*NAR Genomics and Bioinformatics*, 2024, **6**, lqae108
https://doi.org/10.1093/nargab/lqae108
Advance access publication date: 20 August 2024
**Standard Article**

# Using recurrent neural networks to detect supernumerary chromosomes in fungal strains causing blast diseases

**Nikesh Gyawali[1], Yangfan Hao[2], Guifang Lin[2], Jun Huang[2], Ravi Bika[2], Lidia Calderon Daza[2], Huakun Zheng[2], Giovana Cruppe[2], Doina Caragea[1], David Cook[2], Barbara Valent[2] and Sanzhen Liu[2,*]**

[1]Department of Computer Science, Kansas State University, Manhattan, KS 66506, USA
[2]Department of Plant Pathology, Kansas State University, Manhattan, KS 66506, USA

[*]To whom correspondence should be addressed. Tel: +1 785 532 1379; Email: liu3zhen@ksu.edu
Present addresses:
Guifang Lin, Basic Forestry and Plant Proteomics Research Center, Fujian Agriculture and Forestry University, Fuzhou, China.
Jun Huang, Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, NC 27710, USA.
Huakun Zheng, College of Life Science, Fujian Agriculture and Forestry University, Fuzhou, Fujian 350002, China.

## Abstract

The genomes of the fungus *Magnaporthe oryzae* that causes blast diseases on diverse grass species, including major crops, have indispensable core-chromosomes and may contain supernumerary chromosomes, also known as mini-chromosomes. These mini-chromosomes are speculated to provide effector gene mobility, and may transfer between strains. To understand the biology of mini-chromosomes, it is valuable to be able to detect whether a *M. oryzae* strain possesses a mini-chromosome. Here, we applied recurrent neural network models for classifying DNA sequences as arising from core- or mini-chromosomes. The models were trained with sequences from available core- and mini-chromosome assemblies, and then used to predict the presence of mini-chromosomes in a global collection of *M. oryzae* isolates using short-read DNA sequences. The model predicted that mini-chromosomes were prevalent in *M. oryzae* isolates. Interestingly, at least one mini-chromosome was present in all recent wheat isolates, but no mini-chromosomes were found in early isolates collected before 1991, indicating a preferential selection for strains carrying mini-chromosomes in recent years. The model was also used to identify assembled contigs derived from mini-chromosomes. In summary, our study has developed a reliable method for categorizing DNA sequences and showcases an application of recurrent neural networks in predictive genomics.

## Introduction

The fungus *Magnaporthe oryzae*, also known as *Pyricularia oryzae*, is responsible for causing blast diseases on diverse grass species, including major crops (1–3). Rice blast disease, caused by *M. oryzae Oryza* (MoO) pathotype, poses a significant threat to rice production (4). Wheat blast disease, caused by a distinct pathotype, *M. oryzae Triticum* (MoT), emerged in Brazil in 1985 and spread within South America and recently to South Asia and Africa (5–9). Additionally, blast diseases caused by the *Setaria* pathotype (*M. oryzae Setaria*, MoS) on foxtail millet and by the *Eleusine* pathotype (*M. oryzae Eleusine*, MoE) on finger millet are significant diseases of these ancient subsistence crops (10,11). Since the early 1990′s, the *Lolium* pathotype (*M. oryzae Lolium*, MoL) has emerged in the US to cause serious blast diseases on popular turf grass or forage crops, including perennial ryegrass, annual ryegrass, and tall fescue (12). There are other *M. oryzae* strains from other hosts, such as oats (*Avena*), buffelgrass (*Cenchrus*), crabgrass (*Digitaria*) and signalgrass (*Urochloa*) (3,12).

The genome of *M. oryzae* contains seven essential core-chromosomes and many genomes possess one or a few extra, non-essential supernumerary chromosomes (13,14). Besides *M. oryzae*, many plants, animals, and other fungi carry supernumerary chromosomes, which are also known as extra chromosomes, dispensable chromosomes, accessory chromosomes, or B-chromosomes. Supernumerary chromosomes are hypothesized to be an accelerator for fungal adaptive evolution (15). Supernumerary chromosomes in *M. oryzae* are referred to as mini-chromosomes because their sizes are typically smaller than core-chromosomes (14,16,17). As compared to core-chromosomes, mini-chromosomes in *M. oryzae* are more repetitive, containing more transposable elements and fewer genes. The repeat-rich characteristic provides ample intrachromosomal homology for DNA duplication, loss, and rearrangements, creating conducive environments to accelerate genome evolution (14,18). Indeed, mini-chromosomes are highly variable among *M. oryzae* strains (8,14,17). Mini-chromosomes carry effector genes that can be found in core-chromosomes in different strains, suggesting crosstalk between mini- and core-chromosomes (14,17,19). Therefore, mini-chromosomes are thought to be capable of mediating the mobility of effector genes, facilitating fungal adaptation.

To confirm and further understand the evolutionary role of mini-chromosomes, it is critical to be able to determine if a particular *M. oryzae* strain carries a mini-chromosome. Contour-clamped homogeneous electric field (CHEF) electrophoresis of intact chromosomes is the means to provide conclusive evidence for the presence or absence of chromo-

**Received:** October 10, 2023. **Revised:** June 27, 2024. **Editorial Decision:** August 1, 2024. **Accepted:** August 6, 2024

somes with sizes smaller than core-chromosomes (16,20). However, the technique is laborious and requires specific equipment. This is a significant hurdle as research labs may not have access to all strains published in the literature, especially given that MoT wheat blast causing strains are quarantined in the United States. A reliable method to determine the presence of a mini-chromosome from publicly available or newly generated sequencing data would be fast, cost-effective, and decentralized. A simple strategy is to align sequencing reads to known mini-chromosome genomes for the determination of the proportion of mini-chromosome genomes supported by reads. The lack of knowledge about critical elements required for mini-chromosomes, the high-level of variability among mini-chromosomes, and the potential exchanges between core- and mini-chromosomes complicate the analysis. Fortunately, multiple complete core-chromosome and mini-chromosome genomes are currently available, providing the opportunity to deploy deep learning algorithms to learn features of core- and mini-chromosome sequences for prediction.

Use of neural network based deep learning techniques has rapidly increased due to availability of large data, and their ability to find complex patterns. Recurrent Neural Network (RNN) and, specifically, Long Short-Term Memory (LSTM) networks have been used in genomics to utilize the sequential property of DNA sequences for making various predictions (21–23). An RNN represents a group of artificial neural networks that incorporate feedback connections to retain and utilize information from prior input events as activation. These networks leverage their internal state to process input sequences of varying lengths. However, training RNNs to effectively capture long-term dependencies poses challenges as the error signals flowing backward often suffer from issues of either explosive amplification or rapid attenuation, a.k.a. exploding or vanishing gradients (24,25). To address this problem, the LSTM architecture was introduced as an extension to the vanilla RNN, a simple form of RNN (24). Another enhancement is the Bidirectional LSTM (Bi-LSTM), which considers sequential context from both directions and can improve performance (26,27). In our study, we apply Bi-LSTM deep learning to predict the presence of mini-chromosome sequences based on the genomic sequence data. Experimental results show that a Bi-LSTM neural network model can accurately infer the presence of mini-chromosomes in strains of *M. oryzae*.

