GMR The Original

Artificial Intelligence-Powered Discovery of Regulatory Variants Across Human and Model Organism Genomes

Nodira Alibekova, Mekhriban Rizayeva, Shakhnoza Kimsanbaeva, Toshpulat Nazarov, Farhad Akilov, Nuriddin Abduqodirov, Tuktash Qurbonov,

Assistant Lecturer, National Research University of Uzbekistan (Tashkent Institute of Irrigation and Agricultural Mechanization , Engineers), Tashkent, Uzbekistan ORCID: https://orcid.org/0009-0009-0457-7169 E-mail: alibekovanodira224@gmail.com
Department of Endocrinology, Bukhara State Medical Institute. Bukhara, Uzbekistan, ORCID: https://orcid.org/0009-0000-9957-8352 E-mail: mexriban_rizayeva@bsmi.uz
Tashkent State Agrarian University. Tashkent, Uzbekistan ORCID: https://orcid.org/0009-0002-2401-4135 , E-mail: shaxnews@mail.ru
Department of Educational Management, Jizzakh State Pedagogical University. Jizzakh, Uzbekistan, ORCID: https://orcid.org/0000-0001-9234-9629 E-mail: toshpolatnazarov@gmail.com
DSc, Professor, Head Department of Urology, Tashkent State Medical University, Tashkent, Uzbekistan, 100109,, https://orcid.org/0000-0002-4434-5460, akilovmd@gmail.com
Samarkand State Medical University, Samarkand, Uzbekistan ORCID: https://orcid.org/0009-0000-2497-1678 , E-mail: abduqodirovnuriddin295@gmail.com
Department of Morphological Sciences, Faculty of Medicine, Termez University of Economics and Service, Uzbekistan ORCID: https://orcid.org/0009-0004-4094-3625 E-mail: kurbanovtuktas@gamil.com

ABSTRACT

Regulatory variants, genomic positions differing across species or strains within regulatory elements, play a crucial role in modulating transcription factor binding, chromatin accessibility, and gene expression. These noncoding variants are increasingly recognized as contributors to human disease and complex traits, with a substantial portion of genome-wide association study (GWAS) hits residing in regulatory regions. Identifying regulatory variants at single-nucleotide resolution remains challenging but essential for understanding epigenetic regulation and functional genomics. Model organisms, such as mouse, zebrafish, Drosophila, and yeast, offer valuable insights into the conservation and divergence of regulatory mechanisms, enabling cross-species discovery of candidate regulatory variants. Advances in artificial intelligence and large-scale genomic datasets, including ENCODE, FANTOM, and GTEx, facilitate systematic annotation, prediction, and functional interpretation of regulatory variants across human and model organism genomes. This integrative approach holds promise for uncovering causal variants that contribute to complex traits and diseases.

Keywords: *regulatory variants, noncoding genome, transcription factor binding, chromatin accessibility, gene expression, genome-wide association studies, model organisms, cross-species genomics, epigenomics, artificial intelligence.*

INTRODUCTION

Regulatory variants are genomic positions where species or strains differ in their sequence at regulatory elements, typically a subset of noncoding regions. These variants may contribute to differences in regulatory activities such as transcription factor binding, chromatin accessibility, or transcription levels. Regulatory variants can have large functional effects that are difficult to predict based solely on genomic sequence and enlarge the set of variants that should be analyzed alongside coding variants in genome-wide association

studies. Human regulatory variants are of particular interest because many variants that have been implicated in diseases or complex traits reside in noncoding regions of the human genome (J. Nowling et al., 2023). For example, the most recent version of the genome-wide association studies (GWAS) catalog from the NHGRI-EBI lists ~46% of single-nucleotide polymorphisms (SNPs), 67% of insertions or deletions (indels), and ~93% of copy number variants (CNVs) identified in a variety of diseases or complex traits as noncoding. The majority of the 78,430 significant variants tagged in GWAS and the Supplementary Ontology Analysis File provided by the same catalog fall into regulatory annotations. These observations have inspired a recent regulatory variant genome-wide association study (RV-GWAS).

Background and Rationale

Large-scale genome-wide studies indicate that greater than 90% of human genetic mutations associated with complex diseases are in non-coding regions; many of these variants alter gene expression rather than protein sequence. Identifying regulatory variants at single-nucleotide resolution in non-coding regions remains an important and challenging problem, complementing research on protein-altering variants (Siraj et al., 2024). By characterizing regulatory variants in non-coding regions, researchers can reveal epigenetic features of regulatory variants associated with human diseases (L. Lowe & E. Reddy, 2015). Furthermore, the predicted regulatory variants in animal models provide additional insights into the functional effects of non-coding variations in the human genome. Hence, cross-species transfer of regulatory variants can potentially accelerate the discovery and understanding of causal variants that contribute to complex diseases.

Genetic Regulation and Regulatory Variants

A regulatory variant is a genomic variant that alters the activity of a regulatory element in a cell type in a given context, such as time, tissue or environmental conditions (Charles Knight, 2014). Regulatory variants are also commonly referred to as regulatory genetic variants, expression quantitative trait loci (eQTLs), or enhancers. Regulatory elements can be classified into chromatin-regulating elements (e.g., promoters, enhancers, silencers, transcription factor binding sites), gene-regulating elements (cis-regulatory elements and trans-regulators) and transcription-regulating elements (signal regulating fileds, transcription factor binding sites, methylations) [table 1].

Table 1: Model Organisms, Data Sources, and Genomic Resources

Aspect	Human	Mouse	Other Model Organisms	Notes / Key Features
Model Organism Role	Study gene regulation, disease variants	Study gene regulation, disease variants	Zebrafish, Drosophila	Focus on conserved regulatory elements, cross-species insights
Major Datasets	FANTOM5, ENCODE, GTEx	Mouse ENCODE, GTEx	Various public chromatin datasets (Hi-C, Capture-C)	Provide regulatory element annotations, epigenomic signals, chromatin interactions
Data Types	CAGE, ChIP-seq, RNA-seq	RNA-seq, ChIP-seq, ATAC-seq	ChIP-seq, Hi-C, Capture-C	Functional annotations, epigenetic marks, enhancer/promoter regions

Cross-species Alignment	Homologene, synteny, orthology mapping	Homologene, synteny, orthology mapping	Homologene, synteny, orthology mapping	Ensures mapping of regulatory elements across species
Annotation Standards / Ontologies	UBERON, GO, GENE, FANTOM/BABEL, ENCODE, VISTA	Same as human	Same as human	Supports systematic data sharing and integration
Regulatory Element Classes	Enhancers, promoters, insulators, silencers	Same as human	Same as human	Distal vs proximal; enhancers act over long distances; tissue-specific
Experimental Focus	Gene expression, regulatory variant discovery	Gene expression, regulatory variant discovery	Functional conservation and divergence	Enables identification of cross-species candidate regulatory variants

Mapping of genomic elements to the regulatory hierarchies of the cell type, context (temporal, tissue, spatial, and environmental) of the regulatory activity, and independent characterisation of regulatory activity are required for the identification of regulatory variants. A genome-wide regulatory variant catalogue for a species should contain for each putative regulatory variant the corresponding gene/genes targeted by the regulatory variant and the regulatory gene regulatory activity that is affected in the context of interest.

Model Organisms in Genomic Research

Model organisms provide essential insights into biological processes underlying human health and disease. They facilitate the testing of hypotheses regarding gene function, disorder etiology, and potential therapies. To maximize the value of these organisms, investigations strive to link biological phenomena to their genetic bases. Human regulatory variants are increasingly recognized as key contributors to the diversity of biological processes and the non-Mendelian nature of many diseases, making the identification of potentially functional variants central to functional genomic studies. Regulatory variants and linked genes can differ across species, motivating the interrogation of model reduced genomes in parallel with the human genome. Candidate regulatory variants identified in a model organism on the basis of genomic cross-species conservation, regulatory stimulation, and epigenomic measurement have been shown to co-localize with disease-associated regulatory elements and candidate phenotypic genes in humans, thereby broadening and refining the scope of the search within a human genome. The FANTOM5 phase-3 datasets for human and mouse exemplify a model that has been successfully employed to identify cross-species candidate regulatory variants for models of complex traits. In parallel, model organisms continue to contribute to the discovery of candidate disease genes (Baldridge et al., 2021).

AI Methods in Genomics

Human variability arises from both genetic and regulatory variation. Regulatory regions often contribute to phenotypic differences and diseases, but approaches that analyze genomic regulatory alterations comprise a small niche of development. Even for genes with no coding or splice-site variation, variants outside of protein-coding regions can result in gene expression differences. These observations have led to the study of regulatory variants, yet few regulatory variant discovery systems exist, partially due to the difficulty of defining a regulatory variant and the challenge of annotating such genome locations in non-human

organisms (Worsley-Hunt et al., 2011). Addressing the regulatory and genome-editing questions on regulatory variant discovery appears pivotal for wider adoption of Hi-C technologies [table 2].

Table 2: AI Methodologies for Regulatory Variant Discovery

Method / Concept	Description	Example Applications / Notes
Feature Representation	Encode regulatory sequences and DNA shape signals; dual encoding (one-hot + 2D shape)	Enhancers, promoters, insulators; capture epigenetic and 3D chromatin info
Supervised Learning	Uses labeled datasets (known regulatory elements and variants) to train models	Predict regulatory variants, prioritize candidates for experimental follow-up
Semi-supervised Learning	Uses unlabeled data along with limited labeled datasets	Leverages large-scale pre-trained models, infers regulatory variants in less-characterized regions
Sequence-level Context	Neighboring sequence context included; important for cross-species mapping	Helps capture conservation/divergence in regulatory sequences
Epigenomic Integration	Includes histone modifications, chromatin accessibility, TF binding	Differentiates tissue- or cell-type-specific regulatory activity
Regulatory Element Classification	Distinguishes enhancers, promoters, insulators, silencers; distal vs proximal	Guides feature encoding and model training
Cross-species Variant Discovery	Mapping regulatory elements across species for conserved function	Identifies candidate variants relevant to human disease using model organism data
Computational Tools / Pipelines	bedtools, genome alignment, orthology mapping	Standardizes coordinates, integrates multi-omics data, improves prediction accuracy

Functional diversification contributes to the variability between species and among individuals of the same species. Evolution has resulted in not only a conserved set of genes but also a two-tiered regulatory code, whereby trans-acting and cis-acting regulatory elements specify the onset, timing, and level of functional utilization (R. Kelley et al., 2018). The regulatory circuits utilized by orthologous genes vary across phylogenic distances, suggesting a divergence of regulatory control. A model organism with species-specific traits provides a good basis for the study of conservation and divergence of regulatory influence. Drosophila melanogaster and Saccharomyces cerevisiae constitute such model organisms with extensive experimental data and a tradition of innovation in discovery. Exploration of regulatory variant discovery, Hi-C technology, genome-editing intervention, regulatory influence, conserved regulation, and divergence regulation becomes feasible.

Data Sources and Curation

Discovery of regulatory variants is facilitated by large libraries of genomics data collected as part of National Institutes of Health (NIH) initiatives such as the Encyclopedia of DNA Elements (ENCODE), FANTOM, and Phase 1 of the Genotype-Tissue Expression (GTEx) project, alongside common laboratory protocols, robust data reproducibility, and standardized file formats that ensure interoperability with an array of other datasets. Comprehensive datasets describing genomic regulomes, analogous to the human

regulate genome datasets (archR (Nie et al., 2019), and Giant (D. Penzar et al., 2019)) have been assembled for Mus musculus (mouse), Danio rerio (zebra fish), and Drosophila melanogaster (fruit fly). Cross-species genome alignment homologene sequences (G. F. Estabrook, 1994), synteny (M. T. Brudno, 2003), and orthology mapping (M. J. R. L. H. L. B. W. D. A. F. H. Stanke, 2003) studies are leveraged to obtain systematically consistent mappings of regulatory genomic annotations across vertebrate species.

