Bioinformatics, 2024, **40(6)**, btae340
https://doi.org/10.1093/bioinformatics/btae340
Advance Access Publication Date: 25 May 2024
Applications Note

OXFORD

Genome analysis

SuPreMo: a computational tool for streamlining *in silico* perturbation using sequence-based predictive models

Ketrin Gjoni[1,2] **and Katherine S. Pollard**[1,2,3,*]

[1]Institute of Data Science and Biotechnology, Gladstone Institutes, 1650 Owens Street, San Francisco, CA 94158, United States
[2]Department of Epidemiology & Biostatistics, University of California, San Francisco, CA 94158, United States
[3]Chan Zuckerberg Biohub, San Francisco, CA 94158, United States

*Corresponding author. Gladstone Institute of Data Science and Biotechnology, San Francisco, CA 94158, United States. E-mail: katherine.pollard@gladstone.ucsf.edu (K.S.P.)

Associate Editor: Pier Luigi Martelli

Abstract

Summary: The increasing development of sequence-based machine learning models has raised the demand for manipulating sequences for this application. However, existing approaches to edit and evaluate genome sequences using models have limitations, such as incompatibility with structural variants, challenges in identifying responsible sequence perturbations, and the need for vcf file inputs and phased data. To address these bottlenecks, we present Sequence Mutator for Predictive Models (SuPreMo), a scalable and comprehensive tool for performing and supporting *in silico* mutagenesis experiments. We then demonstrate how pairs of reference and perturbed sequences can be used with machine learning models to prioritize pathogenic variants or discover new functional sequences.

Availability and implementation: SuPreMo was written in Python, and can be run using only one line of code to generate both sequences and 3D genome disruption scores. The codebase, instructions for installation and use, and tutorials are on the GitHub page: https://github.com/ketringjoni/SuPreMo.

1 Introduction

Many machine learning (ML) models have been developed that predict cellular profiles from input DNA sequences (Supplementary Table S1). These sequence-to-profile models can predict biological features—including gene expression [Enformer (Avsec *et al.* 2021a), ExPecto (Zhou *et al.* 2018), Xpresso (Agarwal and Shendure 2020)], genome folding [Akita (Fudenberg, Kelley and Pollard 2020), C.origami (Tan *et al.* 2023), DeepC (Schwessinger *et al.* 2020), ORCA (Zhou 2022)], chromatin accessibility [Basenji (Kelley 2020), Basset (Kelley, Snoek and Rinn 2016)], and epigenetic marks [(DeepFIGV (Hoffman *et al.* 2019), HyenaDNA (Nguyen *et al.* 2023), Sei (Chen *et al.* 2022a)]—with incredible accuracy. These approaches are becoming increasingly popular for exploring biological questions at lower cost and higher throughput than experimental methods allow, and for addressing questions that are not possible to test experimentally. One exciting potential is to use sequence-to-profile models in tandem with *in silico* mutagenesis (ISM), in order to investigate how genomic alterations affect cellular profiles. This strategy generates testable, causal hypotheses about genotype-phenotype relationships (Chen *et al.* 2022b). ISM has been applied to the genomes of modern humans, archaic hominins (McArthur, Rinker and Capra 2021), and other species (Keough *et al.* 2023) to prioritize putative pathogenic variants for experimental studies (Benegas, Batra and Song 2023), decode the grammar of noncoding DNA sequences (Deng *et al.* 2023), discover new sequence motifs (Avsec *et al.* 2021b), design tissue-specific enhancers (Gosai *et al.* 2023), and uncover novel roles of sequence elements (Gunsalus, Keiser and Pollard 2023).

In theory, ISM is very high-throughput, making it feasible to quantify the effects of a large set of sequence perturbations, such as all variants in an individual's genome or a cohort of patients. However, the application of ISM at scale is currently limited by the process of generating sequences with and without perturbations. While several tools exist to perform analogous tasks, such as creating synthetic haplotypes [bcftools consensus (Li 2011), GATK FastaAlternateReferenceMaker (Van der Auwera and O'Connor 2020), perEditor (Rivas-Astroza *et al.* 2011), etc] or randomly mutating sequences [SNP mutator (Davis *et al.* 2015), BBMap mutate.sh (Bushnell 2014), etc], they are not compatible with ISM. One of the biggest limitations is that they incorporate all variants from an input variant call format (vcf) (Danecek *et al.* 2011) file into a single output fasta file, making it very difficult to isolate the effects of individual variants. Workarounds, such as generating an independent vcf file for each variant (or variant combination) and looping over these or post-processing the output fasta file to include one variant per locus, are extremely inefficient. Second, existing tools are made largely for SNPs or small insertions or deletions (indels), and cannot accommodate symbolic alleles—annotations in vcf files for structural variants (SVs). A possible workaround is to convert symbolic alleles into sequences by extracting them from a reference genome, but this becomes infeasible with large structural variants due to limitations with both variant complexity and memory allocation. One tool