## Methods

### Near-finished genome assemblies for model training

Near-finished genome assemblies of isolates that are known to carry or not to carry mini-chromosomes were collected for training Bi-LSTM models, which include the assemblies from mini-chromosome-bearing isolates B71 (MoT, Genbank accession: GCA_004785725.2) (8,14), LpKY97 (MoL, Genbank accession: GCA_012272995.1) (28), TF05-1 (MoL, Genbank accession: JAVBIT010000000), and O135 (MoO, https://github.com/PlantG3/miniC/tree/main/data/training_data) (29), as well as assemblies from mini-chromosome-free isolates 70-15 (MoO, Genbank accession: GCA_000002495.2) (13) and MZ5-1-6 (MoE, Genbank accession: GCA_004346965.1) (30).

### Contour-clamped homogeneous electric field (CHEF) electrophoresis of TF05-1

TF05-1 protoplasts were prepared with the procedure slightly modified from the approach used in Orbach *et al.* (16). Briefly, harvested mycelia were washed with 1 M sorbitol and digested with 10 mg/ml Lysing Enzymes from *Trichoderma harzianum* (Sigma Aldrich, CAT#L1412) in 1 M sorbitol at 28°C, 90 rpm for 2.5 h. The digested product was filtered through sterile Nytex nylon mesh l and centrifuged at 4500 rpm at 4°C for 10 min to collect protoplasts. Protoplasts were washed with SE buffer (1 M sorbitol, 50 mM EDTA) and adjusted to $1 \times 10^9$ cells/ml. The CHEF Genomic DNA Plug Kit (Bio-Rad, CAT#1703591) was used for the preparation of protoplast plugs. Protoplasts were mixed with a 2% low melting agarose gel and transferred to modules to form protoplast plugs. After incubating in the proteinase K buffer overnight at 50°C, plugs were washed four times with 1 × wash buffer at 25°C, and then stored in 0.5× TBE at 4°C. A CHEF Mapper XA System (Bio-Rad, CAT#1703671) was used for CHEF gel electrophoresis using 0.7% Certified Megabase Agarose in 0.5 × TBE buffer. The electrophoresis was run at 1.5 V/cm and 6°C, with switch times ranging from 1200 to 4800 seconds for 120 h.

### Identification of common sequences between core- and mini-chromosomes

Alignment was performed between core- and mini-chromosomes for each of B71, O135, LpKY97 and TF05-1 with NUCmer (31). Alignments with at least 105 bp matches and 95% identity were retained. Alignment regions were merged if neighboring alignments were within a 100 bp distance and sequences were extracted from both core- and mini-chromosomes. All common sequences identified from these four genomes and the mitochondrial sequence of B71 were combined to form a database of sequences excluded from the training.

### Bi-LSTM models

The Bi-LSTM model was implemented using Python with TensorFlow and Keras libraries. The architecture consisted of an input layer, hidden LSTM cells layers, and an output layer. The input layer consists of an embedding layer that encodes the input tokens (such as the 11 9-mers tokens of the 99 bp sequence) into vectors of size 128. The hidden layers consisted of two Bidirectional LSTM layers, each with 256 hidden units, stacked on top of each other with the hyperbolic tangent (tanh) activation function. The output of the last Bi-LSTM hidden layer was connected to a dense output layer with sigmoid activation function. A model for 99b bp sequence with eleven 9-mer tokens contained a total of 34 212 225 trainable parameters. The selection of the model architecture and hyperparameters was informed by experimenting with a range of values and selecting those that resulted in the best performance on the validation set.

For training the Bi-LSTM model, backpropagation through time (BPTT) was employed, using binary cross entropy loss as the loss function. The optimization was performed using Adam optimizer with learning rate of 0.001. The training and validation data sets, such as sequences with 99 bp with eleven 9-mers and labeled with either ‘core’ or ‘mini’, were encoded using one-hot encoding and used to train and evaluate the model. The dataset was split into the train, validation, and

test sets with 80/10/10 splits, respectively. To optimize model performance, a large training dataset is required while also ensuring that the validation and test set closely resemble the overall data distribution. We achieve this by utilizing a small percentage of data as the validation and test set when dealing with a large dataset like ours. A mini-batch size of 2048 was used for the training and evaluation at each epoch. To optimize the training and prevent overfitting, an early stopping criterion based on validation loss was implemented. The model was trained for a maximum of 150 epochs with patience of 15 epochs. If the validation loss did not improve over the subsequent 15 consecutive epochs, the training was stopped. The model weights corresponding to the lowest validation loss were restored, representing the best-performing model. Subsequently, the final trained model was then tested on an independent test dataset to evaluate its overall performance.

### Draft genome assemblies of isolates with and without mini-chromosomes

Genome assembly drafts were downloaded from Genbank: GCA_900474655.3 of FR13 (MoO from rice), GCA_900474175.3 of US71 (MoS from foxtail millet), GCA_900474475.3 of CD156 (MoE from goosegrass) and GCA_900474545.3 of BR32 (MoT from wheat). Field isolates FR13, US71 and CD156 all carried mini-chromosomes and BR32 did not contain a mini-chromosome (17).

### Illumina WGS short-reads of 252 *M. oryzae* isolates

WGS reads were downloaded from Sequence Read Archive (SRA). Data of 252 accessions were collected (Supplementary Table S1). Reads were trimmed with software Trimmomatic prior to further analyses (32).

### Subsampled reads for determining miniC proportions

Random seeds were set for sampling reads from the forward reads of the original paired-end Illumina reads of the isolates of P3, B71, T25 and Guy11. Subsampling was implemented using seqtk (version 1.2). Subsampled reads were then used for the prediction with the optimized Bi-LSTM model.

### Indexes of similarity to the B71 mini-chromosome

WGS reads of each strain were used to compare with WGS reads of B71 to infer the genomic regions of the B71 mini-chromosome that were absent in the isolate through Comparative Genomics Read Depth (CGRD) (14,33). Each CGRD analysis may identify B71 mini-chromosome regions that were absent in the analyzed strain. The proportion of the B71 mini-chromosome that was not detected as absence regions represents the portion of sequences similar to the B71 mini-chromosome, referred to as the index of similarity to the B71 mini-chromosome of the strain. A low value of the index indicates the absence of a mini-chromosome.

### Genome sequencing and assembly of an early MoT strain T3

The MoT strain T3 was cultured on oatmeal agar (OMA) plates followed by liquid culture under Biosafety Level 3 (BSL3) laboratory in the Biosecurity Research Institute (BRI) at Kansas State University in Manhattan, KS (8,34). The detailed procedure for genomic DNA extraction was previously described (8). Briefly, mycelial mats were collected, lyophilized, and ground for DNA extraction with a CTAB approach. DNA was stored in the TE buffer containing 1 mg/ml RNase. Approximately 50× paired-end (2 × 150 bp) Illumina data were produced at Novogene USA. Nanopore long reads were generated using the same genomic DNAs per the procedure described previously (8). The genomic DNA was subjected to a size selection (>20 kb) using a BluePippin Gel Cassette (Sage Science, USA, Cat.# BLF7510), followed by a library construction using the SQK-LSK110 kit and sequencing using a R9.4.1 flow cell on a MinION Mk1B device (Oxford Nanopore, UK). Nanopore raw FAST5 data were converted to FASTQ reads using the Guppy Basecaller (version 6.3.2). Reads were assembled with Canu (version 2.2) with the parameters of 'genomeSize = 45 m minReadLength = 10 000 minOverlapLength = 1000 correctedErrorRate = 0.08 rawErrorRate = 0.3 corOutCoverage = 60' (35). The contigs in Canu assemblies were aligned to B71Ref2 to determine the chromosome number and the orientation using NUCmer with the parameters of '-L10000 -I 90' (31). The resulting assembly was polished using Nanopolish (version 0.14.0) and then using Pilon (version 1.24) with Illumina reads (36,37).

