Supplemental Methods

Generation and merging of the sSNVs

To obtain all possible point mutations in coding regions, we first retrieved all hg38 human protein-coding transcripts from BioMart (Kinsella et al. 2011) (Version: Ensembl 110). After removing transcripts containing unknown bases or lacking start/stop codons in their coding sequences, 84,067 transcripts remained. Next, based on their coding sequences, we simulated each base mutating into the other three bases, resulting in all possible point mutations in the coding region. In the second step, we used the Variant Effect Predictor (McLaren et al. 2016) (Version: Ensembl 110) tool to filter out mutations possessing synonymous consequences. Additionally, to provide mutation information for the hg19 reference, we performed genome coordinate conversion using the LiftOver (Kuhn et al. 2013) tool. Due to varying quality control procedures, sequence variation annotation strategies, and reference genome versions, in case of omitted data, we also merged sSNVs from CADD (Schubach et al. 2024), FAVOR (Zhou et al. 2023), and synVep (Zeng et al. 2021), which all contain synonymous mutations across the human genome.

Splice site consensus

In higher eukaryotes, precise splicing is regulated by three weakly conserved cis-elements, 5' and 3' splice sites, and the branch site. According to RegSNPs-splicing (Zhang et al. 2017), if an sSNV falls at the +1, +2, or +3 position of the 5' splice site or the -1 position of the 3' splice site, we classified it as a variant on splice site consensus (VSS). Otherwise, we classified it as variants in internal exons (VIE).

Exonic splice regulatory elements

Referring to sSNVs pathogenic prediction tools such as SliVA (Buske et al. 2013), DDIG-

SN (Livingstone et al. 2017), and regSNPs-splicing (Zhang et al. 2017), exonic splicing regulatory (ESR) sequences are considered to be important features when constructing models. The potential of sSNVs to result in a gain/loss of an ESR may be correlated with its pathogenicity. Hence, a comprehensive ESR motifs set is curated from RESCUE-ESE (Fairbrother et al. 2002), FAS-HEX3 (Wang et al. 2004), SpliceAID, RegRNA2 (These two datasets are retrieved from SynMICdb (Sharma et al. 2019)), Composite-ESR (Ke et al. 2008), NI-ESR (Stadler et al. 2006), and Ast-ESR (Goren et al. 2006). After merging and removing duplicates, we have 701 exonic splicing silencer motifs, 1,048 exonic splicing enhancer motifs, and 285 ESR motifs remaining. The detailed table is available in Supplemental Table S5.

sSNVs from vertebrate species

First, we targeted the species included in the UCSC 100-way vertebrate multiple sequence alignment. Since Ensembl focuses on vertebrate genomes, and Ensembl Variation performs quality control on mutations while providing evidence status and functional consequence annotations, its data quality and reliability are relatively high. Therefore, we obtained sSNV information for all 17 non-human vertebrate species from Ensembl Variation (Hunt et al. 2018) (http://www.ensembl.org/info/genome/variation/index.html, downloaded on 2024-05-09). Additionally, the European Variation Archive (EVA) (Cezard et al. 2022) (https://www.ebi.ac.uk/eva/, downloaded on 2025-01-17), as the most comprehensive platform for genetic mutations across all species, offers extensive information on non-human species. To ensure that the sequence consequences of mutations could be annotated, we filtered vertebrate species mutations that could be annotated using the VEP with available cache files (https://ftp.ensembl.org/pub/release-113/variation/indexed_vep_cache/) and added sSNV

information for 7 additional vertebrate species. Since most annotation resources are designed primarily for humans, we aim to map non-human sSNVs to human reference genomes to enable shared annotation. However, LiftOver may introduce artifacts when mapping genomic coordinates across species. We referred to the method used in PrimateAI (Sundaram et al. 2018). Based on the multiple sequence alignment (MSA, https://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz100way/, downloaded on 2024-10-18), we mapped non-human mutations onto the human genome. Additionally, common variants in other primates are largely benign in humans (Gao et al. 2023; Cheng et al. 2023). Therefore, we selected mutations from five primate species, including those from the Great Ape (https://eichlerlab.gs.washington.edu/greatape/data) and Han (https://figshare.com/articles/dataset/Han_etal_Data_tsv_gz/7855850). Since these mutation datasets were already mapped to hg18 or hg19, we used the LiftOver method to map them to hg38. Detailed statistics can be found in Supplemental Table S4.