Received: 27 October 2023; **Revised:** 4 May 2024; **Editorial Decision:** 17 May 2024; **Accepted:** 24 May 2024

[perEditor (Rivas-Astroza *et al.* 2011)] is compatible with some complex variants but is not comprehensive and has stringent requirements. Finally, existing tools require the perturbations to be in a vcf format, which means that pseudo input files must be generated if one wishes to apply ISM to custom or simulated sequences (e.g. deleting all motifs for a given transcription factor or creating synthetic enhancers).

Due to these limitations, it is common practice for ISM practitioners to write their own code to generate input sequences for ISM studies. Indeed, the codebases for several ML models include code examples or frameworks for performing ISM [Enformer, Sei, Basset], but these are restricted to simple variants (SNPs and indels) and do not generate sequence files for input into other models. SVs make good candidates for ISM since they span larger regions and are more likely to be damaging to the genes, regulatory regions, or active sites they overlap or neighbor. For example, noncoding SVs have been shown to lead to cancer and developmental disorders by disrupting genomic contacts of key genes (Paik *et al.* 2021). SVs also alter more base pairs of the genome than any other type of genetic variation (1000 Genomes Project Consortium et al. 2015). One major challenge with SVs is that, to adhere to the fixed length input requirements of most ML models, input sequences must be padded, and consequently, model outputs require un-padding and masking. Another consideration is that—due to both biological effects and model artifacts related to making predictions for fixed width genomic windows—models can be highly sensitive to small changes in the input, such as masking, padding, and variant position in the window. Therefore, it is important to make predictions for augmented input sequences (shifted and/or reverse complement sequences) and evaluate them consistently across perturbations. Thus, incorporating perturbations into a reference genome becomes increasingly complicated and error-prone as variants get larger and more complex.

2 Tool description

To address these challenges, we developed SuPreMo, a framework for generating perturbed sequences for input into predictive models that is scalable, flexible, and comprehensive (Fig. 1A). SuPreMo, which incorporates variants into the human reference genome one at a time and generates model-ready sequences (Supplementary Fig. S1A), was extended to SuPreMo-Akita, which inputs those sequences into Akita (Fudenberg *et al.* 2020), an ML model that predicts chromatin contact maps, and generates scores that measure variants' disruption to those maps (Supplementary Fig. S1C).

Both tools accept a variety of variant file types—vcf (version 4.1 and 4.2), txt, bed-like, and tsv [generated from AnnotSV (Geoffroy *et al.* 2018)]—making them flexible for use with real or synthetic perturbations (Supplementary Text). The following variant types [marked by their Manta (Chen *et al.* 2016) abbreviations] are supported: SNPs, indels, deletions (DEL), duplications (DUP), inversions (INV), and complex rearrangements with breakends (BNDs). Across a variety of datasets, including the 1K Genome Project, SuPreMo makes it possible to analyze over 50% of SVs that would not be accessible with existing tools (Fig. 1B).

In particular, symbolic alleles are now easily and uniformly processed (Fig. 1B, navy). On the other hand, insertions, which make up <20% of SVs, remain inaccessible for sequence-based models since the precise inserted sequence is not provided by SV calling methods (Fig. 1B, gray).

SuPreMo provides flexibility through various parameters. While the perturbation is by default centered in the generated sequence, the *shift* parameter slides the window around the perturbation by the given number of base pairs, to the right for positive shift and to the left for negative shift (Supplementary Fig. S2). The *seq_len* parameter determines the length of the output sequence, providing compatibility across models (Supplementary Table S1). The *limit* parameter sets a maximum variant length to be processed, with the default set to two thirds of *seq_len*. The *revcomp* parameter takes the reverse complement of the output sequence. Since the position of the perturbation can vary based on its length, the *shift* parameter, or if the perturbation is near chromosome arm ends, generated sequences are accompanied by the relative position of the perturbation in each sequence. This value is relevant because when a variant is too close to the chromosome arm end– meaning near centromeres or telomeres– it will be positioned near the edge of the sequence (Supplementary Fig. S1B), which can have worse prediction accuracy than the rest of the sequence (Kuang and Pollard 2023). Thus, SuPreMo is a flexible tool for performing ISM that can be applied across sequence-based ML models.