## Results

### Training data to assign DNA sequences to core- or mini-chromosomes

The goal of the study was to predict the presence of mini-chromosomes using genomic sequencing data. Although genomic data of hundreds of *M. oryzae* strains are publicly available, there is very little data regarding if an individual strain possesses mini-chromosome(s). We addressed this by building a Bi-LSTM model to classify DNA sequences as originating from core- or mini-chromosomes (Figure 1). The output of the model is used to infer the presence of mini-chromosomes in a strain based on the proportion of mini-chromosome-derived sequences among the total short DNA sequences examined. We collected finished genome assemblies of *M. oryzae* strains with or without mini-chromosomes for model training, from which short sequences were extracted. The strains harboring at least one mini-chromosome include B71 (MoT) (14), LpKY97 (MoL) (38), TF05-1 (MoL) (Supplementary Figure S1) and O135 (MoO) (16), while the strains containing no mini-chromosomes include the MoO reference strain 70-15 (13) and MZ5-1-6 (MoE) (30). Approximately 11.2 and 252.2 Mb from mini- and core-chromosomes were collected for model training (Table 1). Note that the presence of at least one mini-chromosome in B71, TF05-1 and O135 was verified by CHEF (14,16,29). The collected six mini-chromosomes and 42 core-chromosomes were fragmented into short DNA sequences and labeled with either mini or core as the sequence source for model training.

### Training of Bi-LSTM models

Short sequences (e.g. 99 bp) extracted from the six mini-chromosomes and 42 core-chromosomes were termed subsequences (Figure 1A). Each subsequence was then tokenized into non-overlapping k-mers (e.g. 9-mer). Afterwards, the tokenized data were split into train, validation, and test sets with an 80/10/10 split. Models were trained on the train set and evaluated for training performance and hyperparam-

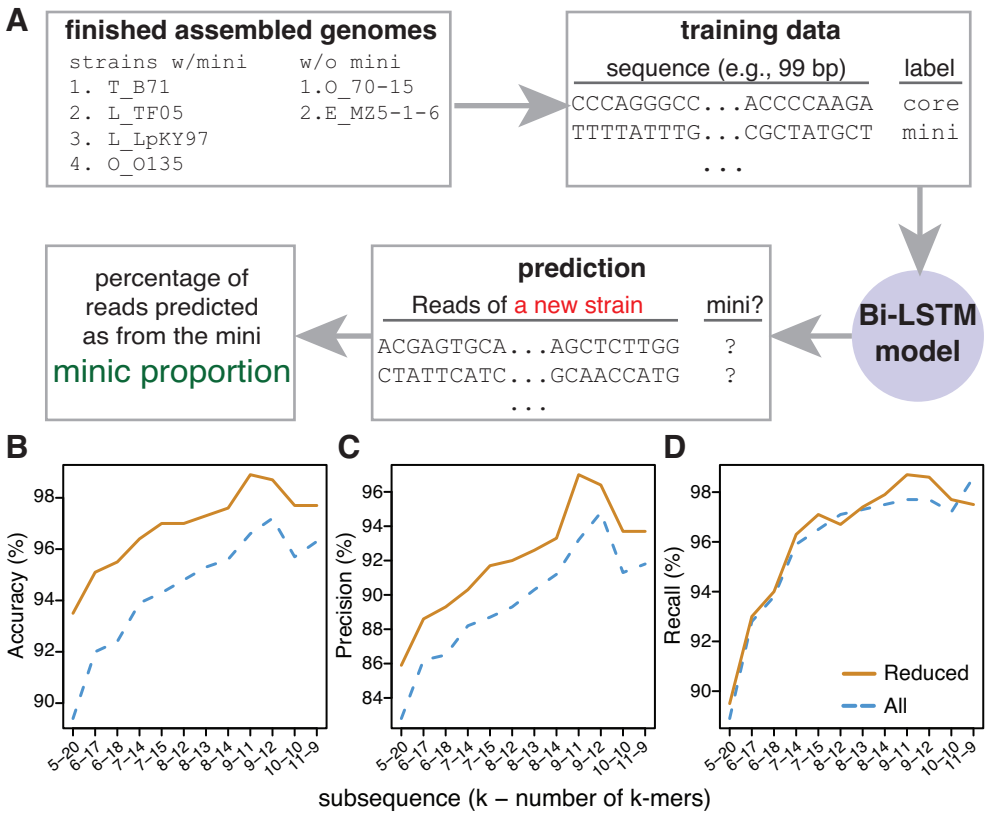


**Figure 1.** Overview of predicting sequences from mini-chromosomes. (**A**) Finished assembled genomes of strains with or without mini-chromosomes were used to generate training data. The training data include DNA fragments with the length around 100 bp and labeled with the origins from either core- or mini-chromosomes. The deep learning model was trained and the optimal model was selected for predicting the origin of each sequencing read from a new strain. The miniC proportion, which is the percentage of reads predicted to originate from mini-chromosomes, is the value for inference of the presence of a mini-chromosome in the strain. (**B–D**) Average performance metrics for models trained using different subsequences, each of which consists of multiple k-mers. Each X-axis label specifies the size of k and the number of k-mers (e.g. 5–20 stands for a 100 bp subsequence with 20 5-mers). Performance was evaluated for all genomic data (All, blue lines) or after removing common sequences shared between core- and mini-chromosomes (reduced, orange dashed lines).

**Table 1.** Summary of genome assembly data for model training

| Strain | Pathotype | Mini (bp) | # mini | Core (bp) | # core |
|---|---|---|---|---|---|
| 70–15 | MoO | 0 | 0 | 41 027 733 | 7 |
| MZ5-1-6 | MoE | 0 | 0 | 42 703 282 | 7 |
| O135 | MoO | 1 747 687 | 1 | 41 933 874 | 7 |
| B71 | MoT | 1 903 245 | 1 | 42 908 164 | 7 |
| LpKY97 | MoL | 3 910 017 | 2 | 41 702 954 | 7 |
| TF05-1 | MoL | 3 598 139 | 2 | 41 915 541 | 7 |
| **Total** | - | **11 159 088** | **6** | **252 191 548** | **42** |

eters selection using the validation set. The selected models were finally evaluated using the test set. When constructing the training dataset, we encountered imbalanced training sequence data from core- and mini-chromosomes in which the total length of core-chromosomes was markedly larger than the total length of mini-chromosomes (Table 1). To create balanced training data from core- and mini-chromosomes, we extracted subsequences with the step size of 1 bp from mini-chromosomes and the step size of 27 bp from core-chromosomes.

Models were trained with DNA sequence data of different k-mer sizes, ranging from 5 to 11 mers, and subsequence lengths. Lengths of subsequences were limited to around 100 bp because lengths of whole genome sequencing (WGS) data, or reads, of most *M. oryzae* strains to be used for the prediction are around 100–150 bp. Overall, the evaluation on the validation data showed that models trained with the 9-mer attained the highest scores in both accuracy and precision, and the recall score was close to the highest score achieved by using 11-mer (Figure 1B–D, Supplementary Table S2). The assessment with the models on the test data set showed the consistent evaluation result (Supplementary Table S3). We previously showed common sequences, particularly transposable elements, occurred in core- and mini-chromosomes (14), which created ambiguous sequence examples that did not have a clear class distinction. To examine if the occurrence of these common sequences impacted the model, we re-trained models using genomic data where common sequences were identified per strain and removed from the training data. Model performance on the validation data was improved when removing these sequences (Figure 1B–D), and the model trained using 9-mers became the best for accuracy, precision and recall (Supplementary Table S4). Within the 9-mer model, the two subsequence lengths of 99 bp and 108 bp did vary for model performance, and the model trained using the 99 bp subsequences (nine 9-mers) attained better scores: 98.9% accuracy, 97.0% precision and 98.7% recall on both the validation and test datasets (Supplementary Figure S2,

Supplementary Tables S4, S5). This model was used for subsequent analysis.

