Evolution, Medicine, and Public Health [2024] pp.67–70
https://doi.org/10.1093/emph/eoae006
Advance access date 3 April 2024

BREVIA

EVOLUTION, MEDICINE, & PUBLIC HEALTH

A multi-million-year natural experiment

Comparative genomics on a massive scale and its implications for human health

Iker Rivas-González[1] and Jenny Tung[1,2,3,4,*]

[1]Department of Primate Behavior and Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany; [2]Department of Evolutionary Anthropology, Duke University, Durham, NC, USA; [3]Department of Biology, Duke University, Durham, NC, USA; [4]Faculty of Life Sciences, Institute of Biology, Leipzig University, Leipzig, Germany
[*]Corresponding author. Department of Primate Behavior and Evolution, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany; Tel: +49 (0)341 3550 - 20; Fax: +49 (341) 3550 - 119; Email: jtung@eva.mpg.de

Received 10 January 2024; revised version accepted 18 March 2024.

ABSTRACT
Improving the diversity and quality of genome assemblies for non-human mammals has been a long-standing goal of comparative genomics. The last year saw substantial progress towards this goal, including the release of genome alignments for 240 mammals and nearly half the primate order. These resources have increased our ability to identify evolutionarily constrained regions of the genome, and together strongly support the importance of these regions to biomedically relevant trait variation in humans. They also provide new strategies for identifying the genetic basis of changes unique to individual lineages, illustrating the value of evolutionary comparative approaches for understanding human health.

LAY SUMMARY A recent analysis of hundreds of mammalian genomes, including many of our closest primate relatives, has identified unusual regions of the genome that have remained nearly unchanged over hundreds of millions of years of evolution. These regions are frequently biomedically relevant, showcasing the importance of evolutionary analysis for understanding human health.

KEYWORDS: comparative genomics; natural selection; mammals; nonhuman primates; genome assembly; phylogenomics

In the children's game of telephone, a phrase gets passed along a chain of participants, and, while some information might be preserved by the end, the original meaning is typically indecipherable by the final version. In contrast, proverbs that are passed from generation to generation are also often modified, but retain a fundamental core meaning. For example, the English saying '*a bird in the hand is worth two in the bush*' parallels similar sayings in other languages: ‘*a bird in the hand is worth more than two in the sky*’ in Portuguese, more than ‘*ten in the sky*’ in Dutch, and more than ‘*a hundred in the sky*’ in Spanish. In German and some Slavic languages, the bird in the hand is a sparrow, and it is worth more than a pigeon on the roof, or a dove on the branch. Small modifications aside, all versions preserve the same structure and function: a guaranteed possession is more valuable than a potential, but uncertain, gain.

Genomes evolve like proverbs, not like games of telephone. Through millions of years of evolution, natural selection constrains some parts of the genome from changing, but not others. By pinpointing these regions, researchers can identify locations in the genome that have a core function and, by extension, are likely indispensable for development, physiology, or behavior. Thirty-five years ago, Tagle and colleagues first leveraged this intuition to show that DNA sequence alignments of distantly related species could help identify gene regulatory elements by highlighting stretches of sequence that remain relatively unchanged across species [1] (Fig. 1). This approach, known as phylogenetic footprinting, became a foundational tool for annotating genes and regulatory elements in the early days of genome sequencing and assembly [4]. Nevertheless, the effectiveness of phylogenetic footprinting relies on the diversity of available genomes. Inadequate sampling, whether because the sample is limited to closely related species or because it is sparsely representative of an evolutionary tree, reduces the power to detect conserved regions and increases the probability that regions that appear to be conserved in a small sample are in fact not of special evolutionary interest.

Last year, the number of available genome assemblies for primates and other mammals skyrocketed. In April, a special issue in

