

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection No software was used to collect the data in this study.

Data analysis The source code for running in silico mutagenesis by MENTR is available at <https://github.com/koido/MENTR>.

For other analyses, we used publicly available software (Python libraries Pytorch (v0.4.0), xgboost (v0.72.1), sklearn (v0.19.1), and SHAP (v0.22.2) on Python 3.6.4; R library pROC (v1.8) on R 3.3.0).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The GWAS trait-associated ncRNA database is available via a user-friendly GUI application (<https://doi.org/10.5281/zenodo.5638259>); basic-usage information is provided in the figure legend of Supplementary Fig. 11. Supplementary files for training MENTR ML models and the pre-trained MENTR ML models are available at <https://github.com/koido/MENTR> (<https://github.com/koido/MENTR/wiki> provides information on how to use them; large files are available at <https://doi.org/10.5281/zenodo.5348471>).

Publicly available datasets that we used in the study are as follows: LCL CAGE transcriptomes22, <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5835>;

GWAS catalogue⁵⁴ (r2019-07-12), <https://www.ebi.ac.uk/gwas/downloads>; 95% credible sets in 139 independent associated regions in IBD, supplementary data from ref. 31; Representative TSS of CAGE peaks (promoters) in FANTOM5, https://fantom.gsc.riken.jp/5/datafiles/phase2.5/extra/CAGE_peaks/hg19.cage_peak_phase1and2combined_coord.bed.gz; Inferred mid position of enhancers in FANTOM5, https://fantom.gsc.riken.jp/5/datafiles/phase2.5/extra/Enhancers/human_permissive_enhancers_phase_1_and_2.bed.gz; CAGE peak annotations in FANTOM5, https://fantom.gsc.riken.jp/5/datafiles/phase2.5/extra/CAGE_peaks_annotation; caQTL from LCL ATAC-seq, supplementary data from ref. 23; 5,376 puQTL and 110 eaQTL from LCL CAGE, data sets from ref. 22 via personal communication; eQTL GTEx v755, <https://www.gtexportal.org/home/>; CAGE and NET-CAGE transcriptome of the five ENCODE cell lines (GM12878, HeLa-S3, HepG2, K562, and MCF-7)¹⁴, <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE118075>; DeepSEA (Beluga)^{15,17}, <https://github.com/FunctionLab/ExPecto>; Basenji¹⁸, <https://github.com/calico/basenji/tree/0.4>; IBD multi-ancestry meta GWAS results^{56,57}, <https://www.ibdgenetics.org>; Asthma multi-ancestry GWAS results³², http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST005001-GCST006000/GCST005212; Results of MPRA: Supplementary data in refs. 24–26; Noncoding credible sets in UK biobank, supplementary data from ref. 39 and <https://www.finucanelab.org/data> (release 1.1); dbSNP 151, https://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606_b151_GRCh37p13/VCF/All_20180423.vcf.gz.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We did not perform sample-size calculations because the main aim of the study was to develop machine-learning-based predictions of mutation effects. We used GWAS summary statistics for validating the approach and for interpreting the GWAS results.
Data exclusions	We downloaded GWAS associations and their ancestry information from the GWAS catalogue (r2019-07-12). From the downloaded 143,963 records, we excluded variants-interaction (epistasis) results, unknown risk allele, and variants whose P-value $\geq 5 \times 10^{-8}$; we also excluded records without rsID (dbSNP 151), and records without SNP or InDels. These steps were needed for performing in silico mutagenesis for trait-associated variants in MENTR.
Replication	The study does not include any experimental results. We used publicly available datasets, and developed or evaluated machine-learning models. We confirmed the high prediction accuracy of in silico mutagenesis via comparisons with results of publicly available eQTL studies and MPRA assays.
Randomization	Randomization was not relevant to the study, because it did not allocate samples, organism or participants into experimental groups.
Blinding	Blinding was not relevant to the study, because it used previously obtained CAGE transcriptome datasets, GWAS results, and other publicly available datasets.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging