Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Plant Biology

Deep learning for plant genomics and crop improvement

Hai Wang[1,2,3], Emre Cimen[2,4], Nisha Singh[2,5] and Edward Buckler[2,6]

Our era has witnessed tremendous advances in plant genomics, characterized by an explosion of high-throughput techniques to identify multi-dimensional genome-wide molecular phenotypes at low costs. More importantly, genomics is not merely acquiring molecular phenotypes, but also leveraging powerful data mining tools to predict and explain them. In recent years, deep learning has been found extremely effective in these tasks. This review highlights two prominent questions at the intersection of genomics and deep learning: 1) how can the flow of information from genomic DNA sequences to molecular phenotypes be modeled; 2) how can we identify functional variants in natural populations using deep learning models? Additionally, we discuss the possibility of unleashing the power of deep learning in synthetic biology to create novel genomic elements with desirable functions. Taken together, we propose a central role of deep learning in future plant genomics research and crop genetic improvement.

Addresses
[1] National Maize Improvement Center, Key Laboratory of Crop Heterosis and Utilization, Joint Laboratory for International Cooperation in Crop Molecular Breeding, China Agricultural University, Beijing 100193, China
[2] Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853, USA
[3] Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China
[4] Computational Intelligence and Optimization Laboratory, Industrial Engineering Department, Eskisehir Technical University, Eskisehir 26000, Turkey
[5] ICAR-National Institute for Plant Biotechnology, New Delhi 110012, India
[6] United States Department of Agriculture, Agricultural Research Service, Ithaca, NY 14853, USA

Corresponding author: Wang, Hai (wanghai@cau.edu.cn)

Current Opinion in Plant Biology 2020, **54**:34–41

This review comes from a themed issue on **Genome studies and molecular genetics**

Edited by **Josh T Cuperus** and **Christine Queitsch**

For a complete overview see the Issue and the Editorial

Available online 24th January 2020

https://doi.org/10.1016/j.pbi.2019.12.010

1369-5266/© 2020 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction

Plants, like all other life forms on earth, can be viewed as a flow of information. This information flow starts from the genomic DNA sequence, and ends at terminal observed phenotypes or for crop species, agronomic traits. In between is a relay of information by transcription and translation, processes summarized as 'the central dogma in molecular biology' by Francis Crick in 1957 [1]. Each step in the central dogma can be viewed as not only transmission, but also transformation of genetic information from an earlier step. The molecular features involved are collectively referred to as 'molecular phenotypes', to set them apart from terminal traits. More importantly, in the genomics era, multifaceted molecular phenotypes involved in the information relay, including the structure, modification, function, and evolution of elements in DNA, RNA, and protein, along with their interactions, are beginning to be revealed at scale and at reduced cost [2], facilitating fine-grained dissection of information transfer and transformation along the central dogma.

Understanding this flow of information is key for both basic research and crop improvement, but the question of how to do this remains. In plant forward genetics, we usually exploit genetic variation at the DNA level (created by either artificial mutagenesis or natural variation) in linkage or association analysis, to identify genomic variation associated with or, ideally, causal to a specific phenotypic variation [3]. However, these approaches are not without their shortcomings: the rich information in the molecular phenotypes is largely unexplored, making an end-to-end mechanistic understanding from DNA sequences to terminal phenotypes difficult.

Notably, this gap is now being closed by advances in two areas of research. One is association analysis linking molecular phenotypes and terminal phenotypes, such as the transcriptome-wide association study (TWAS), which benefits from a shorter path of information relay, and involves fewer steps of information transformation when compared with the genome-wide association study (GWAS) [4]. The other advance is the prediction of molecular phenotypes from their upstream molecular phenotypes, or directly from genomic DNA sequences, by deep learning models [5]. In this review, we introduce recent progress in molecular phenotype modeling by using deep learning approaches, and propose their application to identify or prioritize functional variants potentially valuable for crop genetic improvement. The possibility of using deep learning models in synthetic biology to create novel beneficial alleles is also discussed. We propose that the deep learning framework discussed above, combined with high-throughput genome editing, will prove helpful for the upcoming 'Breeding 4.2' era, in