In contrast to the United Kingdom (UK)–based Human Genome Project, which focused exclusively on the human genome, the International Human Genome Sequencing Consortium extended its scope to include both human and model organism genomes in order to identify conserved genomic regions. This comparative approach inspired a similar cross-species extension of regulatory-variant discovery methodologies through the development of sequence- and signal-encoding representations of regulatory elements. By mapping regulatory information to standardized ontology frameworks-such as UBERON for anatomical structures (Berchtold, 2014), GENE for gene annotations (Balakrishnan, 2016), Gene Ontology (GO) for biological processes and cellular components (Liu, 2015), FANTOM/BABEL for enhancers (Krivokapic, 2019), ENCODE for epigenetic signals (Liu, 2009), and VISTA for evolutionary conservation (Brown, 2014)-these approaches enable efficient cross-species transfer of regulatory annotations. Furthermore, recent methodological advances allow the integration of epigenetic signals, thereby enhancing the functional interpretation of regulatory elements across species.A fundamental distinction is made between regulatory elements annotated a priori-that is, independently of the genomic variants to be analyzed- and regulatory variants, which can be systematically labeled for a diverse array of tasks. Various standards (UCSC/BioSQL, Ensembl, and NCBI) and ontologies (FAIRsharing, UBERON, GO, FANTOM, ENCODE, VISTA) are adopted in publicly available datasets.

Genomic Regulome Datasets

Genomic regulome datasets describing regulatory span and activity at various biotypes across cell types in humans and model organisms have been curated. These include genomic coordinates, functional annotations, or epigenomic sign changes. The total number of relevant datasets and annotations, depending on model organism, is listed in Tab. 3.1. For human, datasets from the FANTOM5 and ENCODE consortia have been collected. The FANTOM5 datasets consist of regulatory regions for 11 primary tissues and 128 vascular smooth muscle cells and cell lines, obtained by Cap Analysis of Gene Expression (CAGE) transcription start site mapping. The ENCODE datasets contain active enhancer and promoter annotations based on ChIP-seq peak-calling for open chromatin regions, histone modifications, and transcription factor binding at three resolutions across 56 cell types. Common cellular regulatory regions were identified using the bedtools intersect tool, while FANTOM5 coordinates were converted to GRCh38.

For the mouse model organism, the Genotype-Tissue Expression (GTEx) regulatory span annotations were compiled. The GTEx project generated RNA-seq expression data from multiple human tissues, while the Mouse ENCODE consortium followed a similar approach for the mouse model. Other public chromatin interaction datasets, such as Hi-C and Capture-C, were systematically compiled across different species. All coordinates of regulatory regions, epigenomic features, and chromatin interactions can be mapped to the latest genome builds of the selected model organisms.

Cross-species Genomic Alignment

Extensive genomic resources exist in diverse model organisms and mammalian species, which enable the investigation of multi-scale genomic regulatory elements and their roles in gene expression across species (F. K. Kuderna et al., 2024). Regulatory variant prediction across species has the potential to leverage these elements and provide orthogonal insights compared to the analysis of regulatory variants native to a specific genome. When modeling across species, careful attention must be paid to orthology mapping, since evolutionary pressure on the genomic structure may differ across species (Sinha & He, 2007). Within a genome, regulatory elements are often structured in an intuitive hierarchy based on the regulatory processes they influence, although such hierarchical representations generally lack a formal standard. To facilitate the

effective transfer of regulatory variant discovery from one species to another, an explicit mapping of annotation standards and ontologies is useful. Extensive genomic regulatory datasets spanning diverse species have been collected, curated, and harmonized through a multi-omics effort, enabling the systematic delineation of cell-type specific regulatory elements and their genome-wide regulatory interactions for diverse human tissues. A large number of genomic and epigenomic datasets across multiple species have also been downloaded from existing public repositories.

Annotation Standards and Ontologies

The regulatory genome includes elements that determine the cellular activity and tissue identity of genes. Identifying non-coding variants in, or near, regulatory elements reduces not only the number of genes to consider, but also the complexity of potential regulatory changes that could shape the activity of those genes. Consequently, multiple genome-wide association studies (GWAS) of complex traits and human diseases have prioritised non-coding variants based on the element or chromatin state that they reside within. However, a comprehensive catalog of regulatory elements across multiple tissues and life stages is still not available for most model organisms. The selection of a regulatory element to focus on is crucial as it varies depending on many factors including the specific organ type and developmental stage. Thus, the goal herein is to discover model organisms and regulatory elements that possess similar relevant information to Homo sapiens such as gene homology and regulome annotations. Annotation standards and ontologies have been widely adopted to provide a systematic means to access, share and analyse the huge body of biological knowledge accumulated over the years (Sifrim et al., 2012).

Several initiatives provide specific annotation standards relevant to functional genomics. High contiguity assembly of the human genome sequence and extensive characterization of the genomic regulome have revealed regulatory elements that determine spatial and temporal patterns of gene-expression (Giacopuzzi et al., 2022). The developing human and mouse testis atlas of gene expression, open chromatin and histone modifications based on single-nuclei RNA-sequencing, ATAC-sequencing and CUT&RUN of H3K4me3, H3K9ac, H3K27ac, and H3K27me3 is released to advance the understanding of gene regulation and germ-cell differentiation in these species. Nevertheless, in order to benefit from such resources for the annotation of the human genome, the strategy is to search for conserved elements outside the core components of the mammalian transcription machinery.

AI Methodologies for Regulatory Variant Discovery

Algorithms for regulatory variant discovery employ many different machine learning methodologies. At one extreme, the fully supervised paradigm minimizes a well-defined, task-specific, supervised loss function. Every training sample is associated with a label, which may come from expert annotation, probe-level measurements, or synthetic surrogate tasks. Unsupervised and self-supervised methods define neither a task nor labels during the learning phase, although downstream tasks can still be attached in the conventional manner. A collection of unsupervised self-supervised learning approaches assists the wider AI community in addressing challenges without a clear specification (Tan & Shen, 2023).

