

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

- | | | |
|-------------------------------------|-------------------------------------|--|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | The statistical test(s) used AND whether they are one- or two-sided <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A description of all covariates tested |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | 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) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted <i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | 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 Data sources (url) are indicated in the Methods section. CTR-seq data were deposited on DNA Data Bank of Japan Sequencing Read Archive (accession number: DRA010491). Processed data are available at <https://gite.lirmm.fr/ibc/deepSTR>.

Data analysis Source code of the models, a readme.txt file and other instructions for installing and running the analyses are available at <https://gite.lirmm.fr/ibc/deepSTR>.

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 urls of the datasets generated during and/or analysed during the current study are indicated in the Methods section and available at <https://gite.lirmm.fr/ibc/deepSTR>.

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 | No sample-size calculation was performed. All STRs listed in the HipSTR catalogs (https://github.com/HipSTR-Tool/HipSTR-references/raw/master/human/hg19.hipstr_reference.bed.gz and https://github.com/HipSTR-Tool/HipSTR-references/blob/master/mouse/mm10.hipstr_reference.bed.gz) were used along with all the data resources used in the current study (see Methods section). |
| Data exclusions | No data was excluded from the study. |
| Replication | Two libraries of CTR-seq were generated (with and without polyA tailing), and the results were consistent as shown Figure 3. For computational analyses, because all data were used, no replication was required. For deep learning, we separated training, testing and validation datasets prior to model training, and these sets were stored on disk (Figure S7). This allowed us to carry analyses on held-out data that has never been seen by the models. We also made sure that our models do not overfit due for instance to homologous sequences present in both train and test sets (see Methods section). All attempts at replicating our models (at least 10) were consistent and successful. |
| Randomization | As indicated in the Methods section, all STR classes of the HipSTR catalogs with > 5,000 elements were analyzed. Assignments to training, testing and validation datasets were done randomly using Pytorch SubsetRandomSampler. |
| Blinding | For deep learning analyses, all analyses were performed on held-out test sets to avoid optimistic bias in accuracy estimation. For other analyses, blinding is not relevant. |

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

Methods

| n/a | Involved in the study | n/a | Involved in the study |
|-------------------------------------|---|-------------------------------------|---|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Antibodies | <input checked="" type="checkbox"/> | <input type="checkbox"/> ChIP-seq |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> Eukaryotic cell lines | <input checked="" type="checkbox"/> | <input type="checkbox"/> Flow cytometry |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Palaeontology and archaeology | <input checked="" type="checkbox"/> | <input type="checkbox"/> MRI-based neuroimaging |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Animals and other organisms | | |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Human research participants | | |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clinical data | | |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Dual use research of concern | | |

Eukaryotic cell lines

Policy information about [cell lines](#)

| | |
|---|---|
| Cell line source(s) | ATCC A549 |
| Authentication | A549 cell line was purchased from ATCC along with a certified authentication. |
| Mycoplasma contamination | Cell lines were tested negative for mycoplasma contamination. |
| Commonly misidentified lines (See ICLAC register) | No commonly misidentified cell line was used. |