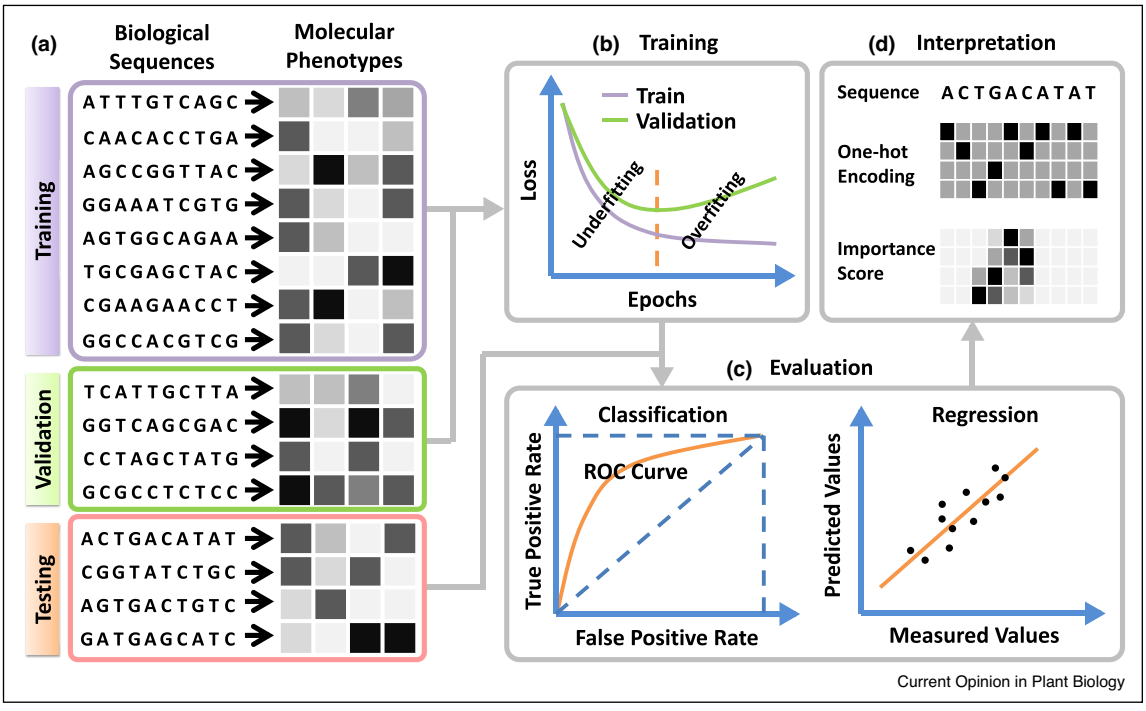
which beneficial variants are rationally combined and created with unprecedented efficiency [6•].

Deep learning: concepts, tools, and caveats

Machine learning is the science of programming computers so they can learn from data [7]. Problems in this field can be divided into two main types: supervised and unsupervised. The aim of supervised learning is to obtain a model which maps its predictors (such as DNA sequences) to target variables (such as histone marks) (Figure 1a). Target variables can be either categorical (classification) or continuous (regression) (Figure 1c). Some examples of supervised learning applications are: Predicting regulatory and non-regulatory regions in the maize genome [8], predicting mRNA expression levels [9••], sequence tagging in rice [10], plant stress phenotyping [11], polyadenylation site prediction in *Arabidopsis thaliana* [12] and predicting macronutrient deficiencies in tomato [13]. If there is no specification about the outcome in the data set, then the problem becomes unsupervised. Clustering and feature extraction [14] are in this group.

Artificial Neural Networks (ANN) are well-known methods to solve machine learning problems, and have been studied since the 1940s, inspired by the animal nervous system. ANNs consist of an input, an output, and several hidden layers. Deep Neural Network (DNN) is a type of ANN and a relatively young branch of machine learning. DNNs distinguish from ANNs by having many more hidden layers. Clearly, as prediction power increases with DNN, data requirement increases as well. As in other conventional learning methods, features in the input vector of a DNN need to be extracted first and are assumed to be independent of each other. A subset of DNNs is Convolutional Neural Networks (CNN). CNNs have at least a convolutional layer, which provides them the ability of automatic feature extraction from a continuous signal (e.g. weather data as a time series, plant image or DNA/RNA sequence). A CNN can be trained with a DNA/RNA sequence that has N base pairs and the sequence can be represented as one-hot encoded $4 \times N$ matrix (Figure 1d) to train the model. CNNs can capture the local motifs even if they appear in different parts of

Figure 1



A workflow for deep learning in genomics.
A deep learning workflow with biological sequences and molecular phenotypes as predictors and target variables, respectively, typically comprises four steps. (a) Preprocessing of predictors and target variables: retrieving and encoding of biological sequences, numerical or categorical representation of molecular phenotypes, and proper splitting of predictor-target pairs into the training, validation and test sets, usually with evolutionary relationships among biological sequences taken into consideration [9••]. (b) Model building and training: selection of model architecture and hyper-parameters as well as training models on the training set. Notably, the performance of the model on the validation set should be monitored continuously during training in order to determine when to stop model training to avoid both under and overfitting. (c) Model evaluation: evaluation of the performance of trained models on another dataset, termed the test set. Metrics used to measure the performance of models depend on the nature of the target variable: area under the ROC Curve (auROC) is one metric used for classification problems, while R-squared is a metric used for regression problems. (d) Interpretation of models by saliency or feature attribution methods to identify functional elements in biological sequences.