Economic traits-related sSNV from cultivated plants and domesticated animals

To understand the relationship between sSNVs and economic traits in major livestock and crops, we integrated 11,963 GWAS associations of sSNVs from 17 crops and 4 animals in the GWAS Atlas (Liu et al. 2023) (https://ngdc.cncb.ac.cn/gwas/, downloaded on 2025-01-06) and CropGS-Hub (Chen et al. 2024) (https://iagr.genomics.cn/CropGS/#/, downloaded on 2024-11-25) databases. These databases are built based on literature-extracted GWAS information. In CropGS-Hub, the literature related to Rice and Sorghum is already included in the GWAS Atlas, so we focused on the literature not covered by it. Detailed statistical information about the data can be found in Supplemental Table S4.

Literature resource in SynMall

PMC/PubMed query statement

SynMall automatically collects literature from PMC and PubMed using the following search query: ("synonymous"[Title/Abstract]) AND ("mutation"[Title/Abstract] OR "variation"[Title/Abstract] OR "variant"[Title/Abstract] OR "mutant"[Title/Abstract]) NOT ("non-synonymous") NOT ("nonsynonymous").

Criterion standard for human sSNVs

For sSNVs in humans, we aim to extract evidence-supporting associations, categorizing them as either benign or pathogenic. Therefore, we referred to the criteria of the ACMG (Richards 2015).

Fields and description of structured information extracted from the literature

In total, 21 fields across three domains are considered when curating each paper. The literature-central domain provides basic publication details and key supporting evidence sentences. The variant-central section offers detailed information about the variant, including allele change, genomic position, strand, coding sequence position, reference single-nucleotide polymorphism ID, and codon change, as well as additional information related to the gene and species. The phenotype-central part contains manually annotated data inferring the phenotypic effects of sSNVs, where the Mechanism field describes how the sSNVs induce the disorder (e.g., through splicing regulation, mRNA structure stability, protein synthesis, etc.). The Trait field is designed for non-human species to capture traits associated with sSNVs, while the Trait Impact field describes the effect of the mutation on the trait, such as promoting or inhibiting, if applicable.

Batch query performance in SynMall

We evaluated the response speed of batch retrieval in SynMall using Apache JMeter (with 10 concurrent threads simulating multiple users), as summarized in Supplemental Table S3. When querying 1,000 records, the average response times were 7.35 s for Genomic Coordinates, 0.99 s for Gene Names, and 2.99 s for RS IDs. Please note that the first request or prolonged inactivity may trigger reinitialization of the database connection pool, leading to slower response times than those shown in the table. Currently, batch queries support up to 1,000 records per request, as larger queries may result in timeout errors. For datasets exceeding this limit (1,000–50,000 records), users are advised to use the Annotation module.

Curation of datasets

We compile a benchmark dataset for machine learning using sSNVs curated from multiple external databases and literature. The dataset includes a balanced training set of 2,362 sSNVs and a balanced test set of 238 sSNVs. First, we retrieve initial data from ClinVar (downloaded April 2025) (Landrum et al. 2020), HGMD (Professional 2023.3) (Stenson et al. 2020) , dbDSM (Wen et al. 2016), and manually reviewed sSNVs from SynMall. In ClinVar, we select variants labeled as "Benign", "Likely Benign", "Likely Benign/Benign", "Likely Pathogenic/Pathogenic", "Likely Pathogenic", and "Pathogenic". Only records with a review status of "criteria provided, multiple submitters, no conflicts", "criteria provided, single submitter", or "reviewed by expert panel" are included. In HGMD, we include DM-classified pathogenic synonymous mutations, and in dbDSM, we select variants from manually curated sources. We remove any variants that appear in both benign and pathogenic categories. To evaluate VEP performance on rare variants, we filter out common variants with AF > 1e-3,

retaining only rare variants. We control sequence similarity using CD-HIT (Fu et al. 2012), ensuring that protein sequences in the training and test sets share less than 40% identity. Finally, we apply a "close-by" strategy (Cheng et al. 2020) to balance pathogenic and benign samples. For each minority pathogenic sample, we select a benign sample with the closest genomic position, creating a dataset with balanced positive and negative samples. The full dataset is available for download on the "Download" page.

Performance evaluation

To evaluate the performance of VEP tools on the synonymous variant test set, we use the Area Under the Receiver Operating Characteristic Curve (AUC) and the Area Under the Precision-Recall Curve (AUPR). The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) across different classification thresholds, while the Precision-Recall (PR) curve plots Precision against Recall. Both metrics provide threshold-independent measures commonly used for assessing binary classification performance.

Notably, some tools produce missing values and fail to provide predictions for certain variants in the independent test set. To address this, we apply both the "subset" and "pairwise" evaluation strategies. Specifically, the subset approach extracts the portion of the test set for which all VEP tools provide prediction scores, and evaluates the performance of all tools simultaneously on this subset. In contrast, the pairwise approach compares synScore against each target VEP individually by selecting the subset of variants without missing values for that tool, and evaluates their relative performance within this set.

Reference

Buske OJ, Manickaraj A, Mital S, Ray PN, Brudno M. 2013. Identification of deleterious synonymous variants in human genomes. *Bioinformatics* **29**: 1843–50. doi:10.1093/bioinformatics/btt308

Cezard T, Cunningham F, Hunt SE, Koylass B, Kumar N, Saunders G, Shen A, Silva AF, Tsukanov K, Venkataraman S, et al. 2022. The European Variation Archive: a FAIR resource of genomic variation for all species. *Nucleic Acids Res* **50**: D1216–D1220. doi:10.1093/nar/gkab960

Chen J, Tan C, Zhu M, Zhang C, Wang Z, Ni X, Liu Y, Wei T, Wei X, Fang X, et al. 2024. CropGS-Hub: a comprehensive database of genotype and phenotype resources for genomic prediction in major crops. *Nucleic Acids Res* **52**: D1519–D1529. doi:10.1093/nar/gkad1062

Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, Pritzel A, Wong LH, Zielinski M, Sargeant T, et al. 2023. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**: eadg7492. doi:10.1126/science.adg7492

Cheng N, Li M, Zhao L, Zhang B, Yang Y, Zheng CH, Xia J. 2020. Comparison and integration of computational methods for deleterious synonymous mutation prediction. *Brief Bioinform* **21**: 970–981. doi:10.1093/bib/bbz047

Fairbrother WG, Yeh R-F, Sharp PA, Burge CB. 2002. Predictive Identification of Exonic Splicing Enhancers in Human Genes. *Science* **297**: 1007–1013. doi:10.1126/science.1073774

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**: 3150–3152. doi:10.1093/bioinformatics/bts565

Gao H, Hamp T, Ede J, Schraiber JG, McRae J, Singer-Berk M, Yang Y, Dietrich ASD, Fiziev PP, Kuderna LFK, et al. 2023. The landscape of tolerated genetic variation in humans and primates. *Science* **380**: eabn8153. doi:10.1126/science.abn8197

Goren A, Ram O, Amit M, Keren H, Lev-Maor G, Vig I, Pupko T, Ast G. 2006. Comparative Analysis Identifies Exonic Splicing Regulatory Sequences—The Complex Definition of Enhancers and Silencers. *Mol Cell* **22**: 769–781. doi:10.1016/j.molcel.2006.05.008

Hunt SE, McLaren W, Gil L, Thormann A, Schuilenburg H, Sheppard D, Parton A, Armean IM, Trevanion SJ, Flicek P, et al. 2018. Ensembl variation resources. *Database (Oxford)* **2018**: bay119. doi:10.1093/database/bay119