SuPreMo-Akita generates an array of 3D genome disruption scores, predicted contact frequency maps for reference (wild type) and alternate (perturbed) sequences, and genomic tracks of disruption scores across the prediction window. Akita predicts contact frequency maps for a ~1 Mb input sequence at a ~2 kb resolution. SuPreMo-Akita inputs variants as described above and optionally also takes in already generated sequences. Since methods for scoring contact maps are biased and sometimes only target certain features, we have made available 13 different predefined metrics (Gunsalus *et al.* 2023) to use with this tool, with the defaults being the most common measures: mean squared error (MSE) and Spearman's rank correlation coefficient (referred to here as just correlation). To assess the robustness of the generated disruption scores, the augmentation parameter optionally provides averages of scores from standard sequences, sequences with −1 bp and +1 bp shifts, and reverse complement sequences, or any other augmentations specified. Each generated map will be accompanied by the start genomic position and the relative bin that the variant lies in.

Lastly, we considered computational efficiency. To enable customization to different hardware, the user can choose the number of rows to be processed at a time from the input file and what outputs to request, keeping in mind storage and memory limitations. We measured the run time, peak memory, and size of outputs on 3 GHz CPUs using a set of 100–1000 SVs of different types from the reference cancer cell line in Fig. 1B. SuPreMo-Akita is fast and easily scaled up—with the augmentation parameter it takes approximately 19 seconds per variant and reaches ~527 Mb of peak memory (Supplementary Table S2).

We implemented SuPreMo using two models, although our framework is extendable to any model utilizing genome sequences as input. First, we used SuPreMo with DeepSEA (Zhou and Troyanskaya 2015) to rank a set of CTCF deletions based on their predicted effect on epigenetic marks. Second, we used SuPreMo-Akita on cancer SVs (Supplementary Fig. S3). SVs were scored using MSE and correlation, and the top 3 scoring variants for each SV type and

