Missing sequences: functionally characterizing species-specific deletions of highly conserved DNA. A major question remaining in biology is how fundamental species differences are encoded in the genome. Genome-sequencing technology has only recently enabled the comparison of high-quality genomes from hundreds of species. However, the interpretation of species-defining genome regions is difficult for three reasons: 1) accurate genome comparisons and alignments are computationally intensive; 2) the search space is vast, with millions of alignable bases differing in mammals alone; and 3) these sequence differences are predominantly in non-coding, potentially gene-regulatory, regions where function is difficult to predict.

One experimentally tractable but understudied set of genomic elements are conserved deletions (ConDels) - regions that show evidence of function due to their strong sequence conservation[1]. ConDels can be uniquely informative because they may cause species-specific functions driven by the deletion. First, I will develop new computational approaches based on high-throughput whole-genome alignment to identify ConDels in hundreds of species, greatly expanding the catalog of species-specific genomic elements. With this newly augmented data set, I will assay the function of 100,000+ ConDels across multiple mammals using a massively parallel reporter assay (MPRA). Finally, I will explore how ConDel function works endogenously by identifying differentially bound transcription factors for a subset of ConDels (**Fig. 1)**. This will allow us and other researchers to begin interrogating the interplay of sequence change and speciation.

Aim 1: Computationally identify ConDels in mammalian genomes and their potential impacts. First, I will create alignments for several diverse vertebrates in order to identify species-specific deletions. While whole-genome, comparative multiZ alignments have been generated for commonly investigated species such as human and mouse, assemblies anchored on diverse taxa which can identify deletions in a variety of focal species are lacking. I will build multiZ alignments using new genomes from the 29 Mammals project and the Vertebrate Genome Project, spanning from coelacanths to humans[2,3]. For each of these 157 species, I will use each species' closest relative to anchor a multiZ alignment of all other genomes, generating a list of conserved elements that are predicted to exist in its most recent common ancestor[4,5]. The target species will be excluded from this analysis so as not to bias which regions are identified as conserved. I will then build a pairwise alignment to identify deletions specific to the species[4]. Cloud computing makes scaling this whole-genome alignment approach to hundreds of newly available genomes feasible.

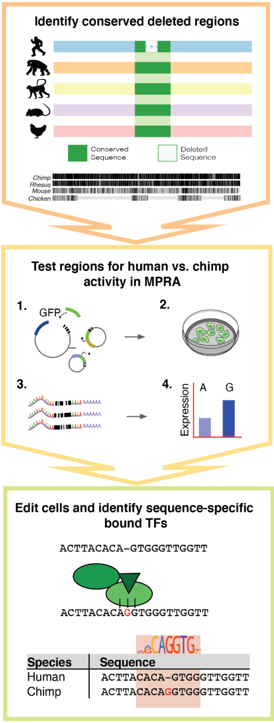

Figure 1. Proposed workflow for species-specific ConDel identification and exploration.

Using this highly detailed catalog of vertebrate ConDels, I will next identify the subset of ConDels impacting gene-regulatory signatures and gene expression, and in turn phenotypes. To identify species-specific regulatory elements overlapping a ConDel, I will first compare existing gene regulatory maps of 20 mammals[6], focusing on liver because this tissue has the most cross-species functional data available. I will also use tissue-matched transcriptomic data[6] to correlate these ConDels with gene expression throughout the genome, as regulatory elements can act at long distances. While most regulatory and expression changes are predicted to cause loss of function, in some cases a change may delete repressive regulatory sequences, leading to gain of function. I will compare ConDels that do, versus do not, show evidence of predicted regulatory impacts in liver, looking for differences in sequence age, complexity, genomic location, or other patterns of functional evolution. Should my computational pipeline fail, I can investigate released[1], smaller ConDel sets' correlation with recently published gene regulation datasets[7]. For subsequent followup, I could identify other tissues enriched for ConDels using preexisting whole-body regulatory maps in humans and mice[7] to expand beyond the liver.