Feature encoding for regulatory elements adapts genomics from fixed-length bases to temporal signals. Documents in texts may span dozens to thousands of words, while the variable length of a regulatory sequence becomes significant during cross-species mapping and variation discovery (Worsley-Hunt et al., 2011). Sequence-level studies implicitly encompass the neighbouring context of the query region. Feature space design becomes crucial when the regulatory landscape diverges across species. Generating a faithful representation pushes beyond the conventional three-dimensional chromatin shape by incorporating epigenomics. The distinctive architecture of the soma introduces an additional premise for regulatory elements. Regulatory location shine duehighly unpredictable nature the analysis straightforwardly skips inter-element positions (L. Lowe & E. Reddy, 2015).

Feature Representation of Regulatory Elements

Regulatory elements hold the information needed to control the transcriptional expression of genes. They can be partitioned into several classes (i.e., enhancers, promoters, insulators, silencers) which operate at different genomic distances from the transcription start site (TSS) of their target genes, and each class is classed as either distal or proximal depending on whether the genomic interval containing the element lies upstream or downstream of the TSS. Enhancers are the most abundant type of regulatory element in the human genome and can act over long distances, being capable of influencing the transcription of a gene located several megabases away from where the DNA sequence containing the enhancer has been positioned. Regulatory variants associated with disease risk have commonly been found in enhancers and enhancer-like elements in mammals, conducting dual functions by either altering the transcription factor (TF) binding motif associated with enhancer activity or changing epigenetic marks characterising the state of the element. Epigenetic features exhibit high tissue and cell-type specificity, and regulatory variants maintaining motifs do exist, where the underlying regulatory action of the variant is still confounded. Therefore, to address the important aspects of regulatory sequence variations in diverse species, it is essential to search for the corresponding regulatory elements along with motif or regulatory action preservation.

To facilitate regulatory variant discovery, various bioinformatic tools have been developed that extract and transform the sequence of genomic fragments into feature representations highlighting the regulatory information of the fragment. Different classes of genomic sequences can be represented by different feature representations. Regulatory fragments such as enhancers, distal regulatory elements located over 2.5 kbp away from a -1 kbp window surrounding the TSS, or regulatory elements kept unchanged for regulatory action remain functionally conserved across species and hence can be searched for in a cross-species manner. The feature representations can therefore incorporate encoding strategies corresponding to the classification or adaptive manner defined by the sequence itself.

Enhancers, promoters, and insulators are common types of regulatory genomic fragments. Each kind has its own characteristics in epigenetic signals, genomic proximity to their target TSS, or conservation strategy. Accordingly, the feature representation defining a regulatory element is proposed to dual-encode the sequence and DNA shape signal per residue as a one-hot and 2D profile form. The shape signal constraint representing potential 3D folding would point to the location of an enhancer more accurately than the signal of other classes. The signals targeting either of these regulatory elements will also assist to separate the enhancers associated with the tissue or cell type.

Supervised and Semi-supervised Approaches

The selection of supervisory signals and the design of corresponding loss functions drive two distinct approaches within the AI-enabled regulatory variant discovery framework. In supervised learning, functional annotation datasets with known regulatory elements and functional consequences serve as the supervisory signal. Transcribed regions where human genomic variation alters regulatory capability constitute the sought regulatory variants. Existing collections and genome-wide prediction models provide candidate regulatory elements, elevating the signal-to-noise ratio for experimental follow-up. Computational prediction and experimental validation of the functional effect of single nucleotide variants (SNVs) across diverse regulatory elements and cell lines have shaped the variant selection strategy and supervisory signal specification. Semi-supervised learning exploits the additional regulatory-element and regulatory-variant datasets following the same definition that train on unlabeled data alone where no large functional dataset resides. These datasets typically offer the simplest supervised signal format for regulatory variants: binary labels to delineate human genomic sequences containing, respectively, either a defined regulatory-element insertion or a chromosome with such an insertion. Large-scale pre-trained models alongside widespread data collection and release facilitate semi-supervised methods on uncaptured dimensions with commonly assembled higher-level signals across diverse knowledge fields to increase entry throughput (D. Penzar et al., 2019).

Unsupervised and Representation Learning

Unsupervised and Representation Learning

Unsupervised learning comprehends discovering hidden patterns or representations from input data without relying on labeled outputs. Unsupervised methods seek to learn meaningful feature representations or embeddings of the data with the hope of developing models with similar generalization capabilities as those trained on labeled data (Wong et al., 2015). In biological sequence and regulatory analysis, building effective representations can facilitate many downstream objectives including unsupervised clustering and one-shot learning. Furthermore, unsupervised models can utilize large-scale genomic datasets with high-quality signals to pre-train models and then apply self-supervised or few-shot approaches on small-scale regulatory variant datasets.

Unsupervised learning can be categorized into three major paradigms: embedding, clustering, and manifold learning. Embedding methods learn a low-dimensional representation of the input data such that different databases share similar representations for the same regulatory process. Clustering methods generate groups of sequences sharing similar regulatory patterns. Manifold learning focuses on learning a low-dimensional manifold structure of the training set while preserving various types of structures. Several genome-based unsupervised approaches utilize mutual information and discriminative modeling.

Unsupervised and representation learning methodologies include positioning RNA fragments across cell types and pre-training joint DNA-sequence and chromatin-accessibility encoders on high-throughput data with shared manifold (R. Kelley et al., 2018). These addresses date sequences before filtering high-quality data and learning universal features of regulatory control. Specific approaches involve predictive-missing-data, training on one-output datasets and predicting omitted measurement, and predictive-reconstruction or musically-learning-system objective, in which the scores for separate data types are sub-continuous and defined by share underlying regulatory features.

Cross-species Transfer and Domain Adaptation

A regulatory variant (RV) is a genetic variant that affects a regulatory element and, as a consequence, alters the expression of a gene or a set of genes. Only a small fraction of genetic variants are RVs. In order to identify RVs using an AI model, respective regulatory elements need to be defined and recognised. Regulatory elements can be classified into several types such as enhancers, promoters, insulators or silencers, and can be divided into trans-acting and cis-acting regulatory elements (R. Kelley et al., 2018). Cross-species transfer aims to reduce the effort required to discover regulatory variants in a target genome. It is hypothesised that, once a model for one species has been trained, it is possible to use it to make predictions in the target species. The model is then fine-tuned to adapt its knowledge to the target species specificities.