## Survey of presences of mini-chromosomes in cereal blast strains

The optimized Bi-LSTM model was used to examine the presence of mini-chromosomes in *M. oryzae* isolates whose WGS data were available. The probability of mini-chromosome origin for each WGS read was estimated, and reads with the prediction probability larger than 0.99 were classified as mini-chromosome reads (Supplementary Figure S3). The proportion of mini-chromosome reads among all examined reads, referred to as the miniC proportion, was determined for each *M. oryzae* isolate. In total, WGS data of 252 *M. oryzae* isolates from multiple pathotypes were analyzed, resulting in miniC proportions ranging from 0.7% to 9.3% (Figure 2A, Supplementary Table S6). Three isolates, B71 (MoT), P3 (MoT) and LpKY97 (MoL), carrying at least one mini-chromosome had miniC proportions of 3.5%, 5.8% and 5.6%, respectively. Note that the P3 and LpKY97 genomes each contained two mini-chromosomes based on the previous reports (14,28). In contrast, the miniC proportions of four isolates with no mini-chromosomes, 70–15, Guy11 (MoO), MZ5-1-6 (MoE) and T25 (MoT), were 0.8%, 0.9%, 1.1% and 0.9%, respectively. Based on miniC proportions of these isolates, we used 1.5% as the miniC proportion threshold to classify isolates as with or without mini-chromosomes.

The Comparative Genomics Read Depth (CGRD) pipeline was employed to identify the genomic regions of the B71 mini-chromosome that were absent in each isolate (14,33). The proportion of the B71 mini-chromosome that was not detected as absence regions represents the similarity of the potential mini-chromosome of an isolate to the B71 mini-chromosome, which was referred to as the index of similarity to the B71 mini-chromosome. Index values of 252 strains ranged from 0.03 to 1. A higher index of similarity indicates a higher possibility that an isolate carries mini-chromosome(s). Based on the index values of isolates known to carry at least a mini-chromosome or none (Supplementary Table S6), the index threshold of 0.2 was used to classify isolates with or without a mini-chromosome.

Comparison between the prediction result from the Bi-LSTM model with the result using the CGRD approach showed that the two methods were highly consistent. Specifically, the prediction of mini-chromosome presence in 98.4% (248/252) isolates were the same. In total, 223 were predicted to contain mini-chromosome(s) using both approaches, indicative of a substantial presence of mini-chromosomes across *M. oryzae* strains. The results also indicated that different pathotypes had varying levels in mini-chromosome prevalence (Figure 2B). More than 90% of both 196 MoO and 25 MoT isolates were predicted to carry mini-chromosomes. All isolates collected from *Avena* spp., *Cenchrus* spp., *Lolium* spp. and *Urochloa* species, and half of isolates from *Digitaria* spp., and *Setaria* spp., were predicted to contain mini-chromosomes. Mini-chromosomes were the least prevalent in isolates from *Eleusine* spp., of which only 29% (2/7) were predicted to contain mini-chromosomes. Note that the number of isolates of each of the pathotypes other than MoO and MoT is relatively small, ranging from 2 to 7. Two MoO isolates, namely IR0095 and JP0091, proved difficult to predict and produced different predictions from the two prediction approaches. The miniC proportions of the two isolates were 1.1%, while the indexes of similarity to the B71 mini-chromosome were 0.211 (IR0095) and 0.278 (JP0091). Both predictions of the two strains were close to the respective thresholds.

Rice isolates (MoO) were classified to four clades (29). The classified strains with whole genome sequencing reads longer than 100 bp were subjected to the miniC analysis. The prediction showed that 75% (9/12) isolates from clade I contain no mini-chromosomes and all isolates ($N = 55$) from clades II, III, IV contain mini-chromosomes with one exception in clade II (Supplementary Table S7).

## Prediction using subsets of reads

To determine the minimal amount of sequencing reads required for reliable prediction, four isolates with known numbers of mini-chromosomes were selected for a simulation. These four isolates included P3 with two mini-chromosomes, B71 with one mini-chromosome, and two mini-chromosome-free isolates: T25 and Guy11 (14,29). Random reads, from 1000 to 300 000, were subsampled from the forward read sets of the original paired-end WGS reads, with subsampling repeated five times per isolate. As expected, the variation of miniC proportions was higher when a low amount of reads were used for the prediction (Figure 3). The simulation from all the four isolates consistently showed that the prediction of the miniC proportion was not very reliable when the number of reads was <20 000. However, even when using such low numbers of reads, the predicted proportion values of mini-chromosomes did not deviate dramatically from the prediction value obtained using the original full read set and none of them caused a misclassification. When 50 000 or more reads were used, the predicted miniC proportions were reliably close to that using the original read set, which included millions of reads. The coefficients of variation, the ratios of standard deviation to the mean of predicted miniC values, from five independent simulations at sampling sizes of 50 000 and above were not higher than 0.082. Based on the simulation result, 100 000 and more reads are conservatively recommended for an accurate prediction of miniC proportions using our Bi-LSTM model.

## Applications to identify mini-chromosome-associated sequences

In addition to predicting if a strain contains mini-chromosomes, we applied the Bi-LSTM model to predict if a DNA sequence from an assembly (termed contig hereafter) represented a mini-chromosome. We split each contig into continuous 99 subsequences and classified each to either core- or mini-chromosome based on the model prediction. The proportion of mini-chromosome subsequence of a contig, referred to the miniC proportion of a contig, indicates the extent to which the contig shares similarity to mini-chromosomes. To test the prediction strategy, the genome assembly of B71, including seven core-chromosomes and one mini-chromosome, was subjected to the analysis. As a result, the miniC proportion of the B71 mini-chromosome was 54.8%, which was markedly higher than miniC proportions of core-chromosomes ranging from 0.1% to 1.7%, (Figure 4A, Supplementary Table S8). A previous study produced draft genome assemblies for MoO FR13 (Figure 4B), MoS US71 (Figure 4C), MoE CD156 (Figure 4D), and demon-

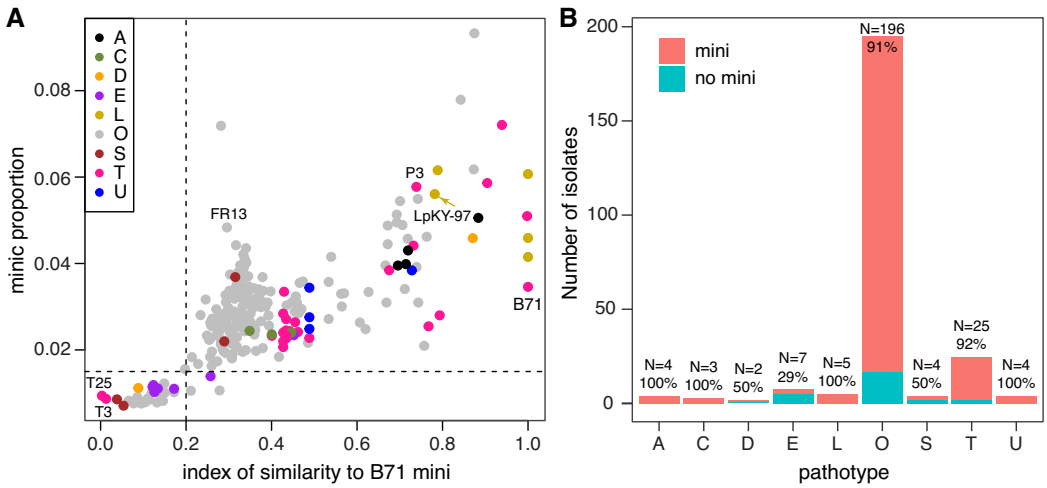


**Figure 2.** Mini-chromosome prediction of isolates from diverse pathotypes. (**A**) The miniC proportion of each isolate was estimated using whole genome sequencing reads. The index of similarity to the B71 mini-chromosome represents the proportion of the B71 mini-chromosome that was not detected as deletion genomic regions using the CGRD pipeline. Dash lines signify the miniC proportion threshold of 1.5% and the similarity index threshold of 0.2 used to determine if an isolate carries a mini-chromosome. Letters stand for host species on which the strains were isolated in the field (e.g. A = *Avena*; C = *Cenchrus*; D = *Digitaria*; E = *Eleusine*; L = *Lolium*; O = *Oryza*; S = *Setaria*; T = *Triticum* and U = *Urochloa*). (**B**) Distribution of the number of isolates with and without mini-chromosomes in each pathotype. Total numbers of isolates and percentages of isolates with mini-chromosomes are labeled on top of bars.

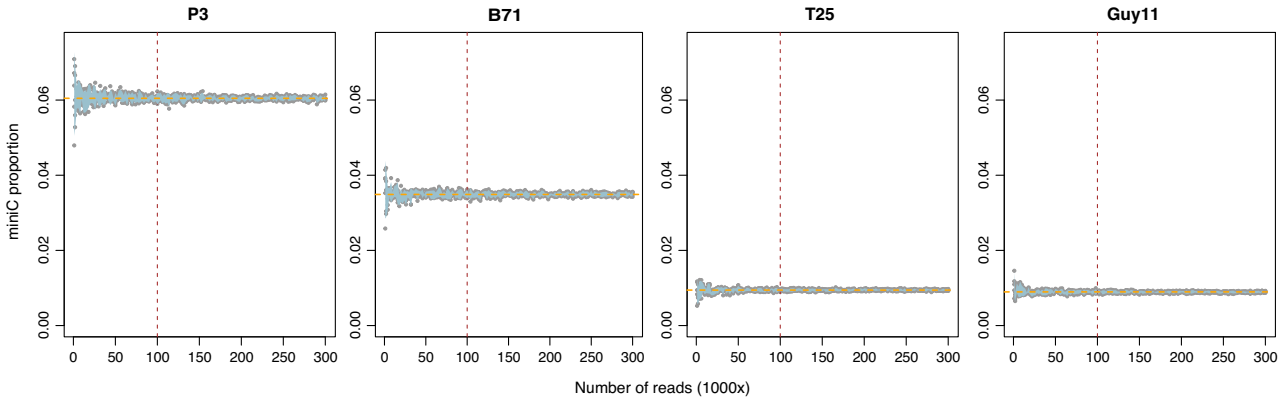


**Figure 3.** Prediction of miniC proportions using subsampled reads. MiniC proportions (Y-axis) from five times of simulations for each of four isolates (P3, B71, T25 and Guy11) were plotted versus numbers of reads used (X-axis). Each gray dot represents a predicted miniC proportion using a certain number of reads randomly extracted from an original read set of the corresponding isolate. The light blue shades were plotted using 95% confidence intervals from five simulations. Orange horizontal dash lines indicate the predicted miniC proportion values using the original full reads. Brown vertical dash lines point the recommended minimum read number for prediction.

strated that all three contained mini-chromosomes by CHEF analysis (17). MiniC proportions of contigs from each drafted assembly were determined and used to infer if a contig was derived from a mini-chromosome. Eight contigs larger than 100 kb previously found to be mini-chromosome derived were supported by the miniC proportion data, which identified an additional five contigs (Supplementary Table S8). The five contigs appeared to possess sequence features related to mini-chromosomes. In the same study, MoT BR32 was found to contain no mini-chromosomes. Consistently, the miniC proportions of all contigs are small, ranging from 0.5% to 3.1% (Figure 4E, Supplementary Table S8). Furthermore, we assembled a new MoT genome from the early isolate T3 (1986) into seven chromosomes, indicative of no mini-chromosomes. All these seven chromosomes had small miniC proportions (0.5–1.4%) and can be assigned to core-chromosomes (Figure 4F, Supplementary Table S8). Collectively, the Bi-LSTM model we constructed can be used to differentiate contigs belonging to core- or mini-chromosomes.

To scan along individual chromosomes, the calculated miniC proportions were determined for 30 kb intervals of each chromosome of B71 and T3, which carried one and zero mini-chromosomes, respectively. Almost all intervals of the B71 mini-chromosome had a miniC proportion larger than 10%. Many intervals on the ends of core-chromosomes showed a relatively high miniC proportion, indicating that they possess sequence features associated with mini-chromosomes. Notably, a region at the end of B71 chromosome 3 contained sequences with a miniC proportion level similar to a mini-chromosome. This region is absent in the genome of T3, which did not contain mini-chromosomes (Figure 4G and H). The region represents a potential translocation event from a mini-chromosome to a core-chromosome.

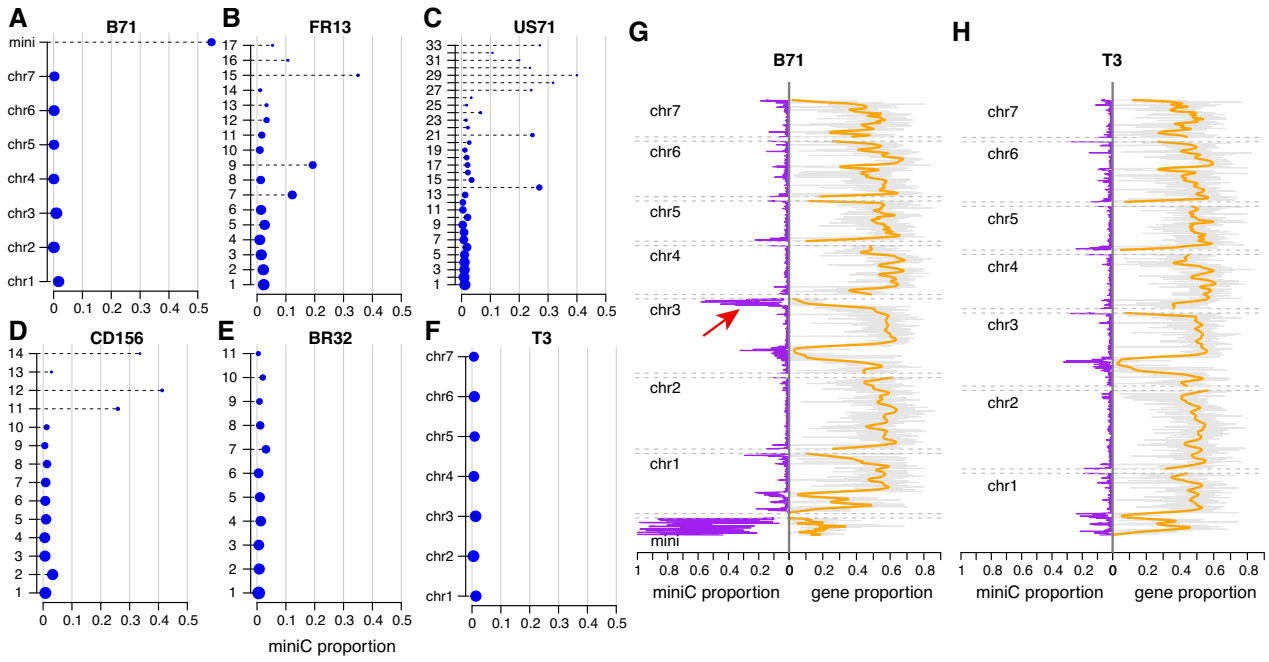


**Figure 4.** Prediction of miniC proportions in assembled contigs or chromosomes. (**A**–**F**) Assembled contigs or chromosomes of six strains, including five strains that are not included in the training data, were subjected to the prediction. The miniC proportion represents the proportion of sequences in each contig/chromosome predicted as miniC sequences. Y-axes signify names of contigs/chromosomes, which are listed on the column of 'contig_id' in Supplementary Table S8. The same rule is applicable for other contig names. Sizes of blue dots indicate the bp length of contigs/chromosomes. (**G, H**) MiniC proportions per 30-kb interval (purple) and proportions of genic sequences per interval along each chromosome (gray). Orange curves represent LOWESS estimates of genic proportions. The red arrow indicates a core-chromosome region with a high miniC proportion value. B71 genome version: B71Ref2; T3 genome version: T3v1.

## Discussion

In this study, we employed a Recurrent Neural Network (RNN) deep learning technique, specifically a Bi-directional Long Short-Term Memory (Bi-LSTM) network, to model the origin of DNA sequences as belonging to core- or mini-chromosomes. The optimized Bi-LSTM model enables examination of the core- or mini-chromosome origin using the input data of WGS reads, assembled contigs or chromosomes, and DNA sequence fragments. The model was trained using multiple genomes with or without mini-chromosomes, learning genomic features from divergent core- and mini-chromosomes. The core- and mini-chromosomes used in training were from different host-adapted pathotypes (MoO, MoT, MoL and MoE) and with differing composition. The prediction result from the Bi-LSTM model was similar to the result from CGRD that was an alignment-based approach and used one reference genome, which indicated that mini-chromosomes from multiple *M. oryzae* pathotypes share certain learnable genomic features. In contrast to CGRD that requires high-depth genome sequencing data, the Bi-LSTM prediction is accurate and reliable even using a very small amount of WGS read data. Also, the Bi-LSTM model is able to analyze both non-repetitive and repetitive sequences, overcoming a common problem of repetitive sequences limiting alignment-based analysis, and thereby allowing regional scanning along chromosomes. Crosstalk between core- and mini-chromosomes in *M. oryzae* was previously hypothesized (14). Our efforts to scan assembled chromosomes identified the end of chromosome 3 in the B71 as being highly similar to mini-chromosomes. Given that this region is absent in the B71 related strain, T3, this region may represent a genome structural variation arising from a translocation event from a mini-chromosome. Future analysis of more high-quality reference level assemblies of more diverse *M. oryzae* strains will further illuminate potential core genome variation influenced by mini-chromosomes.

Analysis of 252 *M. oryzae* isolates reveals the prevalence of mini-chromosomes in at least some field isolates of all *M. oryzae* host-adapted pathotypes that we investigated. Specifically, 91% of 196 rice isolates were predicted to carry mini-chromosomes. The result is consistent with a previous examination of mini-chromosomes conducted using electrophoretic karyotyping, which found 93% of 14 rice isolates harbored mini-chromosomes (16). In the same study, none of seven wheat isolates carried mini-chromosomes. However, our analysis showed that 92% of wheat strains carried mini-chromosomes. The discrepancy appears to be related to the isolation period for these wheat isolates relative to the first report of wheat blast disease in 1985 in Brazil (39). Recent data indicates that both the *Triticum* and *Lolium* pathotypes evolved through two distinct episodes of sexual crosses involving individuals from five different host-adapted pathotypes, including the *Eleusine* pathotype (12,40). After emergence of populations adapted to *Triticum* and to *Lolium* spp., asexual reproduction systems apparently predominated during infections in the field, perhaps allowing mini-chromosome accumulation (41). All isolates (T1 to T7) examined in Orbach et al. (1996) were early wheat strains collected in 1988 or earlier. From our analyses, none of the three early wheat strains (T3, T25 and BR32 collected in 1986, 1988 and 1991, respectively) carried mini-chromosomes. In contrast, all of our wheat isolates collected after 2005 carried mini-chromosomes. The result indicated that wheat strains with mini-chromosome(s)

were preferentially selected in the field since the 1990s. All the *Lolium* isolates examined were collected since the 1990s and carried mini-chromosomes.

Among all host-adapted pathotypes analyzed, a high proportion of strains of the *Eleusine* (MoE) pathotype lacked mini-chromosomes. Combined with two different MoE strains analyzed in the Orbach *et al*. study, 78% of MoE (7/9) isolates contained no mini-chromosomes. MoE strains were classified to the *Eleusine1* and *Eleusine2* lineages previously (3). Our prediction data and previous analyses indicate that an *Eleusine1* strain EI9411 and an *Eleusine2* strain CD156 carried mini-chromosomes (17). Although MoE isolates frequently contained no mini-chromosomes, both lineages could carry mini-chromosomes.

Our data showing a relatively low proportion of MoE strains with mini-chromosomes supports a previous report of an inverse correlation between high levels of sexual fertility and low occurrence of mini-chromosomes (16). For the ascomycetous *M. oryzae*, fully fertile strains are hermaphrodites that are able to serve as a female partner and produce perithecia in sexual crosses with strains of opposite mating type and also serve as male partners in crosses with other hermaphroditic strains. Orbach *et al*. (1996) reported that 18 fertile hermaphroditic strains, including MoE field isolates and derived fertile laboratory strains, uniformly lacked mini-chromosomes. In contrast, mini-chromosomes occur frequently in lower fertility strains, such as most rice pathogens, which either lack any mating ability or cross only as male partners with other hermaphroditic strains (16,20). Independent studies including analysis of complete tetrads showed that mini-chromosomes fail to segregate normally in sexual crosses, typically resulting in fewer ascospore progeny with mini-chromosomes than expected (16,42). Our results support a correlation between lack of mini-chromosomes and full female fertility since MoE strains, in general, are known to possess high levels of female fertility (1,16,43). In addition, the inverse association between sexual fertility and the presence of mini-chromosomes is supported by our MoO data. Our mini-chromosome prediction showed that 75% (9/12) of isolates from clade I are devoid of mini-chromosomes and a mere 2% (1/55) of isolates from clades II, III, and IV lack mini-chromosomes. This aligns consistently with observed reproductive characteristics because clade I includes strains that are fully fertile hermaphrodites (e.g. strain Guy11), and the predominantly asexual clades II, III and IV include infertile strains and strains that only cross as males (29,44,45). Further studies are needed to confirm any correlation and determine the precise role of mini-chromosomes in sexual fertility. Our mini-chromosome prediction model provides a new tool for addressing the question and tracking mini-chromosome presence in evolving populations of the blast fungus.