Aim 2: Functionally test ConDels from multiple species using high-throughput reporter assays. Predicting the potential function of a noncoding element is difficult because there is no known 'grammar' analogous to the protein-coding codon alphabet. However, new high-throughput assays like the massively parallel reporter assay (MPRA) allow us to directly measure the individual effects of >50,000 sequence constructs on gene expression at once. MPRA is an episomal assay that inserts a synthetic sequence upstream of a reporter gene with a unique barcode, generating an expression-based readout where barcode enrichment quantifies each elements' enhancer activity[8]. Because the sequence is synthesized artificially, I can test the regulatory potential of multiple ConDels relative to the intact sequences of its closest relative, in order to identify differences modifying expression *in vitro*. Preliminary analysis finds most ConDels are small, ranging from one to five base pairs, making them well-suited for MPRA sequence synthesis.

To examine the functional consequences of species-specific deletions, I will leverage the wide availability of mammal expression profiles to test species-specific ConDels in an MPRA. I will perform this assay for a representative 5,000 ConDels from a range of 10 mammals that have previous characterization data[6], but only in human liver cells, since RNA-seq work has demonstrated that the cellular environment between the same tissue type from different species is similar enough to act as a suitable control environment[6]. This approach will measure the enhancer ability of each conserved element and the ConDel's impact on that enhancer activity, thoroughly characterizing the potential expression effects of these species-specific ConDels. If my MPRA does not produce enough hits to make genome-wide inferences, the technical approach is still well-powered enough that I will be able to follow up with any individual hits I find. Additional experimentation could test the ConDel library in additional cell types, based on what tissues are most relevant for any interesting hits identified in Aim 1 or the literature.

Aim 3: Endogenous characterization of transcription factor (TF) binding for top expression-altering hits. I will next investigate the underlying molecular mechanisms involved in these species-specific ConDels, prioritizing hits that have clear TF motif disruptions and the large effects in the MPRA and/or previous ConDel studies[1]. Because so few well-characterized examples of species-defining sequences exist, deeply exploring these hits is an important step in understanding how species evolution occurs at a molecular level.

I will first use the full JASPAR[9] motif dataset to determine if some TF binding sites are enriched for disruption by ConDels, which will identify any trends of broad molecular evolvability in mammalian genomes. Then, I will select a feasible 3-5 hits with large effect sizes and clear TF motif disruption to characterize with TF chromatin immunoprecipitation (TF-ChIP)[10]. This assay will identify enrichment of specific TF binding between a ConDel and its intact relative's sequence, confirming whether the ConDel induces a direct molecular change. I will perform this assay in both human and mouse cells as representatives of two highly diverged mammals. Additional followup could include CRISPR-based allelic replacement to test the necessity and sufficiency of the ConDel between cells from different species, characterizing the transcriptomic effects of the changes using RNA-seq and gene ontology terms.

Intellectual Merit. No catalogues of species-specific ConDels exist outside of humans, limiting research potential for the speciation genomics community at large. Additionally, the MPRA method is a cutting-edge, state-of-the-art approach I have specific expertise in. This will provide much-needed functional data in an evolutionary field where many analyses are purely computational. Finally, endogenous binding analysis with TF-ChIP will provide gold-standard models with deeply characterized molecular functions. Together, my approach can provide a genomic paradigm to study species-specific changes more broadly.

Broader Impacts. Evolution is a topic of popular interest for a broad public audience; however, many examples of evolution - such as how we evolved to be different from our closest primate relatives - are explained as simplistic, contrived narratives that often lack evidence linking genome to phenotype to species. I plan to develop detailed, well-characterized examples of how evolution actually works at the molecular as well as phenotypic levels by employing functional genomics methods. I will make the pipelines I use to generate my ConDel set publicly available as annotated learning examples by hosting my scripts on GitHub. I will also collaborate with Yale's natural history museum to include any usable or publicly interesting characterization information in a display on vertebrate evolutionary differences, to make evolution and species-specific differences more accessible to the public.

Citations. [1] McLean, C. Y. *et al.* (2011) *Nature.* [2] Lindblad-Toh, K. *et al.* (2011) *Nature.* [3] Rhie, A. *et al.* (2021) *Nature.* [4] Blanchette, M. W. *et al.* (2004) *Genome Res.* [5] Siepel, A. *et al.* (2005) *Genome Res.* [6] Villar, D. *et al.* (2015) *Cell.* [7] The ENCODE Consortium *et al.* (2020) *Nature.* [8] Tewhey, R. *et al.* (2016) *Cell.* [9] Fornes, O. *et al.* (2020) *Nucleic Acids Res.* [10] Gade, P., and Kalvakolanu, D. V. Springer, 2012.