# A deep learning-based framework for estimating fine-scale germline mutation rates

**Yiyuan Fang[1,*], Shuyi Deng[1,*] & Cai Li[1,#]**

[1]State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, Guangdong, China

[*]Equal contribution.

[#]Correspondance: Dr. Cai Li, licai@mail.sysu.edu.cn.

## Abstract

Germline mutation rates are essential for genetic and evolutionary analyses. Yet, estimating accurate fine-scale mutation rates across the genome is a great challenge, due to relatively few observed mutations and intricate relationships between predictors and mutation rates. Here we present MuRaL (Mutation Rate Learner), a deep learning-based framework to predict fine-scale mutation rates using only genomic sequences as input. Harnessing human germline variants for comprehensive assessment, we show that MuRaL achieves better predictive performance than current state-of-the-art methods. Moreover, MuRaL can build models with relatively few training mutations and a moderate number of sequenced individuals. It can leverage transfer learning to build models with further less training data and time. We apply MuRaL to produce genome-wide mutation rate profiles for four species - *Homo sapiens*, *Macaca mulatta*, *Arabidopsis thaliana* and *Drosophila melanogaster*, demonstrating its high applicability. The generated mutation rate profiles and open source software can greatly facilitate related research.

# Introduction

Germline *de novo* mutations (DNMs), which occur either during gametogenesis or post-zygotically, are crucial for evolution and play important roles in many human diseases [1]. Reported *de novo* mutation rates for single nucleotide variants (SNVs) in the human genome range from 1.0 to 1.8 × $10^{-8}$ per base pair (bp) per generation, corresponding to 44 to 82 *de novo* SNVs per genome per generation [2]. Germline mutation rates exhibit high heterogeneity across the genome, from single-nucleotide level to chromosome level [3]. Mutation rate is important for many genetic and evolutionary analyses, such as inferring population demographic histories [4], detecting genomic regions undergoing natural selection [5], and identifying disease-associated genetic variants [6].

Despite its importance, constructing a fine-scale germline mutation rate map for a eukaryotic genome, such as the human genome, is particularly challenging. One main reason is the rarity of DNMs in each generation, making it costly to obtain a large number of high-quality DNMs using the gold standard family-based sequencing strategy (e.g., sequencing parents and offspring simultaneously). In recent years, the decline of sequencing cost alleviated the problem and enabled large-scale sequencing projects in human populations, leading to rapid accumulation of published DNMs. Nonetheless, the number of published DNMs in humans so far is still relatively small (less than one million) [7], and for most non-human genomes none or few DNMs are available for analysis. Many studies used within-species polymorphisms or interspecies divergence to estimate mutation rates, but a substantial fraction of variants at polymorphic or divergent sites are evolutionarily old and affected by natural selection and (or) nonadaptive processes such as GC-biased gene conversion [8]. Recent studies [9-11] demonstrated that extremely rare variants derived from population polymorphism data can serve as a reasonable proxy for DNMs to predict mutation rates, ameliorating the condition of data insufficiency.

Another challenge in estimating fine-scale mutation rates is the complex relationships between predictor variables and mutation rates. Adjacent nucleotides are significant predictors for SNV mutation rates of a focal nucleotide [3,12], particularly the immediately 5' and 3' nucleotides. Nucleotides more distantly from a focal site are also associated with mutation rate variation, though to a less extent [13,14]. Apart from sequence context, functional genomic features, such as DNA methylation, replication timing and recombination rate [15], were reported to be associated with mutation rate variation and have been included in mutation rate modeling work [10,16]. Existing models have several limitations. First, some models only considered a small number of adjacent nucleotides (typically not longer than 7-mer centered at the focal nucleotide). Second, previous work mainly employed linear or generalized linear models to estimate mutation rates with sequence and functional features, but relationships between mutation rates and these predictors tend to be nonlinear and more complicated. Third, some models required many mutations and (or) additional functional genomic features for training, which limits their use in species lacking published mutations and functional genomic data.

Deep learning methods (e.g., deep convolutional neural networks) have shown outstanding performance in solving difficult predictive problems [17] and have been used to address problems in genomics [18-23]. As the genomic sequence is the predominant factor for estimating mutation rates and many functional genomic features are correlated with the sequence, we reasoned that deep learning would be a promising approach to capturing various signals from genomic sequences to generate improved mutation rate profiles.

With the above considerations, we developed a computational framework based on artificial neural networks (named MuRaL, short for Mutation Rate Learner) to generate single-nucleotide germline mutation rates across the genome. Comprehensive assessment using human variant data showed that predicted mutation rates by MuRaL were highly correlated with observed mutation rates at different scales. Compared to current state-of-the-art models, MuRaL required much less training data and fewer sequenced individuals but exhibited improved performance. We further demonstrated that MuRaL can be easily generalized to generate mutation rate profiles for other species.

# Results

### Design of the MuRaL model

The MuRaL model has two main neural network modules (**Fig. 1; Supplementary Fig. 1;** see Methods), one for learning signals from local genomic regions (e.g., 10bp on each side of the focal nucleotide), the other for learning signals from expanded regions (e.g., 1Kb on each side of the focal nucleotide). The main reason for having both modules is that local and distal sequences likely contribute to the mutability of a focal nucleotide in different ways, thus the signals in them might be better learned by different network architectures.

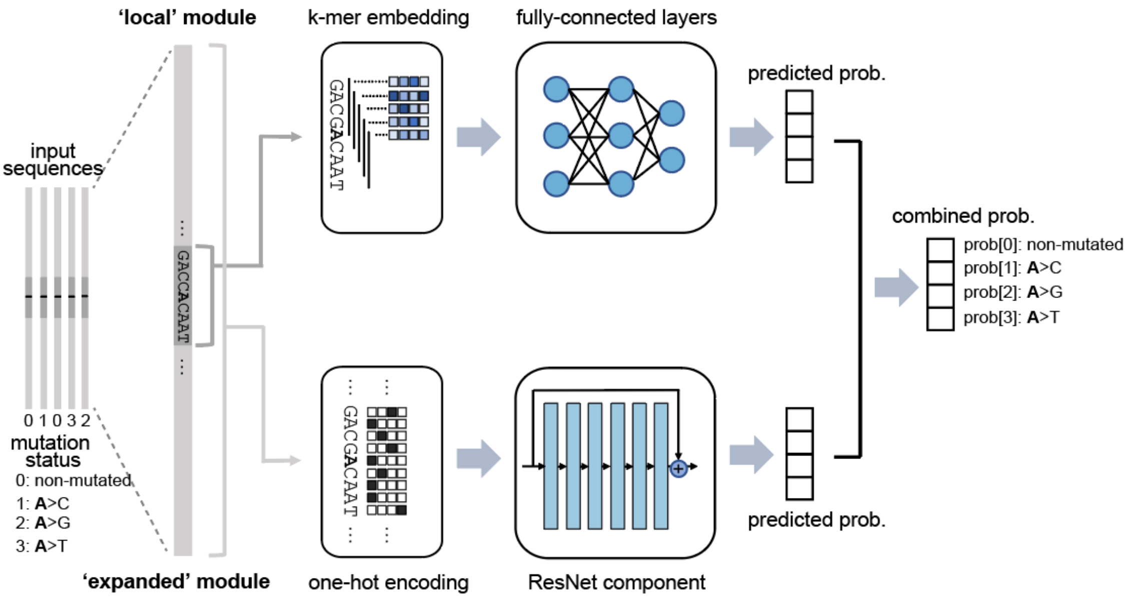

**Figure 1 Schematic of the MuRaL model.** The model consists of a 'local' module and an 'expanded' module. In the 'local' module, the input sequence of the focal nucleotide (e.g., the bold 'A' in the figure) is split into overlapping k-mers which are then mapped into multi-dimensional vectors by the embedding layer. The multi-dimensional vectors are concatenated and passed to three fully-connected (FC) layers. The output of the 'local' module is a probability distribution generated by the softmax function over four predicted classes - non-mutated or one of three possible substitution mutations (e.g., A>C, A>G and A>T). In the 'expanded' module, the input sequence of an expanded region is one-hot encoded. The one-hot encoded matrix is considered as one-dimensional data with four channels and passed to a ResNet component. An additional FC layer and the softmax function following the ResNet component generate a probability distribution over four predicted classes, like that in the 'local' module. The probabilities of 'local' and 'expanded' modules are combined using equal weights (i.e., $0.5*P_{local} + 0.5*P_{expanded}$) to generate the combined probabilities. For training, the mutation status (see bottom left) of each input sequence is also required. More details of the layers are provided in **Supplementary Fig. 1**.

In the 'local' module, we used a k-mer embedding layer and multiple fully-connected (FC) layers to learn signals from the input local sequence surrounding the focal nucleotide (**Fig. 1; Supplementary Fig. 1**). The outputs of the ‘local’ module were probabilities of four-class classification of input samples, which represent the mutational probabilities of a focal nucleotide to another three possible nucleotides and the probability of being non-mutated.

In the ‘expanded’ module, the input sequence of the expanded region was first one-hot encoded and then passed to a series of convolutional neural network (CNN) layers, which form a typical Residual Network (ResNet) architecture (**Fig. 1; Supplementary Fig. 1**). The CNN layers were followed by a FC layer, which produced probabilities of four-class classification of input samples. The meaning of probabilities of ‘expanded’ module is the same as that for the ‘local’ module. The probabilities of ‘local’ and ‘expanded’ modules were combined using equal weights to form a vector of combined probabilities.

Unlike many previous deep learning models in genomics, our model aimed to obtain reliable class probabilities rather than accurate classification (i.e., assign a sample to a specific class). As probabilities derived from neural networks are usually not well calibrated, we further applied a Dirichlet calibration method [24] to obtain calibrated probabilities (**Supplementary Fig. 1**).

For training, we used the cross-entropy loss function and the Adam optimizer [25] for learning model parameters and employed Ray Tune [26] to facilitate hyperparameter tuning (**Supplementary Fig. 2**). We trained separate models for A/T sites and C/G sites, respectively. Moreover, for genomes with exceptionally high mutation rates at CpG sites, we trained models for non-CpG C/G sites and CpG sites separately. The three were called, for short, AT model, non-CpG model and CpG model. We only considered mutation probabilities of single nucleotide substitutions in autosomes because other mutation types and sex chromosomes have specific features that need to be modeled in a different manner.

Regarding the data for model training and evaluation, we generated multiple sets of rare variants in humans based on the large-scale gnomAD data [16] (**Supplementary Fig. 3**; **Supplementary Table 1;** see Methods). We considered the previously reported issue [27] that large sample sizes led to reduced proportions of observed CpG-related mutations

(**Supplementary Fig. 3**). Unless specified elsewhere, we used the '1in2000' data (allele frequency being 1/2000 after downsampling the total allele count to 2000) for training human AT and non-CpG models and the '5in1000' data (allele frequency being ⩽5/1000 after downsampling the total allele count to 1000) for training the CpG model (**Supplementary Table 2**). For detailed evaluation, we mainly used '10in20000' rare variants (as observed mutations) for AT and non-CpG models, and '5in1000' rare variants for CpG models because of their high mutation densities (**Supplementary Table 1**). For the human genome, we used 500,000 mutated and 10,000,000 non-mutated sites for training each MuRaL model (**Supplementary Table 2**), unless specified otherwise. During training, an independent validation dataset consisting of 50,000 mutated and 1,000,000 non-mutated sites was used for evaluating performance and model selection (**Supplementary Fig. 2**).

To evaluate the performance of different models, apart from cross-entropy losses in the validation data, we further considered two metrics - Pearson correlation coefficients between observed and predicted mutation rates for k-mers and binned genomic regions, respectively (see Methods for the detailed definition). We considered k-mer and regional mutation rates for evaluation because observed mutations were sparse across the genome and it was impossible to directly evaluate the accuracy of predicted mutation rates at single-nucleotide resolution.

**Two modules of MuRaL have distinct advantages in learning mutability signals**

Different network architectures and hyperparameters can affect the model performance. To demonstrate that both the 'local' and 'expanded' modules can improve model performance, we constructed 'local-only' and 'expanded-only' models (**Fig. 2a**) and compared them with full models. We set the 'local' region length to be 21bp (10bp on each side of the focal nucleotide; **Fig. 2a**) and the 'expanded' region length to be 2001bp (1Kb on each side). Models of three architectures were trained with the same hyperparameters and same data (see Methods). Predicted mutation rates of sites on human chromosome 20 (Chr20 for short) were used for evaluation. Since the training and validation sites covered only ~1% of the genome, we did not exclude training and validation sites from calculating k-mer/regional mutation rates for model comparison.

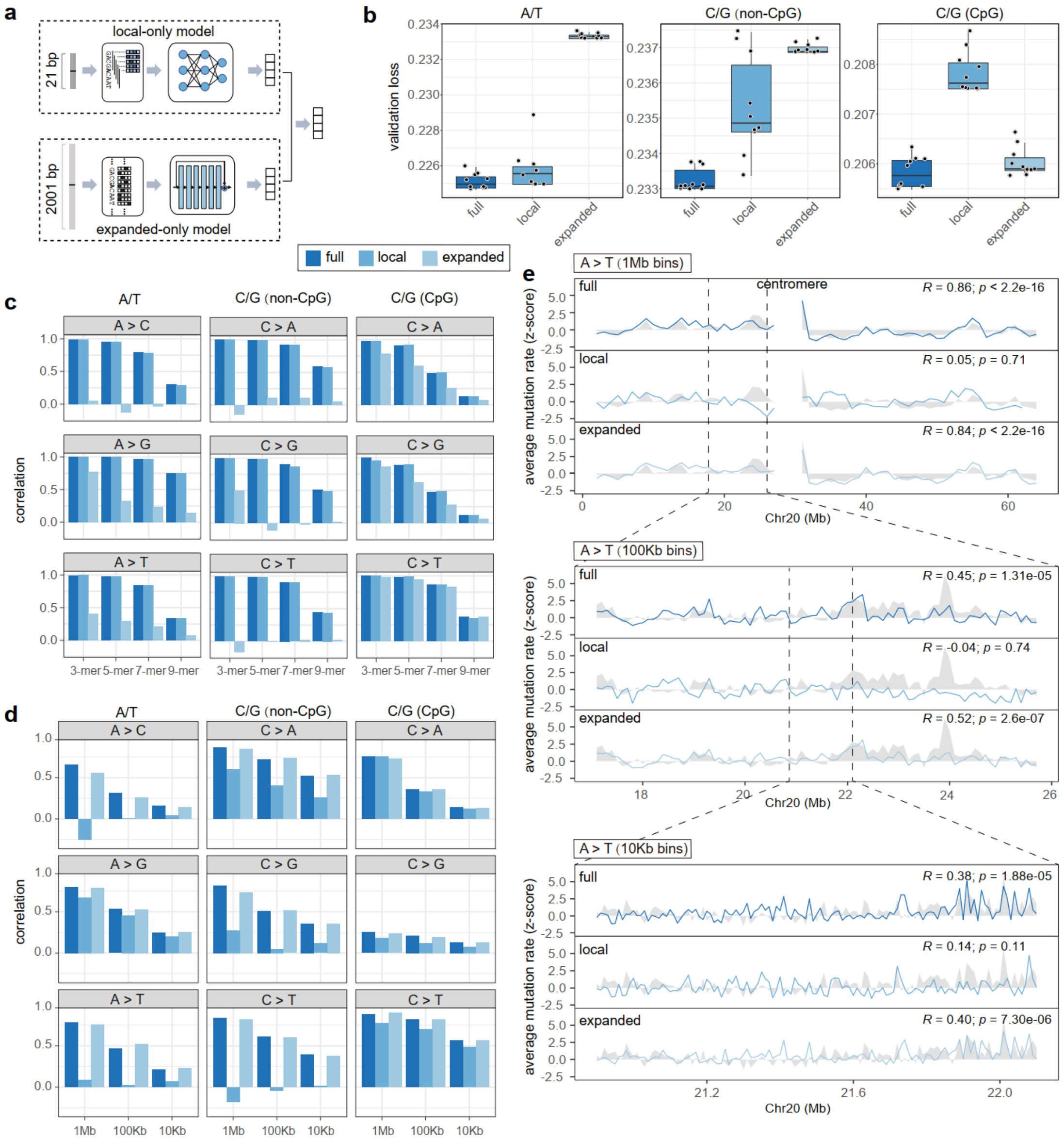

**Figure 2 Two modules of MuRaL learn different mutability signals.** (**a**) Illustration of ‘local-only’ and ‘expanded-only’ models. Diagram elements have same meanings as that in Fig. 1. (**b**) Average validation losses for MuRaL models of three architectures (full, ‘local-only’ and ‘expanded-only’). Separate models were trained for A/T sites, non-CpG C/G sites and CpG sites, respectively. For each model, the lowest loss (mean cross-entropy loss) for each of ten trials was used to generate the boxplots. (**c**) 3-, 5-, 7- and 9-mer mutation rate correlations for different mutation types, based on predicted single-nucleotide mutation rates on human chromosome 20 (Chr20) by models of three different architectures. For each architecture in panel **b**, the best trial with lowest validation loss was used for prediction. The mutations for calculating observed mutation rates were ‘10in20000’ rare variants for AT and non-CpG models, and ‘5in1000’ rare variants for CpG models (see Methods). (**d**) Regional mutation rate correlations with bin sizes of 1Mb, 100Kb and 10Kb on Chr20 for different mutation types. The used observed mutations and meanings of bar colors were the same as that for panel **c**. (**e**) An example showing regional A>T mutation rate correlations at different scales on Chr20 for

three models, with grey shades indicating observed mutation rates and colored lines for predicted rates. The used models and observed mutations were the same as that for panel **d**. As predicted and observed regional mutation rates had different magnitudes, we applied the z-score normalization for visualization. Mutation rates at centromeric regions were not available. Pearson correlation coefficients and p-values for shown regions are provided at the upper right corners. P-values of all correlation tests performed for panels **c** and **d** were provided in **Supplementary Data 1**.

For all three categories of mutation types (mutations related to A/T, non-CpG C/G or CpG sites), the full models always had lowest validation losses among three architectures (**Fig. 2b; Supplementary Fig. 4**). Although the 'local-only' model showed good correlations between observed and predicted k-mer mutation rates (3-, 5-, 7- and 9-mers; **Fig. 2c**), it performed poorly in regional mutation rates of different bin sizes (**Fig. 2d**). The 'expanded-only' model exhibited the opposite patterns. This indicated that 'local-only' and 'expanded-only' models had distinct advantages in capturing signals from input sequences. Notably, the full MuRaL models integrated the advantages of 'local-only' and 'expanded-only' models and performed best among the three. Larger bins generally had higher correlations of regional mutation rates than small bins, which was expected because small bins had more sampling errors. For example, the average A>C mutation rates of different bin sizes across the human Chr20 showed that, full and 'expanded-only' models but not 'local-only' can capture mutation rate variation at different scales (**Fig. 2e**).

**MuRaL can build effective models with relatively few variants from a moderate number of sequenced individuals**

As the number of training mutations required for running MuRaL was not large, we investigated the performance of MuRaL if using training mutations from rare variants of fewer sampled genomes (**Fig. 3a**). We tried training MuRaL models with '1in200' rare variants (allele frequency being 1/200 after downsampling the total allele count to 200), using 500,000 rare variants as training mutations for each of the AT, non-CpG and CpG models. Because the '1in200' data had a small number of rare variants and thus a low mutation density across the genome, we generally got low k-mer and regional correlations if using them for calculating observed mutation rates (**Fig. 3b, c**). However, with more dense rare variant datasets as observed mutations ('10in20000' and '5in1000' data; **Fig. 3b, c**), the '1in200' MuRaL models achieved much increased regional and k-mer mutation rate correlations, suggesting the high predictive performance of these models. This also implies that if only 100 human genomes are available, it is still possible to build reasonably good models by using the rare variants from the 100 individuals.