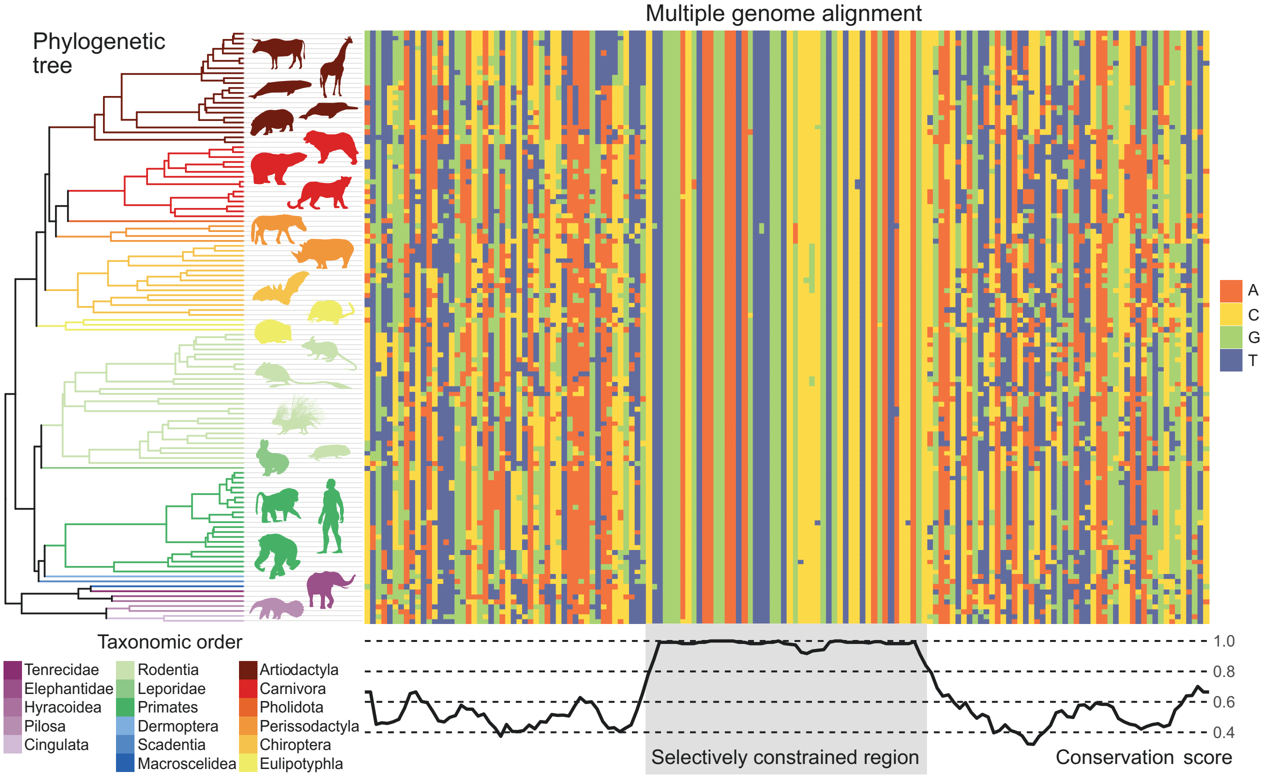

Figure 1. Identification of selectively important genomic regions using large-scale comparative genomics. Regions that have evolved unusually slowly in mammals are a product of selective constraint. The recent expansion of available genome assemblies confers new power to identify such regions. The logic behind these analyses is shown here, using the phylogeny of half of the mammalian species released as part of the Zoonomia Project (left) and an example multiple sequence alignment simulated using *msprime* [2] across the true phylogenetic tree. Each row of the central multiple sequence alignment corresponds to the simulated sequence at this locus for one species; each column corresponds to a single base pair in the sequence. A selectively constrained region appears in the center and is characterized by few to no changes across hundreds of millions of years of mammalian evolution (compared to neutral evolution in the flanking regions). The bottom trace shows per-base-pair conservation scores, calculated by subtracting the coefficient of unalikeability (an approach for estimating the variance of a categorical variable [3]) from 1. A conservation score of 1 corresponds to perfect conservation: the base observed in all aligned genomes is same. A value of 0 corresponds to the highest possible variance, where each of the four possible bases is observed 25% of the time. Note that this metric used here only for illustrative purposes; other conservation scores are typically used in practice for genomic analyses (e.g. phyloP scores, phastCons scores). The analyses presented in this figure could also be extended by adding multiple individuals per species. Selectively constrained regions would then show little within-species polymorphism compared to neutrally evolving regions, which would exhibit more within-species allelic variation.n

Science marked the initial analysis of 240 mammalian genomes generated by the Zoonomia project [5]. In June, this achievement was followed by a second special issue reporting the initial analysis of 233 non-human primate genomes [6, 7]. Then, in November 2023, Kuderna et al. released a whole-genome alignment that includes 239 primates [8], nearly half of all extant primate species. Together, these sequences offer preliminary insight, as well as a set of remarkable resources, for investigating the last 200 million years of evolution in the mammalian branch of the tree of life. In addition to facilitating basic research in evolutionary, conservation, and population genetics, they also have four important implications for evolutionary medicine and research on human health.

First, such a dense sampling of extant genomes substantially increases the power to detect conserved regions across species, expanding the original application of phylogenetic footprinting to a massive scale (Fig. 1). Indeed, multiple contributions from both special issues focus on better estimating phylogenetic constraint to identify selectively important regulatory elements. Annotations resulting from these studies provide comprehensive, base-pair level assessments of constraint across mammals. They also reveal an interesting set of cases in which constraint is either stronger or only detectable, in primates relative to mammals more broadly. Many of these loci appear to have had biochemical functions prior to the expansion of primates but became more selectively important after our order arose [8]. Together, these contributions confirm that a substantial fraction of the mammalian genome is selectively constrained (3.3% of all bases at a false discovery rate of 5%, including 57.6% of coding sites [9]), and that, outside coding regions, constrained regions frequently overlap transcription factor binding sites, DNA hypersensitivity sites (i.e. 'open chromatin' regions), and other indicators of regulatory potential [8, 10].

