Understanding and modeling human traits and diseases: Insights from the comparative genomics resources of Zoonomia

Maosen Ye[1] **and Deng-Feng Zhang**[1,2,3,4,5,*]

[1]Key Laboratory of Animal Models and Human Disease Mechanisms of the Chinese Academy of Sciences and Yunnan Province, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650201, China

[2]Kunming College of Life Science, University of Chinese Academy of Sciences, Kunming 650204, China

[3]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650204, China

[4]National Research Facility for Phenotypic & Genetic Analysis of Model Animals (Primate Facility), Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650107, China

[5]KIZ/CUHK Joint Laboratory of Bioresources and Molecular Research in Common Diseases, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650201, China

*Correspondence: zhangdengfeng@mail.kiz.ac.cn

Received: May 7, 2023; Accepted: May 17, 2023; Published Online: May 20, 2023; https://doi.org/10.1016/j.xinn.2023.100444

Citation: Ye M. and Zhang D.-F. (2023). Understanding and modeling human traits and diseases: Insights from the comparative genomics resources of Zoonomia. The Innovation **4(4)**, 100444.

Understanding the genetic architecture of complex human traits and diseases is one of the major aims of biomedical research. Genetic research, such as genome-wide association studies (GWASs) of large-scale, well-phenotyped cohorts, has identified tens of thousands of genomic variants associated with human traits and diseases. The remaining challenge is how to translate the statistical association of genomic loci to biological mechanisms and clinical strategies. Knowledge of human and mammalian evolution should help in the interpretation, modeling, and targeting of disease-relevant variants. The hypothesis is simple: more conserved genome sequences are more likely to be of greater biological importance. Such sequence conservation across species is defined as evolutionary constraint.

Recently, a special issue of *Science*[1] published a collection of studies from the Zoonomia Project (https://zoonomiaproject.org/), which focused on mammalian evolution. By comparing whole-genome sequences of the 240 placental mammals, researchers identified those genetic elements most constrained along the phylogenic tree or those most rapidly changed among certain evolutionary lineages, both of which are potential indicators of functional importance.[1] These large-scale genomes from hundreds of species not only improve our understanding of evolutionary adaptation and innovation but also contribute to our understanding of human diseases and traits.

SINGLE-BASE CONSTRAINT OF THE HUMAN GENOME

To investigate the evolutionary conservation patterns of the human genome at single-base resolution, Sullivan et al.[2] calculated single-base constraint scores across 240 mammalian genomes. Their analysis revealed that 3.3% of bases in the human genome are constrained or conserved, among which 80.7% are within noncoding regions and 19.3% are within coding sequences (CDSs). The researchers observed that common variants are less likely to occur in constrained bases, with the constraint score found to be negatively correlated with the allele count in human population. Using ClinVar and GWAS datasets, which contain well-characterized disease-related variants, they further found that pathogenic variants show higher constraint scores than benign variants, suggesting that constraint scores may improve causal variant identification. Leveraging the large number of mammalian genomes, they also revealed that constraint scores are base-pair specific, with resolution decreasing to 10–100 base pairs when using primate species only. Additionally, they observed a large-scale clustering trend of constrained bases, indicating the existence of constraints at the gene or element level.

As revealed by single-base analyses,[2] 57.6% of CDS bases are constrained, supporting classical genome annotation of coding genes. Taking advantage of the genomic alignment of hundreds of species, Kirilenko et al.[3] developed the genome annotation software TOGA (tools to infer orthologs from genome alignments) based on a machine learning classifier. They utilized TOGA to annotate