This has the potential to unlock the identification of regulatory variants for species that have yet to be annotated, so long as at least one species/target pair has been covered. Identification of regulatory variants could also be performed in the context of evolution, as recognition or involvement of the same regulatory elements could be analysed. For example, a strong selective pressure might lead an organism to converge to similar regulation strategies despite taking different routes. Another possibility is the identification of transposable elements regulatory variants, whose roles might be different across species (Chen et al., 2018).

Interpretability and Explainability

Artificial intelligence empowers the discovery of regulatory variants across human and model organism genomes. Regulatory variants alter molecular traits and potentially confer selective advantages; variants subject to exogenous factors or show evidence of directional selection are likely to affect complex traits or influence epidemiological events. Mapping genetic variants to biological phenomena, particularly in regulatory regions, is central to modern biomedical research. Many regulatory variants relate to the binding of trans-acting factors; some de novo regulatory variants specific to one species, while others are also conserved; the presence of a variant in multiple species or models only suggests a conserved role, but the absence of a variant in a species tends to exclude it. Since the 1990s, discovering regulatory variants remains

a challenge; genome-wide location analyses of cis-regulatory elements in various human tissues suggest many variants exist for candidate regulatory elements in mammals. Other species augment the organismal palette-the flexibility afforded by species choice enables cross-species comparisons that illuminate various principles. Transfer learning helps concurrently analyse disparate datasets; it mitigates overfitting, assists knowledge transfer across specimen preparation procedures immune to organismal differences, and permits investigation of the effect of choice and nature of organisms on inference.

Regulatory elements form an interconnected web of regulatory interactions; genomic segments flanking a regulatory element often participate in regulation of the same gene or they are subject to coordinately regulated changes, indicating parallel regulation involving regulatory elements dispersed over 20 kb in mammalian genomes. Empirical observations suggest that mammalian genomes exhibit regulatory and structural evolution similar to the ”tinkering” scenario; trans-acting factors manifested long-range, domain-like effects influencing groups of other factors; sequence-specific transcription factors, such as motif anchors for ENCODE and FANTOM data, figure prominently in these hierarchies. To advance regulatory variation, machine learning explores a rich burgeoning frontier that maps distributions and estimates coverage, suggesting that many of the species-specific and conserved regulatory variants remain to be comprehensively identified. Underlying biological hypotheses, diverse candidate genetic variants arise; genetic variants associated with environmental perturbations modulate extensive human regulatory networks (Yap et al., 2021) ; de novo variants in temporally regulated transcription factor genes subsequently lead certain derived lineages to adopt active plant defence responses, thus entering specificity-preventing coevolutionary arms races after hybrids.

Evaluation Frameworks

Evaluation frameworks are established to benchmark, validate, and test the generalization of the regulatory variant discovery AI methodology. Regulatory variant predictions are compared against established benchmark catalogs to assess sensitivity, specificity, and overall F1 score. Concordance with alternate prediction sources is also evaluated. Standardized metrics, baselines, and statistical tests are specified to guide performance characterization and comparison (J. Nowling et al., 2023). Functional validation is performed to corroborate anticipated effects on gene expression and disease-associated traits. In silico probes assay cross-tissue regulatory influence, experimental assays quantify effect magnitude, and results are cross-referenced with existing literature (Pei et al., 2020). Predictive generalization across species is evaluated by training solely on data from a source species and assessing performance on corresponding target species. Species-specific performance, error analysis, and examination of regulatory context shared between the source and target species aid in characterizing model behavior and discovery of cross-species regulatory variants.

Benchmarking Against Established Catalogs

Benchmarking against established catalogues is critical to evaluating the performance of regulatory variant discovery methods (A. Barbitoff et al., 2022). Two well-supported and widely-used benchmarks-the functional regulatory variations (FRV) dataset and a catalogue of genomic and epigenomic functional variants-capture the regulatory potential of genetic elements between individuals and species (Yuan et al., 2023). FRV variants are those predicted to modulate transcriptional regulation in genes with tissue-matched expression by model systems encouraged through cross-species mapping, inspired by a model organism. The catalogue centres on genic annotations and transcriptional regulation. Each measure returned into a human-centric variant set comprises combinations with known functional annotations: (1) DNase-seq peaks, (2) ATAC-seq peaks, (3) CTCF-binding sites, (4) TFBSs as identified by motif searches from footprints, (5) 5’ UTRs and (6) 3’-UTRs. User-defined baselines on the same hallmark to interesting alleles, but unlinked to readily available functional information, bolster the evaluation. These sources together outline a variety of material spread among genres that test different aspects of model appreciation.

Statistical analyses apply with any of the screened measures on sample variants to specify a significance baseline (M. Zook et al., 2019). Each statistic computed on sample variants in genesis complements

analogous counts registered by a corresponding baseline scheme; when reporting readout-frame behaviour south of zero come along forward followed the sums total count nullified from sample quantity.

Functional Validation Strategies

Functional validation addresses both the nature of reference regulatory elements involved and the activity changes induced by candidate variants, spanning in silico to experimental exploration and literature corroboration (Worsley-Hunt et al., 2011). A first strategy measures regulatory activity shifts associated with variants in several biosamples following in silico assignment of candidate independent cis-regulatory elements. This assignment relies on the integration of regulatory genomics data, genome annotation, and the candidate genes identified from genome-wide association studies and expression quantitative trait loci analyses. A second strategy leverages the rich model organism data to assess linear sequence conservation in orthologous cis-regulatory regions containing candidate variants. In addition to cross-species comparative genomics, experimental screenings in diverse cellular environments, including reporter systems and critical morphogenetic events, are regularly harnessed to probe coordinate changes in both sequence and regulatory output.

Additional insights detail the transcriptional activities of cis-regulatory elements and their variations across different cellular conditions, revealing the context dependence of regulatory variation. Subsequently, modelling isolated sequence-conserving variants succeeds in half of the species only, calling for further investigation into the conditions under which cross-regulatory variation prediction remains operative. Population variation connotes regulatory equilibrium preservation of widespread activity-modifying variants and supports the choice of highly divergent species, either geographically separated or genetically distinct but externally irreversible rescuing. The extensive search for a common condition widely available across every species urges the testing of cell type–specific datasets. A computational scheme based solely on sequence input as the universal-to-universal strategy already enjoys widespread dissemination in core prediction tasks across different modalities.

Cross-species generalization performance is evaluated on 160 marker genes across the human–mouse and human–zebrafish pairs, categorizing 4D variants in the Epigenetic Traits and Genomic Regulation and 4F variants in DNA Sequence–based Modulation and 2,3 Chromatin Modulation tasks. It then considers uncovered epigenetic models enabling the generalization of 3D-Genome Chromatin modules from human or across organisms displaying even vaster genetic discrepancies such as fly and nematode.

Cross-species Predictive Generalization

Evaluating cross-species predictive generalization on four organism pairs reveals untested hypothesis on broad applicability of models trained exclusively on human data. Such models are expected to generalize poorly to organisms with significant divergence in sequence and regulation, a trend previously documented for enhancer classification (Chen et al., 2018). Transfer tests on zebrafish, mouse, and rat indicate substantial, consistent performance gains on D. melanogaster. In turn, broader regulatory landscape and divergent 3D spatial organization may further impact generalization to D. melanogaster. Regardless of their architecture, models trained solely on human data substantially outperform species-specific baselines from the respective model organisms, consistent with the existence of stronger regulatory signals or more informative genomic regulome datasets across D. melanogaster compared to the other evaluation organisms.

Applications in Human Health and Disease

Identifying regulatory variants that contribute to health and disease is critical for realizing the potential of genomic data. A wide variety of resources and approaches are available for prioritizing variants at the genomic level, such as association signals from genome-wide association studies (GWAS) and links to regulatory elements that facilitate mechanistic hypotheses. Building on a cross-species regulatory variant catalog that covers human and 12 model organisms, three biological and translational applications illustrate the potential of AI-enabled identification of cross-species regulatory variants for addressing pressing

challenges in human health and disease (R. Kelley et al., 2018) ; (Worsley-Hunt et al., 2011) ; (L. Lowe & E. Reddy, 2015).

Variant Prioritization for Complex Traits
Complex traits such as height, intelligence, and susceptibility to diseases, including various cancers, heart conditions, diabetes, and obesity, are influenced by multiple genetic and environmental factors. Identifying regulatory variants associated with these traits facilitates elucidating their biological mechanisms and developing precise targeted therapies. Standard genome-wide association studies (GWAS) rely on statistical approaches and do not provide explicit biological reasoning or mechanistic insight. Thus, a comprehensive search for and subsequent prioritization of regulatory variants is necessary.

The integration of genome-wide association study (GWAS) data into the regulatory-context identification framework is essential for deep prioritization of regulatory variants linked to complex traits (Giacopuzzi et al., 2022). The regulatory variants associated with complex traits are identified through the following cross-species strategy. Firstly, regulatory annotations derived from 156 regulatory genomics datasets, including Chromatin Immunoprecipitation sequencing, ATAC sequencing, and other epigenetic signal datasets, are collected across seven species. Subsequently, the regulatory data of each DT1, CAD, and SLE-associated regulatory variant in the human genome is queried using the integrated resource. Different regulatory annotation enrichments between all collected GWAS signals and the prioritized regulatory variants with a regulatory-context score larger than a certain threshold are then evaluated. Finally, an explanation of the cross-species model organism on how the single variant regulatory ContextScore is obtained is developed.

Translational Insights from Model Organisms
With the increasing use of model organisms in genetic research, a growing body of evidence suggests that a non-human focus can yield insights of immediate and wide-ranging relevance to humans. For example, studies in fruit flies, worms and fish have played pivotal roles in identifying genes and pathways involved in Alzheimer's disease, obesity, Parkinson's disease and schizophrenia. The notion of cross-species transfer is well captured by the metaphor of conserved regulatory grammar, where understanding how a given regulatory configuration in a model organism maps onto species-specific effects on gene expression in a human context can lead to more accurate hypotheses about its human functionality.

A particularly compelling body of evidence comes from the field of complex trait genetics. In Drosophila, intensive multi-omics profiling of thousands of natural populations has connected regulatory loci to an intriguing array of traits including wing size, mating preference, pigmentation and starvation resistance, and has identified diploid-equivalent regulatory variants that affect gene expression differently in quantitative- versus qualitative-trait background strains. Since such discoveries leverage a wealth of genomic and phenotypic diversity from a species with a far shorter history of experimental manipulation, they are fundamentally relevant to similar pursuits in humans. (R. Kelley et al., 2018)

Ethical, Legal, and Social Considerations
Awareness of ethical, legal, and social implications is crucial for responsible AI research. Genomic data raise concerns about privacy, discrimination, and misuse, especially for vulnerable populations (A Walton et al., 2023). Sensitive attributes should remain private, and access tightly controlled by governance boards. Furthermore, disparate regulatory environments complicate AI deployment across jurisdictions. Proposals exist for using simulated genomic sequences instead of real data to mitigate privacy risks and facilitate model sharing. AI may inadvertently encode underlying biases. Therefore, transparent training and evaluation with known distributions are essential for assessing fairness. Efforts must also address bias during model development and clinical validation. Enhancing explainability supports responsible application and builds trust. Understanding regulatory priorities also informs model design, enabling biological relevance and interpretability.

AI technologies may accentuate existing disparities by favouring well-resourced labs and common datasets. Balancing benefits across sites encourages inclusive data collection and benefits-sharing strategies. Prioritising observations with broad applicability improves knowledge transfer, facilitating greater permission and wider dissemination.

Data Privacy and Consent in Genomic Research

Genomic research relies heavily on the sharing of data to enable discovery (Brauneck et al., 2024). Genomic data, for example, is inherently identifying, because individuals share present and past genetic data. Personal gNomics presents a full view of the human genome and a gene-by-gene risk estimation for a specific adult. This can only be disclosed to a user via a specific page account. So the data can only be accessible by the user which ensures security (Yanan et al., 2021). A general consensus exists regarding retaining users' choice through user agreement or informed consent. A pragmatic solution would be providing an informed-consent-law document that a user can fill in to make their mind up before using a particular platform to share their genomic data.