Our prediction model can be further improved by training using additional core- and mini-chromosome sequencing data for predicting mini-chromosomes from broader *M. oryzae* isolates. Of the four mini-chromosome-bearing isolates used for model training, three strains were from either wheat or *Lolium* hosts. The wheat and *Lolium* strains are genetically close, the samples therefore might be biased in favor of the mini-chromosome of B71, a wheat strain. More mini-chromosome sequencing data in the future model development will allow capturing the high-level diversity among mini-chromosomes, and thereby improving this approach. In addition to the prediction of the presence of mini-chromosomes, the model may predict the number of mini-chromosomes in each isolate. Nevertheless, this study demonstrates the potential of deep learning techniques in genomics for predicting the presence of specific genomic elements. We anticipate that in the near future, using improved explainable deep learning techniques, the critical sequence components of mini-chromosome DNAs may be identified by learning from massive genomic data to further understand the origin and evolution of mini-chromosomes.

## Data availability

Nanopore genomic sequencing data of T3 have been deposited in the Sequence Read Archive (SRA) database under accessions PRJNA1002604, and Illumina sequencing data in PRJNA1002398. Training data and related scripts are available at GitHub (https://github.com/PlantG3/miniC) and Zenodo (https://doi.org/10.5281/zenodo.13177408). The T3 genome assembly is available from Genbank accessions: CP132160–CP132167.

## Supplementary data

Supplementary Data are available at NARGAB Online.

## Acknowledgements

We thank funding provided by the National Institute of Food and Agriculture (NIFA) at the U.S. Department of Agriculture (USDA) award (2018–67013-28511) to S. Liu; USDA NIFA award (2021–68013-33719) to B. Valent; and USDA NIFA award (2021–67013-35724) to S. Liu, B. Valent, D. Cook, and D. Koo; the National Science Foundation (NSF) awards (1741090 and 2311738) to S. Liu; and the NSF award (2011500) to S. Liu, B. Valent, D. Cook, and D. Koo. This is contribution no. 24-030-J from the Kansas Agricultural Experiment Station, Manhattan, Kansas.

*Author contribution*: N.G. and S.L. conceptualized experiments; G.L., J.H., R.B., L.C.D., H.Z., G.C. conducted experiments; N.G., Y.H. and S.L. analyzed data; N.G., D. Caragea, D. Cook, B.V. and S.L. wrote the manuscript; all authors reviewed and revised the manuscript.

## Funding

Directorate for Biological Sciences at the National Science Foundation [1741090, 2311738]; Division of Integrative Organismal Systems at the National Science Foundation [2011500]; National Institute of Food and Agriculture at the U.S. Department of Agriculture [2018-67013-28511, 2021-67013-35724, 2021-68013-33719].

## Conflict of interest statement

S.L. is the co-founder of Data2Bio, LLC. Other authors claim no competing interest.

## References

1. Valent,B., Singh,P.K., He,X., Farman,M., Tosa,Y. and Braun,H.J. (2020) CHAPTER 13: Blast diseases: evolution and challenges of a staple food crop fungal pathogen. In: *Emerging Plant Diseases and*