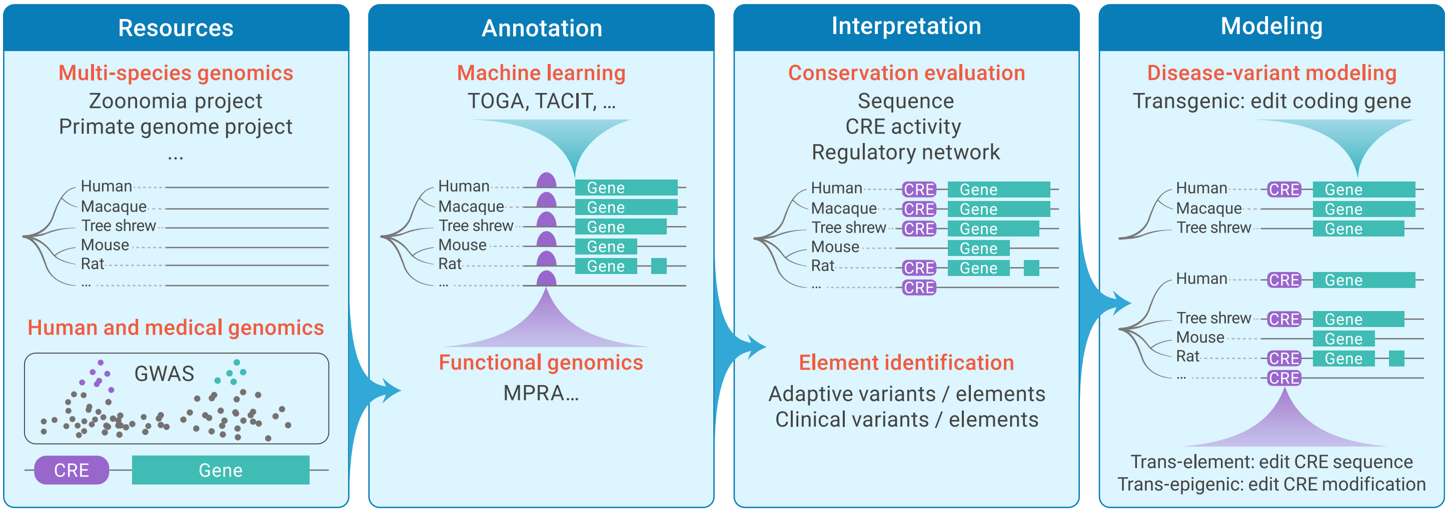

Figure 1. Comparative genomics-based understanding and modeling of human traits and diseases Human and medical genomics focus on identifying genomic variants associated with human traits and diseases, with the functional implications of these variants often remaining unclear. To address this, multi-species genome alignment and annotation tools can be used to annotate and predict the function of genomic sequences and elements where variants are located. These annotated disease-relevant functional alleles and elements can then be targeted in appropriate animals using precise methodologies to model diseases and traits. For disease variants located in constrained coding alleles or conserved genes in a specific animal, constructing a gene knockin model becomes a straightforward process. For those disease variants located in noncoding regions, selecting the appropriate animal and modulating the correct CRE sequence or activity are critical. Starting from a comparative genomics perspective, assessing the sequence conservation and activity of single bases, functional elements, three-dimensional chromatin structures, and transcriptional regulation facilitates the extension of traditional "transgenic" or knockout models to "*trans*-element" and "*trans*-epigenic" models. *For details regarding the Zoonomia Project, please refer to https://zoonomiaproject.org/. CRE, *cis*-regulatory element; GWAS, genome-wide association study; MPRA, massively parallel reporter assay; TOGA, tools to infer orthologs from genome alignments[3]; TACIT, tissue-aware conservation inference toolkit.[5]

the coding features of 488 placental mammal reference genome assemblies, providing a comprehensive comparative genomic resource for both model and nonmodel animals. However, while TOGA is limited to the annotation of conserved coding genes, the conservation and annotation of the major (80.7%) conserved bases—the noncoding region—are also of great value.

EVOLUTIONARY CONSTRAINT OF REGULATORY ELEMENTS

Post-GWAS analyses have suggested that most complex disease-related variants are located in noncoding regions, such as *cis*-regulatory elements (CREs; e.g., enhancers and promoters) and transcriptional factor binding sites (TFBSs). As determined from single-base constraint analyses,[2] constraint scores can provide information for both coding and noncoding regions. Building upon similar methodologies, Andrews et al.[4] observed that ENCODE (www.encodeproject.org/) consortium-defined CREs and TFBSs tend to be conserved in mammals. Among mammalian species, the most constrained CREs (∼439,000, 47.5% of all ENCODE CREs and 4% of the human genome) and TFBSs (∼2 million, 0.8% of the human genome) are presumed to regulate basic biological functions, such as embryonic development. In addition, the less constrained primate-specific CREs, which constitute about 10% of human CREs, appear to regulate genes involved in organismal interactions with the environment. Interestingly, almost all primate-specific TFBSs (20% of all TFBSs) overlap with transposable elements (TEs), indicating the important roles of TEs in primate evolution and expression regulation. Furthermore, through the integration of human GWAS variants (mostly noncoding), Andrews et al.[4] reported that variants in constrained regions tend to explain a greater proportion of heritability of human traits, suggesting the importance of sequence conservation in noncoding regions.