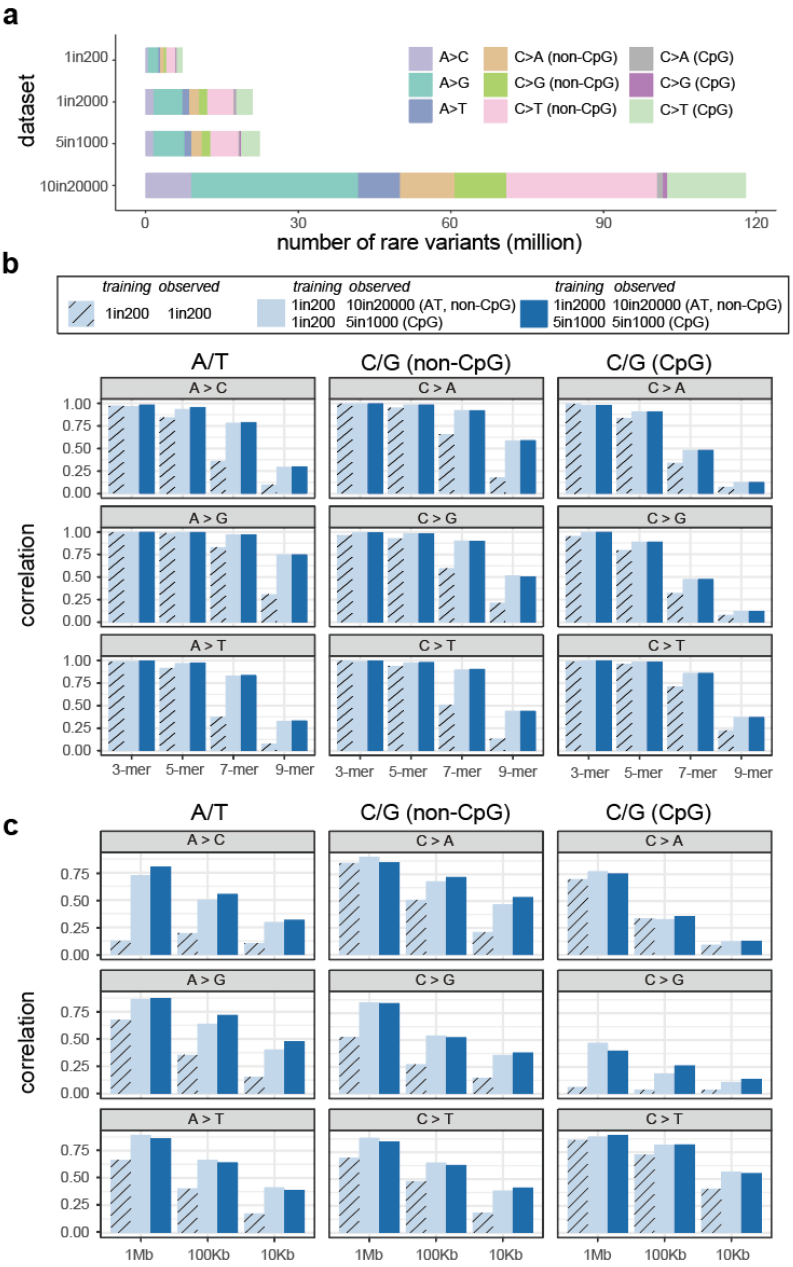

**Figure 3 Comparison of MuRaL models trained with different rare variant data.** (**a**) Numbers of mutations of different types in different rare variant datasets. (**b**) 3-, 5-, 7- and 9-mer mutation rate correlations for different mutation types, based on predicted single-nucleotide mutation rates of Chr20 by different trained models and different observed mutation data (indicated on top of the panel). (**c**) Regional mutation rate correlations with bin sizes of 1Mb, 100Kb and 10Kb on Chr20 for different mutation types and different observed mutation data. The color scheme is the same as that for panel **b**. P-values of all correlation tests performed for panels **b** and **c** were provided in **Supplementary Data 1**.

Moreover, we showed that AT and non-CpG models trained with singleton variants derived from the ‘1in200’ data had similar performance as those trained with the same amount of singleton variants from the ‘1in2000’ data (**Fig. 3b, c**). The mutation rate correlations of the CpG model trained with ‘1in200’ data were also close to that of the model trained with ‘5in1000’ data (**Fig. 3b, c**). These results further corroborated that MuRaL can train effective models with a relatively small number of rare variants from a moderate number of sequenced individuals. Since such requirements can be met by many sequenced species, this opens opportunities for generating fine-scale mutation rate profiles for many species.

### Other factors that affect the performance of MuRaL

As sequencing read coverage can affect mutation calling and is usually accessible for mutation data, we further tried incorporating read coverage into the MuRaL model (**Supplementary Fig. 5;** see Methods). In high-mappability regions, MuRaL models with coverage slightly improved correlations between observed and predicted mutation rates for A/T sites, but not for C/G sites (**Supplementary Fig. 6**). In the poor-mappability regions such as those near the centromere and telomere of Chr20 (**Supplementary Fig. 7**), the model with coverage showed improved correlations between predicted and observed mutation rates. As our work focused on high-mappability regions, MuRaL models without coverage were used for downstream analysis.

We noticed that several chromosomes, such as Chr7, Chr9, Chr15 and Chr16, showed smaller regional correlations than other chromosomes (**Supplementary Fig. 6**), which could be due to their enrichment for recent segmental duplications [28]. The poor regional correlations of Chr8 was ascribable to the under-estimated mutation rates in the region from 0Mb to 25Mb (**Supplementary Fig. 8**), a region reported to have a strikingly high mutation rate [29]. The relatively small learning space (2Kb) of MuRaL models may not efficiently capture distinct region-specific mutability signals in these complicated regions.

In theory, we can increase the lengths of ‘local’ and ‘expanded’ regions in MuRaL models to learn signals from a larger sequence space, yet at the cost of potential overfitting and more computational burden. By testing multiple values (see Methods), we found that for the ‘local’ region length, 5~10 bp on each side appeared to be a proper range (**Supplementary Fig. 9**), as larger lengths didn't confer benefits in reducing the validation loss. Larger lengths of the 'expanded' region generally led to better validation losses (**Supplementary Fig. 9**), but the improvement in the validation loss appeared to diminish when the length is larger than 1Kb (⩾500bp on each side).

Another critical factor affecting model performance is the training data size. We found that increasing training data sizes continuously reduced the validation loss and led to better k-mer/regional mutation rate correlations (**Supplementary Fig. 10**), but the computational burden increased substantially in turn. Posing strict requirements on training data would also limit the application to other species. To balance these, for the human genome, we kept using the data of models trained with 500,000 mutated and 10,000,000 non-mutated sites for downstream analyses.

### MuRaL outperforms existing models

We compared MuRaL with several recently published models for estimating mutation rates across the human genome (**Fig. 4**). Among those, the 'Carlson 7-mer+features' model, which combined 7-mer mutation rates derived from ~36 million singleton variants of 3560 individuals and 14 genomic features for modeling mutation rates, was reported to produce the most accurate map of germline mutation variation in humans [10]. The 'Carlson 7-mer' model in the same study used only the 7-mer mutation rates estimated from singleton variants for prediction [10]. The 'Aggarwala 7-mer' model used 7-mer mutation rates estimated from intergenic SNVs (6~11 million SNVs for each of three populations) of 1000 Genomes Project for prediction [12]. The 'Karczewski 3-mer' model used 3-mer mutation rates estimated with ~24 million rare variants from gnomAD for prediction and took account of DNA methylation levels when predicting mutation rates for CpG sites [16]. Among compared models, MuRaL used the smallest number of training mutations (1.5 million in total) and did not rely on any functional genomic data.

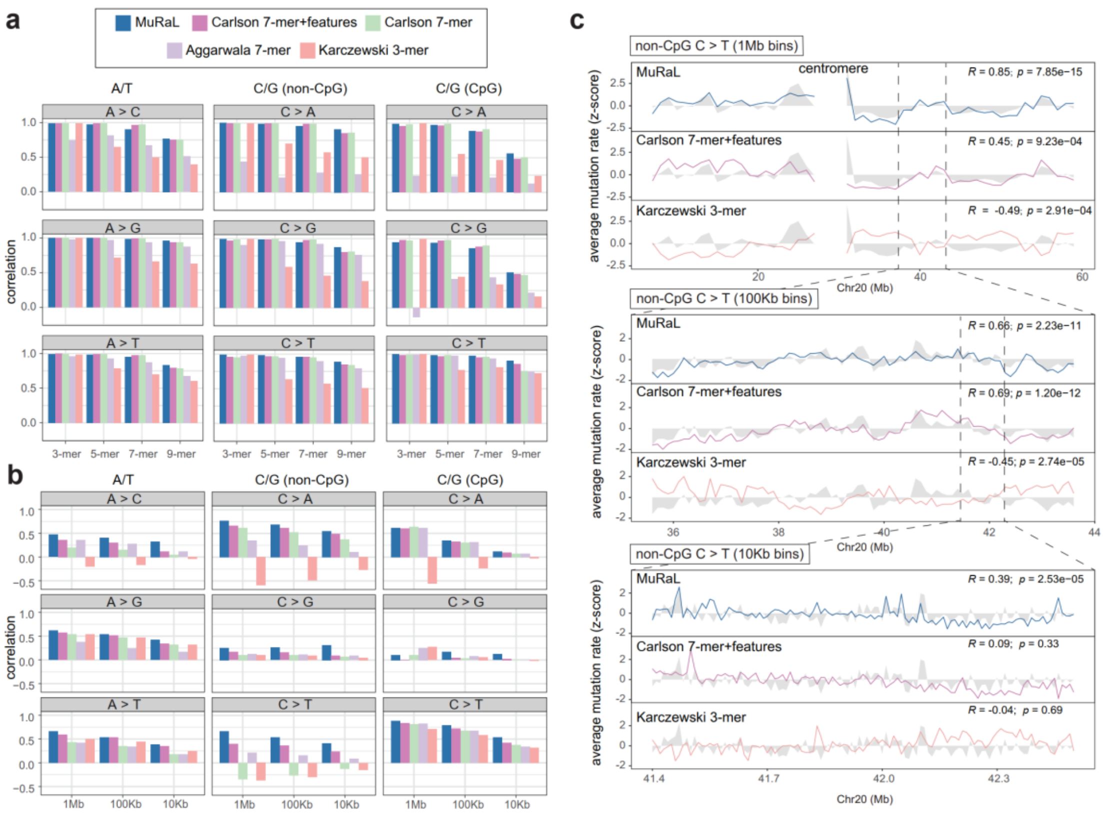

**Figure 4 Comparison of MuRaL and existing models.** (**a**) 3-, 5-, 7- and 9-mer mutation rate correlations for different mutation types, based on predicted single-nucleotide mutation rates of the autosomal genome by different models. The mutations for calculating observed mutation rates were ‘10in20000’ rare variants for AT and non-CpG models, and ‘5in1000’ rare variants for CpG models. (**b**) Regional mutation rate correlations with bin sizes of 1Mb, 100Kb and 10Kb on the autosomal genome for different mutation types. The color scheme was the same as that for panel **a**. (**c**) An example showing regional mutation rate correlations at different scales on Chr20 for three models (MuRaL, ‘Carlson 7-mer+features’ and ‘Karczewski 3-mer’), with grey shades indicating observed mutation rates and colored lines for predicted rates. As predicted and observed regional mutation rates had different magnitudes, we applied z-score normalization for visualization. Mutation rates at centromeric regions were not available. Pearson correlation coefficients and p-values for shown regions are provided at the upper right corners. P-values of all correlation tests performed for panels **a** and **b** were provided in **Supplementary Data 1**.