the input. Moreover, convolutional layers reduce the number of weights to learn related to fully connected layers. There are multiple examples of CNN applications in plant biology [9••,11–13,15]. Zou *et al.* provided an interactive tutorial to build a convolutional neural network to discover DNA-binding motifs [16]. Recurrent Neural Networks (RNNs) (and Convolutional Recurrent Neural Networks) are another subset of DNNs. In RNNs, outputs of some layers are fed back into the inputs of a previous layer. This operation provides memory capability to RNNs. Furthermore, RNNs can handle inputs with different size and they have advantages when the input is time series. There are various examples in the literature that applies RNNs to plant biology [10,17–19].

When machine learning methods are employed to solve problems in genomics, several important caveats should be considered. Most of them are general to machine learning, while others are genomics-specific. When building machine learning models, observations are usually randomly divided into the training set, used to train the model, the validation set, used to determine model architecture and hyper-parameters, and the test set, used to evaluate the performance of the model (Figure 1a). The model is expected to generalize well, which means it should perform similarly on the test set and on the training set. However, sometimes models perform significantly worse on the test set than on the training set, a phenomenon called *overfitting* (Figure 1b). Several scenarios lead to overfitting namely model complexity, high dimensionality, and so on. Dimensionality in the feature space sometimes greatly surpasses the number of observations. For example, the number of genomic SNPs assayed almost always exceeds the number of plant genotypes when predicting a phenotype from genomic variants, as large-scale phenotyping projects are still quite expensive. Moreover, there are also cases where overfitting is hidden and unnoticeable when dealing with problems in genomics. For example, when members of a gene family are split between the training set and the validation/test set, there is a risk that models would learn family specific molecular patterns and report overly optimistic predictive accuracies [9••]. Such hidden overfitting may also be observed when observations in the training set and the test set share common molecular features [20] or genomic loci [21].

When the main purpose is to not only accurately predict but also to explain the biological rules, interpretability [22] of the machine learning model and quantifying feature importance become essential for plant biologists (Figure 1d). For instance, while predicting a phenotype accurately from a plant's genome, additionally a scientist would like to know the effect of each nucleotide. While deep learning provides high accuracy in predictions, sometimes the deep learning models are hard to interpret, which is crucial to explore the inference of biological process. In order to build more interpretable models, SHAP (SHapley Additive exPlanations) [23] assigns each feature an importance value for a particular prediction. DeepLIFT (Deep Learning Important FeaTures) [24] decomposes the output prediction of a neural network on a specific input to define important features. For a similar purpose, integrated gradients [25] aim to attribute the prediction of a deep network to its input features. On the other hand Moreover, the choice of encoded biological features also plays a key role in interpretability. Finally, it is also important to consider measurement errors or errors made during data set submission before running the model or interpreting the results.

Deep learning along the central dogma of molecular biology

DNA and gene properties

Deep learning has been applied in several areas of large-scale data analysis to resolve complex biological problems in genomics, transcriptomics, proteomics, metabolomics and systems biology [26]. Several studies revealed that DNA shape plays an important role in determining transcription factor (TF) DNA-binding specificity [27]. A large range of data types are available, including chromatin accessibility assays (e.g. MNase-seq, DNase-seq, FAIRE) and other genomic assays (e.g. microarray, RNA-seq expression). Similarly, for transcription factor (TF) binding, there exists ChIP-seq data, gene expression profiles, DAP-seq (DNA affinity purification sequencing) and ampDAP-seq, which uses amplified and thus, demethylated DNA as substrates and histone modifications to understand the mechanisms underlying gene expression [28]. To analyze these large-scale data sets, several deep learning methods were developed to model TF DNA-binding specificity. In order to predict *in vivo* TF binding, several methods have been developed based on deep learning. For example, DeepBind can learn several motifs to predict binding sites of DNA and RNA binding proteins [29]. TFImpute predict cell-specific TF binding trained [30]. The effects of functional noncoding variants were evaluated in DeepSEA [31], DeFind [32] and DFIM [33]. To differentiate between DNA and RNA-binding residues DRNApred was developed [34]. All of these above described methods are mostly trained and tested on human tissues or cell lines due to easy availability of data sets. In species such as maize, which has lots of repetitive elements and wide intergenic regions, it is challenging to identify the key genomic regulatory regions. To address these challenges, approaches such as k-mer grammars based on natural language processing have been used to annotate regulatory regions in maize lines in a cost-effective and precise manner [8]. Machine learning approaches have played a significant role in modeling transcription factor binding sites. Machine learning models have proven powerful in several aspects of plant biology. They can be trained from various types of sequencing data either alone or in combined manner, and also further integrate other

information, such as DNase I hypersensitivity data, for better *in vivo* transcription binding sites (TFBSs) prediction [30].