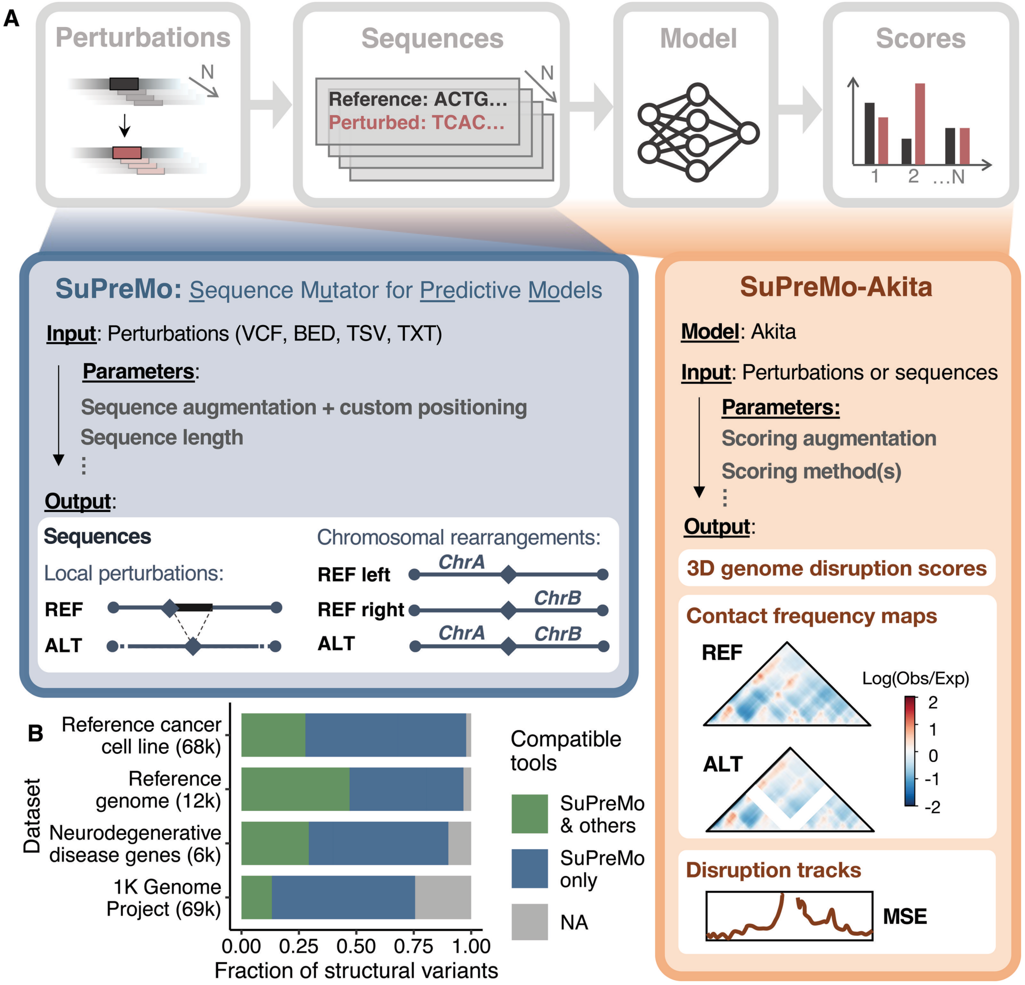

Figure 1. (A) Schematic representation of SuPreMo. SuPreMo generates sequences by incorporating perturbations into the hg38 human reference genome. SuPreMo-Akita applies Akita to those sequences and generates 3D genome disruption scores (effect size of each perturbation) and, optionally, disruption tracks and predicted contact frequency maps. Parameters and outputs are specified. REF: derived from reference allele; ALT: derived from alternate allele; Log(Obs/Exp): log of observed over expected contacts; MSE: mean squared error. (B) Categorization of SVs based on the ability of SuPreMo and other existing tools to incorporate them into a reference genome. SVs that other tools can already process include small indels (green); SVs that only SuPreMo can process include deletions, duplications, inversions, and chromosomal rearrangements (navy); SVs that no tool can process include insertions and copy number variants (CNVs) because the exact sequence is not provided by upstream variant calling pipelines (gray). Datasets are WGS/WES from healthy and disease individuals: a reference cancer cell line (Talsania *et al.* 2022), a reference genome (Zook *et al.* 2020), neurodegenerative disease gene sequences (Kaivola *et al.* 2023), and the 1K Genome Project (Mahmoud *et al.* 2019).

scoring method were selected (Supplementary Fig. S3A–B). We separately ranked variants by their type because their 3D genome disruption scores vary, and by the scoring method because each has unique biases. Using SuPreMo-Akita, contact frequency maps and disruption tracks were generated for these selected SVs and the most interesting variants, based on the structures they disrupt, were chosen (Supplementary Fig. S3C). This method prioritized a deletion of an insulated site that is predicted to cause increased contact frequency between neighboring regions (Supplementary Fig. S3C, left panel). Step-by-step instructions for both implementations are available on Github.

3 Conclusion

SuPreMo is a software tool that facilitates ISM with predictive models and extends this principle with Akita to predict scores for 3D genome folding disruption. Potential use cases include scoring all variants in an individual or cohort for disruption to genome folding, generating predicted contact frequency maps to explore the effects of noncoding variants on regulatory interactions, performing ISM to evaluate or discover sequence motifs using Akita, and, more broadly, generating sequences for input into predictive models of interest to evaluate variant effects. SuPreMo is scalable to a large number of variants and only limited by the storage capacity the user has for the expected outputs. Overall, SuPreMo allows for easy, fast, and broadly applicable analysis of simple variants, SVs, and chromosomal rearrangements in the context of sequence-based predictive models.

Acknowledgements

We thank Shu Zhang for helpful discussion and feedback and for reviewing the manuscript. We thank Daniel Miller and others for helpful advice and test running the tool as well as Miguel Brown for useful suggestions on SuPreMo input compatibilities. We thank Maureen Pittman, Laura Gunsalus, and Shuzhen Kuang for helpful insight on using Akita.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

None declared.

Funding

This work was supported by the National Institutes of Health [grant number R03OD034499 and grant number U01HL157989], Additional Ventures, and Gladstone Institutes.

Data availability

All code and data used for this study is posted on GitHub in the provided repository.

References

Auton A, Brooks LD, Durbin RM, 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* 2015; **526**:68–74.

Agarwal V, Shendure J. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. *Cell Rep* 2020;**31**:107663.

Avsec Ž, Agarwal V, Visentin D *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* 2021a;**18**:1196–203.

Avsec Ž, Weilert M, Shrikumar A *et al.* Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat Genet* 2021b;**53**:354–66.

Benegas G, Batra SS, Song YS. DNA language models are powerful predictors of genome-wide variant effects. *bioRxiv*, 2022.08.22.504706, 2023.

Bushnell B. BBMap. SourceForge2014. sourceforge.net/projects/bbmap/

Chen KM, Wong AK, Troyanskaya OG *et al.* A sequence-based global map of regulatory activity for deciphering human genetics. *Nat Genet* 2022a;**54**:940–9.

Chen V, Yang M, Cui W *et al.* Best practices for interpretable machine learning in computational biology. *bioRxiv*, 2022.10.28.513978, 2022b.