For genome-wide correlations between observed and predicted k-mer mutation rates, MuRaL, ‘Carlson 7-mer+features’ and ‘Carlson 7-mer’ models performed similarly and were much better than the other two models (**Fig. 4a**). Though for specific mutation types such as A>C and CpG>GpG, 5-mer and 7-mer mutation rate correlations of the Carlson models were slightly better than MuRaL, MuRaL always showed better performance in correlations of 9-mer mutation rates, probably because MuRaL considered sequence context beyond 7-mers. At the chromosome level, the patterns were similar to the genome-wide patterns (**Supplementary Figs 11-13**).

For correlations of regional mutation rates, MuRaL performed better than any other model for bin sizes of 1Mb, 100Kb and 10Kb at the genome-wide level (**Fig. 4b)**. In general, the superiority of MuRaL was more pronounced when the bin sizes were smaller, suggesting that MuRaL predicted improved mutation rates at finer scales compared to previous models. (**Fig. 4b, c; Supplementary Figs 11-13**). It is worth noting that, if aggregating three mutation types associated with the same reference base (e.g., merging A>C, A>G and A>T mutations), at the 1Kb scale MuRaL models still achieved regional correlations of ~0.3 for most mutation types (**Supplementary Fig. 14**). At the chromosome level, MuRaL performed best for most chromosomes and most mutation types (**Supplementary Figs 11-14**). For chromosomes that MuRaL had relatively low regional correlations, other models showed similar trends in most cases (**Supplementary Figs 11-14**).

Although previous models using only local sequence context (3-mers or 7-mers) generally had positive correlations for regional mutation rates, for specific mutation types (especially non-CpG C/G mutations), they had poor or even negative correlations (**Fig. 4b; Supplementary Figs 11-14**). This indicates that a short adjacent sequence cannot fully capture the signal related to the mutability of a focal nucleotide.

We also compared coefficients of variation (CVs) of observed regional mutation rates and those of regional mutation rates from different models. We found that CVs of regional mutation rates from all the models were much smaller than that of observed regional mutation rates at different scales (1Mb, 100Kb and 10Kb; **Supplementary Fig. 15**). Among all models, ‘Carlson 7-mer+features’ showed the highest CVs of regional mutation rates, followed by MuRaL. Although larger CVs of observed mutation rates could be partly due to sampling errors (especially for small bin sizes), the big differences between CVs of observed and predicted mutation rates suggested that predicted mutation rates have less dispersion than real ones, an aspect that needs to be improved in future.

### Training with DNMs and transfer learning

The number of published DNMs in humans is much smaller than that of rare variants. However, because MuRaL can be applied with relatively few training mutations, we tried training AT and nonCpG MuRaL models using 150,000 DNMs and the CpG model using 50,000 DNMs (**Supplementary Table 3**; see Methods).

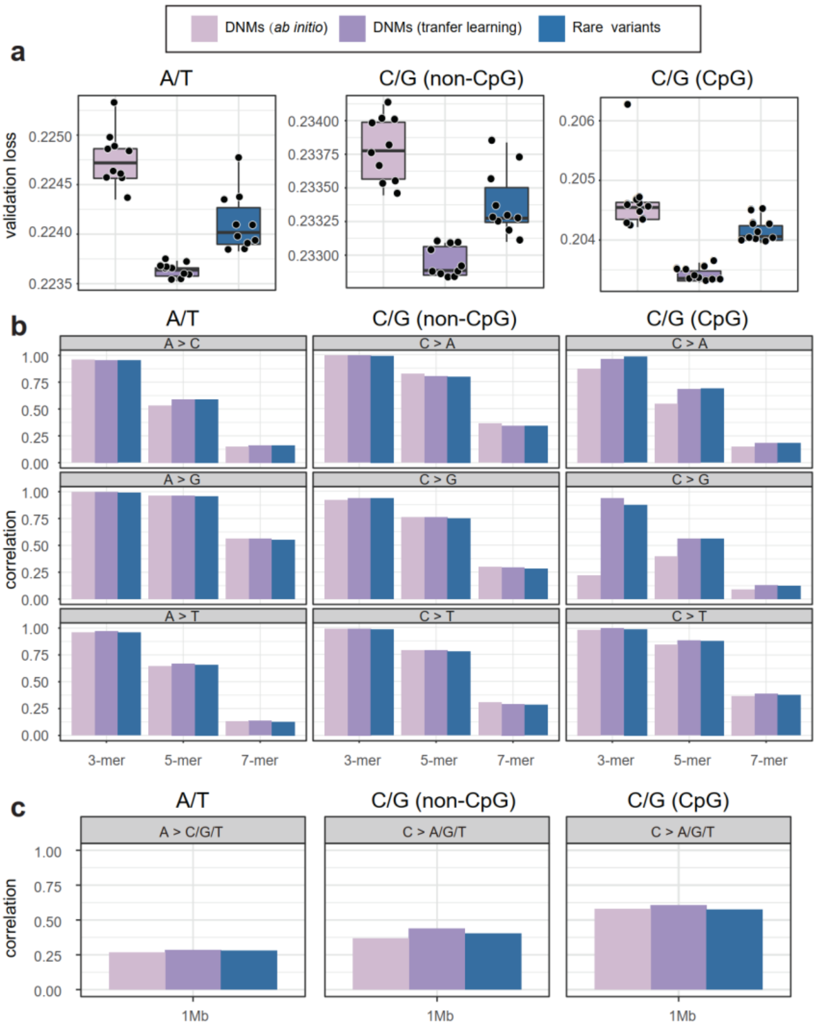

**Figure 5 Training DNM models and transfer learning.** (**a**) Average validation losses on the validation DNMs for three types of models: DNM *ab initio* models, DNM transfer learning models, and rare-variant models. For each model, the lowest loss (mean cross-entropy loss) for each of ten trials was used to generate the boxplots. (**b**) 3-, 5-, and 7-mer mutation rate correlations for different mutation types, based on predicted single-nucleotide mutation rates on human Chr1. Bar colors depict three types of models like that in panel **a**. For each model in panel **a**, the best trial with the lowest validation loss was used for predicting mutation rates on Chr1. The mutations for calculating observed mutation rates were human DNMs. (**c**) Regional mutation rate correlations with a 1Mb bin size on Chr1. The predicted mutation rates of multiple mutation types (e.g. A>C/A>G/A>T) were aggregated for calculating regional correlations, as some mutation types had very few observed DNMs in the data. Smaller bin sizes were not assessed due to few DNMs. Bar colors depict three types of models like that in panel **a**. P-values of all correlation tests performed for panels **b** and **c** were provided in **Supplementary Data 1**.

Transfer learning is widely used in deep learning for scenarios in which the prediction tasks are similar but less training data is available. To study the effectiveness of transfer learning in the MuRaL framework, we trained transfer learning models with the same DNMs, using the pre-trained weights from aforementioned rare-variant models for model initialization. With independent validation DNMs (see Methods), we found that models with transfer learning achieved significantly lower validation losses than those without transfer learning (*ab initio* DNM models; **Fig. 5a**). Transfer learning models also showed better k-mer and regional mutation rate correlations which were calculated with DNMs as observed mutations (**Fig. 5b, c**).

Furthermore, we computed validation losses of the validation DNMs using the rare-variant models described in previous sections. Compared to the *ab initio* DNM models, the rare-variant models achieved significantly lower validation losses for all three categories of mutations **(Fig. 5a**). When looking at k-mer and regional mutation rate correlations, rare-variant models generally performed better than DNM *ab initio* models, and similarly to the DNM transfer learning models (**Fig. 5b, c**). This indicates that if DNMs are unavailable, we can reasonably use mutation rates predicted by rare-variant models to approximate *de novo* mutation rates.

When DNMs are available but limited, it might be beneficial to train transfer learning models with DNMs using pre-trained weights of rare-variant models. However, we note that DNMs collected from different studies could be called in different ways, which could introduce biases when constructing training data and needs to be considered for transfer learning. For example, we found that the collected DNMs were substantially depleted in low-complexity regions and segmental duplications (**Supplementary Fig. 16**), probably due to conservative variant calling procedures.

## Generating mutation rates profile for other species

We further applied MuRaL to estimate mutation rates for three other species. For species that are evolutionarily close to humans, their genomes have high sequence similarities with the human genome, and many mutational processes are likely shared between them. Hence transfer learning can be leveraged for those species. The rhesus macaque (*Macaca mulatta*) is a close relative of humans and a widely used primate model organism. We trained *ab initio* MuRaL models as well as transfer learning models for *M. mulatta* using the rare variants from a dataset of 853 individuals [30]. The training data size of transfer learning models was 30% of that for *ab initio* models (**Supplementary Table 4;** see Methods). We found that transfer learning models showed similar performance to that of *ab initio* models (**Fig. 6a, b**), though transfer learning models used less training data and computation time (**Supplementary Fig. 17**).