When comparing CNNs and k-mer methods, CNNs are more effective in feature extraction. However, CNNs are often considered black boxes because interpretation of their output is challenging and can involve high computational cost. Also, how much of their performance is derived from learning fundamental biological rules such as key motifs, motif relationships, and the general sequence perspective is quite uncertain. For the purpose of interpretation of DNA, k-mer approaches are preferable over CNNs and RNNs. Classification of sequences using frequencies of k-mers (or k-tuples, k-grams) is fast, accurate, reference-free, and alignment-free. A k-mer is gene-based approach to identify sequence signatures. Typically, k-mer frequency vectors are paired together with a distance function in order to measure the quantitative similarity between any pair of sequences. These methods are easily interpretable and based on word statistics to recover semantic and syntactic cues but, determining why a sequence is classified a certain way is not as straight-forward as a more traditional alignment-based approach. However, using a k-mer representation seems to be a good balance for accurate and rapid classification. Notably, there are also examples combining both k-mer approaches and deep learning models [35], although the impact of this approach on precision or interpretability has not been systemically evaluated.

Protein properties

The function of any protein directly depends on its tertiary structure. The tertiary structure of protein can be revealed by synthetically analyzing various protein properties, such as secondary structure, transmembrane topology, signal peptides, solvent accessibility, backbone dihedrals, disorder-to-order transition, contact maps, model qualities, inter-residual contact, protein interaction sites, protein disorders, and enzyme dynamics. To extract important amino acid features from *de novo* peptide sequence DeepNovo was developed using a CNN approach [36]. Recently, Google's AlphaFold has generated tremendous excitement by using advances in artificial intelligence to predict a protein's tertiary structure [37]. In order to predict secondary structure, relative solvent accessibility and inter-residue contact maps raw-MSA was used in deep learning models [38]. However, deep learning algorithms have achieved successful results in diverse areas, but their effectiveness for PPI prediction was quite low due to low coverage and noisy data. In this context, DPPI, a new model able to predict PPIs and homodimeric interactions from sequence information [39]. DEEPre is a sequence-based enzyme EC number prediction by deep learning to annotate enzyme functions in metagenomics, industrials biotechnology and diseases [40].

Model and data sharing by model zoos and data repositories

Although a large number of deep learning models have been developed to solve problems in human or animal genomics, they are often developed in different frameworks that require a myriad of different dependencies, making it difficult for researchers to test published models on new data or adapt existing models in new ensemble or transfer learning tasks. Following FAIR (Findable, Accessible, Interoperable, and Reusable) principles [41], the Kipoi repository has recently been developed to accelerate community exchange and reuse of predictive models for genomics [42••]. Most models in Kipoi developed for animal or human genomes can be easily retrained using plant genomics datasets, or even applied directly on plants (such as models predicting biochemical properties of proteins); however, care must be taken when the biological question being studied involves plant-specific problems. For example, when modeling relative gene expression levels in maize and sorghum, the tetraploidy of maize may cause some challenges [9••]. Polyploidy and extensive tandem duplication of genes in plant species may also lead to biased quantification of gene expression, resulting in lower-quality training and test datasets. Moreover, as genome elements (such as introns, exons, or distances between enhancers and promoters) often differ significantly in their sizes between animal and plant species, re-optimization of model architectures and hyper-parameters may be crucial before models in animals can be retrained for plant species.

In addition to model zoos, databases such as CyVerse (http://www.cyverse.org/) are needed to accommodate omics data on which models are developed. This would alleviate the lack of high-quality large-scale data sets in genomics, and also offer the opportunity to build smart approaches to fuse heterogeneous datasets to further facilitate transfer learning.

Making sense of genomic variation: from association to causality and molecular mechanisms

As discussed above, deep learning models can be used to predict molecular phenotypes (such as transcription factor binding, epigenetic marks, chromatin state, and gene expression levels) given biological sequences as predictors. The most powerful part of deep learning models is their ability to make *ab initio* predictions on new, previously unseen sequence data (i.e. data not in the training set), which has several important implications.