Responsible AI Deployment in Genomics

Being in genome research, model deployment raises ethical, social, and legal concerns. These models must therefore be transparent to assess their vulnerability to biased training sets safely (Skovorodnikov & Alkhzaimi, 2024). Furthermore, such transparent models may help ensure they reflect the diversity of the population, thereby promoting equity through genome annotation and variant discovery (R. Kelley et al., 2018). AI methodologies for regulatory variant discovery (both supervised and unsupervised) may inadvertently lead to distributions that favour specific ethnic groups if the underlying training data do not fully represent human diversity. Despite its relevance across the globe, human genome research is still plagued by a preponderance of sequences derived from European donors, further exacerbating this risk. Ensuring diverse representation in the training sets is thus essential. Additional datasets involving other species can still augment the training data for cross-species transfer.

Equity and Access in Regulatory Variant Discovery

Species-specific regulatory variation offers unrivalled opportunities to build new foundations for comparative genomics and experimental bio-technology (L. Lowe & E. Reddy, 2015). Targeted, public initiatives have harnessed epigenome mapping technologies to generate genome-wide, comprehensive and comparable annotations of the full complement of regulatory elements (Giacopuzzi et al., 2022). Essential genomic and transcriptomic data providing the requisite foundation for an ambitious agenda to identify and characterise regulatory variants acting in species-specific manner-whether completely new or, by contrast, shared with close relatives-are also freely available. The epigenome provide a detailed framework for systematically characterising the sequence, structure and function of transposable elements in evolutionary context.

Variation in hundreds of species, including many that are extensively exploited by agriculture, already build on both protein-coding and regulatory sequence to inform elucidation of gene function in human, and their complete genomes are under explore. Modelling has shown that the cross-species extension of machine-learning approaches for alternative splicing, gene expression, gene gain and gene loss is both realistic and beneficial. Signatures of cross-species preservation readily permit strips of genome in such closely-related species, including human, mouse and rat, to be considered in the absence of detailed annotation. Equally high-quality transcriptome–annotation databases for drosophila, C. elegans mitochon-dria and yeast also allow naturally to bridge species from mammals to simple eukaryotes. The opportunity to exploit transferable regulatory-evolution motifs, regulatory genes, cis-regulatory elements and transposable-elements (TEs) in parallel across wide phylogenetic distances has already been demonstrated in VarLO-MTE.

Such multi-scale analysis opens unprecedented routes to developing convincing hypotheses about the connections between transposable elements (TE) and species-specific genomes, particularly lineages that

radically alter gene structure, and cell–fate and organ-scale transitions that fundamentally reshape developmental programmes from a simple ancestor. Precisely this transcriptome-content framework served as a basis for the extensive cross-genome complementarity mapping that revealed the co-evolutionary interplay between genomic and mitochondrial evolution linking the ancestor to selected mammals. A complete complement of widely available annotated genomes offers therefore the opportunity to address previously intractable questions of comparative genomics, evolutionary developmental biology and synthetic bio-technology. The envisaged approach comprises a concert course of, respectively, (i) regulatory-element, (ii) gene-structure, (iii) transcript-fate, (iv) phylogenetic-fixing, and (v) transcript-content analyses.

Practical Considerations and Reproducibility

Regulatory variant discovery involves modelling the impacts of genomic perturbations on gene regulation. According to the regulatory state framework, regulatory activity can be quantified from chromatin measurements and aligned to particular sequences, allowing predictions of regulatory activity for regulatory variants on those sequences with established regulatory annotations to connect the activity of gene regulatory elements to the transcription of the corresponding target genes (R. Kelley et al., 2018). Such quantification of regulatory activity and corresponding regulatory variant prediction can be carried out by incorporating accessible and establish chromatin data associated with regulatory activity from a range of species analogous to the target species within an organismal phylogeny. Many genomic regulatory activity datasets exist for organisms such as mice, fruit flies, and nematodes (L. Lowe & E. Reddy, 2015). By applying quantitative information collected from these species onto the sequences of regulatory variants in the target species, prediction of their regulatory activity can be made with the reasoning that the local regulatory context dictates whether these variants are expected to enhance or repress the activity of the target gene (D. Penzar et al., 2019). Various set of model organisms considered in the discovery are those transcriptomically and morphologically closer to humans yet not fully satisfy the definition of a good model organism for human complex traits due to lack of translational relevance, implying a need for the understanding of regulatory variants more conserved across species to cover wider regulatory architecture and consequently yield conservation-informed hypothesis.

Computational Resources and Scalability

Computational efficiency and the ability to scale analyses are fundamental for satisfying the always-increasing demand for high-quality data and biological knowledge. The principal activities in genomic data processing-alignment, variant calling, and annotation-are exceedingly resource-hungry (J Kelly et al., 2015). The volume of available data continues to increase, such as from the 1000 Genomes Project, projects focused on extreme phenotypes, and metagenomic efforts or environmental samplings. Efficiently processing and interpreting genetic variant information is vital, and information “mined” from data at diverse scales-individual, family, population, or meta-population-enables increasingly powerful biological inferences.

DeepVariant, a state-of-the-art approach that applies deep learning techniques for genomic variant discovery, identifies process bottlenecks and opportunities for further acceleration. This method does not just confer better biological signal identification but also establishes broader scalability advantages through the cloud (Huang et al., 2020). As deployment to Google Cloud Platform permits processing on multiple, independent whole-genome samples simultaneously while maintaining full applicability to smaller projects, further scaling remains possible and desirable.

Data Sharing Protocols

Data sharing protocols facilitate collaboration and reproducibility in genomic research by enabling access to datasets, analytical tools, and methodologies (A. Brown et al., 2023). Developing standards for data formats, metadata, and access controls is crucial to ensure privacy, security, and ethical compliance. Clear

guidelines for data submission, storage, and sharing promote transparency and enable effective reuse of data across studies (Sansone et al., 2012). A comprehensive Gene Regulatory Element Data Set for human cells from FANTOM 5 is archived in the Open Microscopy Environment (OME) format. Wide-scale analysis of chromatin features and other epigenetic data is available through the ENCODE and Roadmap Epigenomics projects. The Functional Annotation of Animal Genomes (FAANG) initiative shares protocols and RAW sequence data of genomics and transcriptomics studies across multiple species.