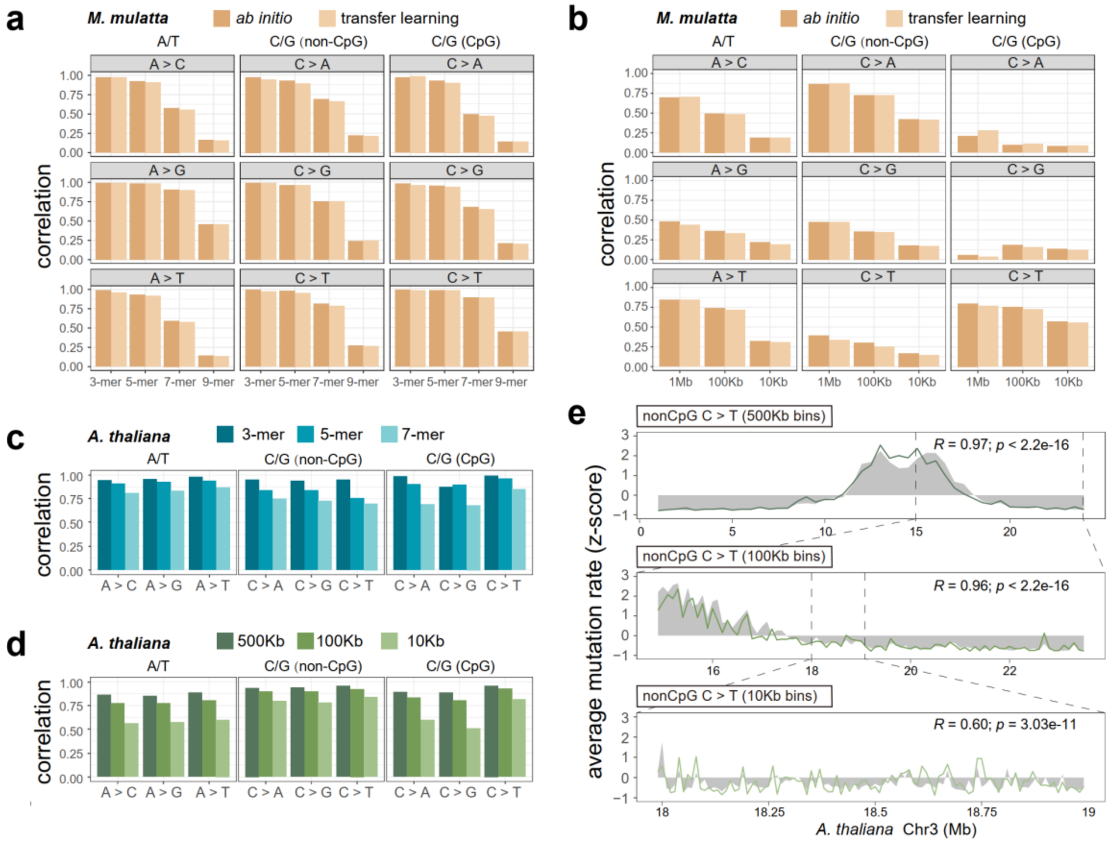

**Figure 6 Application of MuRaL to other species.** (**a**) 3-, 5-, 7- and 9-mer mutation rate correlations for different mutation types, based on predicted single-nucleotide mutation rates on rheMac10 Chr20 by two kinds of models for *M. mulatta*: *ab initio* models and transfer learning models. Rare variants of *M. mulatta* were used for calculating observed mutation rates. Separate models were trained for A/T sites, non-CpG C/G sites and CpG sites, respectively. (**b**) Regional mutation rate correlations with bin sizes of 1Mb, 100Kb and 10Kb on rheMac10 Chr20 for different mutation types. (**c**) 3-, 5-, and 7-mer mutation rate correlations for different mutation types, based on predicted single-nucleotide mutation rates for the *A. thaliana* genome. Rare variants of *A. thaliana* were used for calculating observed mutation rates. Separate models were trained for A/T sites, non-CpG C/G sites and CpG sites, respectively. (**d**) Regional mutation rate correlations with bin sizes of 500Kb, 100Kb and 10Kb for the *A. thaliana* genome. (**e**) An example showing regional mutation rate correlations at different scales on Chr3 for *A. thaliana*, with grey shades indicating observed mutation rates and colored lines for predicted rates. As predicted and observed regional mutation rates had different magnitudes, we applied z-score normalization for

visualization. Pearson correlation coefficients and p-values for shown regions are provided at the upper right corners. P-values of all correlation tests performed for panels **a, b, c** and **d** were provided in **Supplementary Data 1.**

We also trained *ab initio* MuRaL models for two model organisms that are evolutionarily distant to humans - *Drosophila melanogaster* and *Arabidopsis thaliana*. As these genomes (<200Mb) are much smaller than the human genome, we used only 100,000 mutations for training each of the models (**Supplementary Tables 5-6**; see Methods). Despite a relatively small amount of training data, predicted results of the trained models suggested that MuRaL worked well in these species in terms of k-mer and regional mutation rate correlations (**Fig. 6c-e; Supplementary Fig. 18**). For example, for *A. thaliana*, at the scale of 10Kb bins, the correlations of regional mutation rates were >0.5 for all mutation types (**Fig. 6d, e**), indicating the high effectiveness of our method in this species. For *D. melanogaster*, we used singleton variants from only 205 inbred lines for training and still obtained promising results (**Supplementary Fig. 18**), further demonstrating that MuRaL can be applied to the scenarios with a relatively small number of sequenced genomes.

# Discussion

Estimation of sequence mutation rates in the genome can be traced back to the very early period of molecular evolution research [31]. Being hindered by the lack of genomic data, early work only obtained rough estimates of mutation rates for specific genes or genomes. The advent of high-throughput sequencing led to the rapid accumulation of mutation data and enabled more accurate and finer-grained estimation of mutation rates. While several methods have been proposed to infer fine-scale mutation rates across a genome, there was much room for improvement.

In this work, we developed the MuRaL framework to address the challenge in estimating fine-scale mutation rates based on sequences. Compared with the previously best-performed model by Carlson et al. [10], MuRaL learned signals from a much larger sequence space and allowed for more complicated nonlinear modeling. In addition, MuRaL required many fewer training mutations and did not rely on functional genomic data. The human MuRaL models used for most analyses were trained with only 1.5

million rare variants in total, less than 5% of the number of rare variants used for Carlson et al. models (~36 million singletons). Our successful application of MuRaL to three representative species for primates, insects and plants demonstrated its high applicability. We envision that MuRaL will help generate mutation rate profiles for many sequenced species.

Several aspects can be investigated or improved in the near future. A few genomic regions, such as those overlapping recent segmental duplications and the highly mutated regions on human chromosome 8, showed relatively poor mutation rate estimates. To address this, using longer input sequences and (or) function genomic data for training may offer more signals, at the expense of more computational load. Given the rapid development of deep learning hardware, we believe computational load will become a minor obstacle soon. The MuRaL framework can be extended to estimate mutation rates for sex chromosomes and organelle genomes, though more specific assessments are required. As there are already many sequenced genomes for different human populations, it should not be difficult to generate population-specific mutation rate profiles with MuRaL. Similar computational methods could also be developed to predict fine-scale mutation rates for other mutations such as small insertions and deletions.

To our knowledge, this is the first time that deep learning is used to estimate fine-scale mutation rates. Unlike many deep learning models in genomics designed for typical classification or regression problems, our method aimed to predict accurate class probabilities. This work provided an exemplary case for addressing similar problems in genomics. How to obtain reliable class probabilities in deep learning models is still a hot research topic in computer science [32]. The smaller CVs of predicted regional mutation rates than that of observed rates might be partly due to the large class imbalance in the training data, which often causes issues in deep learning models. Future progress on these topics in computer science may help improve our model.

The generated mutation rates and software are relevant for many studies. For example, our predicted mutation rates can help improve previous models of calculating mutation intolerance scores [12,16,33] for prioritizing disease candidate genes or variants.

They can be incorporated into existing phylogenetic models to perform more accurate phylogenetic analysis. They are also informative for detecting regions undergoing selection or introgression in recent evolution. Comparison of mutation rate profiles between species or populations can advance our understanding of mutation rate evolution as well as underlying mutational mechanisms. Since the MuRaL framework has been implemented in an open source package, researchers can train their own models or predict mutation rates for custom sequences using pre-trained models. Although MuRaL is designed for germline mutation rates, it might be adapted for estimating fine-scale somatic mutation rates if mutagenic factors are relatively constant and a considerable number of mutations are available.

In summary, we believe this work represents an important step towards predicting accurate single-nucleotide mutation rates across a genome, facilitating and stimulating future research in related fields.

# Methods

## Design of the MuRaL model

Previous studies revealed that adjacent nucleotides of a specific site predominantly affect its mutation rate and properties of a larger sequence context (e.g., GC content, replication timing) are also associated with mutation rate variation. As local and distal sequences likely affect mutation rates in different ways, we constructed two different neural network modules to learn the signals from the two aspects. One module (termed 'local' module) was designed for learning signals from a local sequence of the focal nucleotide, the other (termed 'expanded' module) for learning signals from an expanded sequence (**Supplementary Fig. 1**).

The 'local' module consists of an embedding layer and three fully-connected (FC) layers to learn signals from the sequence. We used k-mer embedding because it was reported to offer benefits for deep learning models in genomics [34]. The input local sequence was firstly split into overlapping k-mers and the embedding layer maps these k-mers into multi-dimensional vectors. The multi-dimensional vectors from an input sequence were then concatenated to form the input for subsequent two hidden layers and one output layers. For each FC layer, ReLU (Rectified Linear Unit) activation function was used with the output of the FC layer, followed by batch normalization and dropout layers which would facilitate learning and avoid overfitting. The outputs of the 'local' module were probabilities of four-class classification of input samples, representing the probabilities of the focal nucleotide mutated to another three possible nucleotides or being non-mutated.