First, although there are a huge number of genetic variants in a natural population, deep learning models can be trained on a small subset of them to predict the effect of all other variants (i.e. the whole mutation space) [43••]. For example, models trained on some genes can be used to make predictions on other genes. These include

not only common alleles, but also low-frequency and rare variants, irrespective of the magnitude of their effects. As the biology governing molecular processes in closely related species are conserved, models trained in one species can be applied directly on closely related species [44]. Or these models can be used as teacher models in transfer learning tasks in closely related species, facilitating the migration of knowledge from well-studied species (such as Arabidopsis) to related but poorly characterized species (such as other species in the Brassicaceae).

Second, when several variants within an important loci (such as a QTL for a certain trait) are in close linkage disequilibrium (Figure 2a), we may introduce the variants from one haplotype to another one by one by *in silico* mutagenesis, and then evaluate their impacts on the molecular phenotype individually, thereby prioritizing causal variants (Figure 2b,c). Such a breakage of linkage disequilibrium would be labor-intensive and difficult to scale up in wet lab experiments, and virtually impossible in nature.

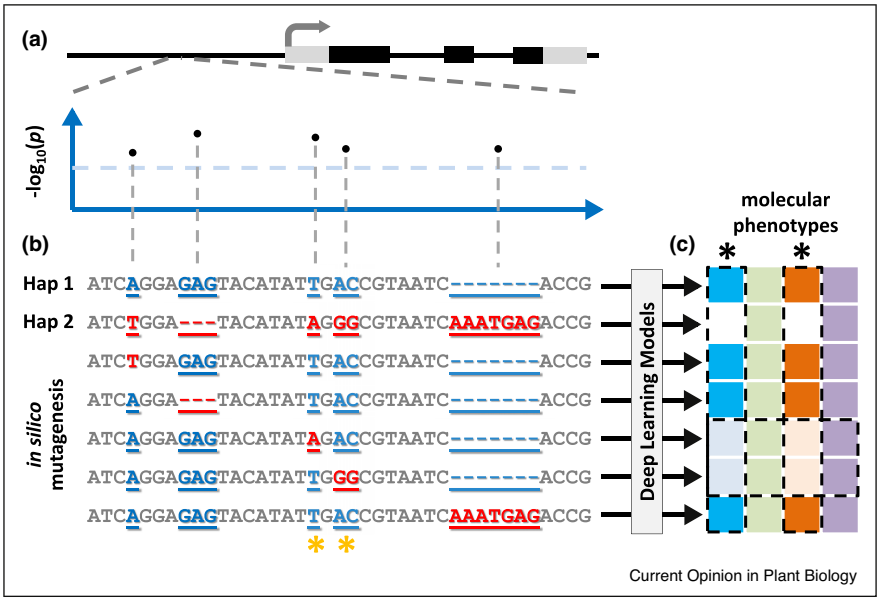
Third, with a rich repertoire of deep learning models each targeting a distinct molecular phenotype, or a multi-task learning model targeting multiple molecular phenotypes simultaneously, it is possible to predict not only the causal variant underlying a QTL, but also its potential molecular mechanism (Figure 2c). Taken together, deep learning models can greatly push forward our understanding of genomic variation underlying terminal phenotypes.

Deep learning for breeding 4: breeding-by-editing

An important component of crop breeding is the purging of deleterious alleles in the context of environmental adaptation and modern management practices. The past 30 years, previously summarized as the Breeding 3.0 era [6•], have seen a great triumph in marker-assisted selection, association analysis, and genomic prediction.

It is worth noting that the genetic variants used in marker-assisted breeding during the Breeding 3.0 era are not necessarily variants causal to agronomic traits. What can breeders do when they have the power to predict causal beneficial and deleterious variants at scale? One answer is the breaking of linkage drag by genome editing: beneficial alleles can be directly introduced into elite germplasms by editing, rather than by backcrossing from another donor parent carrying deleterious alleles at linked loci. Similarly, deleterious alleles can be purged from the genome effectively by editing. Indeed, simulations have shown that breeding in livestock can be significantly

Figure 2



Application of deep learning models on sequence variants.
(a) In natural populations, association analyses often identify groups of variants in close linkage disequilibrium associated with a terminal trait or a molecular phenotype, but the identification of causal variants, as well as the molecular mechanism underlying each variant, is difficult to achieve with association analysis alone. To evaluate the effects of variants in linkage disequilibrium individually, variants from one haplotype can be introduced into another haplotype one at a time by *in silico* mutagenesis. **(b)** Then the effects of each mutation are predicted with a group of deep learning models, each targeting a distinct molecular phenotype. **(c)** Thus, deep learning in genomics provides a powerful tool to prioritize causal variants (denoted by yellow asterisks) by breaking linkage disequilibrium *in silico* and to identify potential molecular mechanisms (denoted by black asterisks) for each putative causal variant.

accelerated by using genome editing to introduce beneficial variants into the genome [45•], or to remove deleterious alleles [46•]. However, because of more prominent interactions between genotypes and environments in crop species than in livestock, it is conceivable that prediction of allele effects (whether it is deleterious, beneficial or adaptive) would be more challenging in crop species. Ideally, environment-specific models or models taking environmental factors as additional inputs would alleviate this problem. Thus, it is reasonable to conceptualize that functional variants predicted by deep learning models will be the key in the next breeding era, termed Breeding 4.0, in which genetic improvement of crop species largely depends on genome editing.