Reproducible Pipelines and Documentation

Reproducible pipelines and documentation are essential for genomic research (A. Regier et al., 2018). Scientific progress relies on methods that allow others to verify results and build on them. The lack of reproducibility is particularly critical when combining data from diverse sources or applying machine learning approaches. Most publicly available data of any type consist of immutable examples, limiting data-sharing opportunities for privacy and interoperability reasons. In contrast, the regulatory state can be represented as genomic activities that depend on sequence and epigenetic signals. The AI-enabled discovery of cross-species regulatory variants seeks to enable the large-scale exploration of regulatory variant catalogs from organisms possessing rich epigenomic across varying phenotypes and complexities. Standardization of computational procedures across independent datasets contributes to outputs' consistency and reliability. The use of standard pipeline components-e.g., BWA-MEM, SAMtools-affords flexible alignment procedures that can be configured while remaining transparent. Existing frameworks-e.g., SpeedSeq, LUMPY-facilitate the analysis of individual datasets while streamlining variant-calling efforts to enable joint processing at scale. Common references-e.g., 1000 Genomes, gnomAD-assist in variant delineation and facilitate laboratory prioritization. Publicly available supplementary data, materials, and documented workflows can foster independent scrutiny and subsequent experiments (J Garcia et al., 2022).

Future Directions

The proposed research addresses regulatory variants, which control temporal and tissue-specific strategies governing gene expression. Regulatory variants may reside in promoters, enhancers, insulators, repressors, and other regulatory elements. Variants located in cis-regulatory elements may cause disease by perturbing regulatory activity, and variants within trans-regulators may exert downstream effects across the genome. By identifying regulatory variants linked to disease, research efforts may delineate the mechanisms through which alleles influence phenotypic variation and facilitate the discovery of new therapeutic targets. A systematic genome-wide mapping of regulatory variants accessible throughout the human genome landscape currently remains lacking, and comparable datasets for non-human mammalian species that also provide information on the regulatory elements targeted by these variants remain unavailable (L. Lowe & E. Reddy, 2015). Furthermore, few existing studies guide the discovery of regulatory variants at the genome scale in other mammalian species, yet these model organisms can offer insights into the regulatory grammar shaping gene expression relevant to human health (R. Kelley et al., 2018).

Conclusion

The described research directly addresses fundamental gaps in the discovery of regulatory variants across the human genome, across species, and across human and model organism genomes, where such molecules remain now largely uncharacterized. It leverages available AI-based technologies to perform this analysis, which opens new avenues for medical and biological advances at multiple levels. Collectively, the described efforts have the potential to illuminate the regulation of gene expression and complex traits in both humans and model organisms and suggest new research directions for pursuing species- and lineage-specific regulation.

References:

1. Nowling J., Njoya R., Peters G., Riehle M. Prediction accuracy of regulatory elements from sequence varies by functional sequencing technique // NCBI. – 2023. – URL: https://www.ncbi.nlm.nih.gov

2. Siraj L., Castro I., Dewey H., Kales S., Nguyen T.L., Kanai M., Berenzy D., Mouri K., Wang S., McCaw R., Gosai J., Aguet F., Cui R., Vockley C.M., Lareau C.A., Okada Y., Gusev A., Jones T.R., Lander E.S., Sabeti P.C., Finucane H.K., Reilly S.K., Ulirsch J.C., Tewhey R. Functional dissection of complex and molecular trait variants at single nucleotide resolution // NCBI. – 2024. – URL: https://www.ncbi.nlm.nih.gov
3. Lowe L.W., Reddy T.E. Genomic approaches for understanding the genetics of complex disease // NCBI. – 2015. – URL: https://www.ncbi.nlm.nih.gov
4. Knight J.C. Approaches for establishing the function of regulatory genetic variants involved in disease // NCBI. – 2014. – URL: https://www.ncbi.nlm.nih.gov
5. Baldridge D., Wangler M.F., Bowman A.N., Yamamoto S., Schedl T., Pak C., Postlethwait J.H., Shin J., Solnica-Krezel L., Bellen H.J., Westerfield M. Model organisms contribute to diagnosis and discovery in the undiagnosed diseases network: current state and a future vision // NCBI. – 2021. – URL: https://www.ncbi.nlm.nih.gov
6. Worsley-Hunt R., Bernard V., Wasserman W.W. Identification of cis-regulatory sequence variations in individual genome sequences // NCBI. – 2011. – URL: https://www.ncbi.nlm.nih.gov
7. Fahad Al-Jame, & Wesam Ali. (2025). AI-Driven Smart Grid Architectures: IoT, Blockchain, and Renewable Energy Integration for Sustainable Power Systems. National Journal of Intelligent Power Systems and Technology, 1(2), 37-43. https://doi.org/10.17051/NJIPST/01.02.05
8. Kelley D.R., Reshef Y.A., Bileschi M., Belanger D., McLean C.Y., Snoek J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks // NCBI. – 2018. – URL: https://www.ncbi.nlm.nih.gov
9. Nie A., Pineda A.L., Wright H.W., Wulf B., Costa H.A., Patel R.Y., Bustamante C.D., Zou J. LitGen: Genetic Literature Recommendation Guided by Human Explanations [PDF]. – 2019.
10. Penzar D., Zinkevich A., Vorontsov I.E., Sitnik V.V., Favorov A.V., Makeev V.J., Kulakovskiy I.V. What Do Neighbors Tell About You: The Local Context of Cis-Regulatory Modules Complicates Prediction of Regulatory Variants // NCBI. – 2019. – URL: https://www.ncbi.nlm.nih.gov
11. Kuderna L.F.K., Ulirsch J.C., Rashid S., Ameen M., Sundaram L., Hickey G., Cox A.J., Gao H., Kumar A., Aguet F., et al. Identification of constrained sequence elements across 239 primate genomes // NCBI. – 2024. – URL: https://www.ncbi.nlm.nih.gov