Second, the recent wave of studies takes advantage of parallel progress in human genetics and functional genomics to directly investigate the degree to which constraint indicates regulatory function and/or differences in disease risk (notably, biochemically functional sequence is not always constrained [11]). For example, by integrating data from transgenic mice [8], massively parallel reporter assays [8, 12], and chromatin accessibility profiles [8, 13], these studies demonstrate that constrained sites are enriched in genes and regulatory elements active in the brain [8, 12, 13], including regions with the capacity to drive gene expression in primary cortical cells sampled in midgestation [13]. Meanwhile, by integrating evolutionary constraint metrics with results from genome-wide association studies (GWAS), Andrews et al. show that GWAS hits that overlap with highly constrained regions across mammals explain far greater (up to 20-fold more) trait heritability than expected by chance [10]. Sullivan et al. go one step further to show that base-pair level constraint scores can be directly integrated into GWAS analysis to improve fine-mapping (i.e. higher-resolution searches to pinpoint potential causal variants), including for health-relevant traits like body mass index and thyroid function [9]. Consequently, evolutionarily constrained regions not only have phylogenetic relevance for understanding mammalian evolution but also have outsized importance in explaining human phenotypic variation today. These results therefore emphasize the utility of comparative evolutionary approaches for identifying specific genes and pathways that account for trait heritability.

Third, several papers not only consider single reference genomes (as in classical phylogenetic footprinting) but also genetic polymorphism data within species. In doing so, they extend the original logic of phylogenetic constraint to shallower, population genetic timescales. The idea here is that variants that impose major fitness costs should not only be infrequent in cross-species comparisons but also appear at low allele frequencies within species. In support of this possibility, the integration of polymorphism data to highlight variants that are also rare in other primates (and hence likely selectively constrained) improves pathogenicity predictions for rare protein-coding variants in humans [14]. It also increases the overall predictive power of rare variant-based polygenic risk scores (i.e. composite summaries of disease risk based on genotypes across many variants) [15]. For example, people who carry a high burden of rare variants in lipid biosynthesis-related genes, particularly at loci where variation is unusual in other primates, represent a disproportionate fraction of clinically at-risk populations for diabetes and dyslipidemia [15]. Because of their low frequency, such variants are difficult to study using classical association approaches: they occur too infrequently to provide the statistical power for correlating genotype to phenotype. Data on genetic variation in other species therefore provide a complementary source of insight beyond what is possible in humans alone. Specifically, they both increase the resolution with which we can identify constrained sequences and contribute to the power of rare variant polygenetic risk scores—a class of variants that are otherwise very challenging to study.

Finally, identifying selectively constrained loci is of interest because it can also point to exceptions to the rule: cases in which a region has evolved under strong constraint through much of evolutionary history but has experienced unusually fast sequence turnover on one branch of the tree [12, 13]. Such a pattern can indicate a switch in evolutionary pressure from constraint to positive selection, and hence suggest changes that underlie the emergence of lineage-specific traits. This logic has already been used to highlight regions of the genome that may contribute to uniquely human traits [16]. For example, a developmental enhancer that regulates the gene *engrailed-1* is highly constrained in other primates but has evolved rapidly in the human lineage to alter the density of eccrine sweat glands in the skin [17]. However, the large set of recently released genome assemblies mean that the same procedure now can be applied in many more lineages. For instance, colobine monkeys—an Asian and African lineage about 30 million years diverged from our own—are the only primates that exhibit foregut fermentation, a dietary adaptation to folivory. By drawing on 49 of the highest-quality primate genome assemblies, Bi et al. showed that colobines also exhibit lineage-specific accelerated evolution in regions linked to metabolite detoxification and possibly maintenance of microbiota that help digest plant fiber [18]. As the number and quality of genome assemblies for vertebrates improve, this strategy may help resolve how other species have evolved traits outside the range observed in our own species—many of which are biomedically relevant (e.g. low cancer incidence in elephants and whales [19, 20], resistance to viral infection in bats [21], or extended longevity in Greenland sharks, which can live over 250 years [22]). Better assemblies—such as gapless telomere-to-telomere assemblies already coming online for humans [23] and a handful of other species—will also facilitate studies of constraint and adaptive evolution in structural variants (e.g. copy-number variants, indels, duplications, inversions, and translocations), which have known implications for disease risk and trait variation in humans [24] but cannot be effectively studied in most current-generation assemblies because of gaps in those sequences.

Together, the assemblies and resequencing data sets released in the past year thus provide an unprecedented view of mammalian biodiversity at the genetic level. They also contribute key evidence that identifying selectively relevant variation in our close living relatives overlaps with the mission of identifying genomic features relevant to human health. The next grand challenge for comparative genomics therefore lies in understanding the phenotypic impact of the candidate regions identified via sequence analysis—an undertaking that will require additional expertise in functional genomics, experimental animal models, and mammalian biology.

AUTHOR CONTRIBUTIONS

Iker Rivas-González (Conceptualization [Equal], Formal analysis [Lead], Visualization [Lead], Writing—original draft [Equal], Writing—review & editing [Equal]), and Jenny Tung (Conceptualization [Equal], Writing—original draft [Equal], Writing—review & editing [Equal])

CONFLICT OF INTEREST

None declared.

REFERENCES

1. Tagle DA, Koop BF, Goodman M *et al*. Embryonic ε and γ globin genes of a prosimian primate (*Galago crassicaudatus*): nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol* 1988;**203**:439–55.
2. Baumdicker F, Bisschop G, Goldstein D *et al*. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics* 2022;**220**:iyab229.
3. Kader GD, Perry M. Variability for categorical variables. *J Stat Educ* 2007;**15**:2. DOI: 10.1080/10691898.2007.11889465
4. Blanchette M, Tompa M. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res* 2002;**12**:739–48.
5. Genereux DP, Serres A, Armstrong J *et al*. A comparative genomics multitool for scientific discovery and conservation. *Nature* 2020;**587**:240–5.
6. Shao Y, Zhou L, Li F *et al*. Phylogenomic analyses provide insights into primate genomic and phenotypic evolution. 2022.
7. Kuderna LFK, Gao H, Janiak MC *et al*. A global catalog of whole-genome diversity from 233 primate species. *Science* 2023;**380**:906–13.
8. Kuderna LFK, Ulirsch JC, Rashid S *et al*. Identification of constrained sequence elements across 239 primate genomes. *Nature* 2023;**625**:735–42.
9. Sullivan PF, Meadows JRS, Gazal S *et al*.; Zoonomia Consortium§. Leveraging base-pair mammalian constraint to understand genetic variation and human disease. *Science* 2023;**380**:eabn2937.
10. Andrews G, Fan K, Pratt HE *et al*.; Zoonomia Consortium§. Mammalian evolution of human cis-regulatory elements and transcription factor binding sites. *Science* 2023;**380**:eabn7930.
11. Dunham I, Kundaje A, Aldred SF *et al*. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74.
12. Xue JR, Mackay-Smith A, Mouri K *et al*.; Zoonomia Consortium†. The functional and evolutionary impacts of human-specific deletions in conserved elements. *Science* 2023;**380**:eabn2253.
13. Keough KC, Whalen S, Inoue F *et al*.; Zoonomia Consortium§. Three-dimensional genome rewiring in loci with human accelerated regions. *Science* 2023;**380**:eabm1696.
14. Gao H, Hamp T, Ede J *et al*. The landscape of tolerated genetic variation in humans and primates. *Science* 2023;**380**:eabn8153.
15. Fiziev PP, McRae J, Ulirsch JC *et al*. Rare penetrant mutations confer severe risk of common diseases. *Science* 2023;**380**:eabo1131.
16. Whalen S, Pollard KS. Enhancer function and evolutionary roles of human accelerated regions. *Annu Rev Genet* 2022;**56**:423–39.
17. Aldea D, Atsuta Y, Kokalari B *et al*. Repeated mutation of a developmental enhancer contributed to human thermoregulatory evolution. *Proc Natl Acad Sci USA* 2021;**118**:e2021722118.
18. Bi X, Zhou L, Zhang J-J *et al*. Lineage-specific accelerated sequences underlying primate evolution. *Sci Adv* 2023;**9**:eadc9507.
19. Vazquez JM, Sulak M, Chigurupati S *et al*. A zombie LIF gene in elephants is upregulated by TP53 to induce apoptosis in response to DNA damage. *Cell Rep* 2018;**24**:1765–76.
20. Nagy JD, Victor EM, Cropper JH. Why don't all whales have cancer? a novel hypothesis resolving Peto's paradox. *Integr Comp Biol* 2007;**47**:317–28.
21. Irving AT, Ahn M, Goh G *et al*. Lessons from the host defences of bats, a unique viral reservoir. *Nature* 2021;**589**:363–70.
22. Nielsen J, Hedeholm RB, Heinemeier J *et al*. Eye lens radiocarbon reveals centuries of longevity in the Greenland shark (*Somniosus microcephalus*). *Science* 2016;**353**:702–4.
23. Nurk S, Koren S, Rhie A *et al*. The complete sequence of a human genome. *Science* 2022;**376**:44–53.
24. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. *Nat Rev Genet* 2020;**21**:171–89.