More importantly, we are not restricted to known beneficial variants existing in nature when carrying out this breeding-by-editing approach. Instead, we enjoy complete freedom to create novel beneficial alleles based on our deep learning models' 'understanding' of the biological processes of interest. For example, Rodriguez-Leal *et al.* edited the tomato *CLAVATA3* gene (*SlCLV3*) promoter to increase the fruit size and optimize the inflorescence branching [47•]. Because of a lack of functional annotations in the *SlCLV3* promoter, saturated promoter mutagenesis by the CRISPR/Cas9 system was employed, followed by selection of mutants with desirable fruit and inflorescence traits. In the future, with a deep learning model predicting gene expression levels from promoter sequences, it is possible to identify key *cis*-elements on the *SlCLV3* promoter by saliency scores at single-nucleotide resolution, predict their loss-of-function effects on *SlCLV3* gene expression, and then implement model-guided promoter editing.

Another way to create novel genomic elements with specific functions is to apply generative models in synthetic biology. For example, it is possible to train models to create new promoters with spatiotemporal specificity after learning the mutation space of existing promoters. However, although generative models such as variational autoencoders and generative adversarial networks have drawn much attention recently, their potential applications in synthetic biology are still quite limited. One such example is to apply GANs to generate synthetic DNA sequences coding for antimicrobial peptides [48]. It will be promising to use generative models to create new DNA elements, genes, or even regulatory circuits with desirable functions, and apply them to crop improvement.

Conclusion

In natural plant populations, association mapping has been successfully exploited to reveal genetic loci associated with molecular phenotypes or terminal traits. However, due to prevalent linkage disequilibrium among nearby variants, causal variants underlying phenotypic variation are still difficult to pinpoint, hampering the genetic improvement of plants by genome editing. On the other hand, progress in molecular biology in the past half century has discovered many of the molecular mechanisms governing the flow of information from DNA to molecular phenotypes such as RNA and protein, and the accumulation of such data has recently been accelerated by various omics approaches based on advanced sequencing techniques. Thus, it is natural to hypothesize that prioritization of causal variant should be achievable by combining models that can 'understand' the flow of information from DNA to molecular phenotypes, as well as association mapping studies linking molecular phenotypes to terminal traits. Indeed, such a framework has been proven not only feasible, but also powerful in human genetics to reveal variants (including rare alleles) underlying certain genetic diseases [43••,49••]. This trend, however, remains not fully exploited by the plant community. The tremendous progress in the development of deep learning models is molecular phenotype prediction, as well as application of these models in functional variant discovery by *in silico* breakage of linkage disequilibrium. We propose that such a framework is a promising approach for genome-wide identification of deleterious and adaptive variants, a prerequisite for editing-based genetic improvement of crops in future agriculture.

Conflict of interest statement

Nothing declared.

Acknowledgements

We thank Travis Wrightsman and Sara Miller for their helpful discussions and suggestions. We apologize to all colleagues whose work was not referenced due to space constraints. This work was supported by the USDA-ARS and the Bill and Melinda Gates Foundation, and the Tang Cornell-China Scholars Program.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Crick F: **Central dogma of molecular biology**. *Nature* 1970, **227**:561-563.

2. Joyce AR, Palsson BØ: **The model organism as a system: integrating "omics" data sets**. *Nat Rev Mol Cell Biol* 2006, **7**:198-210.

3. Sham PC, Cherny SS, Purcell S, Hewitt JK: **Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data**. *Am J Hum Genet* 2000, **66**:1616-1630.

4. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, Ermel R, Ruusalepp A, Quertermous T, Hao K *et al.*: **Opportunities and challenges for transcriptome-wide association studies**. *Nat Genet* 2019, **51**:592-599.

5. Eraslan G, Avsec Ž, Gagneur J, Theis FJ: **Deep learning: new computational modelling techniques for genomics**. *Nat Rev Genet* 2019, **20**:389-403.

6. • Ramstein GP, Jensen SE, Buckler ES: **Breaking the curse of dimensionality to identify causal variants in breeding 4**. *Theor Appl Genet* 2019, **132**:559-567.

The authors discuss approaches to avoid the curse of dimensionality, by involving intermediate phenotypes such as molecular traits and component traits related to plant morphology or physiology.

