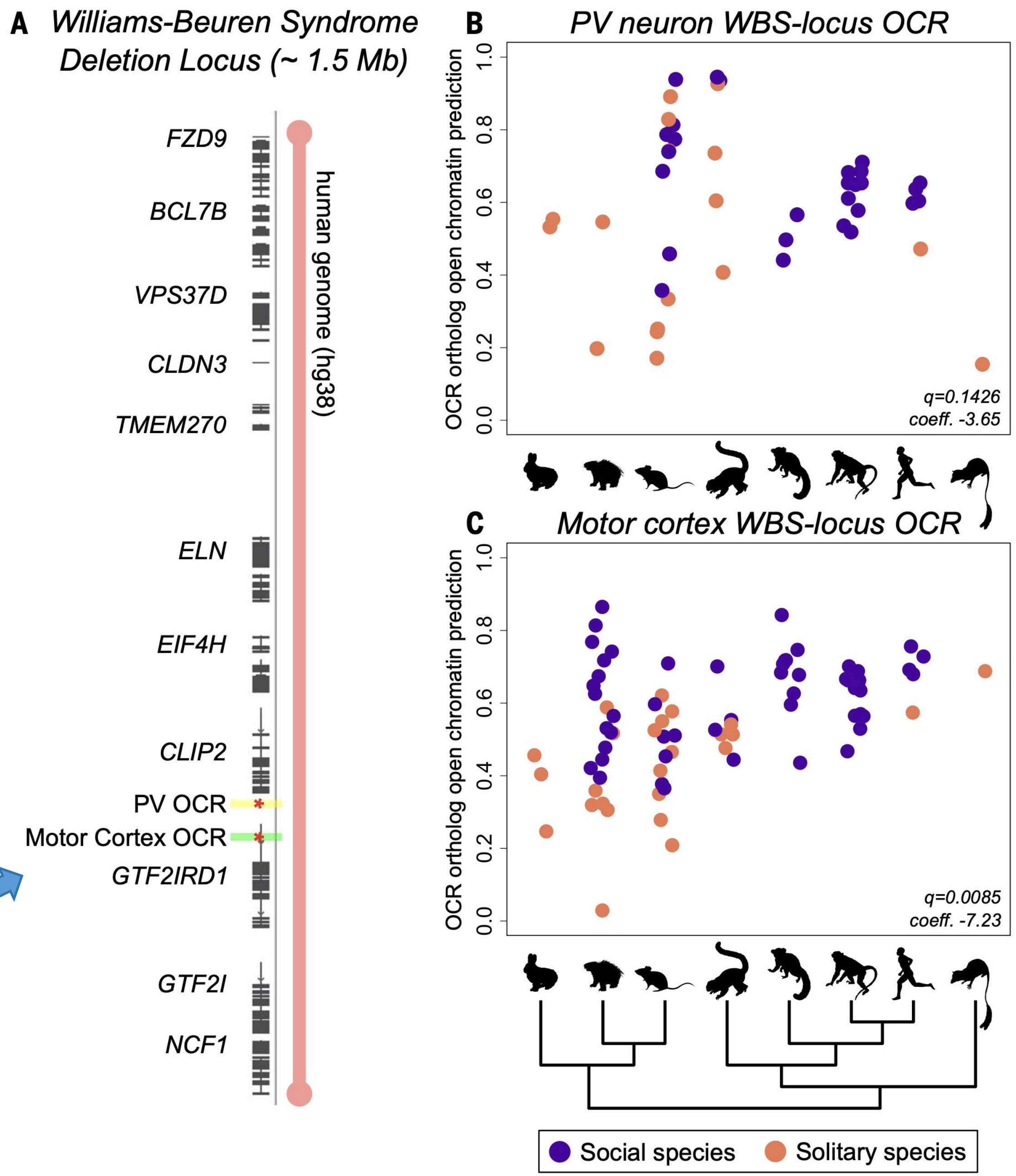


1.5 Mb human Williams-Beuren Syndrome (WBS) deletion locus highlights the locations of specific OCRs identified in PV+ interneurons (yellow) and motor cortex (green)

Genetic basis of social behavior



Williams-Beuren Syndrome (WBS)



1.5 Mb human Williams-Beuren Syndrome (WBS) deletion locus highlights the locations of specific OCRs identified in PV+ interneurons (yellow) and motor cortex (green)

a PV+ interneuron OCR located near the genes GTF2IRD1 and GTF2I shows a marginal negative association between OCR and solitary living across different mammal species.