Ke S, Zhang XH-F, Chasin LA. 2008. Positive selection acting on splicing motifs reflects compensatory evolution. *Genome Res* **18**: 533–543. doi:10.1101/gr.070268.107

Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, et al. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford)* **2011**: bar030. doi:10.1093/database/bar030

Kuhn RM, Haussler D, Kent WJ. 2013. The UCSC genome browser and associated tools. *Brief Bioinform* **14**: 144–161. doi:10.1093/bib/bbs038

Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, Hoffman D, Jang W, Kaur K, Liu C, et al. 2020. ClinVar: improvements to accessing data. *Nucleic Acids Res* **48**: D835–D844. doi:10.1093/nar/gkz972

Liu X, Tian D, Li C, Tang B, Wang Z, Zhang R, Pan Y, Wang Y, Zou D, Zhang Z, et al. 2023. GWAS Atlas: an updated knowledgebase integrating more curated associations in plants and animals. *Nucleic Acids Res* **51**: D969–D976. doi:10.1093/nar/gkac924

Livingstone M, Folkman L, Yang Y, Zhang P, Mort M, Cooper DN, Liu Y, Stantic B, Zhou Y. 2017. Investigating DNA-, RNA-, and protein-based features as a means to discriminate pathogenic synonymous variants. *Hum Mutat* **38**: 1336–1347. doi:10.1002/humu.23283

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. *Genome Biol* **17**: 122. doi:10.1186/s13059-016-0974-4

Richards S. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17**. doi:10.1038/gim.2015.30

Schubach M, Maass T, Nazaretyan L, Röner S, Kircher M. 2024. CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic Acids Res* **52**: D1143–D1154. doi:10.1093/nar/gkad989

Sharma Y, Miladi M, Dukare S, Boulay K, Caudron-Herger M, Groß M, Backofen R, Diederichs S. 2019. A pan-cancer analysis of synonymous mutations. *Nat Commun* **10**: 2569. doi:10.1038/s41467-019-10489-2

Stadler MB, Shomron N, Yeo GW, Schneider A, Xiao X, Burge CB. 2006. Inference of Splicing Regulatory Activities by Sequence Neighborhood Analysis. *PLoS Genet* **2**: e191. doi:10.1371/journal.pgen.0020191

Stenson PD, Mort M, Ball EV, Chapman M, Evans K, Azevedo L, Hayden M, Heywood S, Millar DS, Phillips AD, et al. 2020. The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. *Hum Genet* **139**: 1197–1207. doi:10.1007/s00439-020-02199-3

Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, Fritzilas N, Hakenberg J, Dutta A, Shon J, et al. 2018. Predicting the clinical impact of human mutation with deep neural networks. *Nat Genet* **50**: 1161–1170. doi:10.1038/s41588-018-0167-z

Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. 2004. Systematic Identification and Analysis of Exonic Splicing Silencers. *Cell* **119**: 831–845. doi:10.1016/j.cell.2004.11.010

Wen P, Xiao P, Xia J. 2016. dbDSM: a manually curated database for deleterious synonymous mutations. *Bioinformatics* **32**: 1914–1916. doi:10.1093/bioinformatics/btw086

Zeng Z, Aptekmann AA, Bromberg Y. 2021. Decoding the effects of synonymous variants. *Nucleic Acids Res* **49**: 12673–12691. doi:10.1093/nar/gkab1159

Zhang X, Li M, Lin H, Rao X, Feng W, Yang Y, Mort M, Cooper DN, Wang Y, Wang Y, et al. 2017. regSNPs-splicing: a tool for prioritizing synonymous single-nucleotide substitution. *Hum Genet* **136**: 1279–1289. doi:10.1007/s00439-017-1783-x

Zhou H, Arapoglou T, Li X, Li Z, Zheng X, Moore J, Asok A, Kumar S, Blue EE, Buyske S, et al. 2023. FAVOR: functional annotation of variants online resource and annotator for variation across the human genome. *Nucleic Acids Res* **51**: D1300–D1311. doi:10.1093/nar/gkac966