7. Géron A: *Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc.; 2017.

8. Mejía-Guerra MK, Buckler ES: **A k-mer grammar analysis to uncover maize regulatory architecture**. *BMC Plant Biol* 2019, **19**:103.

9. •• Washburn JD, Mejia-Guerra MK, Ramstein G, Kremling KA, Valluru R, Buckler ES, Wang H: **Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence**. *Proc Natl Acad Sci U S A* 2019, **116**:5542-5549.
Standard training and testing methods are not designed to control for evolutionary relatedness. This paper presents two methods to control for and utilize evolutionary relatedness within a predictive deep learning framework.

10. Do H, Than K, Larmande P: *Evaluating Named-Entity Recognition approaches in plant molecular biology* [date unknown], doi:https://doi.org/10.1101/360966.

11. Ghosal S, Blystone D, Singh AK, Ganapathysubramanian B, Singh A, Sarkar S: **An explainable deep machine vision framework for plant stress phenotyping**. *Proc Natl Acad Sci U S A* 2018, **115**:4613-4618.

12. Gao X, Zhang J, Wei Z, Hakonarson H: **DeepPolyA: a convolutional neural network approach for polyadenylation site prediction**. *IEEE Access* 2018, **6**:24340-24349.

13. Tran T-T, Choi J-W, Le T-T, Kim J-W: **A comparative study of deep CNN in forecasting and classifying the macronutrient deficiencies on development of tomato plant**. *Appl Sci* 2019, **9**:1601.

14. Wu B, Zhang H, Lin L, Wang H, Gao Y, Zhao L, Chen Y-PP, Chen R, Gu L: **A similarity searching system for biological phenotype images using deep convolutional encoder-decoder architecture**. *Curr Bioinf* 2019, **14**:628-639.

15. Zhao Q, Mao Q, Zhao Z, Dou T, Wang Z, Cui X, Liu Y, Fan X: **Prediction of plant-derived xenomiRs from plant miRNA sequences using random forest and one-dimensional convolutional neural network models**. *BMC Genomics* 2018, **19**:839.

16. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A: **A primer on deep learning in genomics**. *Nat Genet* 2019, **51**:12-18.

17. Kulkarni S, Mandal SN, Srivatsa Sharma G, Mundada MR, Meeradevi: **Predictive analysis to improve crop yield using a neural network model**. *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* 2018 http://dx.doi.org/10.1109/icacci.2018.8554851.

18. Meng J, Chang Z, Zhang P, Shi W, Luan Y: **lncRNA-LSTM: prediction of plant long non-coding RNAs using long short-term memory based on p-nts encoding**. *Intell Comput Methodol* 2019, **11645**:347-357.

19. Li H, Yin Z, Manley P, Burken JG, Shakoor, Fahlgren N, Mockler T: **Early drought plant stress detection with bi-directional long-term memory networks**. *Photogramm Eng Remote Sens* 2018, **84**:459-468.

20. Xi W, Beer MA: **Local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy**. *PLoS Comput Biol* 2018, **14**:e1006625.

21. Schreiber Jacob, Singh Ritambhara, Bilmes Jeffrey, Stafford Noble William: **A pitfall for machine learning methods aiming to predict across cell types**. *bioRxiv* 2019, **512434** http://dx.doi.org/10.1101/512434.

22. James Murdoch W, Singh C, Kumbier K, Abbasi-Asl R, Yu B: **Interpretable machine learning: definitions, methods, and applications**. *arXiv [statML]* 2019.

23. Lundberg S, Lee SI: **A unified approach to interpreting model predictions**. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. 2017.

24. Avanti S, Greenside P, Kundaje A: **Learning important features through propagating activation differences**. *Proceedings of the 34th International Conference on Machine Learning* 2017:3145-3153.

25. Sundararajan M, Taly A, Yan Q: **Axiomatic attribution for deep networks**. *arXiv [csLG]* 2017.

26. Xu C, Jackson SA: **Machine learning and complex biological data**. *Genome Biol* 2019, **20**:76.

27. Lai X, Stigliani A, Vachon G, Carles C, Smaczniak C, Zubieta C, Kaufmann K, Parcy F: **Building transcription factor binding site models to understand gene regulation in plants**. *Mol Plant* 2019, **12**:743-763.

28. Zampieri G, Vijayakumar S, Yaneske E, Angione C: **Machine and deep learning meet genome-scale metabolic modeling**. *PLoS Comput Biol* 2019, **15**:e1007084.

29. Alipanahi B, Delong A, Weirauch MT, Frey BJ: **Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning**. *Nat Biotechnol* 2015, **33**:831-838.