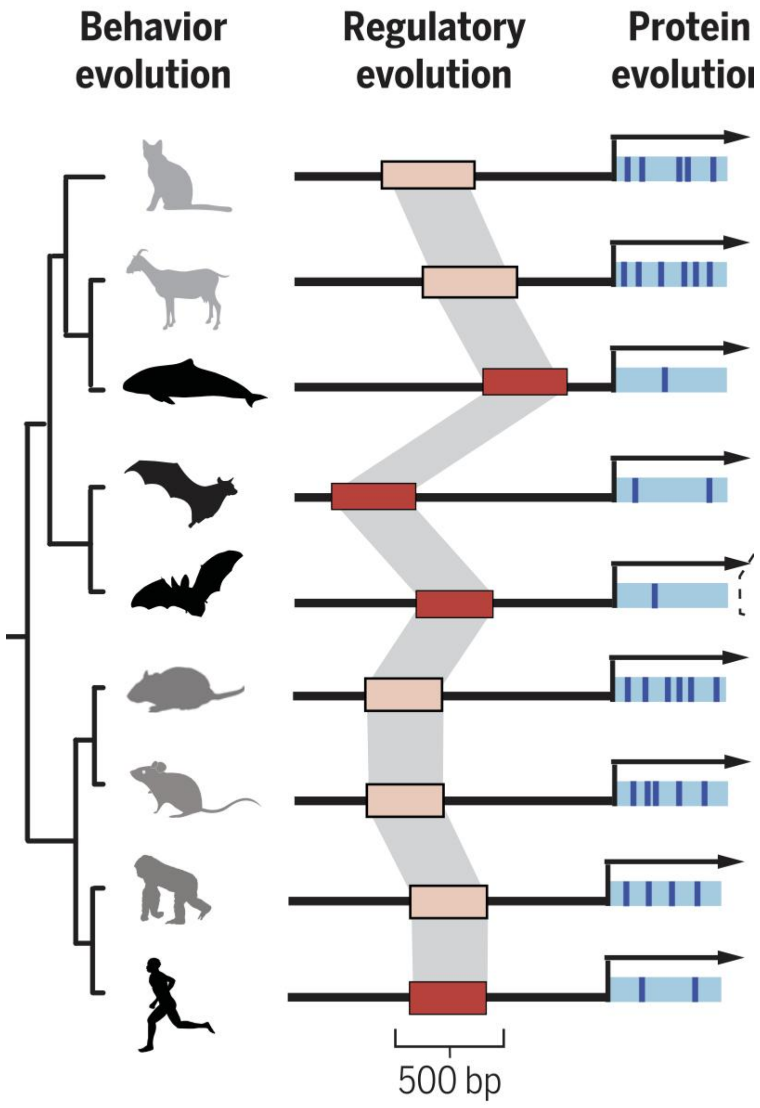
a motor cortex OCR near GTF2IRD1 and GTF2I demonstrates a more statistically significant negative association between predicted motor cortex open chromatin at this site and solitary living.

Takeaways from TACIT

Core Idea: TACIT leverages machine learning predictions of enhancer activity conservation rather than relying solely on nucleotide-level sequence conservation.

1. Advancing the Study of Non-Coding DNA in Evolution by Focus on Enhancer Function
2. TACIT approach could be extended to other genomic regions involved in gene regulation, such as promoters, and splicing enhancers and silencers

- Assumption-Related Limitations
    1. Conserved Regulatory Code - Assumes that the regulatory code for a specific tissue or cell type is conserved across all species in which predictions are made.
    2. Limited to Orthologs
- Experimental Validation:
    1. Confirming Enhancer Activity
    2. Linking Enhancers to Target Genes
    3. Phenotype



2024

Q&A

Thank you for your attention!