In the 'expanded' module, the sequence of an expanded region was first converted into four-dimensional vectors using one-hot encoding. Regarding one-hot encoding, each of the four bases ('A', 'C', 'G' and 'T') was converted to a four-element vector, in which all the elements were 0 except for one (e.g., 'A' converted to the vector [1, 0, 0, 0], 'C' converted to [0, 1, 0, 0]). The resulting matrix of the sequence was considered as one-dimensional data with four channels and then passed to a series of one-dimensional convolutional neural network (CNN) layers. The CNN layers were designed following a typical Residual Network (ResNet) architecture. ResNet was previously demonstrated to have outstanding performance in deep neural networks [35]. The CNN layers were followed by an FC output layer, which produced probabilities of four-class classification of input samples. The probabilities of the 'expanded' module had the same meanings as that for the 'local' module.

Next, probabilities of 'local' and 'expanded' modules were combined using equal weights (i.e., $0.5*P_{local} + 0.5*P_{expanded}$) to form a vector of combined probabilities. We also tried using an additional FC layer to combine the outputs of two modules, but such models were not well trained.

The key hyperparameters in the MuRaL model are the length of local sequences, the length of expanded sequences, the length of k-mers in the embedding layer, sizes of two hidden FC layers in the 'local' module, the kernel size and the number of channels for convolutional networks in the 'expanded' module (see **Supplementary Fig. 1**).

### Model implementation

We implemented the MuRaL model with PyTorch framework [36], along with APIs from pybedtools [37] and Janggu [38]. For model training, we used the cross-entropy loss function and the Adam optimizer [25] for learning model parameters, and employed Ray Tune [26] to facilitate hyperparameter tuning (**Supplementary Fig. 2**). The scheduler 'ASHAScheduler' in Ray Tune was used to coordinate trials and execute early stopping before reaching the specified maximum number of training epochs (e.g., 10), which can substantially reduce the training time. The mean cross-entropy loss of the validation sites (i.e., validation loss) was calculated at the end of each training epoch. We further set a stopping rule to terminate a trial if three consecutive epochs did not obtain a validation loss smaller than the current minimum validation loss. The 'learning rate' and 'weight decay' of Adam optimizer were two hyperparameters that could affect the learning performance significantly. Instead of using fixed values, we set specific intervals for values of 'learning rate' and 'weight decay' and used Ray Tune to run trials with different sampled values for the two hyperparameters. To have better convergence, we used the learning rate scheduler 'lr_scheduler.StepLR' in PyTorch to decay the learning rate after each epoch by a specified factor.

## Human mutation data for model training and evaluation

### *Rare variants from gnomAD*

Rare variants generally arose recently in the genome and were less affected by natural selection and nonadaptive evolutionary processes than common variants. Previous studies [9-11] have established that rare variants can be used for estimating mutation rates. For model training and evaluation, we took advantage of the gnomAD database (v2.1.1) which contained genetic variation of 15,708 whole genomes [16]. Only single nucleotide substitutions in autosomes were considered, as other mutation types and sex chromosomes have specific features that need to be modeled separately. We extracted rare variants from gnomAD to approximate DNMs.

When the sample size is as large as that of gnomAD, some mutation types (e.g., CpG>TpG) with high mutation rates could be close to saturation, and the probability of multiple independent mutations (recurrence) at a same position increases. Therefore, we downsampled the gnomAD data into a specified total allele count using a hypergeometric distribution (see the probability density function below), and generated the random alternative allele counts from the hypergeometric distribution:

$$P(AC_{down}) = \frac{\binom{AC}{AC_{down}}\binom{AN-AC}{AN_{down}-AC_{down}}}{\binom{AN}{AN_{down}}} \qquad (1)$$

where AN and AC are the total allele count and the alternative allele count in original data, respectively, $AN_{down}$ and $AC_{down}$ are the total allele count and the alternative allele count after downsampling, respectively. For each polymorphic position, given values of AN, AC and $AN_{down}$, we generated a random number for $AC_{down}$ using the hypergeometric distribution of equation (1). We then extracted the variants with a specific alternative allele frequency (i.e., $AC_{down}$/ $AN_{down}$) in the downsampled data to form the rare variant datasets for subsequent analyses.

First, we downsampled the gnomAD data to total allele counts of 200, 2000 and 7000 (corresponding to 100, 1000 and 3500 diploid genomes) respectively and extracted the singleton variants (corresponding to AFs of 1/200, 1/2000 and 1/7000) in the three downsampled datasets. The three rare variant datasets were named '1in200', '1in2000' and '1in7000', respectively. We considered the sample size of 7000 because Carlson et al. [10] recently used singleton variants from a population of ~3500 individuals for modeling mutation rates. In addition, Karczewski et al. [16] used variants with ≤5 copies in a downsampled set of 1000 haploid genomes for mutation rate estimation, so we also did downsampling for the sample size of 1000 and extracted the variants of $AC_{down}$ ≤5 (termed '5in1000'). To increase the mutation density for smaller-scale evaluation, we further did downsampling for the sample size of 20000 and extracted the variants of $AC_{down}$ ≤10 (termed '10in20000'). For SNVs with two or more different alternative alleles in the original gnomAD data, we did hypergeometric sampling for each alternative allele. In the downsampled dataset, if more than one alternative allele satisfied the rare variant criterion for a specific position, only one alternative allele was randomly selected for downstream analyses (i.e., not allowing multiple rare variants at one position). The numbers of rare variants in different datasets were summarized in **Supplementary Table 1**.

*De novo mutations*

We collected DNMs from the gene4denovo database [7] for analysis. Because some data sources in gene4denovo database contributed only a small number of DNMs and different studies used distinct methods for variant calling, we used only the DNMs from three large-scale studies[39-41] for our analysis, which consisted of 445,467 unique *de novo* SNVs.

To check whether the extracted rare variants can well represent properties of DNMs, we compared mutation spectra of rare variants and DNMs. We counted the occurrences of 1-mer and 3-mer mutation types for each dataset and calculated the relative proportion of each mutation type in the specific dataset. We found that mutation spectra of rare variants were highly similar with that of DNMs (**Supplementary Fig. 3**). However, when the sample size increased, the proportion difference in CpG>TpG mutation subtypes between rare variants and DNMs became larger (**Supplementary Fig. 3**). This was not surprising as mutation rates of CpG>TpG mutation subtypes were highest among all.

Because genomic regions with too low or too high read coverage could have a high probability of false positives/negatives for mutation calls, we utilized the coverage information from gnomAD to exclude the positions with too low or too high read coverage. The genome-wide mean coverage per individual in the gnomAD data was 30.5, and genomic positions within the coverage range of from 15 to 45 (2,626,258,019 bp in autosome retained and considered as high-mappability sites) were used for downstream analyses.

### *Training and validation data for human MuRaL models*

For the human data, we trained separate models for A/T sites, non-CpG C/G sites and CpG C/G sites, respectively. For training each MuRaL model, we randomly chose 500,000 mutations and 10,000,000 non-mutated sites. During training, we used an independent validation dataset consisting of 50,000 mutations and 1,000,000 non-mutated sites for evaluating training performance. The configuration of key hyperparameters for human MuRaL models was provided in **Supplementary Table 7**. As shown above, rare variants derived from a large sample of population could lead to depletion of mutation types of high mutability. On the other hand, rare variants derived from a small sample were relatively ancient and more affected by selection or other confounding processes. To balance the two constraints, we used the '1in2000' data for training models of A/T sites and non-CpG C/G sites in human. Because '1in2000' data showed more depletion of CpG>TpG mutations than '1in200' and '5in1000' data, it was not the ideal data for training the model of CpG sites. Although both '1in200' and '5in1000' datasets were rare variants of $AC_{down}/AN_{down} \leq 0.005$, we chose '5in1000' data for training the CpG model because of its larger number of mutations. For detailed model evaluation, we mainly used '10in20000' rare variants for A/T and non-CpG models, and '5in1000' rare variants for CpG models due to their high mutation densities (**Supplementary Table 1**).

## Calibrating predicted probabilities

The main aim of our work is to obtain reliable class probabilities rather than accurate classification (e.g., predicting ones or zeros). Because probabilities from neural networks are usually not well calibrated, after training a MuRaL model, we applied a Dirichlet calibration method [24] on the output combined probabilities to obtain better calibrated probabilities. Parameters of a Dirichlet calibrator were estimated by fitting the calibrator to the predicted probabilities of the validation data. Metrics such as Expected Calibration Error (ECE), classwise-ECE and Brier score [24] were used for evaluating the performance of Dirichlet calibration. By comparing predicted mutation rates of validation data before and after calibration, we found that Dirichlet calibration indeed resulted in better ECE, classwise-ECE and Brier scores (**Supplementary Fig. 4**), although the improvements appeared to be relatively small. Small values of ECE and classwise-ECE scores before calibration (**Supplementary Fig. 4**) suggested that the original predicted mutation probabilities were already quite well calibrated in terms of such metrics.

The absolute values of above combined probabilities were not mutation rates per bp per generation. To obtain a mutation rate per bp per generation for each nucleotide, one can further scale the calibrated probabilities based on previously reported genome-wide DNM mutation rate per bp per generation. We note that whether to do or not do this scaling does not affect the calculation of k-mer and regional mutation rate correlations in this study.

### Extending MuRaL with read coverage

Read coverage (or read depth) of alignments can affect variant calling and thus observed mutation densities. We tried extending the MuRaL model to incorporate the coverage information (**Supplementary Fig. 5**). We used the pre-compiled coverage track from gnomAD (v2.1.1). In the 'local' module, we calculated mean coverage of the local sequence of the focal nucleotide, and added it as an additional element to the concatenated vector of embeddings of the local sequence. In the 'expanded' module, we extracted a coverage vector for the nucleotides of the expanded sequence, and merged it with the one-hot encoded matrix of the expanded sequence to form a five-channel input for subsequent convolutional networks. Such a design can also easily incorporate other genome-wide tracks (e.g., replication timing, recombination rate, etc.) to extend the

MuRaL model.

### Correlation analysis of k-mer mutation rates

We classified mutations into six mutation types according to the reference and alternative allele: A>C, A>G, A>T, C>A, C>G, and C>T. Mutations with reference nucleotides T and G were reverse-complemented to that with A and C, respectively. For each mutation type, the k-mer subtypes were defined by the upstream and downstream bases flanking the variant site. For example, there are four possible bases at both the upstream -1 position and downstream +1 position, respectively, so there are $6 \times 4^2 = 96$ 3-mer subtypes, 16 3-mer subtypes for each basic mutation type. Similarly, for 5-mers and 7-mers, there are $6 \times 4^4 = 1{,}536$ and $6 \times 4^6 = 24{,}576$ subtypes respectively. In some analyses, we also considered 9-mers ($6 \times 4^8 = 393{,}216$ 9-mer subtypes) if the number of mutations was large enough. For example, for the mutation type A>G, G[A>G]C and AG[A>G]CT are a 3-mer subtype and 5-mer mutation subtype associated with it respectively.