*Global Food Security*. Epidemiology. The American Phytopathological Society, pp. 267–292.
2. Ristaino,J.B., Anderson,P.K., Bebber,D.P., Brauman,K.A., Cunniffe,N.J., Fedoroff,N.V., Finegold,C., Garrett,K.A., Gilligan,C.A., Jones,C.M., *et al.* (2021) The persistent threat of emerging plant disease pandemics to global food security. *Proc. Natl. Acad. Sci. U.S.A.*, **118**, e2022239118.
3. Gladieux,P., Condon,B., Ravel,S., Soanes,D., Maciel,J.L.N., Nhani,A. Jr, Chen,L., Terauchi,R., Lebrun,M.-H., Tharreau,D., *et al.* (2018) Gene flow between divergent cereal- and grass-specific lineages of the rice blast fungus Magnaporthe oryzae. *mBio*, **9**, e01219-17.
4. Fernandez,J. and Orth,K. (2018) Rise of a cereal killer: the biology of Magnaporthe oryzae biotrophic growth. *Trends Microbiol.*, **26**, 582–597.
5. Islam,M.T., Croll,D., Gladieux,P., Soanes,D.M., Persoons,A., Bhattacharjee,P., Hossain,M.S., Gupta,D.R., Rahman,M.M., Mahboob,M.G., *et al.* (2016) Emergence of wheat blast in Bangladesh was caused by a South American lineage of Magnaporthe oryzae. *BMC Biol.*, **14**, 84.
6. Malaker,P.K., Barma,N.C., Tiwary,T.P., Collis,W.J., Duveiller,E.P., Singh,K., Joshi,A.K., Singh,R.P., Braun,H.-J., Peterson,G.L., *et al.* (2016) First report of wheat blast caused by Magnaporthe oryzae pathotype triticum in Bangladesh. *Plant Dis.*, **100**, 2330.
7. Tembo,B., Mulenga,R.M., Sichilima,S., M'siska,K.K., Mwale,M., Chikoti,P.C., Singh,P.K., He,X., Pedley,K.F., Peterson,G.L., *et al.* (2020) Detection and characterization of fungus (Magnaporthe oryzae pathotype Triticum) causing wheat blast disease on rain-fed grown wheat (Triticum aestivum L.) in Zambia. *PLoS One*, **15**, e0238724.
8. Liu,S., Lin,G., Ramachandran,S.R., Daza,L.C., Cruppe,G., Tembo,B., Singh,P.K., Cook,D., Pedley,K.F. and Valent,B. (2023) Rapid mini-chromosome divergence among fungal isolates causing wheat blast outbreaks in Bangladesh and Zambia. *New Phytol.*, **241**, 1266–1276.
9. Latorre,S.M., Were,V.M., Foster,A.J., Langner,T., Malmgren,A., Harant,A., Asuke,S., Reyes-Avila,S., Gupta,D.R., Jensen,C., *et al.* (2023) Genomic surveillance uncovers a pandemic clonal lineage of the wheat blast fungus. *PLoS Biol.*, **21**, e3002052.
10. Lenne,J.M., Takan,J.P., Mgonja,M.A., Manyasa,E.O., Kaloki,P., Wanyera,N., Okwadi,J., Muthumeenakshi,S., Brown,A.E., Tamale,M., *et al.* (2007) Finger millet blast disease management: A key entry point for fighting malnutrition and poverty in East Africa. *Outlook Agric*, **36**, 101–108.
11. Sharma,R., Girish,A.G., Upadhyaya,H.D., Humayun,P., Babu,T.K., Rao,V.P. and Thakur,R.P. (2014) Identification of blast resistance in a core collection of foxtail millet germplasm. *Plant Dis.*, **98**, 519–524.
12. Rahnama,M., Condon,B., Ascari,J.P., Dupuis,J.R., Del Ponte,E.M., Pedley,K.F., Martinez,S., Valent,B. and Farman,M.L. (2023) Recent co-evolution of two pandemic plant diseases in a multi-hybrid swarm. *Nat. Ecol. Evol.*, **7**, 2055–2066.
13. Dean,R.A., Talbot,N.J., Ebbole,D.J., Farman,M.L., Mitchell,T.K., Orbach,M.J., Thon,M., Kulkarni,R., Xu,J.-R., Pan,H., *et al.* (2005) The genome sequence of the rice blast fungus Magnaporthe grisea. *Nature*, **434**, 980–986.
14. Peng,Z., Oliveira-Garcia,E., Lin,G., Hu,Y., Dalby,M., Migeon,P., Tang,H., Farman,M., Cook,D., White,F.F., *et al.* (2019) Effector gene reshuffling involves dispensable mini-chromosomes in the wheat blast fungus. *PLoS Genet.*, **15**, e1008272.
15. Croll,D., Zala,M. and McDonald,B.A. (2013) Breakage-fusion-bridge cycles and large insertions contribute to the rapid evolution of accessory chromosomes in a fungal pathogen. *PLoS Genet.*, **9**, e1003567.
16. Orbach,M.J., Chumley,F.G. and Valent,B. (1996) Electrophoretic karyotypes of Magnaporthe grisea pathogens of diverse grasses. *Mol. Plant Microbe Interact.*, **9**, 261–271.
17. Langner,T., Harant,A., Gomez-Luciano,L.B., Shrestha,R.K., Malmgren,A., Latorre,S.M., Burbano,H.A., Win,J. and Kamoun,S. (2021) Genomic rearrangements generate hypervariable mini-chromosomes in host-specific isolates of the blast fungus. *PLoS Genet.*, **17**, e1009386.
18. Huang,J., Liu,S. and Cook,D.E. (2023) Dynamic Genomes - Mechanisms and consequences of genomic diversity impacting plant-fungal interactions. *Physiol. Mol. Plant Pathol.*, **125**, 102006.
19. Chuma,I., Isobe,C., Hotta,Y., Ibaragi,K., Futamata,N., Kusaba,M., Yoshida,K., Terauchi,R., Fujita,Y., Nakayashiki,H., *et al.* (2011) Multiple translocation of the AVR-Pita effector gene among chromosomes of the rice blast fungus Magnaporthe oryzae and related species. *PLoS Pathog.*, **7**, e1002147.
20. Talbot,N.J., Salch,Y.P., Ma,M. and Hamer,J.E. (1993) Karyotypic Variation within Clonal Lineages of the Rice Blast Fungus, Magnaporthe grisea. *Appl. Environ. Microbiol.*, **59**, 585–593.
21. Shen,Z., Bao,W. and Huang,D.-S. (2018) Recurrent neural network for predicting transcription factor binding sites. *Sci. Rep.*, **8**, 15270.
22. Zhang,Y., Qiao,S., Ji,S. and Li,Y. (2020) DeepSite: bidirectional LSTM and CNN models for predicting DNA–protein binding. *Int. J. Mach. Learn. Cybernet.*, **11**, 841–851.
23. Liu,Q., Fang,L., Yu,G., Wang,D., Xiao,C.-L. and Wang,K. (2019) Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nat. Commun.*, **10**, 2449.
24. Hochreiter,S. and Schmidhuber,J. (1997) Long short-term memory. *Neural Comput.*, **9**, 1735–1780.
25. Bengio,Y., Simard,P. and Frasconi,P. (1994) Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.*, **5**, 157–166.
26. Schuster,M. and Paliwal,K.K. (1997) Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, **45**, 2673–2681.
27. Baldi,P., Brunak,S., Frasconi,P., Soda,G. and Pollastri,G. (1999) Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, **15**, 937–946.
28. Rahnama,M., Wang,B., Dostart,J., Novikova,O., Yackzan,D., Yackzan,A., Bruss,H., Baker,M., Jacob,H., Zhang,X., *et al.* (2021) Telomere roles in fungal genome evolution and adaptation. *Front. Genet.*, **12**, 676751.
29. Rowe,D., Huang,J., Zhang,W., Mishra,D., Jordan,K., Valent,B., Liu,S. and Cook,D.E. (2023) Natural genomic variation in rice blast genomes is associated with specific heterochromatin modifications. bioRxiv doi: https://doi.org/10.1101/2023.08.30.555587, 01 September 2023, preprint: not peer reviewed.
30. Gómez Luciano,L.B., Jason Tsai,I., Chuma,I., Tosa,Y., Chen,Y.-H., Li,J.-Y., Li,M.-Y., Jade Lu,M.-Y., Nakayashiki,H. and Li,W.-H. (2019) Blast fungal genomes show frequent chromosomal changes, gene gains and losses, and effector gene turnover. *Mol. Biol. Evol.*, **36**, 1148–1161.
31. Marçais,G., Delcher,A.L., Phillippy,A.M., Coston,R., Salzberg,S.L. and Zimin,A. (2018) MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.*, **14**, e1005944.
32. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
33. Lin,G., He,C., Zheng,J., Koo,D.-H., Le,H., Zheng,H., Tamang,T.M., Lin,J., Liu,Y., Zhao,M., *et al.* (2021) Chromosome-level genome assembly of a regenerable maize inbred line A188. *Genome Biol.*, **22**, 175.
34. Valent,C.,,Weaver and,Chumley (1986) Genetic studies of fertility and pathogenicity in Magnaporthe grisea (Pyricularia oryzae). *Iowa State J. Res.*, **60**, 569–594.
35. Koren,S., Walenz,B.P., Berlin,K., Miller,J.R., Bergman,N.H. and Phillippy,A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.*, **27**, 722–736.
36. Walker,B.J., Abeel,T., Shea,T., Priest,M., Abouelliel,A., Sakthikumar,S., Cuomo,C.A., Zeng,Q., Wortman,J., Young,S.K., *et al.* (2014) Pilon: an integrated tool for comprehensive microbial

variant detection and genome assembly improvement. *PLoS One*, **9**, e112963.
37. Loman,N.J., Quick,J. and Simpson,J.T. (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods*, **12**, 733–735.
38. Rahnama,M., Novikova,O., Starnes,J.H., Zhang,S., Chen,L. and Farman,M.L. (2020) Transposon-mediated telomere destabilization: A driver of genome evolution in the blast fungus. *Nucleic Acids Res.*, **48**, 7197–7217.
39. Igarashi,S. (1986) Pyricularia em trigo. 1. Ocorrencia de Pyricularia sp. noestado do Parana. *Fitopatol.Bras*, **11**, 351–352.
40. Inoue,Y., Vy,T.T.P., Yoshida,K., Asano,H., Mitsuoka,C., Asuke,S., Anh,V.L., Cumagun,C.J.R., Chuma,I., Terauchi,R., *et al.* (2017) Evolution of the wheat blast fungus through functional losses in a host specificity determinant. *Science*, **357**, 80–83.
41. Maciel,J.L.N., Ceresini,P.C., Castroagudin,V.L., Zala,M., Kema,G.H.J. and McDonald,B.A. (2014) Population structure and pathotype diversity of the wheat blast pathogen Magnaporthe oryzae 25 years after its emergence in Brazil. *Phytopathology*, **104**, 95–107.
42. Chuma,I., Tosa,Y., Taga,M., Nakayashiki,H. and Mayama,S. (2003) Meiotic behavior of a supernumerary chromosome in Magnaporthe oryzae. *Curr. Genet.*, **43**, 191–198.
43. Valent,B., Farrall,L. and Chumley,F.G. (1991) Magnaporthe grisea genes for pathogenicity and virulence identified through a series of backcrosses. *Genetics*, **127**, 87–101.
44. Latorre,S.M., Reyes-Avila,C.S., Malmgren,A., Win,J., Kamoun,S. and Burbano,H.A. (2020) Differential loss of effector genes in three recently expanded pandemic clonal lineages of the rice blast fungus. *BMC Biol.*, **18**, 88.
45. Thierry,M., Charriat,F., Milazzo,J., Adreit,H., Ravel,S., Cros-Arteil,S., Borron,S., Sella,V., Kroj,T., Ioos,R., *et al.* (2022) Maintenance of divergent lineages of the Rice Blast Fungus Pyricularia oryzae through niche separation, loss of sex and post-mating genetic incompatibilities. *PLoS Pathog.*, **18**, e1010687.



**Received:** October 10, 2023. **Revised:** June 27, 2024. **Editorial Decision:** August 1, 2024. **Accepted:** August 6, 2024