30. Qin Q, Feng J: **Imputation for transcription factor binding predictions based on deep learning**. *PLoS Comput Biol* 2017, **13**:e1005403.

31. Zhou J, Troyanskaya OG: **Predicting effects of noncoding variants with deep learning-based sequence model**. *Nat Methods* 2015, **12**:931-934.

32. Wang M, Tai C, Weinan E, Wei L: **DeFine: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants**. *Nucleic Acids Res* 2018, **46**:e69.

33. Greenside P, Shimko T, Fordyce P, Kundaje A: **Discovering epistatic feature interactions from neural network models of regulatory DNA sequences**. *Bioinformatics* 2018, **34**:i629-i637.

34. Yan J, Kurgan L: **DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues**. *Nucleic Acids Res* 2017, **45**:e84.

35. Shen Zhen, Bao Wenzheng, Huang De-Shuang: **Recurrent neural network for predicting transcription factor binding sites**. *Sci Rep* 2017, **8**:15270.

36. Tran NH, Zhang X, Xin L, Shan B, Li M: **De novo peptide sequencing by deep learning**. *Proc Natl Acad Sci U S A* 2017, **114**:8247-8252.

37. Evans R, Jumper J, Kirkpatrick J, Sifre L, Green TFG, Qin C, Zidek A, Nelson A, Bridgland A, Penedones H *et al.*: **De novo structure prediction with deep-learning based scoring**. *In Thirteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstracts)*. 2018.

38. Mirabello C, Wallner B: *rawMSA: End-to-end Deep Learning Makes Protein Sequence Profiles and Feature Extraction obsolete*. [date unknown], doi:https://doi.org/10.1101/394437.

39. Hashemifar S, Neyshabur B, Khan AA, Xu J: **Predicting protein–protein interactions through sequence-based deep learning**. *Bioinformatics* 2018, **34**:802-810.

40. Li Y, Wang S, Umarov R, Xie B, Fan M, Li L, Gao X: **DEEPre: sequence-based enzyme EC number prediction by deep learning**. *Bioinformatics* 2018, **34**:760-769.

41. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE *et al.*: **The FAIR guiding principles for scientific data management and stewardship**. *Sci Data* 2016, **3**:160018.

42. •• Avsec Ž, Kreuzhuber R, Israeli J, Xu N, Cheng J, Shrikumar A, Banerjee A, Kim DS, Beier T, Urban L *et al.*: **The Kipoi repository accelerates community exchange and reuse of predictive models for genomics**. *Nat Biotechnol* 2019, **37**:592-600.
A comprehensive and easy-to-use platform for researchers to share deep learning models in genomics.

43. •• Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG: **Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk**. *Nat Genet* 2018, **50**:1171-1179.

The authors present, for the first time, the feasibility to discover causal variants for genetic diseases, by using deep learning models that predict molecular phenotypes from genomic sequences.

44. David R. Kelley: **Cross-species regulatory sequence activity prediction.** *bioRxiv* 660563; doi: https://doi.org/https://doi.org/10.1101/660563.

45. • Jenko J, Gorjanc G, Cleveland MA, Varshney RK, Whitelaw CBA, Woolliams JA, Hickey JM: **Potential of promotion of alleles by genome editing to improve quantitative traits in livestock breeding programs**. *Genet Sel Evol* 2015, **47**:55.
The authors simulated the benefit of using genome editing to improve quantitative traits.

46. • Johnsson M, Gaynor RC, Jenko J, Gorjanc G, de Koning D-J, Hickey JM: **Removal of alleles by genome editing (RAGE) against deleterious load**. *Genet Sel Evol* 2019, **51**:14.
Model breeding can largely be viewed as a depletion of deleterious alleles. In this work, the authors proposed that depletion of deleterious alleles by genome editing would significantly accelerate breeding.

47. • Rodríguez-Leal D, Lemmon ZH, Man J, Bartlett ME, Lippman ZB: **Engineering quantitative trait variation for crop improvement by genome editing**. *Cell* 2017, **171**:470-480.e8.
The authors showed that it is possible to create novel beneficial alleles by fine-tuning the expression levels of genes controlling important agronomic traits.

48. Gupta A, Zhou J: ***Feedback GAN (FBGAN) for DNA: a NovelFeedback-Loop Architecture for Optimizing Protein Functions***. *arXiv*:1804.01694 [q-bio.GN].

49. •• Zhou J, Park CY, Theesfeld CL, Wong AK, Yuan Y, Scheckel C, Fak JJ, Funk J, Yao K, Tajima Y *et al.*: **Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk**. *Nat Genet* 2019, **51**:973-980.
The authors presented a concrete example of using deep learning models to reveal causal non-coding variants underlying complex human diseases.