For the *i*th k-mer subtype, we calculated the observed mutated rate $K_i^{obs}$ and the predicted mutation rate $K_i^{pred}$ in the considered regions:

$$K_i^{obs} = \frac{m_i}{N_i} \quad (2)$$

$$K_i^{pred} = \frac{\sum_{j=1}^{N_i} p_j}{N_i} \quad (3)$$

where $m_i$ is the observed number of mutated sites belonging to that k-mer subtype, $N_i$ is the total number of sites harboring the reference k-mer motif (e.g., all AAT 3-mers for the subtype A[A>C]T), and $p_j$ is the predicted mutation probability of the *j*th valid site.

Based on the calculated observed and predicted k-mer mutation rates, we can calculate the Pearson correlation coefficient for any set of k-mer subtypes (one subtype as a datapoint). For example, we can calculate a correlation coefficient for 16 3-mer subtypes associated with the mutation type A>C in a specific chromosome. Note that for CpG sites, there are only four 3-mer subtypes for a basic mutation type as the +1 position is fixed to be ‘G’. The Dirichlet calibrated mutation rates were used for calculating k-mer mutation rates unless otherwise specified.

**Correlation analysis of regional mutation rates**

Since it was impossible to evaluate accuracy of predicted mutation rates on the single-nucleotide level, we compared average mutation rates in binned regions and calculated correlations between observed and predicted regional mutation rates to evaluate the performance of predicting models.

More specifically, for a specific mutation type, we calculated the observed and predicted mutation rates as below.

First, we divided a specified region (e.g., a chromosome) into non-overlapping bins with a given bin size (e.g. 10kb, 100kb, etc.). For the *i*th binned region, we calculated the observed mutated rate $R_i^{obs}$ and the predicted mutation rate $R_i^{pred}$,

$$R_i^{obs} = \frac{m_i}{N_i} \tag{4}$$

$$R_i^{pred} = \frac{\sum_{j=1}^{N_i} p_j}{N_i} \tag{5}$$

where $m_i$ is the number of observed mutations of the specific mutation type (e.g. A>C), $N_i$ is the total number of sites with same base as the reference base of that mutation type (e.g. all A/T sites for the mutation type A>C), and $p_j$ is the predicted mutation probability of the *j*th valid site in the binned region.

Based on the calculated observed and predicted regional mutation rates, we can calculate the Pearson correlation coefficient for any set of binned regions (one bin as a datapoint). For example, we can calculate the correlation coefficient for regional mutation rates of the mutation type A>C in all 100kb bins in a chromosome. As there are gaps and low-mappability regions in the genome, to avoid using regions with few valid sites for correlation analysis, we only used the bins that fit the criterion $N > 20\% * N_{median}$, where $N_{median}$ is the median of numbers of valid sites in all bins for a chromosome. The Dirichlet calibrated mutation rates were used for calculating regional mutation rates unless otherwise specified.

**Comparison of models with different network architectures**

To see how the 'local' and 'expanded' modules contribute to the model, we considered three models – the 'local-only' model containing only the 'local' module, the 'expanded-only' model containing only the 'expanded' module and the full model with both modules. We trained the models with a training dataset of 500,000 mutated and 10,000,000 non-mutated sites randomly selected from autosomes. For each network architecture, ten trials were trained with Ray Tune. A validation dataset consisting of 50,000 mutated and 1,000,000 non-mutated sites was used to compare performance of three architectures based on the validation losses of trained trials. We also used the best trained trial with lowest validation loss for each of three architectures to predict mutation probabilities on the whole chromosome of human Chr20, results of which were then passed to compute k-mer/regional mutation rates for model comparison.

### Comparison of models with different hyperparameters

Due to the high demand for GPU memory and computing, it is impossible to test the behaviors of all model hyperparameters comprehensively. At the beginning, we set a relatively large search space for hyperparameters and used Ray Tune to run dozens of trials to get more narrowed ranges. We further detailly investigated the impact of two input-related hyperparameters – 'local radius' (length of the local sequence on each side of the focal nucleotide) and 'distal radius' (length of the expanded sequence on each side of the focal nucleotide). For each hyperparameter, we set five different values for it and fixed the setting of other hyperparameters. We used a training dataset consisting of 100,000 mutated and 2,000,000 non-mutated A/T sites and a validation dataset consisting of 50,000 mutated and 1,000,000 non-mutated A/T sites. For each setting of hyperparameters, we ran ten trials and used the model with lowest validation loss of each trial for comparison (see **Supplementary Fig. 9**).

We also investigated the impact of different training data sizes. We tried four different numbers of A/T sites as training data: 1) 50,000 mutated+1,000,000 non-mutated; 2) 100,000 mutated+2,000,000 non-mutated; 3) 200,000 mutated+4,000,000 non-mutated; and 4) 500,000 mutated+10,000,000 non-mutated. For model evaluation, we used a validation dataset consisting of 50,000 mutated and 1,000,000 non-mutated A/T sites to calculate validation losses for comparison.

### Comparison of MuRaL models with previously published models

We considered the following four published models in our comparative analysis:

1) 'Aggarwala 7-mer' model [12]: this model estimated 7-mer mutation rates based on intergenic polymorphic sites from 1000 Genomes Project (~11 million variants in the African populations, ~7 million variants in the European populations, and ~6 million variants in the East Asian populations.). The original study provided 7-mer mutation rates for three populations ('Supplementary Table 7'). We used the averaged mutilation rate among three populations for each 7-mer to generate mutation rates of all bases in human autosomes.

2) 'Carlson 7-mer' model [10]: this model used 7-mer mutation rates estimated from 36 million singleton variants from 3560 individuals. Note that some 7-mers didn't have any observed mutations and thus had mutation rates of zero, which was a limitation of this method. We downloaded the 7-mer mutation rates from 'Supplementary Data1' of the study and generated mutation rates of all bases in human autosomes.

3) 'Carlson 7-mer+features' model [10]: this model used 7-mer mutation rates of the 'Carlson 7-mer' model and 14 genomic features for modeling. We noticed that some sites had zero mutation rates for specific mutation types. In addition, this model did not generate predicted rates for sites within 5 Mb of the start/end of a chromosome because of lacking corresponding recombination rate data. We downloaded the whole genome mutation rate profile of this model from the original study (http://mutation.sph.umich.edu/hg19/) for analysis.

4) 'Karczewski 3-mer' model [16]: this model estimated 3-mer mutation rates based on rare variants in gnomAD database. For CpG sites, this model divided the methylation levels into three classes (high, medium and low) and applied separate mutation rates for CpG sites with different methylation levels. We downloaded the 3-mer mutation rates from 'Supplementary Dataset 10' of the study and generated mutation rates of all bases in human autosomes. We used the same methylation data as that described in the study for predicting mutation rates at CpG sites.

When performing comparative analyses between our models and other existing models, we excluded the genomic sites without predictive values in at least one model. In total, 2,390,435,721 bases of the autosome genome were used in comparison. Note that among the four existing models, the 'Carlson 7-mer+features' model had strongest data requirements for prediction and its mutation rate profile contains the smallest number of predicted sites. We calculated k-mer and regional mutation rate correlations for the four models using the same method as that for MuRaL models.

The MuRaL models used in comparative analysis were those trained with 500,000 mutated and 10,000,000 non-mutated sites. The numbers of trainable parameters for AT, non-CpG and CpG models were 180,257, 175,557, and 169,682, respectively. The total number of trainable parameters (180,257 + 175,557 + 169,682 = 525,496) was close to that of the 'Carlson 7-mer+features' model (392,128) [10].

## Transfer learning

Transfer learning is widely used in deep learning for scenarios in which the prediction tasks are similar. After training MuRaL models with rare variants from gnomAD, we took advantage of published human DNMs to perform transfer learning. For each of the AT and non-CpG models, we compiled a training dataset consisting of 150,000 DNMs and 3,000,000 non-mutated sites and an independent validation dataset consisting of 20,000 DNMs and 400,000 non-mutated sites. For the CpG models, we compiled a training dataset consisting of 50,000 DNMs and 1,000,000 non-mutated sites and a validation dataset consisting of 20,000 DNMs and 400,000 non-mutated sites. We tried two transfer learning strategies: 1) using all pre-trained weights for model initialization and re-training all weights and 2) use all pre-trained weights for model initialization but only re-training the weights of last FC layers of two modules. We chose the results of the first strategy for later comparative analysis, as the second strategy led to poor performance.  We also trained *ab initio* models using the same DNM training datasets, with the same hyperparameter setting as that for the rare-variant models. Because we found that the collected DNMs were highly depleted in low-complexity regions and segmental duplications, we excluded DNMs located in these regions from training and evaluating models.

## Apply MuRaL to other species

We used MuRaL to train mutation rate models for three other species: *Macaca mulatta, Drosophila melanogaster* and *Arabidopsis thaliana. M. mulatta* is a widely used primate model organism with similar genome size as that of the human genome. *D. melanogaster* and *A. thaliana* are widely used model organisms but with much smaller genomes (169 Mb and 119 Mb, respectively).

The variants of *M. mulatta* were from a recent study [30] and downloaded from https://hgdownload.soe.ucsc.edu/gbdb/rheMac10/rhesusSNVs/. This dataset included 853 sequenced genomes and 85.7 million variants. We extracted 19,553,394 singleton variants (requiring AC=1 and AN>=1500) of autosomes for training AT and non-CpG models. For training CpG models, we did downsampling to the total allele count (AN) of 1000 and extracted the variants with $AC_{down} \leq 5$ (6,422,014 CpG-related rare variants on autosomes). To identify regions with poor mappability in the *M. mulatta* genome, we downloaded raw reads of three individuals (accession numbers: SRR11999190, SRR11999224 and SRR12070989) and mapped them to the rheMac10 assembly using bwa-mem2 [42]. The peak read depth for alignments of the three libraries was 127, and we kept genomic sites with read depth within the range of 63-190 (2,620,098,971 bp in autosomes in total ) for downstream analyses.