Chen X, Schulz-Trieglaff O, Shaw R *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 2016;**32**:1220–2.

Danecek P, Auton A, Abecasis G, 1000 Genomes Project Analysis Group *et al.* The variant call format and VCFtools. *Bioinformatics* 2011;**27**:2156–8.

Davis S, Pettengill JB, Luo Y *et al.* CFSAN SNP pipeline: an automated method for constructing SNP matrices from next-generation sequence data. *PeerJ Comput Sci* 2015;**1**:e20.

Deng C, Whalen S, Steyert M *et al.* Massively parallel characterization of regulatory elements in the developing human cortex. *Science*, 2024; 384: eadh0559. https://doi.org/10.1126/science.adh0559.

Fudenberg G, Kelley DR, Pollard KS. Predicting 3D genome folding from DNA sequence with Akita. *Nat Methods* 2020;**17**:1111–7.

Geoffroy V, Herenger Y, Kress A *et al.* AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics* 2018; **34**:3572–4.

Gosai SJ, Castro RI, Fuentes N *et al.* Machine-guided design of synthetic cell type-specific cis -regulatory elements. *bioRxiv*, 2023. https://doi.org/10.1101/2023.08.08.552077.

Gunsalus LM, Keiser MJ, Pollard KS. In silico discovery of repetitive elements as key sequence determinants of 3D genome folding. *Cell Genom* 2023;**3**:100410.

Gunsalus LM, McArthur E, Gjoni K *et al.* Comparing chromatin contact maps at scale: methods and insights. *bioRxiv*, 2023. https://doi.org/10.1101/2023.04.04.535480.

Hoffman GE, Bendl J, Girdhar K *et al.* Functional interpretation of genetic variants using deep learning predicts impact on chromatin accessibility and histone modification. *Nucleic Acids Res* 2019; **47**:10597–611.

Kaivola K, Chia R, Ding J *et al.* Genome-wide structural variant analysis identifies risk loci for non-Alzheimer's dementias. *Cell Genom* 2023;**3**:100316.

Kelley DR. Cross-species regulatory sequence activity prediction. *PLoS Comput Biol* 2020;**16**:e1008050.

Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* 2016;**26**:990–9.

Keough KC, Whalen S, Inoue F, Zoonomia Consortium *et al.* Three-dimensional genome rewiring in loci with human accelerated regions. *Science* 2023;**380**:eabm1696.

Kuang S, Pollard KS. Exploring the roles of rnas in chromatin architecture using deep learning. *Biorxiv*, 2023. https://doi.org/10.1101/2023.10.22.563498.

Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 2011;**27**:2987–93.

Mahmoud M, Gobet N, Cruz-Dávalos DI *et al.* Structural variant calling: the long and the short of it. *Genome Biol* 2019;**20**:246.

McArthur E, Rinker DC, Capra JA. Quantifying the contribution of neanderthal introgression to the heritability of complex traits. *Nat Commun* 2021;**12**:4481.

Nguyen E, Poli M, Faizi M *et al.* HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution. *ArXiv*, 2023.

Paik S, Maule F, Gallo M. Dysregulation of chromatin organization in pediatric and adult brain tumors: oncoepigenomic contributions to tumorigenesis and cancer stem cell properties. *Genome* 2021; **64**:326–36.

Rivas-Astroza M, Xie D, Cao X *et al.* Mapping personal functional data to personal genomes. *Bioinformatics*2011;**27**:3427–9.

Schwessinger R, Gosden M, Downes D *et al.* DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nat Methods* 2020;**17**:1118–24.

Talsania K, Shen T-W, Chen X *et al.* Structural variant analysis of a cancer reference cell line sample using multiple sequencing technologies. *Genome Biol* 2022;**23**:255.

Tan J, Shenker-Tauris N, Rodriguez-Hernaez J *et al.* Cell-type-specific prediction of 3D chromatin organization enables high-throughput in silico genetic screening. *Nat Biotechnol* 2023;**41**:1140–50.

Van der Auwera GA, O'Connor BD. *Genomics in the Cloud: Using docker, GATK, and WDL in Terra*. Sebastopol, CA: O'Reilly Media, Inc.. 2020.

Zhou J. Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nat Genet* 2022; **54**:725–34.

Zhou J, Theesfeld CL, Yao K *et al.* Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet* 2018;**50**:1171–9.

Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat Methods* 2015; **12**:931–4.

Zook JM, Hansen NF, Olson ND *et al.* A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol* 2020;**38**:1347–55.

Bioinformatics, 2024, 40, 1–4
https://doi.org/10.1093/bioinformatics/btae340
Applications Note