In addition to constraint score-based sequence conservation analysis, Kaplow et al.[5] developed TACIT (tissue-aware conservation inference toolkit), which uses machine learning to assess activity conservation in noncoding regions of the genome. Trained with tissue/cell type-specific enhancers identified in several species, TACIT can predict enhancer activity for any tissue of any species by considering phylogenetic relatedness. By identifying candidate enhancers associated with traits such as brain size, Kaplow et al. showed that TACIT can be applied to identify candidate enhancers or other CREs associated with various phenotypes.

The sequencing of animal genomes across hundreds of species, such as the Zoonomia Project, provides the opportunity to more closely examine animal evolution. With the integration of machine learning approaches, such as TACIT and TOGA, and high-throughput functional genomic assays, comparative genomic resources will undoubtably contribute to our understanding of human traits. Furthermore, an evolutionary constraint-based understanding of coding and noncoding regions is essential for modeling complex human diseases.

MODELING NONCODING ELEMENTS IN EXPERIMENTAL ANIMALS

Mimicking human traits and diseases in experimental animals is an important step following mechanistic investigations. However, current animal models predominantly concentrate on targeting coding genes. This is attributed to the fact that noncoding bases generally exhibit less conserved sequences, and their conservation and functional consequences remain unclear. The availability of multi-species genome alignments and annotation algorithms tailored for gene annotation and CRE activity prediction provide promising tools for evaluating the conservation of disease-relevant genetic elements in animals of interest, which is crucial for modeling. Although some animal models have been developed for noncoding elements, their scope is limited to noncoding RNAs and rodents. Using Zoonomia data and tools, one can select the most suitable animal for characterizing the function of certain disease-related genetic elements, especially for rarely characterized noncoding regulatory elements in less-studied animals. For a given trait-related allele or element, determining its sequence and activity conservation in specific specie is of practical importance. Therefore, the generation of comparative functional genomic resources, especially for noncoding regions with complex regulatory patterns, including gene co-regulation, gene-gene interactions, and gene-environment interactions, is crucial for modeling complex diseases.

Constructing an animal model for a disease variant that involves a constrained coding allele or is situated in a conserved gene can be achieved through direct gene editing of the chosen animal, e.g., knockin mouse models that replicate human mutations (Figure 1). However, for modeling disease variants located in noncoding regions like CREs, careful selection of the appropriate animal species and modulation of the correct CRE sequence or activity are critical (Figure 1). In such cases, several factors must be considered for the chosen animal and genetic element, including the conservation of the noncoding allele at the single-base level, the presence of constrained core sequences and regulatory activities in the located CREs, the conservation of the regulatory pattern of target genes, and the optimal animal species for the specific variant and element.

Taking comparative genomics as the starting point, and comprehensively assessing sequence conservation and activity of single bases, functional elements, three-dimensional chromatin structures, and transcriptional regulation, facilitates the extension of traditional “transgenic” or knockout models to the “*trans*-element” or “*trans*-epigenic” models in the functional genomic era (Figure 1). Hopefully, additional resources like the Zoonomia Project will further expand our understanding of the relationship between genotype and phenotype and benefit the modeling and targeting of complex diseases and traits using animals.

REFERENCES

1. Vignieri, S. (2023). Zoonomia. Science **380**, 356–357.
2. Sullivan, P.F., Meadows, J.R.S., Gazal, S., et al. (2023). Leveraging base-pair mammalian constraint to understand genetic variation and human disease. Science **380**, eabn2937.
3. Kirilenko, B.M., Munegowda, C., Osipova, E., et al. (2023). Integrating gene annotation with orthology inference at scale. Science **380**, eabn3107.
4. Andrews, G., Fan, K., Pratt, H.E., et al. (2023). Mammalian evolution of human cis-regulatory elements and transcription factor binding sites. Science **380**, eabn7930.
5. Kaplow, I.M., Lawler, A.J., Schäffer, D.E., et al. (2023). Relating enhancer genetic variation across mammals to complex phenotypes using machine learning. Science **380**, eabm7993.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their helpful suggestions. This study was supported by the Ministry of Science and Technology of China (STI2030-Major Projects-2022ZD0213500), the National Natural Science Foundation of China (82022017 and 31970965), and the Youth Innovation Promotion Association of Chinese Academy of Sciences.

DECLARATION OF INTERESTS

The authors declare no competing interests.