We trained *ab initio* models as well as transfer learning models for *M. mulatta*. For *ab initio* models, we compiled a training dataset consisting of 500,000 mutated and 10,000,000 non-mutated sites, and an independent validation dataset consisting of 50,000 mutated and 1,000,000 non-mutated sites. We used the same hyperparameter setting as that for human *ab initio* models. For transfer learning models, we compiled a training dataset consisting of 150,000 mutated and 3,000,000 non-mutated sites and an independent validation dataset consisting of 50,000 mutated and 1,000,000 non-mutated sites. For each model, ten trials were run and the checkpointed model with lowest validation loss among all trials was used for prediction.

The variant file of *A. thaliana* was downloaded from 1001 Genomes project (https://1001genomes.org/) [43], which included 12,883,854 polymorphic sites for 1135 inbred lines. The variants of each individual in the VCF file were all homozygotes because of long-term inbreeding and thus the lowest AC is 2. We excluded the poorly mapped genomic regions by using the coverage information from the 1001 Genomes project. We first calculated the average read depth across 1135 lines for each nucleotide and the mode of the rounded average depths across the genome was 21. We retained the positions whose average read depth was within the range of 10-30 (102,069,978 sites in total). For training and validating the AT model, we used singleton variants by requiring AC to be 2 and AN of >= 1000. Because there is a high mutation rate of C>T at both CpG and non-CpG C/G sites, the C>T mutations were depleted in the singleton rare variants. We further compiled a rare variant dataset by requiring AC <= 10 and AN >= 1000 for training and validating non-CpG and CpG models. For each of AT, non-CpG and CpG models, we randomly selected 100,000 rare variants and 2,000,000 non-mutated sites for training, and 10,000 rare variants and 200,000 non-mutated sites for validation. For the CpG model, we randomly selected 50,000 rare variants and 1,000,000 non-mutated sites for training, and 5,000 rare variants and 100,000 non-mutated sites for validation.

The variant dataset of *D. melanogaster* used in our analysis was from Drosophila Genetic Reference Panel (DGRP) [44], which sequenced 205 inbred lines. The original variant file in VCF format contained 3,837,601 polymorphic sites (excluding sex chromosomes and heterochromatic sequences). Because of being derived from inbred lines, each polymorphic site was homozygous for each individual and original AN and AC tags in the variant file were corresponding to counts of individuals rather than alleles. We extracted 702,864 singleton rare variants by requiring AC to be 1 and AN of >= 100. Then the dataset of rare variants was divided into A/T sites (285,374) and C/G sites (418,713), respectively. Because there is little methylation at CpG sites in the genome of *D. melanogaster* [45] and the mutation rate of CpG>TpG is not exceptionally high in this species, we did not separate non-CpG and CpG C/G sites for training. For each of the AT and CG models, we randomly selected 100,000 rare variants and 2,000,000 non-mutated sites for training, and 10,000 rare variants and 200,000 non-mutated sites for validation.

The configurations of hyperparameters for MuRaL models of the three species were provided in **Supplementary Tables 8-10**. For each training task, the checkpointed model with lowest validation loss among all trials was used for predicting base-wise mutation rates in the whole genome. The calculation of k-mer and regional mutation rate correlations was the same as that for the human data.

## References

1 Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. *Nat Rev Genet* **13**, 565-575, doi:10.1038/nrg3241 (2012).

2 Acuna-Hidalgo, R., Veltman, J. A. & Hoischen, A. New insights into the generation and role of de novo mutations in health and disease. *Genome Biol* **17**, doi:10.1186/s13059-016-1110-1 (2016).

3 Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* **12**, 756-766, doi:10.1038/nrg3098 (2011).

4 Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**, 919-925, doi:10.1038/ng.3015 (2014).

5 Pavlidis, P. & Alachiotis, N. A survey of methods and tools to detect recent and strong positive selection. *J Biol Res (Thessalon)* **24**, 7, doi:10.1186/s40709-017-0064-0 (2017).

6 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315, doi:10.1038/ng.2892 (2014).

7 Zhao, G. *et al.* Gene4Denovo: an integrated database and analytic platform for de novo mutations in humans. *Nucleic Acids Res* **48**, D913-D926, doi:10.1093/nar/gkz923 (2020).

8 Messer, P. W. Measuring the rates of spontaneous mutation from deep and large-scale polymorphism data. *Genetics* **182**, 1219-1232, doi:10.1534/genetics.109.105692 (2009).

9 Zhu, Y. O., Sherlock, G. & Petrov, D. A. Extremely Rare Polymorphisms in Saccharomyces cerevisiae Allow Inference of the Mutational Spectrum. *PLoS Genet* **13**, e1006455, doi:10.1371/journal.pgen.1006455 (2017).

10 Carlson, J. *et al.* Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nat Commun* **9**, 3753, doi:10.1038/s41467-018-05936-5 (2018).

11 Agarwal, I. & Przeworski, M. Signatures of replication timing, recombination, and sex in the spectrum of rare variants on the human X chromosome and autosomes. *Proc Natl Acad Sci U S A* **116**, 17916-17924, doi:10.1073/pnas.1900714116 (2019).

12 Aggarwala, V. & Voight, B. F. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat Genet* **48**, 349-355, doi:10.1038/ng.3511 (2016).

13 Zhao, Z. & Boerwinkle, E. Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Res* **12**, 1679-1686, doi:10.1101/gr.287302 (2002).

14 Li, C. & Luscombe, N. M. Nucleosome positioning stability is a modulator of germline

mutation rate variation across the human genome. *Nat Commun* **11**, 1363, doi:10.1038/s41467-020-15185-0 (2020).

15 Segurel, L., Wyman, M. J. & Przeworski, M. Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet* **15**, 47-70, doi:10.1146/annurev-genom-031714-125740 (2014).

16 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443, doi:10.1038/s41586-020-2308-7 (2020).

17 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444, doi:10.1038/nature14539 (2015).

18 Eraslan, G., Avsec, Z., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet* **20**, 389-403, doi:10.1038/s41576-019-0122-6 (2019).

19 Avsec, Z. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* **18**, 1196-1203, doi:10.1038/s41592-021-01252-x (2021).

20 Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* **12**, 931-934, doi:10.1038/nmeth.3547 (2015).

21 Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* **33**, 831-838, doi:10.1038/nbt.3300 (2015).

22 Schwessinger, R. *et al.* DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nat Methods* **17**, 1118-1124, doi:10.1038/s41592-020-0960-3 (2020).

23 Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548 e524, doi:10.1016/j.cell.2018.12.015 (2019).

24 Kull, M., Perello-Nieto, M., Kängsepp, M., Song, H. & Flach, P. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. *arXiv preprint arXiv:1910.12656* (2019).

25 Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

26 Liaw, R. *et al.* Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118* (2018).

27 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).

28 Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003-1007, doi:10.1126/science.1072047 (2002).

29 Nusbaum, C. *et al.* DNA sequence and analysis of human chromosome 8. *Nature* **439**, 331-335, doi:10.1038/nature04406 (2006).

30 Warren, W. C. *et al.* Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. *Science* **370**, doi:10.1126/science.abc6617 (2020).

31 Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624-626, doi:10.1038/217624a0 (1968).

32 Ovadia, Y. *et al.* Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *arXiv preprint arXiv:1906.02530* (2019).

33 di Iulio, J. *et al.* The human noncoding genome defined by genetic diversity. *Nat Genet* **50**, 333-337, doi:10.1038/s41588-018-0062-7 (2018).

34 Trabelsi, A., Chaabane, M. & Ben-Hur, A. Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities. *Bioinformatics* **35**, i269-i277, doi:10.1093/bioinformatics/btz339 (2019).

35 He, K., Zhang, X., Ren, S. & Sun, J. in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770-778.

36 Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32**, 8026-8037 (2019).

37 Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423-3424, doi:10.1093/bioinformatics/btr539 (2011).

38 Kopp, W., Monti, R., Tamburrini, A., Ohler, U. & Akalin, A. Deep learning for genomics using Janggu. *Nat Commun* **11**, 3488, doi:10.1038/s41467-020-17155-y (2020).

39 Jonsson, H. *et al.* Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519-522, doi:10.1038/nature24018 (2017).

40 Yuen, R. *et al.* Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat Neurosci* **20**, 602-611, doi:10.1038/nn.4524 (2017).

41 An, J. Y. *et al.* Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**, doi:10.1126/science.aat6576 (2018).

42 Vasimuddin, M., Misra, S., Li, H. & Aluru, S. in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* 314-324 (IEEE).

43 Consortium, T. G. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166**, 481-491, doi:10.1016/j.cell.2016.05.063 (2016).

44 Huang, W. *et al.* Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. *Genome Res* **24**, 1193-1208, doi:10.1101/gr.171546.113 (2014).

45 Lyko, F., Ramsahoye, B. H. & Jaenisch, R. DNA methylation in Drosophila melanogaster. *Nature* **408**, 538-540, doi:10.1038/35046205 (2000).

# Data availability

All the analyses in this study were based on published data. The predicted mutation rate profiles for genomes of human, *M. mulatta*, *A. thaliana* and *D. melanogaster* are available at the ScienceDB repository: https://www.doi.org/10.11922/sciencedb.01173. Trained models of the four species, which can be used for prediction or transfer learning tasks, are provided in the MuRaL package (https://github.com/CaiLiLab/MuRaL).

# Code availability

The Python package implementing the MuRaL framework is available at: https://github.com/CaiLiLab/MuRaL.

## Acknowledgements

We thank Xionglei He, Xia Shen and Nicholas Luscombe for giving insightful comments on the manuscript. We thank all lab members for discussion and help throughout this project. This work was supported by National Natural Science Foundation of China (32070593), Science and Technology Planning Project of Guangzhou (202102020816).

## Author contributions

C.L. designed and supervised the project. C.L. developed the MuRaL framework, with input from Y.F. and S.D. for detailed evaluation. Y.F. and S.D. performed comparative analyses and generated mutation rate profiles. C.L, Y.F. and S.D. wrote the manuscript.

## Competing interests

All authors declare no competing interests.